Behavior Research Methods
2008, ?? (?), ???-???
doi: 10.3758/X

Identifying topic sentencehood

Philip M. McCarthy, Adam M. Renner, Michael G. Duncan,
Nicholas D. Duran, Erin J. Lightman, and Danielle S. McNamara
University of Memphis, Memphis, Tennessee

Four experiments were conducted to assess two models of topic sentencehood identification: the *derived model* and the *free model*. According to the derived model, topic sentences are identified in the context of the paragraph and in terms of how well each sentence in the paragraph captures the paragraph's theme. In contrast, according to the free model, topic sentences can be identified on the basis of sentential features without reference to other sentences in the paragraph (i.e., without context). The results of the experiments suggest that human raters can identify topic sentences both with and without the context of the other sentences in the paragraph. Another goal of this study was to develop computational measures that approximated each of these models. When computational versions were assessed, the results for the free model were promising; however, the derived model results were poor. These results collectively imply that humans' identification of topic sentences in context may rely more heavily on sentential features than on the relationships between sentences in a paragraph.

Topic sentences have been a target of academic debate and study for nearly 150 years (e.g., Angus, 1862). Researchers and writers, spanning the fields of composition, linguistics, and psychology, generally have agreed that topic sentences help readers to remember text better (Aulls, 1975) and facilitate comprehension (Kieras, 1978), particularly when the text is challenging and when the reader lacks domain-specific knowledge (Goldman, Graesser, & van den Broek, 1999; McNamara, Kintsch, Songer, & Kintsch, 1996; see Duncan, in press, for an extensive review). Although countless definitions of topic sentences have been published, there is widespread general agreement that topic sentences tend to consist of all, or most, of the following features. First, topic sentences tend to be structured as a *claim* as to the main theme or topic of the paragraph. Second, they tend to occur in the first-sentence, or *paragraph-initial*, position. Third, they are generally supported and elaborated by other sentences in the paragraph. Finally, topic sentences are more likely to appear in expository texts (i.e., as compared with narrative texts). Along these lines, Graesser, McNamara, and Louwerse (2003) provided the following advice to writers:

> It is good policy for expository text writers to follow a Topic Sentence + Elaboration rhetorical format. The first sentence identifies the main topic or theme of the paragraph, whereas the subsequent sentences supply additional detail that is relevant to the topic sentence. (p. 87)

Despite its undoubted usefulness, however, a series of empirical studies across a wide range of genres (scientific, academic, technical, and periodical writing) have shown that topic sentences often appear in only 50% of paragraphs (Popken, 1987, 1988, 1991a, 1991b). Popken's broad findings support earlier research that drew similar conclusions (e.g., Braddock, 1974). This lack of topic sentencehood (much lamented by Braddock, 1974) may be less important for texts for which writers and readers share the same discourse community and, therefore, shared knowledge can be assumed. However, low-knowledge and/or less skilled readers may be in particular need of explicit cues in the text, such as topic sentences, to help them organize the information in the text (Aulls, 1975; Fishman, 1978; Goldman, Saul, & Coté, 1995; Graesser et al., 2003; McNamara et al., 1996). Thus, the apparent utility of topic sentences suggests that using them more often may be beneficial. It is with this in mind that the developers of the computational tool Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004), which assesses text on over 600 indices of cohesion and difficulty, have sought to develop a measure that evaluates the quality and distribution of topic sentencehood across texts. We describe here the work that we conducted to develop and test theoretical and computational models of topic sentencehood.

The term *topic sentence* first appeared in a textbook in 1885 by John McElroy, but the concept originates earlier (Angus, 1862) and most clearly in Alexander Bain's (1866) *English Composition and Rhetoric.* Empirical studies of topic sentences began with Meade and Ellis (1970), who observed that writers often ignore traditional instruction and use alternative structures. As a result, 50% of the paragraphs in top English journals do not contain clear evidence of topic sentences. Similarly, Braddock (1974) found that only 13% of the paragraphs in a corpus of popular magazines contained *explicit* topic sentences, although he cautioned that the readability of many para-

P. M. McCarthy, pmccarthy@mail.psyc.memphis.edu

graphs would have been improved by greater explicit use of topic sentences.

Braddock (1974) was the first to define distinct kinds of topic sentence constructions (*simple*, *assembled*, *implied*, and *major*), each with different demands and effects on the reader. The simple topic sentence is a traditional first-position sentence that clearly encompasses the meaning of the entire paragraph. The assembled topic sentence is similar to the simple one but requires reading at least one more sentence to understand the theme of the paragraph. The implied topic sentence is similar to the assembled one, but all of the sentences in the paragraph are required to understand the theme. The final type, the major topic sentence, is structured like the simple kind but serves also to summarize other paragraphs.

Popken (1987, 1988, 1991a, 1991b) built on Braddock's (1974) work by assessing topic sentencehood frequency across a variety of registers. Popken reported that 55% of the paragraphs in scientific writing, 54% in academic writing, 32% in technical writing, and 30% in periodical writing contained topic sentences. Popken's findings suggested that accomplished writers frequently leave paragraphs without topic sentences, presumably because the topic of their text is familiar to the intended audience, rendering explicit theme sentences unnecessary. Following such studies, composition studies tended to view the topic sentence as optional and elementary, rather than as a necessary component of writing (e.g., D'Angelo, 1986; Eden & Mitchell, 1986). As such, the assumption appears to be that instructing beginning writers to use topic sentences frequently is misleading if accomplished writers use explicit topic sentences in only half of their paragraphs.

Psychological research on topic sentences focuses on readers. For example, Aulls (1975) examined the effects of manipulating the presence of topic sentences in paragraphs with sixth-grade readers: Aulls found that the paragraphs with topic sentences were recalled better by the students than were the paragraphs without the topic sentences. Similarly, Kieras (1978) presented participants with a series of paragraphs that either followed or violated the convention that the topic sentence should appear first in the paragraph. The results suggested that violation of the topic-sentence-first paradigm increased the readers' processing load, presumably because the readers needed to hold more information in immediate memory. Kieras (1978) concluded that the role of the paragraph and the initial topic sentence was to minimize memory load (cf. McNamara et al., 1996).

In other studies, the initial topic sentence position has been compared with other positions (e.g., Clements, 1979; Richards, 1975–1976), showing that the earlier the topic sentence was placed within the paragraph, the more likely it was that the participants would remember the text. Both studies led to the conclusion that topic sentences prepare or prime readers' memories, facilitating easier integration of subsequent information.

Individual differences must also be considered. Goldman et al. (1995) showed that topic sentences primarily affect readers who have little subject knowledge. Similarly, McNamara et al. (1996) manipulated the presence of topic sentences in a study that varied both local and global cohesion. Local cohesion manipulations included increasing overlap between sentences, reducing anaphor, and defining unfamiliar terms. Global manipulations included adding headers and topic sentences. The results indicated that adding headers and topic sentences benefited low-knowledge readers' text comprehension. Similarly, studies by both León and Carretero (1995) and Lorch and Lorch (1995) demonstrated that *headings* assist readers' mental organization of upcoming paragraphs, presumably by preparing the reader for the forthcoming information's topic, thus serving a role similar to that of topic sentences.

Topic sentences appear to have a critical capacity in facilitating readability. Numerous studies have shown that even minor topic shifts between sentences can be detrimental to readability, whereas the facilitative organization of topic sentences can be beneficial (e.g., Haviland & Clark, 1974; Kieras, 1981; Kintsch & van Dijk, 1983; Lesgold, Roth, & Curtis, 1979; McNamara et al., 1996; J. R. Miller & Kintsch, 1980). Accordingly, our goal in this study was to better understand topic sentences and their role in the paragraph.

Models for the Identification of Topic Sentencehood

The derived model. A prominent model of topic sentencehood identification assumes that a single sentence emerges as the topic sentence by consequence of its coreference with neighboring sentences and its communication of the theme (or *topicality*) of the paragraph. In this article, we will use the term *derived model* to refer to this theory. The derived model can trace its history through philosophers as diverse as Aristotle (trans. 1954), Wittgenstein (1953), and Toulmin (1969), through to discourse analysts as diverse as Hoey (1991), Halliday and Hassan (1976), and Kintsch (2002). The argument is that the role of sentences emerges from the interrelationship of smaller sections of text. More important, the argument implies that the role of these sentences cannot be sufficiently understood or identified without incorporating the information provided by the surrounding sentences.

In Toulmin's (1969) model, the interrelationship of sentences results in three primary elements: the *claim*, the *evidence*, and the *warrant*. The claim (essentially, the topic sentence) is supported by evidential sentences. The warrant (a relatively rare feature) shows the relevance of the evidence in light of the claim and will occur near or at the end of the paragraph. The point made by Toulmin, therefore, was that elements of texts serve distinct functions that become meaningful if understood in context. These relationships serve no function without the context of the paragraph.

Empirical research (e.g., Graesser et al., 2003; Kieras, 1978) suggests that topic sentences occur in the paragraph-initial position and are supported by sentences that supply additional detail relevant to the stated topic sentence. Such findings support the work of Toulmin (1969), inasmuch as the argument is that sentences fulfill roles in text but that those roles are realized through an interdependency that serves no function outside of the paragraph context.

Such an interdependent focus has led computational linguists to seek out topicality through the comparison of one element of a text with another (e.g., Kintsch, 2002; Rosch & Mervis, 1975; Sardinha, 2001; cf. Olney & Cai, 2005). Sardinha, for example, adopts Hoey's (1991) model of cohesion to locate topically related segments. More closely related to the topic sentence and its computationally identifiable role in the paragraph, however, is the concept of *family resemblance evaluations* (Rosch & Mervis, 1975). This concept was developed by Kintsch into a computational approach that we will describe in this article as *section topicality values*. This measure, and its theoretical underpinning, will be discussed below.

If elaborative sentences are thematically related to topic sentences, we can expect a high degree of semantic overlap between the two kinds of sentences, and semantic overlap can be evaluated with latent semantic analysis (LSA; Landauer & Dumais, 1997; Landauer, McNamara, Dennis, & Kintsch, 2007). LSA is a method through which the conceptual overlap between textual sections (e.g., sentences, paragraphs, or whole texts) can be evaluated. LSA evaluations are based on co-occurrences of lexical items across a large corpus of similarly themed texts. The principle underlying LSA is that a word's meaning can be approximated by the words with which it tends to co-occur. Thus, the word *table* tends to co-occur more often with words such as *chair*, *cup*, and *wood* than it does with words such as *car*, *octopus*, or *friend*. LSA generates a cosine value, usually between 0 and 1, for each textual comparison. A value closer to 1 indicates greater semantic similarity.

The ability of LSA to assess the relative similarity between texts makes this measure a reasonable point of departure for estimating the topic sentencehood of a paragraph. Indeed, such a proposition was discussed by Kintsch (2002), who argued that we can derive a precise mathematical representation of the *topic* or *theme* of a text by using LSA: "LSA can be used to select from each [paragraph] the most typical and most important sentence" (p. 166). This representation can be achieved in two ways. First, each sentence can be compared with the complete section of text (e.g., a paragraph). This approach Kintsch terms the *section importance value*. Second, each sentence can be compared with every other sentence. This second approach (the section topicality value) relates to the Rosch and Mervis (1975) family resemblance evaluations mentioned above.

In discussing these two approaches to evaluating the theme of a paragraph, Kintsch (2002) argued that comparing longer sections with shorter sections (the sentence importance value) is not a good solution "because long sections necessarily are advantaged" (p. 164). The *advantage* that Kintsch mentions is that the longest sentence in any section will have more words in common with the paragraph as a whole and, as a result, generate a higher cosine. Instead, Kintsch advocates comparing each section (or sentence) of text with each other section (or sentence) of the text and computing the average cosine from the results to derive what Kintsch calls a "reasonable approximation" of the "most typical sentence in [a] section" (p. 166). Although Kintsch's concern over the section importance value seems reasonable, Kintsch does not seem equally concerned with the section topicality value, which is likely to face a problem similar to the one observed with the section importance value. That is, in any comparison of text sections, one section will tend to contain more words than will others and, consequently, these longer sections will have the inherent advantage of having more words in common with other sections. That is, there is a greater likelihood of overlap when there are more words in the sentence to provide overlap. The result, therefore, is that section topicality values (like section importance values) are likely to correlate with text length. Of course, since even very long sentences tend to be shorter than paragraphs, the section topicality value is likely to correlate with text length to a lesser degree than does the section importance value. Indeed, Kintsch demonstrated that section topicality value is not quite as influenced by sentence length as is the section importance value by showing that section importance value and length of text correlate at $r = .87$, whereas section topicality value and length correlate at $r = .39$. Despite this possible text length confound, in this study, we followed Kintsch and considered the section topicality value as an approximation of the derived model.

The free model. In this article, we introduce the free model of computational identification of topic sentences and nontopic sentences. The free model differs from the derived model in that the latter identifies topicality by the relationship of one sentence to other sentences in its textual vicinity, whereas the free model treats all sentences as independent of context (i.e., the role of a sentence does not substantially change, whether other sentences from the text are present or not). Specifically, the fundamental claim underlying the free model is that topic sentences and nontopic sentences can be identified by paragraph-independent features, such as *sentence length*, *connectives*, and *sentence initial phrase structure*.

To some degree, the free model can be related to *speech acts theory* (Austin, 1962; Searle, 1969). That is, a sentence is of a certain recognizable *type* that can be identified as that type, independently of a context. Of course, a sentence's role may vary depending on its context (pragmatics); for example, the sentence "Can you pass the salt?" may be generally thought of as *request sentences* but, under certain circumstances, could be a genuine question as to the recipient's ability. Such counterexamples, however, do not negate the claim that the *probable* role of *most* sentences does not rely on context.

In this study, we compared topic sentences and nontopic sentences with respect to an array of sentence features. Given the function of topic sentences and nontopic sentences, a variety of sentence features emerged as probable sentence type identifiers. Some of these features would tend to be more prominent in topic sentences, whereas others would be more indicative of nontopic sentences. For example, we predicted that *adjectives* would be more likely to appear in topic sentences because sentences that make claims (i.e., topic sentences) often do so with the support of adjectives (e.g., "Exams are *stupid*"; "The proposal was both *original* and *creative*"; "The investigation reported *surprising* findings"). We also predicted that because topic sentences *make a claim* rather than *explain a claim*, topic sentences would, consequently, tend to be shorter in length than nontopic sentences. Furthermore, we expected that the number of words before a main verb would be likely to be higher in topic sentences, because new claims (i.e., topic sentences) often begin with a bridging contrast from a previous claim.

In contrast, we hypothesized that the presence of *pronouns* would be more likely to be a feature of nontopic sentences. That is, pronouns are used to refer anaphorically to concepts that have been stated beforehand. Thus, pronouns are more likely to occur in sentences other than the first sentence of a paragraph (and therefore, presumably, the topic sentence). And just as we predicted that topic sentences would tend to be shorter because they make claims rather than explain them, we expected that nontopic sentences were likely to feature more *connectives*, such as *because*, *but*, and *in order to*, which serve to link textual ideas and provide fuller explanations to a reading audience.

Experimental Design

To test the performance of the two models of topic sentence identification (the derived model and the free model), we performed four experiments. The first experiment was designed to establish whether coreferential indices (e.g., LSA) tend to identify paragraph-initial sentences. That is, the literature on paragraphs overwhelmingly argues that topic sentences appear at or near the beginning of paragraphs (e.g., Graesser et al., 2003; Kieras, 1978) and in approximately 50% of paragraphs (e.g., Popken, 1987, 1988, 1991a, 1991b). Thus, if coreferential indices tend also to identify paragraph-initial sentences, we would have substantial evidence that the derived model successfully identifies topic sentences. More precisely, such an outcome would lend support to the claim of Kintsch (2002) that overlap indices produce a "reasonable approximation" of the "most typical sentence in [a] section" (p. 166). On the other hand, if the coreferential indices do not appear to identify any single sentence position (because topic sentences do not have to be the first sentence in a paragraph), a deeper analysis becomes necessary.

In the second experiment, we used a corpus of 800 independently identified and published topic sentences and nontopic sentences to test the hypothesis that expert human raters can distinguish topic sentences from nontopic sentences in isolation, without the use of context from surrounding sentences. This experiment thus tested the free model by examining whether topic sentences and nontopic sentences exhibit independent features that allow humans to identify their sentential role without context. We included experts in our experiment because we sought to train our computational measures on the basis of the evidence and data supplied by experts. That is, we can claim the efficacy of a computational approach as a function of its ability to produce results that are comparable to those of trained human judges.

In Experiment 2, we also examined whether the free model sentential features, including *sentence length*, *incidence of connectives*, and *pronoun frequency*, were sufficient for a computational distinction between topic sentences and nontopic sentences. If a computational model using these variables successfully identifies the two sentence types comparably to the human raters, the free model will be able to assess text and evaluate topic sentencehood. Such an algorithm will provide writers and publishers with a better approximation of the cohesion of a particular text.

In the third experiment, we used a corpus that was a large subset of that used in Experiment 1. These were natural paragraphs that might or might not contain topic sentences. We asked the same expert human raters as those in Experiment 2 to rate the sentences in each paragraph for topic sentencehood. Thus, in contrast to Experiment 2, this time the raters evaluated sentences *with* context, rather than isolated sentences. In other words, the human raters judged sentences not only for their individual features, but also for how the sentences functioned in the context of the paragraph.

In Experiment 4, the expert raters were asked to evaluate sentences from a subset of the paragraphs from Experiment 3, along with a new set of sentences. Experiment 4 differed from Experiment 3 because, in Experiment 4, the experts rated the sentences *without* context. The same sentences were presented in isolation, and not in the context of the paragraph. Experiment 4 was conducted to establish whether the experts would rate topic sentencehood similarly for sentences without context and for those with context, such as those in Experiment 3.

EXPERIMENT 1

Along with *narrative* texts, the corpus for Experiment 1 included a large number of expository texts from the domains of *history* and *science*. Therefore, we assumed that these expository texts would feature a large number of topic sentences.[1] Topic sentences tend to occur in the sentence-initial position. Therefore, we predicted that the derived model of topic sentencehood identification (i.e., the section topicality values) would produce higher values for the first sentence of paragraphs. We further predicted that because history texts are expository in nature, the topic sentencehood evaluations of these texts would more closely match the results of the science texts, as compared with the narrative texts (Lightman, McCarthy, Dufty, & McNamara, 2007a; McCarthy, Graesser, & McNamara, 2006; cf. Duran, McCarthy, Graesser, & McNamara, 2007).

Method

The corpus in our analysis contained 150 academic texts and was compiled by Duran et al. (2007). Each text was approximately 400 words in length, with the aggregated texts comprising a total of 1,205 paragraphs (3,116 sentences). The strength of the corpus lay in the rigorous constraints applied to text selection. Thousands of 400-word paragraph-to-paragraph slices were randomly sampled from 27 published textbooks provided by the MetaMetrics repository of electronic duplicates. The textbooks covered a range of domains (i.e., science, history, and narrative), and within each domain there were multiple grade levels, including 7th–9th grade (i.e., junior high) and 10th–12th grade (i.e., high school). An automated process, selecting from this comprehensive source, allowed a large, unbiased representation of topics. However, because the text samples were removed from the overall context, human raters had to evaluate each text to ensure *self-contained* topic continuity. Texts that did not meet these criteria were discarded. Human raters also processed the texts for typographical and content error. For example, the electronic duplicates were void of the original graphics (e.g., maps and figures) but retained the original captions to the graphics. This discrepant material was removed. In addition, some texts contained poorly formatted sentence and paragraph breaks. If there was uncertainty on how to demarcate a sentence or paragraph, the text was discarded.

After the extensive "cleaning" of the corpus, Duran et al. (2007) selected texts to maximize the uniform representation of domain (i.e., science, history, and narrative) and grade level (i.e., junior high and high school). Within each representative domain, 25 texts from the junior high grades and 25 texts from the high school grades were sampled. Within each grade level, three or more unique textbooks were sampled. The latter constraint ensured variability across authorship.

We next separated from these texts all paragraphs that contained between three and five sentences. Because most definitions of a paragraph with a topic sentence state that the topic sentence (singular) is supported by evidential sentences (plural), it was reasonable to assume that candidate paragraph sentences contained at least three sentences in total. As such, one- and two-sentence paragraphs were excluded, since they did not constitute a paragraph according to this definition. Because paragraphs containing six sentences for narratives amounted to only 3% of the total corpus, we used only three-, four-, and five-sentence paragraphs from the corpus. In total, therefore, our corpus contained 403 paragraphs, or 33% of the total paragraphs in the corpus.

To obtain the *section topicality values* for these paragraphs, we used two coreference measures: LSA and lemma overlap. These indices are available through Coh-Metrix (Graesser et al., 2004), and although both are measures of coreference, they differ in terms of sophistication. For example, lemma overlap assesses sentence pairs that share common roots (e.g., *table*/*table* or *table*/*tables*). In contrast, LSA assesses conceptual overlap between sentences so that the degree of similarity between two words can be approximated. Thus, for this sophisticated coreference index, *table*/*tables* would be judged as more similar than *table*/*chair*, which, in turn, would be judged more similar than *table*/*octopus*.

Results

The results for Experiment 1 indicated that there was an approximately equal distribution of paragraph serial positions attaining the highest coreference values (see Tables 1, 2, and 3). That is, no sentence position for any of the three domains, for any of the paragraph lengths, appeared dominant. To confirm this appearance, we conducted a χ^2 analysis. The results suggested that no serial paragraph position garnered a greater frequency of higher coreference evaluations.

If we assume that topic sentences typically appear in paragraph-initial positions, Experiment 1 provides no evidence to suggest that coreferential indices (including LSA) provide an effective method for identifying topic sentencehood. We posit two possible causes for this outcome. The first explanation is that topic sentences are *not* generally in sentence-initial positions. Such an explanation would run counter to over a century of empirical and theoretical research; however, in Experiments 3 and 4 we examined this possibility by having expert raters code the sentences from the corpus used in Experiment 1 for topic sentencehood. If the raters were *not* more likely to identify paragraph-initial sentences as topic sentences, the possibility that the derived model can identify topic sentences remains. If, on the other hand, the results indicated that the raters were more likely to identify paragraph-initial sentences as topic sentences, this would raise doubts as to the validity of the derived model.

Table 1
Quantity of Sentence Positions With Highest Final Mean Cosine Values for Three-Sentence Narrative, History, and Science Paragraphs

					Cosine	
Domain	Measure	Sentence	Quantity	Percentage	*M*	*SD*
Narrative	LSA	1	10	28.571	0.314	0.170
		2	13	37.143	0.453	0.209
		3	12	34.286	0.397	0.153
	Lemma	1	6	17.143	0.253	0.127
		2	13	37.143	0.315	0.088
		3	16	45.714	0.197	0.125
History	LSA	1	17	30.357	0.535	0.199
		2	25	44.643	0.547	0.200
		3	14	25.000	0.601	0.168
	Lemma	1	16	28.571	0.220	0.088
		2	21	37.500	0.273	0.117
		3	18	32.143	0.277	0.100
Science	LSA	1	14	21.875	0.588	0.206
		2	24	37.500	0.608	0.185
		3	26	40.625	0.572	0.195
	Lemma	1	16	25.000	0.282	0.133
		2	25	39.063	0.341	0.139
		3	23	35.938	0.314	0.115

Note—LSA, latent semantic analysis.

Table 2
Quantity of Sentence Positions With Highest Final Mean Cosine Values for Four-Sentence Narrative, History, and Science Paragraphs

					Cosine	
Domain	Measure	Sentence	Quantity	Percentage	*M*	*SD*
Narrative	LSA	1	3	13.043	0.292	0.228
		2	4	17.391	0.350	0.208
		3	8	34.783	0.364	0.210
		4	8	34.783	0.411	0.226
	Lemma	1	4	17.391	0.183	0.008
		2	3	13.043	0.273	0.076
		3	8	34.783	0.219	0.072
		4	8	34.783	0.200	0.105
History	LSA	1	17	26.563	0.535	0.201
		2	15	23.438	0.517	0.218
		3	15	23.438	0.435	0.209
		4	17	26.563	0.516	0.190
	Lemma	1	15	23.438	0.264	0.054
		2	13	20.313	0.278	0.100
		3	23	35.938	0.201	0.075
		4	13	20.313	0.212	0.082
Science	LSA	1	15	24.590	0.564	0.142
		2	16	26.230	0.563	0.192
		3	17	27.869	0.605	0.231
		4	13	21.311	0.597	0.166
	Lemma	1	17	27.869	0.307	0.090
		2	13	21.311	0.260	0.121
		3	12	19.672	0.337	0.159
		4	19	31.148	0.267	0.101

Note—LSA, latent semantic analysis.

Table 3
Quantity of Sentence Positions With Highest Final Mean Cosine Values for Five-Sentence Narrative, History, and Science Paragraphs

					Cosine	
Domain	Measure	Sentence	Quantity	Percentage	*M*	*SD*
Narrative	LSA	1	7	22.581	0.399	0.139
		2	6	19.355	0.515	0.114
		3	7	22.581	0.483	0.093
		4	4	12.903	0.450	0.063
		5	7	22.581	0.464	0.170
	Lemma	1	5	16.129	0.211	0.117
		2	5	16.129	0.294	0.118
		3	3	9.677	0.316	0.091
		4	8	25.806	0.256	0.048
		5	10	32.258	0.238	0.082
History	LSA	1	10	29.412	0.562	0.191
		2	3	8.824	0.674	0.082
		3	7	20.588	0.499	0.171
		4	7	20.588	0.492	0.113
		5	6	17.647	0.591	0.132
	Lemma	1	5	14.706	0.209	0.099
		2	10	29.412	0.261	0.059
		3	3	8.824	0.250	0.132
		4	9	26.471	0.294	0.072
		5	7	20.588	0.239	0.096
Science	LSA	1	10	28.571	0.501	0.203
		2	4	11.429	0.477	0.199
		3	11	31.429	0.553	0.134
		4	2	5.714	0.525	0.092
		5	8	22.857	0.536	0.248
	Lemma	1	9	25.714	0.233	0.113
		2	6	17.143	0.272	0.072
		3	9	25.714	0.280	0.060
		4	7	20.000	0.257	0.078
		5	4	11.429	0.264	0.141

Note—LSA, latent semantic analysis.

Table 4
Correlations of Section Topicality Values With Sentence Length for Lemma Overlap and Latent Semantic Analysis (LSA)

Domain	Length	Lemma	LSA
Science	3	.200	**.092**
	4	.269	.244
	5	.441	.583
History	3	.449	**.163**
	4	.236	.258
	5	.427	.445
Narrative	3	.412	.333
	4	.244	.300
	5	.422	.270

Note—All correlations except those in boldface are significant at $p < .01$.

Another plausible cause for the lack of coreferential identification of probable topic sentences in Experiment 1 stems from the widely acknowledged confound that overlap indices tend to be overly influenced by text length (Dennis, 2007; McCarthy, Rus, et al., 2007; McNamara, Ozuru, Graesser, & Louwerse, 2006; Penumatsa et al., 2004; Rehder et al., 1998). That is, when longer texts (in this case, sentences) are compared, there is greater likelihood that the same (or similar) words will be identified. Thus, longer sentences are more likely to generate higher coreference values, which, in this case, means that they are more likely to be identified as the topic sentence. Indeed, as was mentioned above, results produced in Kintsch (2002) showed that the section topicality value correlated with text length.[2] If this second explanation was a factor in Experiment 1, we would expect similar correlations between sentence length and coreference indices. And indeed, this prediction was confirmed (see Table 4).

Conclusion

In Experiment 1, we tested the coreference model of topic sentencehood identification, using two coreferential indices across a corpus of three-, four-, and five-sentence 7th–12th grade school texts taken from the domains of history, narrative, and science. The results suggested that no particular sentence was more likely to produce a higher coreference value as a function of serial position in the paragraph. Further analyses suggested that the measures had a tendency to favor longer sentences within the text. If we assume that topic sentences tend to appear in paragraph-initial positions, the experiment yielded no evidence to support the claim that the derived model identifies topic sentences.

EXPERIMENT 2

The primary purpose of Experiment 2 was to test the free model hypothesis—namely, that expert raters can reliably distinguish topic sentences from nontopic sentences without the benefit of the presence of context.

Method

For Experiment 2, we collected a corpus of over 400 independently identified topic sentences. These sentences were all published in textbooks and on Web sites and were identified by other researchers and writers as being prototypical examples of topic sentences (see the Appendix). When the topic sentences we collected were identified within a paragraph, we used the co-occurring sentences that were not identified as topic sentences as our examples of nontopic sentences.

Three experts, all with first-author publications in the field of discourse processing, were trained to distinguish topic sentences from nontopic sentences. Training involved reading examples of topic sentences and nontopic sentences, as well as completing various exercises, such as choosing the most appropriate sentence for a paragraph from a selection of candidate sentences. The exercises for training were taken from the books and Web sites identified in the Appendix. Care was taken to ensure that the examples given did not appear in the experimental evaluations. In addition, the experts were not given insight into the measures and approaches used to computationally approximate topic sentencehood. Upon completion of training, the experts were asked to assign values of 1 through 6 for each sentence in our prepared corpus. Raters were informed that a score of 1–3 indicated that the sentence was a *nontopic* sentence, whereas a score of 4–6 indicated that the sentence was a *topic* sentence. The ranges between the scores (1–3 and 4–6) reflected the confidence of the raters, so that a score of 3 would indicate a nontopic sentence with *little confidence*, whereas a score of 6 would indicate a topic sentence with *maximum confidence*. Our reasoning for using such a scale was that we required both a binary distinction (topic sentence/nontopic sentence) and also a flexible, continuous scale that would facilitate computational training (see Experiment 3).

Results

A programming error resulted in three evaluations from each rater being lost. Consequently, a total of 791 rated

Table 5
Pearson, Spearman, and Similarity Correlations for 3 Human Raters and a Gold Standard

	Rater 2			Rater 3			Rater Gold		
	Pearson	Spearman	Similarity	Pearson	Spearman	Similarity	Pearson	Spearman	Similarity
Rater 1	.591	.589	.751	.661	.662	.814	.883	.882	.985
Rater 2				.525	.523	.694	.844	.836	.967
Rater 3							.830	.826	.980

Note—All correlations are significant at $p < .001$.

responses were assessed. Since the Likert-like scale used for the assessments might be viewed as either an interval or an ordinal scale, to calculate the reliability of the raters, our results show Pearson correlations supported by Spearman nonparametric correlations (Table 5). One further, *similarity correlation* (Sanz, 2005) is also provided. The similarity correlation is a proportion score based on assessing all judgment ratings within ±1 point as being a *hit* and all other ratings as being a *miss*. A *rater gold* value has also been added to the correlation table, formed from the means for the 3 expert raters. As can be seen from Table 5, there was significant interrater reliability.[3]

It is not a simple task to interpret the strength of interrater correlations. A range of Pearson correlations between $r = .525$ and $r = .661$ might be described as moderate (Shrout, 1998); however, as Hatch and Lazaraton (1991) reminded us, "The more raters, the more we must trust the ratings and so the more raters, the higher the reliability" (p. 533). Hatch and Lazaraton supplied the following formula for converting multiple raters' correlations into a single effective gold interrater value:

$$R_{\mathrm{tt}} = \frac{{}^{nr}AB}{1 + (n-1)\,{}^{r}AB}.$$

In this formula, R_{tt} is the reliability of all the judges, n corresponds to the number of raters, and ${}^{r}AB$ is the average correlation across the raters. Thus, the effective interrater reliability for the Pearson correlations given in Table 5 is $r = .813$: a high agreement, and one in line with the value reflecting the similarity correlations.

We also argue that when such interrater reliability data are interpreted, the task at hand needs to be taken into consideration. Our own interpretation of the correlations is that they should be taken at face value; that is, even when sentences are viewed with no other paragraph context present, there is significant agreement between experts as to *what is* and *what is not* a topic sentence. However, to further establish the validity of the raters' assessments, a further series of analyses was conducted.

First, to assess the accuracy of the expert raters in terms of percentage of correct judgments, we marked all sentence evaluations of <3.5 (the midway point) as a nontopic sentence, and all ratings of ≥3.5 as a topic sentence. As was described in the Method section, the raters were instructed to consider evaluations of 1–3 as nontopic sentences and ratings of 4–6 as topic sentences; thus, a rating of 3.5 from the mean of the ratings was judged to be the most appropriate cutoff point. The results ranged from the lowest at 73% accuracy (Rater 2) to the highest at 77% accuracy (Rater 1). The rater gold accuracy was 78% (see Table 6). For contrast, Table 6 also shows the ratings for judging nontopic sentences as <2.5 and <4.5. Adjusting the division of topic sentence and nontopic sentence did not improve the rater accuracy. These results offer support that the expert ratings of 1–3 for nontopic sentences and 4–6 for topic sentences are reliable.

To further establish the reliability of the expert raters, χ^2 values for rater gold were also recorded. As is shown in Table 7, setting the cutoff point at 3.5 offers the most accurate results overall. Table 7 also shows that rater gold is a better judge of topic sentence (84% accuracy) than it is of nontopic sentence (73% accuracy). The lower accuracy suggests that nontopic sentences resemble topic sentences more often than vice versa. This result is not surprising if we consider that warrant sentences (Toulmin, 1969), although relatively rare, are briefly stated claims and, therefore, more likely to resemble the topic sentence type. Assessing the accuracy of the raters' judgments in terms of recall, precision, F_1, and d' (see Table 8), we again see that the accuracy of the raters is high.

The raters were required to rate sentences on a Likert-like scale of 1–6, in which values of 3 and 4 indicated *lower confidence*. Therefore, we also wanted to access the degree of accuracy of human raters, relative to this confidence (see Table 9). The issue of rater confidence is important because both Kavanaugh (1989) and Fox, Bizman, Hoffman, and Oren (1995) found that rater confidence is positively correlated with rater accuracy, so that high and low ratings tend to be more accurate than ratings in the midrange (i.e., those around the 3–4 point area in this study).

Thus, to better assess the reliability of the raters, we adjusted the lower and upper bounds of the *topic sentence/*

Table 6
Accuracy of Raters in Predicting
Topic and Nontopic Sentences (NTSs)

NTS Midpoint	Rater 1	Rater 2	Rater 3	Rater Gold
<2.5	.719	.709	.630	.692
<3.5	**.770**	**.728**	**.736**	**.781**
<4.5	.709	.594	.692	.679

Table 7
Accuracy for Predicting Topic Sentences (TSs)
and Nontopic Sentences (NTSs)

NTS Midpoint	NTS Accuracy	TS Accuracy	χ^2	Significance
.25	.439	.942	154.130	<.001
.35	**.726**	**.836**	**253.177**	**<.001**
.45	.919	.441	132.340	<.001

Table 8
Recall, Precision, and F_1 Values for Topic Sentences and Nontopic Sentences

Sentences	Hits	Misses	False Alarms	Recall	Precision	F_1
Topic	332	65	108	.836	.755	0.793
Nontopic	286	108	65	.726	.815	0.768

Note—$d' = 1.58$.

Table 9
Accuracy and Confidence Ratings for Gold Standard

Lower Bound	Upper Bound	Records	Percent of Records	Accuracy
1.5	5.5	59	.075	.932
2.5	5.5	212	.268	.882
0.5	5.5	16	.020	.875
1.5	4.5	250	.316	.864
2.5	4.5	403	.509	.864
0.5	4.5	207	.262	.845
3.5	4.5	558	.705	.826
3.5	5.5	367	.464	.817
2.5	3.5	636	.804	.794
2.5	3.5	636	.804	.794
3.5	**3.5**	**791**	**1.000**	**.781**
1.5	3.5	483	.611	.772
0.5	3.5	440	.556	.755

nontopic sentence division so that an intermediate category of *not sure* emerged. For example, when the lower threshold for nontopic sentences is 1.5 and the upper threshold for topic sentences is 5.5, all rater gold values between 1.5 and 5.5 are ignored. Obviously, when the division between lower and upper bounds increases, the amount of data assessed decreases. As such, there is a trade-off between accuracy and amount of data considered. Table 9 shows that the greatest accuracy for rater gold is 93%. However, to obtain this level of accuracy, only 7% of the data are evaluated. The division point of 3.5 for topic sentences, therefore, achieves a high degree of accuracy, relative to other divisions, without having to ignore any data. The results suggest that the raters' evaluations in the middle of the scale are reasonably accurate.

The results of Experiment 2 demonstrate that raters can reliably distinguish topic sentences from nontopic sentences without the benefit of context. Such a result offers support to the free model. However, a major goal of this study was to create a computational measure that approximated the free model. In Experiment 1, we used two coreference indices to create a measure called the *section topicality value*. This measure was used to approximate the derived model. In our next analysis, we sought to establish whether a computational measure approximating the free model would identify topic sentences with a level of accuracy comparable to that of our expert raters.

EXPERIMENT 2A

The objective of Experiment 2A was to create and test an algorithm that could distinguish topic sentences from nontopic sentences. To achieve this goal, we used the corpus from Experiment 2 and 15 predictor variables generated from Coh-Metrix. Coh-Metrix variables are based on over 30 years of psychological and computational linguistic theory (Graesser et al., 2004). Coh-Metrix variables have been validated (e.g., Duran et al., 2007; Hempelmann et al., 2005; McNamara et al., 2006), and they have formed the basis of numerous text analysis studies (e.g., Crossley, Louwerse, McCarthy, & McNamara, 2007; Hall, McCarthy, Lewis, Lee, & McNamara, 2007; Lightman, McCarthy, Dufty, & McNamara, 2007b; McCarthy, Graesser, & McNamara, 2006; McCarthy, Rus, et al., 2007). To assess the accuracy of the predictor variables, we used discriminant analysis and followed procedures similar to those in earlier Coh-Metrix studies (e.g., Hall et al., 2007; McCarthy, Lehenbauer, et al., 2007; McCarthy, Lewis, Dufty, & McNamara, 2006).

Free Model Predictor Variables

As was mentioned earlier, Coh-Metrix provides over 600 indices of cohesion, difficulty, and language. Many of these variables approximate the values of textual features that we hypothesized would be predictive of either topic sentencehood or nontopic sentencehood. In total, we selected 15 such variables. We will provide descriptions of these variables and our reasons for their selection below.

As our first predictor of topic sentencehood, we selected *adjectives incidence*. Since topic sentences make claims, we hypothesized that adjectives would tend to be used to support those claims. Indeed, in Lorch, Lorch, and Matthews (1985), 75% of the topic sentences provided in their Appendix A contained adjectives (e.g., "The geography of Morinthia is particularly *rugged*"; "Culatta's geography is quite *ordinary*"; "Morinthia has a *strong, democratic* government"). We next selected the variable *number of words before a main verb*. Topic sentences often begin with a bridging contrast from a previous claim (a transition). Thus, we hypothesized that topic sentences would tend to contain a greater number of pre-main-verb lexical items. Our third predictor of topic sentencehood was the *hypernymy* variable. The hypernym value refers to the number of levels that a word has in a conceptual, taxonomic hierarchy: the higher the number, the greater the hypernymy. A low hypernym value indicates word abstractness, because the word has few distinctive features. Because topic sentences make general claims (rather than providing specific evidence), we hypothesized that higher values would be indicative of topic sentences. The fourth predictor was the *polysemy* variable. The polysemy value reflects the number of WordNet (G. A. Miller, 1995) synonym sets assigned to any given word. Polysemy scores provide information concerning a word's potential ambiguity. Because topic sentences may provide greater generality, the terms used may be vaguer and, therefore, more ambiguous. We predicted higher values for cases of topic sentence. The fifth predictor was the *frequency* variable. Once again, topic sentences make broad claims, and therefore, we predicted higher average frequency values for the topic sentence case. The sixth topic sentence predictor we selected was *existential there incidence*. We predicted that topic sentences would be more likely to contain a general claim in the form of *there is*/*there are* than would nontopic sentences. Our final predictor for topic sentencehood was *incidence of third-person singular grammatical form*. We predicted that this variable would be a good indicator of topic sentencehood because the construct occurs only in the present tense. Since science texts tend to be written in present tense and narratives tend to be written in the past tense, and since science texts are expository and, therefore, more likely to have topic sentences, we predicted that the third-person singular variable was likely to have a higher incidence in the topic sentence condition.

We selected eight variables as predictors of nontopic sentences. First, we selected *pronoun incidence*. Pronouns tend to refer to previously mentioned concepts. As such, pronouns are more likely to occur later in the paragraph than earlier. Thus, they are less likely to be in the first (or topic) sentence. Next, since nontopic sentences explain previously made claims, such sentences need to show how discrete concepts are interrelated. To achieve this, writers use *coordinating conjunctions*, such as *and*, *but*, and *because*, and *connectives*, such as *in order to*, *due to*, and *in addition to*. Thus, we predicted greater frequencies of these two variables in the nontopic sentence cases. The *clarifications incidence* variable was our fourth choice for predicting nontopic sentencehood. Clarifications, such as *for example*, tend to follow claims, rather than being a part of a claim. Our fifth predictor variable of nontopic sentencehood was the *concreteness* index. The concreteness variable is provided through the MRC database (Coltheart, 1981) and approximates values for words that are nonabstract. Since nontopic sentences must explain general claims, we hypothesized that nontopic sentences would feature greater values of highly concrete words. The sixth predictor of nontopic sentencehood was *sentence length*. Because nontopic sentences explain ideas (often with the use of connectives and multiple clauses), they are predicted to require more words. Thus, we predicted that longer sentences would be indicative of nontopic sentences. Our seventh variable for nontopic sentencehood was *cardinal number incidence*. Again, numbers are specifics, rather than general claims, and therefore, we predicted that they would have greater presence in nontopic sentences. Finally, we selected *incidence of past tense endings*. Past tense is predominantly used in narratives, and narratives are less likely to have topic sentences.

It should be noted that the predictions made above are derived largely from the theory of the *function* of topic sentences, rather than from the *form* of topic sentences. That is, although the literature has much to offer as to what a topic sentence *should be* or *should do*, there is little advice as to how to actually *form* the sentence in order to manifest these functions. As such, some of our predictions are derived from researchers' hypotheses regarding how such functions would be represented formally. For instance, Eden and Mitchell (1986) claimed that readers have expectations that emerge with a paragraph's opening statement and that these expectations include demonstrating how the paragraph connects with what has previously been stated. At the same time, Eden and Mitchell also viewed the opening sentence as an "instruction" (p. 418). Both points are reasonable from a functional point of view; however, formally, a connection to *what has previously been stated* suggests that opening sentences may feature transitional phrases (hence, our prediction of *number of words before the main clause*), whereas an *instruction* suggests a brief command, rather than an elaborative explanation (hence, our prediction of fewer words for topic sentences). Obviously, where transitions are present in opening sentences, such sentences are unlikely to be among the shortest in this text.

Method

The corpus from Experiment 2 was randomly divided into a *training set* ($n = 396$) and a *test set* ($n = 395$). Using the training set, we conducted an ANOVA to establish the reliability of the 15 predictor variables (see Table 10). Three of the variables (incidence of clarifications, polysemy, and frequency) were not reliable ($p > .100$) and were subsequently excluded from the analysis. All variables that were at least marginally significant ($p < .100$) were retained and used as predictors (independent variables) in the discriminant analysis. The purpose of the discriminant analysis was to create a computational model to be representative and comparable to the free model of topic sentence identification.

Results

A discriminant analysis was conducted with sentence type (topic sentence/nontopic sentence) as the dependent

Table 10
Results for ANOVA Conducted on Training Set ($n = 396$), Showing the 12 Variables With Differences Between Nontopic Sentences (NTSs) and Topic Sentences (TSs) at the $p < .1$ Level

	NTS		TS				
Variable	*M*	*SD*	*M*	*SD*	*F*	Significance	η^2
Meaningfulness	232.09	54.60	199.78	57.60	32.72	<.001	.08
Pronouns	15.56*	33.71	40.24*	52.01	31.32	<.001	.07
Word hypernym	1.44**	0.46	1.68**	0.51	23.79	<.001	.06
Third-person verbs	18.78*	35.27	41.76*	56.72	23.43	<.001	.06
Conjunctions	39.17*	43.97	20.54*	36.40	21.08	<.001	.05
Connectives	74.85*	63.37	47.55*	56.39	20.51	<.001	.05
Words per sentence	17.43*	8.86	14.06*	6.41	18.90	<.001	.05
Past tense	20.99*	44.06	11.65*	32.34	5.78	.017	.01
Adjectives	92.41*	78.32	110.29	81.92	4.93	.027	.01
Existential *there*	3.06**	14.77	7.07**	25.31	3.70	.055	.01
Cardinal numbers	13.68*	43.27	7.33**	24.50	3.23	.073	.01
Words before verb	5.58	4.20	6.30	4.39	2.80	.095	.007

Table 11
Accuracy of Discriminant Analysis for Distinguishing Topic Sentences (TSs) and Nontopic Sentences (NTSs) Across the Whole Corpus (All) and the Test Set (Test)

	Recall			Precision		
Sets	NTS	TS	Mean	NTS	TS	Mean
All	.726	.695	.711	.704	.718	.711
Test	.730	.673	.701	.677	.726	.702

variable. A total of 12 indices, identified as at least marginally significant in the aforementioned ANOVA, were used as predictor variables. The results of the discriminant analysis suggested an accuracy of .701 for recall and .702 for precision (see Table 11). To obtain these results, we conducted a series of five discriminant analyses, using the random split-half groups' method (Witten & Frank, 2005). On each occasion, half the group was used to generate the discriminant functions, and the other half (test set) was used to calculate the accuracy of the model. The results in Table 11 represent the average across all five analyses and show results both for the corpus as a whole and for the test set alone.

The results of the discriminant analysis are comparable to those of the expert raters, which ranged in accuracy from 73% to 78%. We suspected that the lower accuracies in the case of the computational model were the result of indices that were not completely representative of our theoretical position on the free model. For example, the pronoun density index included first-person singular pronouns (e.g., *I*/*we*), whereas third-person pronouns (e.g., *it*) may have been more indicative. We can also speculate that key words, indicative of register, may have assisted the experts in their evaluations of the sentences, whereas no attempt was made in our model to select key terms. Despite these differences, the computational free model makes a reasonable approximation of the expert evaluations.

Conclusions

In Experiment 2, three experts in the field of discourse processing assessed a corpus of topic and nontopic sentences. The experts' agreement was reliable, and a golden mean value created from the experts' ratings recorded 78% accuracy for identifying the independently assessed topic sentences. Using 12 Coh-Metrix sentence-level variables, we then created an approximation of the free model of topic sentencehood identification. The results for the model were significant and comparable to the human ratings. The results of Experiment 2 and Experiment 2A offer support for the free model of topic sentencehood identification.

EXPERIMENT 3

In Experiment 1, we showed that coreference indices (and the section topicality value that was formed from them) did not provide a solution to the problem of computationally identifying topic sentencehood. That is, overlap coreference indices such as LSA did not show that paragraph-initial sentences featured significantly higher values than did any other sentence positions. In Experiment 2, we examined whether topic sentences themselves consisted of internal features that expert raters and a computational system, such as Coh-Metrix, could identify. The results suggested that both expert raters (78%) and Coh-Metrix (70%) could, indeed, accurately distinguish the two sentence types. The results of Experiment 2 offered evidence regarding the accuracy of the expert raters—in terms of both interrater reliability and accuracy in identifying independently observed topic sentences. As such, in Experiment 3, we asked the same raters to evaluate a subsection of the paragraphs used in Experiment 1. The purpose of such evaluations was to better establish whether overlap indices (such as LSA) or a free model algorithm (as generated in Experiment 2) better identifies expert judgments of topic sentencehood in published texts.

Method

For the corpus of Experiment 3, we used a subset of the paragraphs used in Experiment 1. Specifically, our goal was to select a representative and equal sample of all types of paragraphs used in Experiment 1. Because there were only 23 examples of three-sentence paragraphs from the narrative domain, we selected 23 samples from each paragraph size of each domain. Thus, each domain (history, narrative, and science) was represented by 23 paragraphs for the paragraph sizes of three, four, and five sentences. The expert raters for Experiment 3 were the same as those used in Experiment 2. However, the raters were given a review of their previous training, because several weeks had passed since Experiment 2. For Experiment 3, we also had to slightly alter the sentence evaluation method. Specifically, in Experiment 2, each rater was asked to score each sentence independently of any context. In Experiment 3, the raters were asked to score each sentence of each paragraph in full view of the entire paragraph. This difference was necessary because we wanted to evaluate sentences in context and, therefore, the context was necessarily provided to the raters.

It was also necessary to guard against the first sentences of each paragraph's having the privilege of being evaluated first. Therefore, the automated scoring program was modified so that the sentences for scoring were presented in a random order. As such, there were two boxes on the screen. In the top box, the raters were shown the entire paragraph for 30 sec. In the bottom box, the program then showed each sentence of the paragraph (in correct order) individually below the entire paragraph block. The order, as to which sentence would be evaluated in which order, was randomized. As such, Sentence 4 might be evaluated first, Sentence 2 second, and so forth; however, all the sentences were always in full view in the top box and always appeared in order. Once all the sentences had been scored, the raters were allowed to adjust their scoring for any sentence as they saw necessary.

For Experiment 3, we predicted that the expert raters would overwhelmingly identify the paragraph-initial sentence as the topic sentence, especially in the case of the two expository domains. In addition, if the algorithm for identifying topic sentencehood generated in Experiment 2 is reliable, we could further expect that the algorithm would identify the same topic sentences as those identified by the expert raters.

Results

Interrater reliability for Experiment 3 was comparable to that for Experiment 2 ($M = .600$, $SD = .054$). Pearson correlations ranged from .542 ($p < .001$) to .650 ($p < .001$). As was previously mentioned, we considered a mean topic sentence score of greater than 3.5 to be a topic sentence. Using this division, the results showed that of the total 207

Table 12
Incidence of Topic Sentences and Their Serial Position in the Paragraph (Sentence 1 vs. Other) As Evaluated by Experts Over 69 Paragraphs per Domain

	Topic Sentences		Sentence 1		Other	
Domain	No.	Proportion	No.	Proportion	No.	Proportion
History	51	.739	46	.667	5	.072
Narrative	12	.174	11	.159	1	.014
Science	51	.739	49	.710	2	.029
All	114	.551	106	.512	8	.039

paragraphs evaluated, 114 (55%) were deemed to contain a topic sentence. This percentage of topic sentence presence largely mirrors the findings of Popken (1987, 1988), who reported topic sentence presence at 55% and 54%, respectively.

As is shown in Table 12, the 114 identified topic sentences overwhelmingly occurred in the expository texts of the history and science domains (89.5%; $\chi^2 = 48.76$) and overwhelmingly occurred in the paragraph-initial position (history, $\chi^2 = 141.82$, $p < .001$; science, $\chi^2 = 168.57$, $p < .001$; narrative, $\chi^2 = 29.74$, $p < .001$). In total, 93% of all identified topic sentences occurred in the paragraph-initial position.

Recalling the results from Experiment 1 (see Tables 1, 2, and 3), we can note that the coreference methods (e.g., LSA) showed no evidence of any sentence position's being privileged. That is, overlap indices were as likely to select a mid- or final-sentence of a paragraph as the topic sentence as they were the first sentence. The results of Experiment 3 (which used a subset of the same paragraphs as those used in Experiment 1) provide evidence supporting the widely held belief that when topic sentences do occur, they occur in expository texts and in the paragraph-initial position.

The application of the free model algorithm generated in Experiment 2 produced mixed results. Overall, the accuracy of the model ranged from 79.3% (history) through 80.4% (science) to 94.6% (narrative) accuracy. In total, the model correctly assessed 84.8% of the data.

At first blush, such results appear to be good; however, closer examination suggests that the model is not sufficiently accurate in critical areas of topic sentence detection. Specifically, experts assessed 714 of the 828 sentences to be nontopic sentences, correctly identifying 691 of the nontopic sentences (recall, 96.8%; precision, 87.0%). However, of the 114 sentences that the experts assessed as topic sentences, the free model detected only 11 and made 23 further false alarms (recall, 9.6%; precision, 32.4%). Thus, although the model appears extremely effective at identifying nontopic sentences, topic sentence identification was extremely weak.

To better assess the contribution of factors that may have led to the evaluations of the topic sentences, we ran a multiple regression using the rater mean values (golden mean) as the dependent variable and the 12 predictor indices from Experiment 2 as the independent variables. We further added two binary variables as predictors. These variables were *sentence-initial position*/*other position* and *expository*/*nonexpository* text. Using the stepwise method, a significant model emerged [$F(5,822) = 179.847$, $p < .001$]. The model explains 52.0% of the variance (adjusted $R^2 = .520$). Table 13 provides information for the predictor variables entered into the model. The predictor variables *meaningfulness* (mean), *pronoun incidence*, *word hypernym* (mean), *third-person verbs* (incidence), *coordinating conjunctions* (incidence), *adjectives* (incidence), *existential "there"* (incidence), *cardinal numbers* (incidence), and *number of words before main verb* (mean) were *not* significant predictors.

The results of the regression suggest that paragraph-initial position and domain (expository/nonexpository) were the key factors for the experts in their decision as to whether or not a sentence was a topic sentence. Specifically, if the sentence was first in the paragraph and was also from an expository domain, the raters were more likely to evaluate the sentence as a topic sentence. The predictors of *past tenses* and *connectives* were predictably negative; that is, they were more likely to occur in nontopic sentences. The *average sentence length* variable, however, suggested that longer sentences may be indicative of topic sentencehood when the other variables are taken into consideration.

Conclusions

The differences between the results of Experiment 2 and those of Experiment 3 suggest two concerns: (1) a problem with the expert ratings and (2) a problem with the algorithm generated in Experiment 2. Given that the correlations for the expert ratings in Experiment 3 were consistent with those in Experiment 2 (the accuracy of which was 78%), we could argue that we have sufficient reason to accept the expert evaluations as reliable. In addition, close examination of the texts and their corresponding evaluations showed no obvious errors or reasons to doubt the evaluations. As such, we did not consider the

Table 13
Results of Multiple Regression, Showing the Unstandardized and Standardized Regression Coefficients for the Variables Entered Into the Model

	B				
Variable	*M*	*SE*	*β*	*t*	Significance
Sentence position	1.774	0.063	.682	28.181	<.001
Domain	0.466	0.062	.195	7.488	<.001
Past tenses	−0.001	0.000	−.056	−2.120	.034
Connectives	−0.001	0.000	−.066	−2.682	.007
Average sentence length	0.009	0.004	.055	2.227	.026
Constant	1.366	0.091		14.965	<.001

ratings to be the cause of the inconsistent results. Turning to the algorithm, the high accuracy of *non*topic sentence identification suggested that the problem of classification for topic sentences may lie with the *types* of topic sentences that the expert raters had identified. That is, we can argue that the topic sentences used in Experiment 2 were *ideal topic sentences*. We refer to these topic sentences as ideal because we can presume that they had been specially written or selected as ideal examples for the concept of topic sentence. That is, if the goal of a textbook is to provide examples of topic sentences, the examples are likely to be clearly topic sentences, rather than ambiguous. In contrast, the rated topic sentences in Experiment 3 are *naturalistic topic sentences*. These sentences naturally occurred in high school textbooks in which there was little or no particular attention (we assume) given to topic sentence construction (because the texts were not about composition skills). Therefore, we concluded that there appear to be two types of topic sentences: *ideal topic sentences*—those that have been specially selected, or specially written, for their prototypical qualities—and *naturalistic topic sentences* that emerge naturally in texts without explicit attention to their construction.

On the basis of the explanation above, the results of Experiments 2 and 3 can be summarized as follows. Ideal type topic sentences can be identified reliably by a free model algorithm that requires nothing more than sentence-level features. Naturalistic type topic sentences, on the other hand, occur predominantly in the first-sentence position of expository texts and *cannot* be reliably identified by either the free model or the derived model.

These results led us to a new question: If naturalistic type topic sentences can be reliably identified by experts but cannot be identified by either computational model (whether through *free* features or through *coreference*), what is it that makes such sentences topic sentences? We hypothesized that the answer to that question may be that it is not textual features per se that are most responsible for rendering a sentence a topic sentence. Instead, it may simply be a combination of the fact that the text is designated a *paragraph* in conjunction with the application of the knowledge and skill of the reader. Longacre (1979) argued that readers process paragraphs as hierarchically organized structural units: The first sentence of the paragraph signifies to the reader the topic of the structure. Similarly, Heurly (1997) argued that the very existence of marking a paragraph can be a signal from the writer that a shift of topic is taking place. Such studies, along with a considerable amount of related research (e.g., Bestgen, 1992; Britton, 1994; Fayol & Schneuwly, 1987; Hinds, 1980; Hofmann, 1989; Le Ny, 1985), led us to conclude that an opening sentence in a paragraph may hold a *privileged position* beyond the features (free or derived) that constitute the sentence. That is, when *high-skilled*/*high-knowledge* readers (as in the case of our raters) are asked to read professionally constructed text (as in the case of our corpus), those readers will be able to quickly comprehend the type of text they are dealing with (McCarthy & McNamara, 2007), to have the knowledge that expository text commonly features topic sentences, and to know that topic sentences tend to fall in the paragraph-initial position. Thus, given a paragraph, we hypothesized that readers assume that the first sentence will be a topic sentence unless there is evidence otherwise.

To test this hypothesis, we conducted a fourth experiment in which a subset of the paragraphs used in Experiment 3 were divided into individual sentences (as with Experiment 2) to be once more evaluated by our experts. We predicted that under such circumstances, the paragraph-initial sentences would lose some of their privilege and would, consequently, receive lower ratings. Similarly, sentences that were *topic-sentence-like*, but not originally placed in the paragraph-initial position, would receive higher ratings when viewed in free form. Such findings, if confirmed, would indicate that topic sentences do contain free features but that our raters (who were high-ability readers) placed greater value on the paragraph construct itself, giving privilege to the sentence-initial position.

EXPERIMENT 4

In Experiment 4, we tested the hypothesis that topic sentencehood evaluations would differ depending on whether the sentences appeared out of context (free model) or in context (derived model). The results of Experiment 3 suggested that expository paragraph-initial sentences were overwhelmingly selected as topic sentences. However, we could not be confident whether it was the sentence itself that led to these evaluations, the surrounding sentences that influenced such ratings, or merely the position and domain (in conjunction with the skills/knowledge of the readers) that influenced the ratings. In Experiment 4, therefore, we asked our experts to rate the expository sentences for topic sentencehood by evaluating the sentences free of context. We hypothesized that *non*-paragraph-initial sentences, when evaluated as free sentences, would increase in topic sentencehood evaluation. That is, we hypothesized that paragraph-initial sentences were given special status by evaluators as the *topic sentence elect*, meaning that any other sentence would have to prove itself to be a better topic sentence than was the paragraph-initial sentence. In effect, we hypothesized that this paragraph-initial primacy was leading raters to believe that the first sentence of a paragraph (especially an expository paragraph) *was* the topic sentence. We further hypothesized that paragraph-initial sentences from Experiment 3, when viewed context free, would decrease in topic sentencehood evaluation. That is, we hypothesized that the evaluation of paragraph-initial sentences from Experiment 3 tended to boost scores for paragraph-initial sentences on the basis of the rater interpretation of the paragraph as being expository and the position as being primary.

For the Experiment 4 corpus, we used all the 4-sentence expository paragraphs (history and science) in Experiment 3, together with an equal number of 4-sentence paragraphs (history and science) that had not been used in Experiment 3. In total, therefore, the corpus comprised 184 sentences (92 from history and 92 from science), with half the corpus having ratings that were *old* (rated in Experiment 3 within context) and half the corpus having ratings that were *new* (rated only in Experiment 4 in free form).

One concern with Experiment 4 was a possible confound caused by the raters' having already viewed the sentences for this experiment when they evaluated the sentences in Experiment 3. It was for this reason that a 4-month period was allowed to elapse between the experiments and that the further *new* but equivalent corpus was added to the experiment for comparison purposes.

Results

The results of Experiment 4 showed that interrater reliability was once more significant and comparable to previous ratings ($M = .602$, $SD = .110$). Pearson correlations ranged from .460 ($p < .001$) to .759 ($p < .001$). Our results also showed that there were no significant differences between the evaluations of the data from paragraphs used in Experiment 3 and from those sentences that were completely new to the raters (meaning that the fact that the raters had viewed the sentences 4 months earlier was not a significant factor). To assess the experts' evaluations of the paragraph data with values from derived and free context conditions, we conducted a paired *t* test on the golden mean score of the expert evaluations, including the within-subjects factors of *free evaluation* ($M = 2.95$, $SD = 1.20$) and *derived evaluation* ($M = 2.295$, $SD = 1.22$). As was predicted, the ratings of the experts were significantly higher for sentences evaluated in the free condition, with a main effect of context condition [$t(183) = 7.425$, $p < .001$]. When each sentence position was identified individually, the results largely confirmed our predictions. Sentence Positions 2, 3, and 4 (for both the science and the history domains) were all significantly higher when considered in the free condition. Also as was predicted, Sentence Position 1 was significantly lower for the science domain when evaluated in free context. In contrast, there was no significant effect for the history domain (see Table 14). These results suggest that expert raters do, indeed, evaluate sentences differently depending on whether the sentences are presented out of versus in context.

Conclusions

In Experiment 4, the *in-context* sentences from Experiment 3 were rated free of context. The results suggested that expert raters evaluate sentences out of context differently from how they are evaluated in context. These results suggest that the actual entity of the paragraph, the readers' interaction with the paragraph text as a whole, and the readers' skills and knowledge may all play a part in judging topic sentencehood. This may be especially the case for topic sentences that do not significantly exhibit the features of the ideal topic sentences identified in Experiment 2.

GENERAL DISCUSSION AND CONCLUSIONS

Although some studies have suggested that topic sentences are relatively rare (e.g., Braddock, 1974), psychological studies have suggested that the presence in text of topic sentences is beneficial to comprehension (e.g., Aulls, 1975; Kieras, 1978; Lorch & Lorch, 1996; McNamara et al., 1996). The importance of identification and assessment of topic sentences within text (topic sentencehood) is also raised by the recent increase in technology, bringing about such computational systems as Coh-Metrix (Graesser et al., 2004), the goal of which is to better assess text in order to optimally match text to reader.

In the present study, we assessed two theoretical models of topic sentencehood identification, along with two corresponding computational models. We called these the derived model and the free model. The derived model posits that topic sentences are dependent on their context and can be identified only when assessed against their context. Thus, a topic sentence is a topic sentence because the other sentences in the paragraph *make it* a topic sentence. Furthermore, the individual features of this derived topic sentence are indistinguishable from other sentences when considered independently. The free model posits that a topic sentence is a topic sentence independently of its context. As such, the free model holds that topic sentences have lexical and syntactic features that can be identified both computationally and by human raters without the need for any contextual information.

The present study included four experiments in which the validity of the models described above was assessed. The coreferential indices of LSA and lemma overlap were used to approximate the derived model. The free model was assessed using Coh-Metrix sentence- and word-level indices.

In Experiment 1, we assessed two coreference indices as approximations for the derived model of topic sen-

Table 14
Results of Paired *t* Test Including Within-Subjects Factor of Derived/Free Condition for History and Science Four-Sentence Paragraphs

		Mean		*SD*			
Domain	Sentence	Derived	Free	Derived	Free	*t* (*df* = 22)	Significance
Science	1	4.044	3.493	1.130	1.154	2.224	.037
	2	2.159	2.855	0.909	1.118	−2.448	.023
	3	1.551	2.754	0.410	1.248	−5.719	<.001
	4	1.565	2.449	0.639	1.043	−3.923	<.001
History	1	3.609	3.942	1.149	1.266	−1.452	n.s.
	2	1.942	2.942	0.625	0.993	−4.870	<.001
	3	1.638	2.493	0.401	0.840	−4.686	<.001
	4	1.855	2.551	0.931	1.196	−3.761	<.001

tencehood identification. The best-known method for identifying the derived model is the section topicality value (Kintsch, 2002). In this model, all the sentences in a paragraph are compared with all the other sentences in the paragraph. The sentence with the highest mean cosine is credited as being the sentence that has more information, relative to the entire paragraph, than does any other sentence and is, therefore, the topic sentence of the paragraph. Because topic sentences tend to occur in the paragraph-initial position, the model would presumably identify more paragraph-initial sentences as topic sentences than it would sentences in any other sentence position. In Experiment 1, we tested this prediction against a corpus of history, narrative, and science texts, using three-, four-, and five-sentence paragraphs. The results produced no evidence to support the derived model. However, further analysis revealed that the section topicality values tended to be confounded by sentence length. As such, the main conclusion from Experiment 1 was that coreference computational models may have difficulty identifying topic sentences because longer sentences tend to yield higher overlap values.

For Experiment 2, we collected a corpus of independently identified and published topic sentences. We trained 3 experts in discourse processing to distinguish such topic sentences from nontopic sentences. We also trained a computational model to distinguish the sentence types. The results suggested that human raters significantly agreed on the distinctions between the sentence types with a mean accuracy of 78%. The computational model achieved approximately 70% accuracy. We can speculate that the weaker performance (although significant) of the computational model may have been the result of a lack of keyword identification. That is, we believe that many of the sentences that the experts selected as topic sentences were selected because lexical features indicated that the sentence derived from an expository or nonexpository source. Future research should consider developing the computational free model to further improve its accuracy. Nevertheless, our conclusion from Experiment 2 was that both humans and a computational model could distinguish topic sentences from nontopic sentences in a context-free study. This evidence supports the free model of topic sentencehood.

In Experiment 3, we used a large sample of Experiment 1 texts to examine how well the same experts would rate sentences from Experiment 1 for topic sentencehood. Specifically, we randomly selected 69 paragraphs from each of the three domains, with the three-, four-, and five-sentence paragraph examples all being equally represented. The experts rated each sentence in each paragraph but, unlike in Experiment 2, they were able to see the entire paragraph in context. The results did not offer any support for the derived model or, more precisely, for the coreference indices that were used to approximate it. Overwhelmingly, the raters selected topic sentences from expository paragraphs and from the paragraph-initial position. Fewer than 4% of the non-paragraph-initial sentences were selected as clear topic sentences by the raters; over 55% of the paragraph-initial sentences *were* selected. When the free model algorithm from Experiment 2 was applied to the paragraphs in Experiment 3, the model could not identify the expert-selected topic sentences. In fact, the model found extremely few examples that it designated as topic sentences. The coreferential model was also very poor at identifying the expert-rated topic sentences from this corpus. The results from Experiment 3, in conjunction with those from a number of other studies (e.g., Aulls, 1975; Fishman, 1978; Goldman et al., 1995; Graesser et al., 2003; McNamara et al., 1996), cause us to speculate that given a complete paragraph structure, skilled/knowledgeable readers can process the entire information of a paragraph and will tend to assign topic sentencehood to the first sentence of the paragraph.

In Experiment 4, we tested the hypothesis formed from Experiment 3 by taking Experiment 3's sentences out of context. Using only the expository sentences (science and history) from all 46 paragraphs (*old data*), as well as an additional 46 paragraphs (23 science and 23 history; *new data*), our experts rated each sentence for topic sentencehood. The results suggested high agreement between the raters and no significant difference between the old data and the new data. When the data from Experiment 4 (free context) were compared with the values for those sentences from Experiment 3 (in context), significant differences were identified. Specifically, in the free condition, non-paragraph-initial sentences were scored higher, whereas paragraph-initial sentences tended to be scored lower. Thus, our conclusion from Experiment 4 is that sentences that are topic sentences *in context* may not be considered topic sentences when taken *out of context*; and by the same token, sentences that are not considered topic sentences free of context may be considered topic sentences if put in context. In other words, some sentences appear to have consistent features that identify these sentences as topic sentences; however, these features may be less important to readers than is the positioning of the sentence and the domain to which the sentence belongs. The fact that our results are based on the ratings of experts, however, also causes us to speculate that readers with low-knowledge/low-skill may not so readily or easily identify the paragraph domain and process the paragraph's information with the paragraph-initial sentence as the topic sentence elect.

Our overall conclusion, therefore, is that there are (at least) two types of topic sentences. One type of topic sentence may be viewed as *ideal* and can be represented by the free model (i.e., out of context). These ideal topic sentences are often used as examples in textbooks (see Experiment 2), and they are computationally identifiable with a high degree of accuracy. We conjecture that such a form of topic sentence will be helpful to low-skill or low-knowledge readers, because the topic sentencehood features are diverse, consistent, teachable, and recognizable.

The other type of topic sentence may be considered *naturalistic*. Without context, these sentences are far harder to recognize as topic sentences. Our results suggest that these sentences tend to *become* topic sentences as a result of a combination of the readers' recognition of the domain, the positioning of the sentence, the skills/knowledge of the reader, and the interaction of the other sentences in the paragraph. As such, these naturalistic type topic sentences may demand that the reader process the entire paragraph and integrate prior knowledge with the content of the paragraph with some proficiency. We therefore conjecture that this type of topic sentence may be less beneficial to less skilled or low-knowledge readers.

Computationally, our Coh-Metrix-based *sentence-* and *word-index* free model was highly predictive in identifying topic sentences from nontopic sentences for ideal topic sentences. We believe that given sufficient additional indices, such as a corpus of keywords, that the accuracy can be increased further. A similarly accurate measure identifying naturalistic type sentences is a different issue, however, because as long as the text is known to be expository and the sentence position is first, few other factors of the text seem critical. Coreferential indices, such as LSA, were not able to identify any kind of human-rated topic sentences.

Future research should consider comprehension and recall experiments in which texts are manipulated from naturalistic type to ideal type. Our hypothesis, based on the results of this study and the findings of such research as McNamara et al. (1996), is that lower knowledge participants will benefit from the ideal topic sentences form, whereas high-knowledge participants will show no significant benefit from the ideal topic sentence type.

The purpose of this study was to evaluate approaches to identifying topic sentencehood. The identification of such sentence types and their subsequent evaluation in terms of quality has relevance to a number of possible applications. For instance, in composition assessment tools for basic writers, algorithms that identify textual features (or variations from prototypical features) can provide feedback to users so as to facilitate writing development. In summarization tools, where shorter sentences are often viewed as less important because they carry less information, the identification of topic sentencehood can signal the importance of the sentence, leading to key textual information's being retained. And in information retrieval applications, a paragraph that can be identified as containing a topic sentence may be a better candidate response to a user inquiry because the relevance of the text to the enquirer is more explicit.

In this study, however, our interest in topic sentencehood identification was directed at better evaluation of text structure in order to more effectively match text to reader. Given that topic sentences are more likely to provide assistance to low-skilled/low-knowledge readers, and given that such readers would probably benefit more from ideal type topic sentences, the free model of topic sentencehood introduced here offers systems such as Coh-Metrix the opportunity to better assess texts and better fulfill the Coh-Metrix goal of optimally matching text to readers.

AUTHOR NOTE

This research was supported by the Institute for Education Sciences (IES R305G020018-02). Opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES. The authors also acknowledge the contributions made to this article by Art Graesser, Max Louwerse, and Zhiqiang Cai. Correspondence concerning this article should be addressed to P. M. McCarthy, Institute for Intelligent Systems, University of Memphis, 4th Floor, Room 410, Memphis, TN 38152 (e-mail: pmccarthy@mail.psyc.memphis.edu).

REFERENCES

Angus, J. (1862). *Handbook of the English tongue.* London: Religious Tract Society.

Aristotle (1954). *The rhetoric.* In *Rhetoric and the poetics of Aristotle.* (W. Rhys Roberts, Trans.). New York: Random House.

Aulls, M. W. (1975). Expository paragraph properties that influence literal recall. *Journal of Reading Behavior*, **7**, 391-399.

Austin, J. L. (1962). *How to do things with words.* Cambridge, MA: Harvard University Press.

Bain, A. (1866). *English composition and rhetoric.* New York: Appleton.

Bestgen, Y. (1992). *Deux approches du discours narratif: Marqueurs de la segmentation et profil émotionnel.* Unpublished doctoral dissertation, Catholic University of Louvain-La-Neuve, Belgium.

Braddock, R. (1974). The frequency and placement of topic sentences in expository prose. *Research in the Teaching of English*, **8**, 287-302.

Britton, B. K. (1994). Understanding expository text: Building mental structures to induce insights. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 641-674). New York: Academic Press.

Clements, P. (1979). The effects of staging on recall from prose. In R. O. Freedle (Ed.), *New directions in discourse processing* (pp. 297-330). Norwood, NJ: Ablex.

Coltheart, M. (1981). The MRC psycholinguistics database. *Quarterly Journal of Experimental Psychology*, **33A**, 497-505.

Crossley, S. A., Louwerse, M. M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *Modern Language Journal*, **91**, 15-30.

D'Angelo, F. (1986). The topic sentence revisited. *College Composition & Communication*, **37**, 431-441.

Dennis, S. (2007). Introducing word order in an LSA framework. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. xxx-xxx). Mahwah, NJ: Erlbaum.

Duncan, M. G. (in press). What ever became of the paragraph. In *College English*.

Duran, N. D., McCarthy, P. M., Graesser, A. C., & McNamara, D. S. (2007). Using temporal cohesion to predict temporal coherence in narrative and expository texts. *Behavior Research Methods*, **39**, 212-223.

Eden, R., & Mitchell, R. (1986). Paragraphing for the reader. *College Composition & Communication*, **37**, 416-430.

Fayol, M., & Schneuwly, B. (1987). La mise en texte et ses problèmes. In J. L. Chiss, J. P. Laurent, J. C. Meyer, H. Romain, & B. Schneuwly (Eds.), *Apprende/enseigner à produire des texts écrits* (pp. 223-240). Bruxelles: De Broek.

Fishman, A. S. (1978). The effect of anaphoric reference and noun phrase organizers on paragraph comprehension. *Journal of Reading Behavior*, **10**, 159-169.

Fox, S., Bizman, A., Hoffman, M., & Oren, L. (1995). The impact of variability in candidate profiles on rater confidence and judgments regarding stability and job suitability. *Journal of Occupational & Organizational Psychology*, **68**, 13-23.

Goldman, S. R., Graesser, A. C., & van den Broek, P. (1999). Essays in honor of Tom Trabasso. In S. R. Goldman, A. C. Graesser, & P. van den Broek (Eds.), *Narrative comprehension, causality, and coherence* (pp. 1-10). Mahwah, NJ: Erlbaum.

Goldman, S. R., Saul, E. U., & Coté, N. (1995). Paragraphing, reader, and task effects on discourse comprehension. *Discourse Processes*, **20**, 273-305.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text? In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 82-98). New York: Guilford.

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

Hall, C., McCarthy, P. M., Lewis, G. A., Lee, D. S., & McNamara, D. S. (2007). Using Coh-Metrix to assess differences between

English language varieties. In J. Brewer, P. O'Rourke, & P. Richtsmeier (Eds.), *Coyote papers: Working papers in linguistics* (Vol. 15, pp. 40-54). Tucson: University of Arizona Linguistics Circle.

Halliday, M. A. K., & Hassan, R. (1976). *Cohesion in English*. New York: Longman Group.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York: Newbury House.

Haviland, S. E., & Clark, H. H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning & Verbal Behavior*, **13**, 512-521.

Hempelmann, C. F., Dufty, D. F., 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. Barsalau & M. Bucciarelli (Eds.), *Proceedings of the XXVII Annual Meeting of the Cognitive Science Society* (pp. 6941-6946). Austin, TX: Cognitive Science Society.

Heurly, L. (1997). Processing units in written texts: Paragraphs or information blocks? In J. Costermans & M. Fayol (Eds.), *Processing interclausal relationships: Studies in the production and comprehension of text* (pp. 179-200). Mahwah, NJ: Erlbaum.

Hinds, J. (1980). Japanese expository prose. *Papers in Linguistics*, **13**, 117-158.

Hoey, M. (1991). *Patterns of Lexis in text*. Oxford: Oxford University Press.

Hofmann, T. R. (1989). Paragraphs, & anaphora. *Journal of Pragmatics*, **13**, 239-250.

Kavanaugh, M. J. (1989). *Performance rating accuracy improvement through changes in individual and system characteristics*. San Antonio: Texas Maxima Corp. (Sponsored by the Air Force Human Resources Lab).

Kieras, D. E. (1978). Good and bad structure in simple paragraphs: Effects on apparent theme, reading time, and recall. *Journal of Verbal Learning & Verbal Behavior*, **17**, 13-28.

Kieras, D. E. (1981). Component processes in the comprehension of simple prose. *Journal of Verbal Learning & Verbal Behavior*, **20**, 1-23.

Kintsch, W. (2002). On the notions of theme and topic in psychological process models of text comprehension. In M. Louwerse & W. van Peer (Eds.), *Thematics: Interdisciplinary studies* (pp. 157-170). Amsterdam: Benjamins.

Kintsch, W., & van Dijk, T. A. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, **104**, 211-240.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.) (2007). *Handbook of latent semantic analysis.* Mahwah, NJ: Erlbaum.

Le Ny, J. F. (1985). Texte, structure mentale, paragraphe. In J. Châtillon (Ed.), *La notion de paragraphe* (pp. 129-136). Paris: Editions du Centre National de la Recherche Scientifique.

León, J. A., & Carretero, M. (1995). Intervention in comprehension and memory strategies: Knowledge and use of text structure. *Learning & Instruction*, **5**, 203-220.

Lesgold, A. M., Roth, S. F., & Curtis, M. E. (1979). Foregrounding effects in discourse comprehension. *Journal of Verbal Learning & Verbal Behavior*, **18**, 291-308.

Lightman, E. J., McCarthy, P. M., Dufty, D. F., & McNamara, D. S. (2007a). Cohesion and structural organization in high school texts. In *Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS)* (pp. 235-240). Menlo Park, CA: AAAI Press.

Lightman, E. J., McCarthy, P. M., Dufty, D. F., & McNamara, D. S. (2007b). Using computational text analysis tools to compare the lyrics of suicidal and non-suicidal songwriters. In D. S. McNamara & G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 1217-1222). Austin, TX: Cognitive Science Society.

Longacre, R. E. (1979). The paragraph as a grammatical unit. In T. Givón (Ed.), *Syntax and semantics: Vol. 12. Discourse and syntax* (pp. 115-134). New York: Academic Press.

Lorch, R. F., Jr., & Lorch, E. P. (1995). Effects of organizational signals on text processing strategies. *Journal of Educational Psychology*, **87**, 537-544.

Lorch, R. F., Jr., & Lorch, E. P. (1996). Effects of headings on text recall and summarization. *Contemporary Educational Psychology*, **21**, 261-278.

Lorch, R. F., Jr., Lorch, E. P., & Matthews, P. D. (1985). On-line processing of the topic structure of a text. *Journal of Memory & Language*, **24**, 350-362.

McCarthy, P. M., Graesser, A. C., & McNamara, D. S. (2006, July). *Distinguishing genre using Coh-Metrix indices of cohesion.* Paper presented at the Society for Text and Discourse conference, Minneapolis, MN.

McCarthy, P. M., Lehenbauer, B. M., Hall, C., Duran, N. D., Fujiwara, Y., & McNamara, D. S. (2007). A Coh-Metrix analysis of discourse variation in the texts of Japanese, American, and British scientists. *Foreign Languages for Specific Purposes*, **6**, 46-77.

McCarthy, P. M., Lewis, G. A., Dufty, D. F., & McNamara, D. S. (2006). Analyzing writing styles with Coh-Metrix. In *Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS)* (pp. 764-769). Menlo Park, CA: AAAI Press.

McCarthy, P. M., & McNamara, D. S. (2007). Are seven words all we need? Recognizing genre at the sub-sentential level. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (pp. 1295-1300). Austin, TX: Cognitive Science Society.

McCarthy, P. M., Rus, V., Crossley, S. A., Bigham, S. C., Graesser, A. C., & McNamara, D. S. (2007). Assessing entailer with a corpus of natural language. In *Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS)* (pp. 247-252). Menlo Park, CA: AAAI Press.

McElroy, J. G. R. (1885). *The structure of English prose: A manual of composition and rhetoric*. New York: Armstrong.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Text coherence, background knowledge, and levels of understanding in learning from text. *Cognition & Instruction*, **14**, 1-43.

McNamara, D. S., Ozuru, Y., Graesser, A. C., & Louwerse, M. (2006). Validating Coh-Metrix. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 573-578). Austin, TX: Cognitive Science Society.

Meade, R., & Ellis, W. G. (1970). Paragraph development in the modern age of rhetoric. *English Journal*, **59**, 219-226.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, **38**, 39-41.

Miller, J. R., & Kintsch, W. (1980). Readability and recall of short prose passages: A theoretical analysis. *Journal of Experimental Behavior: Human Learning & Memory*, **6**, 335-354.

Olney, A., & Cai, Z. (2005). An orthonormal basis for topic segmentation in tutorial dialogue. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (pp. 971-978). Philadelphia: Association for Computational Linguistics.

Penumatsa, P., Ventura, M., Graesser, A. C., Franceschetti, D. R., Louwerse, M., Hu, X., et al. (2004). The right threshold value: What is the right threshold of cosine measure when using latent semantic analysis for evaluating student answers? *International Journal of Artificial Intelligence Tools*, **12**, 257-279.

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

Popken, R. L. (1988). A study of topic sentence use in scientific writing. *Journal of Technical Writing & Communication*, **18**, 75-86.

Popken, R. L. (1991a). A study of topic sentence use in technical writing. *The Technical Writing Teacher*, **18**, 49-58.

Popken, R. L. (1991b). A study of topic sentence use in the familiar essay. *CCTE Studies*, **56**, 47-56.

Rehder, B., Schreiner, M., Laham, D., Wolfe, M., Landauer, T., & Kintsch, W. (1998). Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes*, **25**, 337-354.

Richards, J. P. (1975-1976). Processing effects of advance organizers interspersed in text. *Reading Research Quarterly*, **11**, 599-622.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, **7**, 573-605.

Sanz, C. (Ed.) (2005). *Mind and context in adult second language acquisition: Methods, theory and practice.* Washington, DC: Georgetown University Press.

Sardinha, T. B. (2001). Lexical segments in text. In M. Scott & G. Thompson (Eds.), *Patterns of text: In honor of Michael Hoey* (pp. 213-237). Amsterdam: Benjamins.

Searle, J. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, **7**, 301-317.

Toulmin, S. (1969). *The uses of argument*. Cambridge: Cambridge University Press.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques.* San Francisco: Morgan Kaufmann.

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Blackwell.

NOTES

1. Most research points toward topic sentences' occurring in approximately 50% of expository paragraphs.

2. Kintsch (2002) still recommended this measure as the best approximation for identifying the topic of a sentence.

3. Naturally, since the rater gold result is formed from the raters' individual results, the individual-rater to rater gold correlation will be high.

APPENDIX
Books and Web Sites Used to Identify Topic Sentence Data

Books

Bracher, F., & Elsbree, L. (1967). *College handbook of composition*. Boston: Heath.

Charvat, W., Mead, C. D., & Legget, G. (1965). *Prentice-Hall handbook for writers* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Moore, R. H. (1965). *Effective writing* (3rd ed.). New York: Holt, Rinehart & Winston.

Crews, F. (1980). *The Random House handbook* (3rd ed.). New York: Random House.

Bergman, C. A., & Senn, J. A. (1986). *Heath grammar and composition*. Toronto: Heath.

Blanchard, K., & Root, C. (1997). *Ready to write more: From paragraph to essay*. White Plains, NY: Addison Wesley Longman.

Baker, S. (1984). *The complete stylist and handbook* (3rd ed.). New York: Harper Collins.

Rosenberg, H. M., & Suberman, J. (1968). *Basic composition* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Heffernan, J. A. W. (1986). *Writing: A college handbook* (2nd ed.). New York: Norton.

Stephens, R. D. (1986). *Sequence: A basic writing course* (2nd ed.). New York: Holt, Rinehart & Winston.

Guth, H. P. (1975). *Words and ideas: A handbook for college writing* (4th ed.). Belmont, CA: Wadsworth.

Canavan, P. J. (1975). *Paragraphs and themes* (2nd ed.). Toronto: Heath.

Sullivan, K. E. (1976). *Paragraph practice* (3rd ed.). New York: Macmillan.

Christ, M. F., & Himes, C. A. (1938). *A laboratory course in reading and writing*. New York: Crofts.

Web Sites

www.utoronto.ca/ucwriting/topic.html. Accessed March 13, 2006.

english.glendale.edu/topic.html. Accessed March 13, 2006.

web.mit.edu/writing/Writing_Process/topicsentence.html. Accessed March 13, 2006.

www.lausd.k12.ca.us/lausd/resources/composition/compmenu.html. Accessed March 13, 2006.

www.mce.k12tn.net/reading39/paragraph_unit/lesson4.htm. Accessed March 13, 2006.

www.utoronto.ca/ucwriting/topic.html. Accessed March 13, 2006.

www2.actden.com/WRIT_DEN/tips/paragrap/topic.htm. Accessed March 13, 2006.

english.glendale.edu/topic.html. Accessed March 13, 2006.

web.mit.edu/writing/Writing_Process/topicsentence.html. Accessed March 13, 2006.

www.fas.harvard.edu/~wricntr/documents/TopicSentences.html. Accessed March 13, 2006.

www.lausd.k12.ca.us/lausd/resources/composition/compmenu.html. Accessed March 13, 2006.

www.mce.k12tn.net/reading39/paragraph_unit/lesson4.htm. Accessed March 13, 2006.

grammar.ccc.commnet.edu/grammar/paragraphs.htm. Accessed March 13, 2006.

www.arts.uottawa.ca/writcent/hypergrammar/partopic.html. Accessed March 8, 2006.

cla.univ-fcomte.fr/english/paragraph/tutorial_5/organization.html. Accessed March 8, 2006.

www.indiana.edu/~wts/pamphlets/paragraphs.shtml. Accessed March 9, 2006.

www.kent.k12.wa.us/KSD/KR/WRITE/GEN/topic_sent.html. Accessed March 9, 2006.

www.utoronto.ca/ucwriting/topic.html. Accessed March 13, 2006.

www2.actden.com/WRIT_DEN/tips/paragrap/index.htm. Accessed March 13, 2006.

209.61.208.100/exe/writing-effectively/Module%201/m01.079.htm. Accessed January 20, 2006.

www.uefap.co.uk/writing/parag/partopic.htm. Accessed January 20, 2006.

www.busyteacherscafe.com/units/paragraph.htm. Accessed January 20, 2006.

academic.cuesta.edu/acasupp/AS/308.HTM. Accessed January 20, 2006.

www.bfwpub.com/pdfs/Anker_RealWriting/Chapter3.pdf. Accessed January 20, 2006.

ec.hku.hk/writingmachine/bin5/topic_dev.htm. Accessed January 20, 2006.

www.hhpublishing.com/_onlinecourses/clastdemo/clast/readingskills/A3.html. Accessed January 20, 2006.

mr_sedivy.tripod.com/essay.html. Accessed January 20, 2006.

www.learnnc.org/lessons/MarthaOwens6182002064. Accessed January 20, 2006.

www.uottawa.ca/academic/arts/writcent/hypergrammar/partopic.html. Accessed January 20, 2006.

www2.actden.com/writ_den/tips/paragrap/topic.htm. Accessed January 20, 2006.

www.utoronto.ca/ucwriting/topic.html. Accessed January 20, 2006.

APPENDIX (Continued)

www.internationalstudent.com/termpaper_writing/topic.shtml. Accessed January 20, 2006.
www.mce.k12tn.net/reading39/paragraph_unit/lesson4.htm. Accessed January 20, 2006.
www.dvc.edu/english/Learning_Resources/TopicSentenceDefandEx.htm. Accessed January 20, 2006.
learnline.cdu.edu.au/studyskills/wr/wr_pa_ts.html. Accessed January 20, 2006.
www.mhhe.com/mayfieldpub/tsw/topic-s.htm. Accessed January 20, 2006.
www.physics.ohio-state.edu/~wilkins/writing/Assign/so/baltimore.html#endso. Accessed January 20, 2006.
www.uhv.edu/ac/research/write/topicsentences.html. Accessed January 20, 2006.
www.eslbee.com/topic_sentences.htm. Accessed January 30, 2006.
wps.ablongman.com/long_johnsonshe_tct_1/0,10006,1798190-content,00.html. Accessed January 30, 2006.
wps.ablongman.com/long_long_rw_1/0,8256,1041180-,00.html. Accessed January 30, 2006.

(Manuscript received December 11, 2006;
revision accepted for publication June 25, 2007.)