**Philip M. McCarthy (pmmccrth@memphis.edu)**
Department of English, The University of Memphis, Memphis. TN 38152
**John C. Myers (jcmyers@memphis.edu)**
Department of Psychology, The University of Memphis, Memphis. TN 38152
**Stephen W. Briner (sbriner@depaul.edu)**
Department of Psychology, DePaul University, Chicago, IL 60614
**Arthur C. Graesser (graesser@memphis.edu)**
Department of Psychology, The University of Memphis, Memphis. TN 38152
**Danielle S. McNamara (dsmcnamr@memphis.edu)**
Department of Psychology, The University of Memphis, Memphis. TN 38152

# A Psychological and Computational Study of Sub-Sentential Genre Recognition

## Abstract

Genre recognition is a critical facet of text comprehension and text classification. In three experiments, we assessed the minimum number of words in a sentence needed for genre recognition to occur, the distribution of genres across text, and the relationship between reading ability and genre recognition. We also propose and demonstrate a computational model for genre recognition. Using corpora of narrative, history, and science sentences, we found that readers could recognize the genre of over 80% of the sentences and that recognition generally occurred within the first three words of sentences; in fact, 51% of the sentences could be correctly identified by the first word alone. We also report findings that many texts are heterogeneous in terms of genre. That is, around 20% of text appears to include sentences from other genres. In addition, our computational models fit closely the judgments of human result. This study offers a novel approach to genre identification at the sub-sentential level and has important implications for fields as diverse as reading comprehension and computational text classification.

Key words: Genre recognition, reading comprehension, text classification.

## Introduction

The term *genre* designates a category of text (Graesser, Olde, & Klettke 2002). As with all categories, a genre cannot be specified by a qualitative analysis of a single exemplar (Davies & Elder 2004), but rather reflects the characteristics of a family of exemplars. A genre has an underlying set of norms that are mutually understood (either consciously or unconsciously) by the audience for whom the text was created (Downs 1998; Hymes 1972). Thus, it is the presence, prevalence, and prominence of the norms characterizing the genre of a text that allow the text to be recognized as an *interview*, a *lecture*, a *conversation*, a *story*, a *home page*, a *blog*, an *exposition* of some aspect of science, history, or art, or any other genre from a myriad of possibilities.

Any definition of genre would assume that the text in question is of a sufficient length for it to be classified on the bases of the features that accrue. Interestingly, we know of no study below the paragraph level in which genre has been deemed recognizable. Further, because genre has traditionally been viewed as a characteristic of the text (Biber 1988, Graesser et al. 2002), there is the implicit assumption that the texts in a genre have some degree of homogeneity. The common features of genre may be either absolute invariance (that is, the feature is necessary for a genre), but more frequently they are statistical regularities (i.e., the feature occurs more frequently in one genre than alternative genres). Whether the features are absolute or statistically distinctive, however, there is the question of how much and what type of information is needed to make a classification decision that a text is in a genre. That is, if a text T belongs to a genre G then the text itself is composed of sub-textual features (i.e. phrases, clauses, sentences) that are always or frequently diagnostic of genre G.

In this study, we investigate these assumptions by collecting data that explore six primary questions:

(1) How short (in terms of number of words) can a text be for its genre to be accurately recognized?

(2) What types of errors (i.e., genre misclassifications) do readers make when identifying genres?

(3) To what degree are texts heterogeneous (i.e., have characteristics of multiple genres)?

(4) Does the process of genre identification depend on reading skill?

(5) What textual features (e.g., syntax, lexical choice) influence genre identification?

(6) Can a computational model categorize genre using only as much text as humans appear to need?

**Psychological and Computational Goals of Study**

This study serves two primary purposes: one psychological, relating to reading comprehension; and one computational, relating to text classification.

**Reading comprehension.** Readers' comprehension of a text can be facilitated or otherwise influenced by the text genre, which is identified on the basis of the textual characteristics (Bhatia 1997; Graesser et al. 2002; Zwaan 1993). Given that familiarity with textual structure is an important facet of reading skill, training struggling readers to recognize text structure can help students improve their comprehension (Meyer & Wijekumar 2007; Oakhill & Cain 2007; Williams 2007).

Available research in discourse processing indicates that skilled readers utilize different comprehension strategies that are sensitive to text genre (van Dijk & Kintsch 1983; Zwaan 1993). Once a text genre is identified, it guides the reader's memory activations, expectations, inferences, depth of comprehension, evaluation of truth and relevance, pragmatic ground-rules, and other psychological mechanisms. For example, when reading a history text, it is important to scrutinize whether an event actually occurred. In contrast, in most narrative fiction, the truth of the event is not a particularly relevant consideration (Gerrig 1993), presumably because there is a "willing suspension of disbelief" (Coleridge 1985). Further, expository texts are more likely to discuss unfamiliar topics. Consequently, the lack of sufficient prior knowledge forces higher ability readers to process the details of the text at a more local level (e.g., connections between adjacent clauses). In contrast, narratives are more easily mapped onto everyday experience and, as a result, readers tend to process the global and thematic relationships in a passage (Otero, Leon, & Graesser 2002). Empirical evidence supports such claims through recall (Graesser, Hoffman, & Clark 1980) and reading time experiments (Graesser, Hoffman, & Clark 1980), demonstrating that narrative text is recalled approximately twice as well as expository text, and also read approximately twice as fast. Thus, stylistic surface structure attributes of the language and discourse vary in importance dependent upon the genre of the text (Zwaan 1993).

A better understanding of the nature of text genre is important for text comprehension theories as well as interventions to improve comprehension. If readers are using different strategies to process different genres of text, then it is important to understand the processes and information constraints during the course of genre identification. An understanding of the circumstances under which readers make correct or incorrect attributions of genre could expand our knowledge of the reading strategies used for each genre.

**Text classification.** According to the Netcraft Web Survey (December, 2007), the internet consists of *at least* 155,230,051 sites, an increase of 5.4 million sites since the previous month. And with many sites boasting 1000s of web pages, the number of web documents available to browsers is astronomical. With such an abundance of information, locating the desired information is becoming ever more problematic.

Search engines categorize web pages using *spiders*, which crawl through the internet, storing information embedded in web pages. Although each spider is different, the typical information gathered from web pages is based on high frequency words, key words in headers and links, and meta-tags that

specifically indicate terms of relevance. But despite such a broad approach (or maybe *because of* such a broad approach), the majority of web pages located from any given search are not relevant to the user. This overabundance of non-relevant documents is generally caused by such search criteria establishing not the genre of the text (e.g., blogs, home pages, narratives) but the *topic* of the text (e.g., politics; see Boese, 2005, McCarthy, Briner, Rus, & McNamara, 2007; Santini, 2006)

One approach to narrowing user searches is to locate documents based on genres (Stamatatos, Fakotakis, & Kokkinakis, 2001), and particularly web genres (Meyer zu Eissen & Stein, 2004; Roussinov et al., 2001; Santini, 2006). Definitions of *web* genres do not differ substantially from definitions of *text* genres. For instance, Boese (2005) argues that web genres are elements of the presentation of the article, effective analyses of the writing style, the formats or layouts of the documents, and the actual content of the articles. Roussinov et al. (2001) argues that web genres have socially recognized norms of format and purpose that appear in the text. Whereas traditional text genres may include expository, interview, conversation, and children's story, web genres subsume text genres and include others such as home page, opinion, review, course description, and blog.

The advantage of genre categorization over (or in addition to) topic categorization is one of focus. For instance, a Google search for the leading candidates in the 2008 presidential race for the White House (as of March, 2008), returned 1000s of web pages on the relevant *topic* (e.g., current affairs in the presidential race), but included *genres* as diverse as news, blog, review, research group, TV archive, and Q&A site. While all such genres may potentially provide the user with the desired information, it is safe to assume that most searches would be facilitated by the option or the availability of classification by genre.

A better understanding of the nature of web genres is important for search classification approaches. Improved knowledge of what constitutes a genre can lead to improvements in the efficiency of spiders. As a result of such improvements, categorizing searches by genres can help users by limiting and focusing the returned web pages, offering significant savings in time and effort.

**Our approach: Less is more?**

Within our six research questions, our *approach* to genre recognition focuses on the following two questions, previously unaddressed in the genre literature: (1) How long does a text have to be for it to be considered a member of a genre? And (2) To what degree are text genres heterogeneous; that is, is a text of one genre composed entirely of sentences that are also identified as being of that genre?

Regarding the first question, we can safely assume that a person who reads an entire book, article, or web page will have little doubt as to its genre. Similarly, we can assume that the first word alone from such a reading might not inspire great confidence that the correct genre will be identified. Our question is how much text is necessary for most readers to accurately identify the genre of a text.

Our first question is important in any model of genre identification, and comprehension in general. The sooner the reader identifies the genre of the text, the sooner the appropriate background knowledge will be activated and guide comprehension accordingly. We can also hypothesize that readers who recognize a text's genre earlier and more accurately possess more developed reading skills. That is, their experience or knowledge better allows them to recognize genre specific words or structures. Thus, it is conceivable that early and accurate genre recognition may be a diagnostic, practical estimation of reading skill.

Our second question regards the heterogeneity of text. While a given text T may be considered a member of genre G, we cannot assume that G is wholly composed of sentences from G. For instance, a science text may begin with a scene setting *narrative*, or a *history* of the theme to be considered. Similarly, a web blog may comprise (and indeed must comprise) *news* as much as *views*. Such a conjecture is highly related to our first question; that is, if sentences (or sub-sentences) are recognizable as genres, then what is the distribution of such genre-recognizable-fragments across text? A better understanding of the composition of genres may facilitate improving reading comprehension. For example, if lower grade-level science texts contained more narrative sentences (presumably a form more familiar to the readers) then the

expository information in the text may be more easily integrated.

Just as both our main questions address reading comprehension, they also address computational text classification. While categorizing web searches may facilitate users by focusing web page returns, the additional processing of documents may be prohibitive. Franklin (2008) reports that Google search engines operating at peak performance, using four spiders, could crawl at a rate of 100 pages per second. While such a performance is impressive, *peak* performance is not *typical* performance and with many billions of pages to crawl through, many billions of seconds are required. However, processing time can be significantly reduced by limiting the amount of text needed to be analyzed. For instance, early approaches to genre classification assessed the text as a whole (Biber 1988; Karlgren & Cutting 1994; Kessler, Numberg, & Shütze 1997), and this tradition continues into contemporary studies (Boese 2005; Bravslavski & Tselischev 2005; Finn & Kushmerick 2006; Kennedy & Shepherd 2005; Lee & Myaeng 2002, 2004; Meyer zu Eissen & Stein 2004; cf. Lim et al. 2005, for results on titles and meta-tags). But by considering genre as an identifiable feature at the sub-sentential level, perhaps only a small amount of text needs to be processed. If so, identifying the genre of a text and the heterogeneity of the text may be feasible with a relatively small (possibly random) sample.

## The Experiments

This study includes three experiments. Experiments 1 and 2 constitute the psychological portion of the study, focusing on our first four research questions: (1) How short (in terms of number of words) can a text be for its genre to be accurately recognized?; (2) What types of errors (i.e., genre misclassifications) do readers make when identifying genres?; (3) To what degree are texts heterogeneous?; and (4) Does the process of genre identification depend on reading skill? Experiment 3 constitutes the computational portion of the study, focusing on our remaining two research questions: (5) What textual features (e.g., syntax, lexical choice) influence genre identification?; and (6) Can a computational model categorize genre using only as much text as humans appear to need?

## Experiment 1

The goal of Experiment 1 is to investigate experts' ability to recognize the genre of sentence fragments presented out of context. Specifically, we examine whether three experts in discourse psychology agree on the genre classification for isolated sentence beginnings; and, if so, how many words are required for accurate genre classification to occur. Experiment 1 is limited in the number of participants because what might be described as our first *real question* is simply "is the task even possible?" Given that numerous psychological and computational studies have investigated genre using text no shorter than the paragraph, it is appropriate that our initial study is relatively modest in scope.

### The Genres

In this experiment (and throughout the study), we consider three genres: *narrative*, *history*, and *science*. We include science and narrative because they have been the focus of numerous previous psychological studies (e.g., Albrecht, O'Brien, Kendeou & van den Broek 2005; Linderholm & van den Broek 2002; Mason, & Myers 1995; Kaup & Zwaan 2003; Trabassao & Batolone 2003) and, therefore, provide a relatively uncontroversial point of departure. We include history because whereas no one disputes that science texts can be described as expository, there is a question as to whether history is more expository-like or more narrative-like. Some researchers, for example, have recognized that history texts can be similar to narratives, the two genres tending to be presented more as a chronological series of events on topics with which many readers are familiar (Duran, McCarthy, Graesser, & McNamara, 2007; Tonjes, Ray, & Zintz 1999). In contrast, other researchers (e.g., Radvansky, Zwaan, Curiel, & Copeland 2001) have used history texts as examples of expository texts, without any mention that such a genre could be considered narrative-like.

Empirical computational approaches to distinguishing the genres used in this study provide evidence for both categorizations: For instance, McCarthy, Graesser, and McNamara (2006) used an array of cohesion indices showing that history texts were more similar in structure to science texts. That is, both history and science texts were more cohesive than narrative texts. On the other hand, Duran et al. (2007) used temporal indices and found

evidence that history texts were more similar to narratives. That is, both history and narrative texts were structured similarly in terms of temporal development. Meanwhile, Lightman, McCarthy, Dufty, and McNamara (2007) found evidence for all three genres having distinct characteristics. Thus, one question addressed in this study is whether history sentences are correctly classified to a similar degree as narrative and science sentences; and if not, to which genre are they more likely to be assigned. As such, the choice of genres used throughout this study was motivated by two considerations. First, that the genres were sufficiently diverse in terms of structure, style, and purpose that differences in recognition accuracy would be identified; but second, that distinguishing the genres would not be a trivial task.

## Predictions

For the narrative genre, we predicted that incorrectly assessed sentences would more likely be classified as history sentences because both genres typically describe past events. For the history genre, we predicted misclassified sentences to be equally distributed between narrative and science, because history texts are equally likely to be descriptive of an event (thus, narrative-like) or feature explicit lexical cause and effect relationships (thus, science-like). For the science genre, we predicted that misclassified sentences would more likely be assessed as history sentences, because some elements of scientific texts present explanations from a chronological perspective.

We further predicted that our expert raters would correctly identify a high percentage of sentences requiring approximately only half of the words in a sentence to do so. This prediction is based on typical features of verb and pronoun positioning. Verbs, for example, feature early in a sentence, and their tense is indicative of their genre (McCarthy et al., 2008). Similarly, the subjects of sentences are generally positioned at the beginning of sentences. Regardless of whether the subject of the sentence is a pronoun or named entity, the characteristics of the sentence subject are at least somewhat indicative of text genre.

## Corpus

The corpus in our analysis was composed of a subset of sentences taken from the 150 academic text corpus compiled by Duran et al. (2007). In that corpus, the texts were sampled from 27 published textbooks provided by the MetaMetrics repository of electronic duplicates. A subset of the Duran and colleague's corpus (McCarthy et al., in press) further focused the corpus by filtering out an equal number of similarly sized paragraphs. The McCarthy and colleague's sub-corpus featured 207 paragraphs in total (828 sentences): 69 paragraphs in each of the three genres, and 23 paragraphs each of 3, 4, and 5 sentences in length. The approach we adopted for sentence selection from these paragraphs is based on studies indicating that topic sentences are processed differently to other sentences in a paragraph (e.g., Kieras 1978, Clements 1979, McCarthy et al. in press). Because such research also indicates that topic sentences are more likely to occur in the paragraph initial position (Kieras 1978; McCarthy et al. in press), we sampled an equal number of paragraph-initial sentences and paragraph-non-initial sentences. For the paragraph-non-initial sentences, we used the third sentence of each paragraph. This choice was made for two reasons. First, all paragraphs contained a third sentence; and second, third-sentences are presumably less closely related in terms of co-reference to first-sentences than first-sentences are to second-sentences; thus, the effects of a possible confound are reduced. This reduction to first-sentences and third-sentences left 414 candidate sentences in our corpus. To ensure that participants viewed sentences of approximately equal length, we further reduced the size of the corpus by only including all sentences that were within one SD of the average length in terms of number of words of the 414 candidate sentences (mean number of words = 15.437; SD = 7.113). Using this criterion, 298 sentences remained, of which the smallest group was 35 sentences belonging to the genre of narrative-paragraph-non-initial. We thus selected 35 to be the number of sentences from each of the six groups (narrative/history/science by paragraph-initial/paragraph-non-initial). Consequently, our corpus consisted of 210 sentences, equally representing the three genres and the initial/non-initial sentence dichotomy (see Appendix).

## Method

**Participants.** The participants included three researchers in discourse processing (one post-doc,

one graduate student, and one advanced and published under-graduate). Each participant assessed each of the 210 sentences that equally represented the genres of narrative, history, and science.

**Procedure.** A Visual Basic program was created to evaluate genre recognition. The program included three parts: *instructions*, *practice examples*, and *testing*. Following the instructions, participants were provided with six practice sentences. Once the practice was completed, a message informed the participants that the experiment would begin. Each participant evaluated all 210 sentences. The sentence order was randomized for each participant. The program operated by displaying the first word of the first sentence in a text window. Participants were required to assess the genre to which they thought the sentence fragment belonged. Participants registered their choice by clicking on one of four on-screen buttons: *Narrative*, *History*, *Science*, and *Don't Know*. As soon as a genre choice was made, the next word from the sentence appeared in the text window. All punctuation was retained in the display and was attached to the word it adjoined (e.g., in the sentence fragment *Yes, it was a …* the word *Yes* would appear as *Yes* + comma).

After 10 seconds, if the participant made no decision, then a new word automatically appeared in the text window with a message informing the participant of the new word. The variables of *genre choice* and *accuracy* were recorded. Participants evaluated each word of each sentence until they had either given the same decision of the genre of the sentence three consecutive times (whether right or wrong), or until all the words in the sentence were presented. The final choice of participants was recorded as the genre choice, regardless of previous decisions. For the variable *number of words*, the number was determined as the point of the first instance of a choice in a string of three consecutive identical choices. Thus, if a participant's genre selection was *don't know*, *don't know*, *narrative*, *science*, *science*, *science* then the count at the point of the first instance of *science* would be the number of words used: in this case, four words. That is, although the participant viewed six words in total, the participant's final choice occurred at the fourth word and was confirmed by the fifth and sixth selections.

## Results

### Raters

We begin our analyses by demonstrating inter-rater reliability. This reliability establishes confidence in our evaluation of the data as typical of expert ratings and is particularly important when using few raters. On average, the raters correctly identified the genre of the sentences for 90% of the data. Inter-rater agreement between Raters 1 and 2 for correctly assessed sentences was approximately 90% ($X^2$ = 41.077, $p$ < .001). Inter-rater agreement between Raters 1 and 3 was also approximately 90% ($X^2$ = 47.569, $p$ < .001). And the Inter-rater agreement between Raters 2 and 3 was approximately 91% ($X^2$ = 61.145, $p$ < .001).

Of the 210 sentences assessed, *all three* raters classified the correct genre for approximately 69% of data. Two of the three raters correctly classified an additional 17% of the sentences (i.e., 86% of the data). At least one of the three raters correctly identified an additional 6% of the data (i.e., 92% of the data). Also, less than 9% of the data were incorrectly assessed by any of the raters. Thus, the raters' accuracy was quite high. Further reliability of the raters' analyses can be demonstrated in terms of recall and precision (see Table 1). Such accuracy and agreement between the three raters (M=82%) offers support for the forthcoming analyses to be considered representative of genre recognition at the word level by experts in discourse processing.

**********

Table 1 about here

**********

### Genre

In the experiments presented throughout this study, the accuracy of the results is reported in terms of *recall*, *precision*, and *F1*. Such reporting is common when, as in this study, we are concerned with predictions of categories (i.e., narrative, history, science). To briefly explain each term, *recall* (R) shows the number of correct predictions divided by the number of true items in the group. In other words, recall is the number of *hits* over the number of *hits* + *misses*. *Precision* (P) is the number of correct predictions divided by the number of correct and incorrect predictions. In other words, precision is the number of *hits* divided by the number of *hits*

*+ false alarms*. The distinction is important because an algorithm that predicts everything to be a member of a single group will account for all members of that particular group (scoring 100% in terms of recall) but will also falsely claim many members of other group(s), thereby scoring poorly in terms of precision. Reporting both values allows for a better understanding of the accuracy of the model. The F1 value is the harmonic mean of precision and recall. It is calculated as *2PR / (P+R)*.

In terms of genre recognition accuracy, the expert raters correctly classified 516 of the 630 sentences (i.e., 210 sentences * 3 raters): an average accuracy of 82% (see Table 2). This result is in line with our prediction. While the results appear consistent across the genres (Min. F1 = 82, Max. F1 = 84), closer analyses suggest that the genres elicit quite distinct patterns of responses.

**********

Table 2 about here

**********

**Narratives.** The narrative genre received the highest recall value (89%); however the narrative genre was also the least precise (80%), with 47 additional false alarms. Indeed, of all misclassifications, more sentences were incorrectly assigned by the experts as narrative, than either of the two expository genres (narrative = 51%; history = 38%; science = 11%). The misclassifications to the narrative genre suggest that narrative sentence structures may be the most ubiquitous type. The approximately equal division of false alarm narrative sentences to the science (22) and history (25) genres further suggests that the two expository genres may comprise, to a small but notable degree, narrative-like sentences. Indeed, for six sentences (three history and three science) all three-raters categorized the sentences as narratives (see Table 3).

**********

Table 3 about here

**********

Looking more closely at these "misclassified" sentences, we observe that all three raters classified Example 1 as narrative by the 9th word of the sentence. It is only after this point that the words *Blackfoot chief* reveals the sentence more clearly as a history text. For Example 2, all three raters classified the text by the 4th word. Indeed, although the text recounts an historical event, the use of first person pronoun (rare in expository structures) may be indicative of a narrative style of writing. This appears again in Example 3. All three raters classify the sentence in Example 3 by the 5th word. Again, the incorporation of first-person pronouns renders the sentence more narrative-like, even though the text as a whole is taken from a history book. Example 4 is actually a sentence fragment and resulted in one rater having to view the entire sentence before deciding that it was narrative[1]. While the sentence lists symptoms of depression, the text could easily be read as describing a character. For Example 4, all raters agreed on narrative by the 5th word. However, had the raters read a little further, the science-like nature of the sentence (passive construction) may have been more easily recognized. The final example is deemed narrative by the 3rd word. It is possible that the raters saw the subject word *Watson* and considered the text to be from Sherlock Holmes. The results are in line with our predictions that the early presence of key lexical and grammatical features triggers the expert readers' genre recognition.

**History** As predicted, when history sentences were misclassified, they tended to be identified as narratives. This result supports the conclusions of Duran et al. (2007) and Tonjes et al. (1999). The three examples above (see Table 3) demonstrate the type of narrative-like text that appears to be a feature of history texts.

**Science** Only 75% of the science sentences were classified accurately, the lowest of the three genres. However, when raters did label a sentence as from the science genre then they were nearly always correct to do so (precision = 94%, the highest of the three genres). Of the 52 misclassified science items, most were attributed to history (25) and narrative (22). The high history value is as predicted, because much scientific discussion begins from a historical perspective. The equally high narrative value suggests that science texts may be equally viewed as

[1] This sentence was subsequently modified for later experiments.

narrative-like in the description of many of their topics.

**Don't Know** As predicted, the raters correctly identified the vast majority of items. Only 22 sentences remained unclassified with no particular domain attracting more *Don't Know* classifications. Only one sentence was rated as *Don't Know* by all three raters: *Many of those years were harsh and cruel.* Although from a history text, the sentence could equally well be attributed to narrative given that the author seems to be voicing an opinion rather than an objective fact.

### Number of Words Used

High inter-rater reliability is required to establish confidence that the number of words used by raters to assess the genre of sentences is suitably representative of experts' judgments. Following Hatch and Lazarton (1991), the adjusted correlation for three raters was $r = .660$, $p < .001$. For items for which *all three* raters correctly assessed the genre of the sentence, the correlation was $r = .732$, $p < .001$. The consistency across raters means that we can take the average number of words used by raters as the gold-standard representative of experts in assessments of the genre of sentences.

For the corpus as a whole ($N = 210$), the average number of words used by raters was 4.948 ($SD = 2.818$; Mode = 5). As predicted, this is less than half the average length of sentences in the corpus; indeed, it was a *third* of the length. However, when we divide the corpus for the condition of *all raters giving correct judgments*/*other sentences*, the results show that significantly fewer words were required to *correctly* identify the genre (Correct: $N = 144$, $M = 4.419$, $SD = 2.407$; Incorrect: $N = 66$, $M = 6.101$, $SD = 3.256$; $F(1,208) = 31.140$, $p < .001$, $\eta^2 = .130$). This result suggests that a rater judgment of *fewer* than five words is more likely to be correct, and a judgment of *greater* than five words is more likely to be *in*correct. The three sentences for which raters took the most words to arrive at the *wrong* genre are shown in Table 4.

**********

Table 4 about here

**********

To better understand the above result, we considered each genre individually. The results suggested that the five-word average applied only to narratives (Correct: $N = 187$, $M = 4.808$, $SD = 3.029$; Incorrect: $N = 23$, $M = 7.870$, $SD = 4.808$; $F(1, 208) = 18.028$, $p < .001$). There was no significant difference for correctly identifying genre using fewer words for the genres of history or science. The similarity here between the history and science genres and the distinction from narrative genre offers support to the conclusions of Graesser et al. (2002), McCarthy et al. (2008) and McDaniel et al. (1986). The result offers evidence that if an expert reader of a narrative sentence has not become sufficiently aware of the sentence's genre by the fifth word that it is unlikely that subsequent words will make the reader any the more sure of the genre.

### Discussion

In Experiment 1, we asked three experts in discourse processing to identify the genre of isolated sentences culled from a corpus of narrative, history, and science texts. Demonstrating high agreement, the raters showed that expert readers could significantly identify the genre of over 80% of sentences. Further, our raters demonstrated that fewer than five words (less than a third of the sentence) were required to correctly classify these sentences. Indeed, for the narrative sentences, viewing more than five words did not improve the accuracy of identifying the genre. These results suggest that the first third of sentences alone contains sufficient genre characteristics for skilled readers to begin the process of activating knowledge of text structure: a process which facilitates comprehension.

Our results also showed that expert readers viewed many of the history and science sentences as narrative, suggesting that expository texts tend to comprise a notable number of narrative-like sentences. On the other hand, regardless of the genre from which sentences were taken, our raters were least likely to classify sentences as science. This result sheds like on the heterogeneous compositionality of text, providing significant implications for computational research in genre recognition. Specifically, computational approaches to genre recognition have tended to assume that the text as a whole is representative of the genre or text-type to which it has been assigned (e.g., Biber 1988,

Louwerse, McCarthy, McNamara, & Graesser 2004). The results of Experiment 1 suggest that texts of any given genre may typically comprise sentences from many other genres. Understanding this diverse compositionality may lead to changes in how computational tools assess text searches and evaluations.

The compositionality of text is also a factor for research in reading development. Our results here suggest that for a text to be suitably representative of any given genre, it may require that the text contains a notable number of sentences more indicative of other genres. If a text does not contain this mixture of genre sentences, it is possible that a reader may have greater difficulty processing the text, as certain expectations may not be met.

In Experiment 1, we also addressed the question as to whether the genre of history was closer to science or to narrative. Our results suggest that expert readers are as able to identify and distinguish history sentences as they are science and narrative sentences. This result supports the findings of Lightman et al. (2007), who found that history texts were distinct from both science and narrative texts. However, if we consider only the 39 misclassified sentences of the history genre, our results showed that 64% of these sentences were incorrectly assigned by our experts as narratives, whereas only 18% of the sentences were identified as science (and the remainder as *don't know*). Viewed this way, the result suggests that a notable portion of history texts comprise narrative-like structures, a result that supports Duran et al. (2007), who found that history texts were more narrative-like than science-like. The categorization of history texts is important to cognitive science as many experiments have assumed that a history text is an expository text (e.g. Radvansky et al. 2001). Consequently, researchers can often assume that history text will lead to *similar* results as science text and *different* results from narrative texts. The results of Experiment 1 demonstrate that such an assumption could lead to erroneous conclusions.

Above all, the results of Experiment 1 demonstrate that genre recognition at the sub-sentential level is possible. There having been no previous investigations of how much text is required to recognize genre, this first experiment indicates that very little text is actually required and that readers most likely activate information about text structure very early in the reading process. Such recognition might provide a signature of reading ability, and as a consequence, a method of assessing reading ability. The principle results of Experiment 1 certainly provide sufficient initial evidence that such an approach is viable and that this paradigm can be further explored as an assessment of reading skill. In addition, if only the first five words of a sentence is sufficient for experts to recognize the text's genre, then computational approaches to text analyses may need to follow this lead. That is, text assessment for such features as readability, difficulty, cohesion, and genre recognition may also need to be performed on just the first third of sentences because it is here that a significant portion of human evaluation of the text seems to occur. More specifically, computationally evaluating an entire sentence may incorrectly assess the sentences' remaining two-thirds as relevant to the reader's processing. Indeed, this remainder may be redundant or even noise in terms of reader activation of certain processing components. In Experiments 2 and 3 we explore these issues more closely.

## Experiment 2

In Experiment 1, three experts (i.e. published authors) in discourse processing were asked to identify the genre of isolated sentences culled from a corpus of narrative, history, and science texts. The experts had high inter-rater agreement (min = 90%) and required about a third of the words in the sentence to accurately identify genres (accuracy as measured by F1, Narrative = .82; History = .84; Science = .82). The results further showed that these experts often classified history and science sentences as narrative, suggesting that expository texts tend to be composed of a notable number of narrative-like sentences. On the other hand, science-like sentences were the least likely to be misclassified into other genres, suggesting the science-like sentences seldom occur in the non-science genres. The results also showed that these skilled readers required about a third of the sentence to successfully activate sufficient knowledge to recognize textual genres. Presumably, this activation skill is beneficial to reading and comprehension development. As such, we might expect that the number of words necessary to correctly recognize genres to be indicative of reading ability.

The results of Experiment 1 were intriguing. However, the most compelling result was the one informing us that genre recognition at the sub-sentential level was, indeed, possible. To establish greater confidence in our paradigm, Experiment 2 builds on Experiment 1 by including a larger sample of participants, an independent assessment of reading ability, a measure of *time on task*, and recording accuracy in terms of *number of words* used. In this experiment, we ask four main questions. First, how quickly (in terms of number of words) do readers identify the genre of a text? Second, what types of errors (i.e., genre misclassifications) do readers make when identifying genres? Third, does the process of genre identification depend on reading skill? And fourth, how does *time on task* affect the accuracy of genre decisions?

## Corpus

The corpus used in Experiment 2 was the same as that used Experiment 1, with the following modification: We modified one science sentence that was a sentence fragment, changing *Taking no joy in life, looking forward to nothing, wanting to withdraw from people and activities* to *Examples are taking no joy in life, looking forward to nothing, wanting to withdraw from people and activities*.

## Methods

*Participants.* There were 22 participants (Male = 10, Female = 12; *M* = 24.1 years old) who received $50 in exchange for participation in two experiments, of which, this was one. The other experiment was unrelated. All participants were native English speakers. Fifteen participants were undergraduate students, five participants were graduate students, and two participants identified themselves as non-students.

*Assessments.* To assess reading skill, we used the Gates-MacGinitie (GM) reading test, a multiple-choice test consisting of 48 questions designed to measure reading comprehension. We used the level 10/12 version of the test, which has a reliability of .93 (MacGinitie et al, 2002).

Participants' genre recognition was evaluated using a similar Visual Basic program to that used in Experiment 1. Three variables were recorded: *genre choice*, *accuracy*, and *time on task*. To accommodate the *time on task* assessment, the following modification from Experiment 1 was made: As in Experiment 1, participants made their selection by clicking on one of four on-screen buttons: *Narrative*, *History*, *Science*, and *Don't Know*. However, in Experiment 2, the buttons' position was randomized such that the genre choice could appear in any of the four buttons. Upon selecting one of the buttons, the mouse cursor returned to a central position so that each button was always equidistant from the start point of the cursor. As soon as a genre choice had been made, as in Experiment 1, the next word from the sentence appeared in the text window.

## Results

### Subject Analysis

Our results showed that participants typically needed only a sentence's first three words to make their decision on genre (overall words used: $M$ = 3.35, $SD$ = 1.50; words used in correct assessments only: $M$ = 3.33, $SD$ = 1.45).The average accuracy of genre categorization was high (Recall: 0.86; Precision: 0.71; F1: 0.77), and this accuracy was consistent across the three genres (see Table 5). These results are consistent with Experiment 1.

**********
Table 5 about here
**********

While the average number of words used for correct assessments was 3.33, the mode for number of words used in correct assessments was 1.00 (25.02% of the data, see Table 6). The second highest frequency for number of words used was 2.00 (21.88%), followed by 3.00 (15.36%), and so forth such that the distribution of words used for correct assessments described a logarithmic curve (df = 16, $F$ = 244.95, $p < .001$, $r^2$ = .939). Such a result is unlikely to mean that participants blindly hit the same genre choice button, because the genre buttons randomly changed position, meaning that participants had to find their genre choice. Additionally, the result is unlikely to suggest that participants were simply trying to get the task done as quickly as possible because examining *all* final decisions made on the first word (in other words, decisions for which participants had selected a genre on the first word and selected that same genre for

the second and third words), 50.69% of the genre decisions were correct (baseline = 33.34%). As such, there is some evidence here that humans make their genre decision on the very first word of a sentence, and more often than not their decision is correct.

**********
Table 6 about here
**********

The magnitude of the correlation between *reading skill* (GM) and *words used* was moderate ($r$ = .37, $p$ = .09), as was the relationship between *words used* and *accuracy* (in terms of correlations with F1 participant evaluations, Science: $r$ = .43, $p$ < .05; Narrative: $r$ = .37, $p$ = < .09, History: $r$ = .37, $p$ < .09). We examined the results more closely by dividing the participants into two groups based on a mean split of the Gates-MacGinitie test scores ($M$ = 24.00; $SD$ = 9.14). Using these values, 13 participants were designated as lower-skill (LS) and 9 participants were designated as higher-skill (HS). Differences in Gates-MacGinitie test scores were analyzed using Levene's test for equality of error variances. No significant differences between groups were detected ($p$ > 0.5), indicating that the groups are suitable for comparison.

We conducted an exploratory Analysis of Variance (ANOVA) to determine which of 22 variables best distinguished the reading skill groups. The analysis revealed that 7 variables significantly distinguished the two skill groups ($p$ < .05) and 4 variables were marginally significant ($p$ < .10; see Table 7).

**********
Table 7 about here
**********

The *narrative-precision* variable suggests that higher-skilled readers tend to be better at *not* classifying non-narrative sentences as narratives. In other words, skilled readers know better when a sentence is *not* a Narrative. These readers' greater accuracy may be because they are prepared to use more words than the lower-skilled readers. However, a t-test revealed no significant differences between the number of words required by lower-skilled readers ($M$ = 2.97; $SD$ = 1.21) and higher-skilled readers ($M$ = 3.85; $SD$ = 1.68), $t$ > 1.0, $p$ > .1. Despite the lack of a significant difference between the higher-skilled and lower-skilled readers in terms of words used, the direction of the difference suggests that lower-skilled readers may too easily assume the direction or nature of the sentence discourse.

The variable, *time on task for the 3rd word in history sentences,* indicates the time on task for judging the third word of history sentences for correct decisions. Lower-skilled readers took significantly *more* time on this word. Indeed, *time on task* negatively correlated consistently with GM reading skill across all three genres for both 2nd words of sentences (Narrative: $r$ = -.427, $p$ = .05; History: $r$ = -.443, $p$ = .04; Science: $r$ = -.523, $p$ = .01) and 3rd words of sentences (Narrative: $r$ = -.596, $p$ < .01; History: $r$ = -.606, $p$ < .01; Science: $r$ = -.500, $p$ = .02). These results suggest that higher-skilled readers may be able to more quickly integrate new information.

Taken together, the results suggest that higher-skilled readers are more able to quickly and accurately process sentential information, using as few as the first three words. This advantage appears most evident in two features: on the 3rd word of sentences (all other word positions demonstrated weaker results); and in the precision result for the narrative genre. One further variable of interest is that higher-skilled readers may be prepared to use more words before making genre decisions. This final point is consistent with Experiment 1 in which expert readers (and therefore, presumably higher in ability than those who participated in this Experiment) tended to use at least two more words than those who participated here. However, caution should be taken with this conclusion because a step-wise multiple regression revealed that only the time on task for 3rd words of history sentences variable contributed to the model (adjusted R-square = .336).

**Item Analysis**

Of the 210 sentences in Experiment 2, only 4 (2%) failed to be correctly evaluated by any of the participants. For instance, the history sentence "*I had vainly flattered myself that without very much bloodshed it might be done"* was evaluated by all participants as a narrative; and the science sentence "*Hindi is the most widely used, but English is often spoken in government and business"* was evaluated

by 20 participants as history and by 2 as narrative. A further 33 sentences (16%) were correctly categorized by all the participants. For instance, the narrative sentence "*Why, I wouldn't have a child of mine, an impressionable little thing, live in such a room for worlds*" resulted in no misclassifications. For over half the sentences (55%) at least 19 of the 22 participants correctly evaluated the genre. For instance, the science sentence "*In areas with hard water, many consumers use appliances called water softeners to remove the metal ions"* recorded only three misclassifications. Conversely, only 10% of the sentences received less than 6 correct evaluations, an example being the narrative "*The Empress of Russia looked dressed for war, Igor thought.*"

The item analysis also showed that the sentences that received the highest accuracy in terms of categorization were likely to require fewer *words* for such categorization to be made. Thus, there was a negative correlation between the percentage of participants who correctly evaluated a sentence and the number of words needed to correctly categorize the sentence ($r$ = -.639, $p$ < .001). For example, "*Chemical weathering processes change the chemical composition of rocks"* was correctly identified as a science sentence by all of the participants and required an average of only 1.23 words to be identified. In contrast, "*However, this process was too slow to satisfy the Renaissance demand for knowledge and books"* was correctly categorized by only 27% ($n$ = 6) of the participants and required 10 words to be correctly identified as a history sentence.

The results of the *time on task* demonstrated similar results. Specifically, there was a negative correlation between the percentage of participants who correctly assessed a sentence and average *time on task* for assessment ($r$ = -320, $p$ < .001). The results for both *words used* and *time on task* were consistent across the genres of narrative (words: $r$ = -.613, $p$ < .001; time: $r$ = -.466, $p$ < .001); history (words: $r$ = -.701, $p$ < .001; time: $r$ = -.404, $p$ < .001); and science (words: $r$ = -.578, $p$ < .001; time: $r$ = -.257, $p$ = .034).

Thus, consistent with the results of Experiment 1, viewing more words does not lead to greater genre classification accuracy. This result indicates that if a sentence does not contain genre-specific features early in its structure, then it is also unlikely to contain those features later in its structure. The results for *time on task* indicate that sentences that are more accurately classified are also more quickly classified. We can presume that the quicker the decision, the less the processing necessary to make the correct decision. Thus, we did not observe a time/accuracy tradeoff.

Collectively, the results suggest that most sentences from the three genres can be accurately categorized in relatively few words and relatively little time. However, the variation within this accuracy suggests a continuum of *sentence-categorization difficulty*. That is, the first few words of sentences can often be sufficiently non-prototypical or ambiguous to reduce the likelihood of correct reader categorization. As such, it is feasible that the construction of the initial aspects of a sentence may significantly affect sentence processing, with less prototypical constructions causing readers to activate less relevant expectations of prior knowledge.

**Discussion**

In Experiment 2, 22 participants identified the sentence genres of 210 sentences. The results indicated that both higher- and lower-skilled readers used about three words to accurately identify genres. Two primary variables related strongly to participants' reading ability: *Narrative-precision* and *Time on Task* for the 3rd word (i.e., typically the word with which participants make their decision). Thus, higher-skilled readers are less likely to think a sentence is a narrative when it is not, and they also require less time to make their decisions.

Taken together, the results of Experiments 1 and 2 allow us to make the following conclusions. The results suggest that 1) a wide range of readers can accurately categorize genres at the sub-sentential level; 2) as few as the first three words of a sentence may be all that is required for that assessment to occur, and in over half the cases just the very first word; 3) genre recognition may be indicative of reader ability; and 4) variables such as *time on task*, *accuracy*, and *number of words used* may be the indicators of reading ability.

The research presented in these initial two experiments offers an interesting and promising direction toward a better understanding of how genre knowledge is represented in the mind and subsequently activated. We plan to use this

understanding to better establish our *genre identification paradigm* as an assessment of reading skill, and even as a possible intervention for reading development. While much remains to be done in this respect, the results presented here offer an exciting new perspective on the nature of text and the possibilities of reading skill assessment.

## Experiment 3

### Introduction

The results of Experiments 1 and 2 provided evidence that genre recognition could be accomplished with a high degree of accuracy using as few as the first three words of sentences. Given such accuracy from such little discourse information, we can hypothesize that readers are utilizing shallow lexical and syntactic sentential features to identify genre. To address this hypothesis, we examined whether a computational model based on only lexical and syntactic features (i.e., the information apparently used by participants) provided similar results. If the model could replicate the results found with humans, then it potentially provides evidence that participants use such sentential features when processing text.

In Experiment 3, we construct a computational model based on our results from Experiments 1 and 2. We use the model to investigate what information could be present in the initial words of sentences such that it can provide participants with sufficient information to make a genre evaluation. The question of whether or not we could build a computational model is important for two reasons. First, our computational model sheds light on the features of the text that most likely influences readers' genre classifications. And second, if a computational model can categorize genre using minimal sentence information, then such an approach could facilitate text classification systems.

### Computational Approaches to Text Classification

Computational approaches to categorizing genre have tended to treat text as a homogeneous whole. Thus, the whole text is analyzed and, based on the results, the text is categorized as a single genre. Such an approach is as common in traditional text-genre classifications studies (e.g., Biber 1987; Biber 1988; Duran et al. 2007; Hall, McCarthy, Lewis, Lee, & McNamara 2007; Karlgren & Cutting 1994; Louwerse et al. 2004; McCarthy, Graesser, & McNamara 2006; McCarthy, Lewis, Dufy, & McNamara 2007) as it is in web-genre classification studies (e.g., Boese 2005; Bravslavski & Tselischev 2005; Finn & Kushmerick 2006; Kennedy & Shepherd 2005; Lee & Myaeng 2002, 2004; Meyer zu Eissen & Stein 2004).

For example, in traditional genre classification studies, Biber (1987) identified lexical diversity and singular person pronoun use as key predictors in distinguishing British-English from American-English. Kessler, Nunberg, and Schutze (1997) used part of speech tags, lexical cues (e.g., Mr. and Mrs.), punctuation features, and shallow discourse features such as sentence length, to distinguish registers such as editorials, romantic fiction, and biographies. Louwerse et al. (2004) used cohesion values to distinguish both spoken from written texts and narratives from non-narratives. And Stamatatos, Fakotatos, and Kokkinakis (2001) used various style markers such as punctuation features and verb- and noun-phrases frequencies to distinguish between the authors of a variety of newspaper columns. What each of these studies have in common is that the whole text is analyzed and, based on a distribution of features, is labeled as a member of a single category.

Meanwhile, the more contemporary web-genre identification studies typically rely on three categories of features: *style*, *form*, and *content* (Boese & Howe, 2005). Style includes readability formula (e.g. Flesch Kincaid Grade Level), syntactical information (e.g. passives/actives), and various heads of phrases such as the articles or prepositions that precede noun-phrases. Form includes such aspects as frequencies of paragraphs, emphasis tags, images, and links. And content includes such aspects as bags of words, stop-lists, number types, and closed-word sets. Whichever features, or combination of features are used, it is still typical that the whole text is analyzed and subsequently categorized into a single genre.

Whole text approaches tend to be successful because different categories of texts comprise different types and quantities of features. And, to be sure, such approaches have yielded impressive results, finding significant distinctions in categories as diverse as dialect, mode, domain, genre, and author. However, to take some slightly more arcane examples, Miliv and Slane (1994) distinguished

narratives from treatises by way of the letters D and S respectively; Gordon (2004) identified the penultimate chapter of Joyce's *Ulysses* by the incidence of the letter C; and Šatava (2006) explains that the Võro-Seto ethnolect differs from standard Estonian by way of the letters Q and D respectively: the nominative plural in Estonian featuring a glottal stop, which is marked by the letter Q in Võro-Seto and the letter D in standard Estonian. Such examples may seem churlish but they serve to demonstrate that distinguishing texts, in and of itself, is not difficult, given enough texts and enough variables (and, presumably, enough researchers).

**Our Approach**

The possible problems with the approaches listed above are ones of *time* and *compositionality*. With regard to time, search engines operating at *peak* performance can only assess the multiple billions of web documents at the rate of 100 pages per second (Franklin 2008). Assessing whole documents over multiple variables may simply be too computationally expensive; thus, there is a significant trade off between time and accuracy. Of course, technology is constantly improving, and consequently, time may become less of a factor. However, by the same token, it could be equally argued that the expansion of the Internet (around 5 million web sites per month) could easily outpace any advances in technology.

With regard to compositionality, our results from Experiments 1 and 2, suggest that texts are heterogeneous in terms of genres. Indeed, the heterogeneous nature of text is well established (e.g., Kintsch & van Dijk 1978; Mann & Thompson 1988; McCarthy, Briner, Rus, & McNamara 2007; Propp 1968, Teufel & Moens 1991; Swales 1990). And this heterogeneity research extends to multiple-genres within texts (see Bazerman 1995; Crowston & Williams 2000; Orlikowski & Yates 1994). Researchers such as these point to *embedded genres* and *genre systems* wherein a single text may feature multiple genres as in *memos*, which may contain proposals; *trials*, which include examination and cross-examination; *expository* texts, which may include histories, *narratives*, which may include factual claims, and *blogs*, which may include factual accounts, stories, and reviews.

In our approach, both time and compositionality are considered. However, primarily, we base our approach on our psychological findings on genre recognition from Experiments 1 and 2. Our results from these experiments suggest that humans are able to classify narrative, history, and science genres using as few as the first three words of sentences. The results suggest that sentence-level syntax and word-level frequency features may be sufficient for accurate and reliable genre recognition to occur. In our approach, we present a computational model for genre classification based on these findings. That is, we ask: *can a computational model using less than a third of the words in a sentence accurately classify genre using only word-level and syntactical information*? If such an approach is successful, then issues of time and compositionality can be addressed. Our approach would address issues of time because, feasibly, we could imagine a system that samples just a few sentences (or parts of sentences) from the target text. In requiring such a small sample, computational expense is reduced. Our approach would address issues of compositionality because, feasibly, we could imagine the system returning results as to the genre distribution of the samples. That is, perhaps 80% of the samples are science, 15% history, and 5% narrative. Such a result not only informs us of the main genre of the text, it also indicates potential levels of readability or difficulty of the text.

Of course, bringing the discussion above to fruition requires considerable research. And in Experiment 3, we take just the first step towards our goal. Namely, we analyze the sentences from Experiment 2 to assess what degree of accuracy we can expect when using solely the portion of a sentence that humans require for genre recognition.

**Methods**

To address our computational question, we conducted a number of basic assessments, suitable for sentence level analysis, using the first *three words*, *five words*, and *whole sentence* for each sentence in the corpus. For the lower bound of sentence fragment length, we selected the conservative size of the first three words of the sentences because this was the lowest average number of words for any of the groups from Experiment 2: (i.e., the lower-skill group: $M = 2.98$ words, $SD = 1.24$). For the upper bound of sentence

fragment length, we selected the whole sentence to serve as a baseline.

To conduct our analysis, we used as our dependent variable the genre of the sentences as determined from their original source (narrative, history, science). Our independent (or predictor) variables were calculated using the web-based computational tool, Coh-Metrix (Graesser, McNamara, Louwerse, & Cai 2004) and included *word frequency values* (from the Celex data base, Baayen, Piepenbrock, & van Rijn 1993), *word information values* (from the MRC data base (Coltheart 1981), and *parts of speech frequency counts* (Charniak 2000). In addition, we also included a *syllable count (www.wordcalc.com)*.

The object of the analysis was to ascertain how well the independent variables (i.e. information similar to that which humans might have available) were able to predict the categories of the sentence fragments. One way of achieving this goal is to conduct a series of *discriminant analyses*. A discriminate analysis is a statistical procedure, culminating with a prediction of group membership (in this case, genre) based on a series of independent variables (in this case, the word and syntax variables mentioned above). To guard against issues of over-fitting and colinearity caused by applying multiple predictor variables, we followed established procedures of training and testing the algorithm (see Witten & Frank 2005; McCarthy et al. 2007). Thus, the corpus was randomly divided into a training set (67%) and a test set (33%). Using the training set, we conducted an analysis of variance (ANOVA) to identify and retain only those variables that significantly distinguished the genre groups. We then conducted correlations among these variables and eliminated variables that presented problems of colinearity using $r \geq .70$; the variable with the higher univariate F-value was retained and the lower eliminated. Of the 16 remaining variables, the 14 with the highest univariate F-values were used in a discriminate analysis; there was an item to predictor ratio of 10:1. This procedure was then repeated for data collected from the *five words* the *whole sentence* conditions (see Table 8).

**********

Table 8 about here

**********

Having established the predictor variables, we used the *training set* data to generate our discriminant function (the algorithm that calculates the prediction of group membership) and we used those generated predictions on the *test set* data to calculate the accuracy of our analysis. Thus, if the results of the discriminant analysis are statistically significant, then we can claim to have evidence that validates the initial analysis. Such a validation affords application of the model to other text corpora of a similar nature. In this study, as is typical of discriminant analysis studies and as is consistent with previous analyses in this study, the accuracy of the results are reported in terms of recall, precision, and F1.

The results of the discriminant analyses were significant (3-words: $\chi^2 = 33.689$, $p < .001$; 5-words: $\chi^2 = 30.127$, $p < .001$; whole sentence: $\chi^2 = 71.704$, $p < .001$). The accuracy of the models in terms of recall, precision, and F1 were comparable to human results (see Tables 9, 10, and 11). The results suggest that as few as the first three to five words of a sentence contain enough *syntactic* and *word level information* to distinguish between genres.

**********

Tables 9, 10, 11 about here

**********

The three-word model is most impressive at identifying narratives (all data F1 = .70, human = .77) and reasonable at identifying science (all data F1 = .65, human = .76). The three-word model appears weakest at identifying history (all data F1 = .52, human = .72). The five-word model returns similar results although the history identification is improved (all data F1 = .59, human = .72). The whole-sentence model returns human like results for all three genres (narrative: all data F1 = .73, human = .77; history: all data F1 = .66, human = .72; science: all data F1 = .76, human = .76). The results suggest that with some modifications to the model (e.g., genre related frequencies) that a highly accurate sub-sentential genre identification model is feasible.

**Discussion**

In Experiment 3, we developed and tested a computational model of human genre recognition at the sub-sentential level. Our results suggest that basic sub-sentential features such as *parts of speech* and *word frequencies* significantly distinguished between genres. Further, the success of our computational model suggests that the features of only the first three to five words are sufficient for this classification.

The results of our model are particularly impressive when considering humans' advantages when recognizing genre in comparison to our model. For example, the computational models did not contain information about semantics and word knowledge, which humans would likely use when recognizing text genre. Thus, when participants see a number such as 1776 they are presumably more able to interpret this as an historical date. Second, even though word frequency was included as a predictor, the results are based on *frequencies in general* rather than genre specific. We can hypothesize that word information relevant to specific genres would enhance the accuracy of the prediction. For instance, we might assume that participants have knowledge that *cannon* is a word associated with history whereas *nucleus* is a word associated with science. Third, we might further hypothesize that our model could be improved if frequencies were calculated from only the sentence-initial fragment. Thus, words such as *to*, operating as an infinitive marker as in *to understand this process* …, may be more indicative of expository text.

While the results of Experiment 3 suggest that word and syntax variables may be all that humans (and computational models) need to recognize genre, this does not mean that more complex discourse variables such as cohesion variables and temporal features are not a characteristic of genre differences. However, our results do indicate that readers can and do make genre decisions before such features become available. Such a result is important when considering a light and efficient approach to genre categorization where computational expense is an important factor.

Finally, the results of our three-word model are impressive; although we cannot claim that the model is as good as human performance. We have given modifications above for improving our model, but it is still worth noting that some features of our model do match human performance. For instance, the narrative precision evaluation for the test set (.74), all data (.73), and for participants (.71) are highly similar. Given that Experiment 2 showed that the human narrative precision variable correlated highly with reading skill ($r$ = .520, $p$ = .002), it is reasonable to assume that the computational model might reflect some aspects of reader strategy, at least in its propensity to correctly reject non-narrative decisions for narrative sentences. Additionally, the model's false alarms for narratives were similar to those decisions made by humans: that is, false alarms were less likely to be science decisions (History = 11; Science = 6).

## Final Discussion

This study included three experiments designed to address issues concerning genre recognition. Experiments 1 and 2 addressed issues concerning (1) how many words were necessary to constitute human recognition of genre, (2) to what degree were the texts heterogeneous in terms of genre, and (3) to what degree was genre recognition a predictor of reading ability. In Experiment 3, we used the information gathered from the previous two experiments to create a computational model for genre recognition.

## Research Questions

Our study began with six research questions. Here, we briefly summarize the responses to those questions based on the current research.

1. *How short (in terms of number of words) can a text be for its genre to be accurately recognized?* Using a baseline of 33%, most readers (77%) can accurately recognize genre within the first three words of a sentence. Many readers (50%) can accurately recognize genre using just the first word. This result suggests that genre is (also) a sub-sentential feature of text.

2. *What types of errors (i.e., genre misclassifications) do readers make when identifying genres?* The most common type of genre misclassification appears to be the assigning of narrative to non-narrative text. We can presume from this result that readers are more familiar with the features of narratives and tend to make a default assumption that a text is narrative unless shown to

be otherwise. We can hypothesize that explicit training in recognizing non-narrative features may facilitate reader comprehension if it facilitates earlier and more accurate genre recognition.

3. *To what degree are texts heterogeneous?* Our results suggest that texts are about 83% homogenous in terms in genre. For the remainder, narratives tend to comprise mostly history sentences; histories tend to comprise mostly narrative sentences; and science texts tend to comprise an even number of narrative and history sentences. We can hypothesize that variation in the heterogeneity of the text may benefit some readers more than others based on their knowledge or skill level. For instance, we can presume that lower skilled/knowledge readers of science texts would be facilitated by a higher incidence of narrative/history sentences because the sentences features used in this genre are likely to be more familiar.

4. *Does the process of genre identification depend on reading skill?* Our results suggest that higher-skilled readers are less likely to think that a sentence is a narrative when it is not, and they also require less time to accurately recognize genre. These results lead us to believe that a reading skill assessment based on genre recognition is viable.

5. *What textual features (e.g., syntax, lexical choice) influence genre identification?* Our results suggest that such features as presence of *past tense*, *length of words*, and *word frequencies* offer readers substantial indication of genre at the sub-sentential level. This result is important for designing and modifying computational approaches to genre classification, as well as forming part of the training for an intervention approach to helping students with reading skills.

6. *Can a computational model categorize genre using only as much text as humans appear to need?* The results of our computational models were statistically significant and comparable to humans. The three-word and the five-word models were most impressive at identifying narratives and science. The whole-sentence model returned human like results for all three genres. We hypothesize that improvements to our word frequency database and using genre-specific word frequencies would significantly improve the computational model. Overall, the results provide a good deal of confidence that computational genre recognition is achievable using only as much sentential information, as is required by humans.

**Limitations of our study**

While we would argue that the results presented in this study offer a significant contribution to research in genre recognition, the limitations of the study are worth acknowledging. First, having only considered three traditional text genres, we cannot be sure how such analysis would scale up to a finer grained analysis of genres such as those encountered on the internet. In addition, the genres used in the study were presented to participants as their one and only choice. It is possible that participants may have preferred to make multiple choices or categorized sentences in genres other than those we stated. Second, we cannot be sure that the research presented here suitably distinguishes *topic* from *genre*. Addressing this issue is of significant importance to future research. Third, the numbers of sentences and participants in our experiments are relatively small. Such limitations are common in initial forays into new research; however, given such numbers, we must be cautious as to the conclusions we draw. Fourth, while our computational models showed promise, and while our extensions to these models seem reasonable, there is considerable work to be done if we are to establish that such an approach can produce a desirable accuracy while minimizing computational expense.

**Conclusion**

The research presented here offers an interesting and promising direction toward a better understanding of genre recognition. In psychological terms, we plan to use this research to better establish our *genre identification paradigm* as an assessment of reading skill, and even as a possible intervention for reading development. In computational terms, our results suggest that text classification model at the genre level is possible using only a limited selection of text fragments. Such an approach offers the possibility of fast and accurate genre classification as well as information as to the genre distribution within a text. While much remains to be done, the results presented here

offer a new and exciting perspective on the nature of text, the possibilities of new assessments of reading skill, and an intriguing and novel approach to computational text classification.

**References**


Albrecht, J. E., O'Brien, E. J., Mason, R. A., & Myers, J. L. (1995). The role of perspective in the accessibility of goals during reading. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 21, 364-372.

Baayen, R. H., R. Piepenbrock, and H. van Rijn (Eds.) (1993). *The CELEX Lexical Database* (CD-ROM). University of Pennsylvania, Philadelphia (PA): Linguistic Data Consortium.

Bazerman, C. (1995). Systems of genres and the enactment of social intentions. In A. Freedman and P. Medway (Eds.), *Genre and the New Rhetoric*. London: Taylor and Francis.

Bhatia, V. (1997). Applied genre analysis and ESP. In T. Miller (Ed.), Functional approaches to written text: Classroom applications. Washington, DC: USIA.

Biber, D. (1988). Variation across speech and writing. New York: Cambridge University Press.

Boese E. (2005). Stereotyping the web: genre classification of web documents, *M.S. Thesis*, Computer Science Department, Colorado State University. Fort Collins, CO.

Boese, E. S. & Howe, A. E. (2005). Effects of Web Document Evolution on Genre Classification. In *proceedings CIKM 05*.

Charniak, E. A Maximum-Entropy-Inspired Parser. In *Proceedings of the North American Chapter of the Association for Computational Linguistics* (2000), pp. 132-139.

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.

Coleridge, S.T. (1985). Biographia Literaria: Samuel Taylor Coleridge, H.J. Jackson (ed.), Oxford.

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

Crowston, K. & Williams, M. (2000). Reproduced and emergent genres of communication on the World-Wide Web. *The Information Society* 16, 201-215.

Davies, A, & Elder, C. (2004). *The handbook of applied linguistcs*. Blackwell Publishing. Oxford.

Downs, W. (1998). *Language and society*. Cambridge University Press.

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 and Methods, 29*, 212-223.

Finn A. & Kushmerick N. (2006). Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology, Special Issue on Computational Analysis of Style, 7*.

Franklin, C (2008) http://computer.howstuffworks.com/hsw-contact.htm. Retrieved 03/05/2008

Gerrig, R. (1993). Experiencing narrative worlds: On the psychological activities of reading. Cambridge, MA: MIT Press.

Gordon, J. (2004). *Joyce and Reality: The Empirical Strikes Back*. Syracuse, NY: Syracuse University Press.

Graesser, A. C., Hauft-Smith, K., Cohen, A. D., & Pyles, L. D. (1980). Advanced outlines, familiarity, text genre, and retention of prose. *Journal of Experimental Education, 48*, 209-220.

Graesser, A. C., Hoffman, N. L., & Clark, L. F. (1980). Structural components of reading time. Journal of Verbal Learning and Verbal Behavior, 19, 131-151.

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

Graesser, A.C., Olde, B. A., & Klettke, B. (2002). How does the mind construct and represent stories? In M. Green, J. Strange, and T. Brock (Eds.), *Narrative Impact: Social and Cognitive Foundations*. Mahwah: NJ: Erlbaum.

Hatch, E. & Lazarton, A. (1991). Research manual: Design and statistics for applied linguistics. New York: Newbury House.

Hymes, D. (1972). Models of interaction of language and social life. In *Directions of Sociolinguistics: The Ethnography of Communication* (Eds.) J.J. Gumperz & D. Hymes. New York: Holt, Rinehart and Winston.

Karlgren J. and Cutting D. (1994). Recognizing Text Genre with Simple Metrics Using Discriminant Analysis. Proceedings of COLING 1994, Kyoto.

Kaup, B., & Zwaan, R. A. (2003). Effects of negation and situational presence on the accessibility of text information. Journal of Experimental Psychology: Learning, Memory and Cognition, 29, 439–446.

Kennedy A. & Shepherd M. (2005), Automatic identification of home pages on the web. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*.

Kessler, B., Nunberg, G., & Schütze, H. (1997). Automatic detection of text genre. In Proceedings of the 35th annual meeting on Association for Computational Linguistics, p.32-38, Madrid, Spain.

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

Kintsch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 363–394.

Lee J., Kwong O. Y. (Eds.) Natural Language Processing. Springer, Berlin.

Linderholm, T., & vcan den Broek, P. (2002). The effects of reading purpose and working memory capacity on the processing of expository text. Journal of Educational Psychology, 94(4), 778–784.

Kendou, P. & van den Broek, P. (2005). The effects of readers' misconceptions on comprehension of scientific text. *Journal of Educational Psychology, 97*, 235-245.

Radvansky, G. A., Zwaan, R. A., Curiel, J. M., & Copeland, D. E. (2001). Situation models and aging. *Psychology and Aging, 16*, 145-160.

Lee Y. and Myaeng S. (2002). Text Genre Classification with Genre-Revealing and Subject-Revealing Features. *Proceedings of the 25th Annual International ACM SIGIR* : 145-150.

Lee Y. and Myaeng S. (2004). Automatic identification of text genres and their roles in subject-based categorization. *Proceedings of the 37th Hawaii International Conference on System Sciences*.

Lightman, E.J., McCarthy, P.M., Dufty, D.F., & McNamara, D.S. (2007). The structural organization of high school educational Texts. FLAIRS, 2007.

Lim C., Lee K. and Kim G. (2005). Automatic genre detection of web documents. In Su K., Tsujii

Louwerse, M.M., McCarthy, P.M., McNamara, D.S., & Graesser, A. C. (2004). Variation in language and cohesion across written and spoken registers. In K. Forbus, D. Gentner, T. Regier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp. 843-848). Cognitive Science.

McCarthy, P.M., Briner, S.W., Rus, V., & McNamara, D.S. (2007). Textual signatures: Identifying text-types using latent semantic analysis to measure the cohesion of text structures. In A. Kao, & S. Poteet (Eds.), *Natural language processing and text mining* (pp. 107-122) . London: Springer-Verlag

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., Renner, A.M., Duncan, M.G., Duran, N.D., Lightman, E.J., & McNamara. D.S., (2008). Identifying topic sentencehood. *Behaviorial Research and Methods*, 21, 364-372.

McDaniel, M. A., Einstein, G. O., Dunay, P. K., & Cobb, R. E. (1986). Encoding difficulty and memory: Toward a unifying theory. *Journal of Memory and Language, 25*, 645-656.

Meyer zu Eissen S., Stein B. (2004). Genre classification of web pages : User study and feasibility analysis. In Biundo S., Fruhwirth T., Palm G. (Eds.), *Advances in Artificial Intelligence*. Springer, Berlin : 256-269.

Meyer, B. J. F., & Wijekumar, K. (2007). Web-based tutoring of the structure strategy: Theoretical background, deisgn, and findings. In D. S. McNamara (Ed.), *Reading Comprehension Strategies: Theories, Interventions, and Technologies*. Erlbaum.

Miliv, L. T. & Slane, S. (1994). Qualitative aspects of genre in the century of prose corpus. *Style 24*, 42-57.

Oakhill, J., & Cain, K. (2007). Issues of causality in children's reading comprehension. In D. S. McNamara (Ed.), *Reading Comprehension Strategies: Theories, Interventions, and Technologies*. Erlbaum.

Orlikowski, W. J. and Yates, J. (1994). Genre repertoire: The structuring of communicative practices in organizations. *Administrative Sciences Quarterly, 33*, 541-574.

Otero, J., Leon, J. A., & Graesser, A. C. (Eds.), (2002). The psychology of science text comprehension. Mahwah, NJ: Erlbaum.

Radvansky, G. A., Zwaan, R. A., Curiel, J. M., & Copeland, D. E. (2001). Situation models and. aging. *Psychology and Aging, 16*, 145-160.

Rosso, M. A. (2005). Using genre to improve web search. *PhD. Thesis*. University of North Carolina, Chapel Hill.

Roussinov, D., Crowston, K., Nilan, M., Kwasnik B., Cai, J., & Liu, X. (2001). Genre based navigation on the web. In *Proceedings of the 34th Hawaiian International Conference on System Sciences*, Hawaii. IEEE ComputerPress.

Santini M. (2006). Some issues in Automatic Genre Classification of Web Pages, JADT 2006 - 8èmes Journées internationales d'analyse statistique des données textuelles du 19 au 21 avril 2006 à l'université de Besançon (France).

Šatava, L. (2006). *"Regional languages" as emancipation strategy*. Czech lands in the middle of Europe in the past. *MSM* 0021620827.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G.(2001), Automatic text categorization in terms of genre and author, *Computational Linguistics, 26(4)*, 471–495.

Swales, J. (1990) *Genre Analysis.* Cambridge: Cambridge University Press.

Teufel, S. & Moens, M. (1999). Argumentative classification of extracted sentences as a first step towards flexible abstracting. In: I. Mani, M. Maybury (Eds.), *Advances in automatic text summarization*, MIT Press, 1999.

Tonjes, M.J., Ray, W., & Zintz, M.V. (1999). Integrated content literacy. New York: The McGraw-Hill Publishers.

Trabasso, T., & Bartolone, J. (2003). Story understanding and counterfactual reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 904-923.

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

Vidal-abarca, E., Martinez, G., & Gilabert, R. (2000). Two procedures to improve instructional text: Effects on memory and learning. *Journal of Educational Psychology, 92,* 107-116.

Williams, P. J. (2007). Literacy in the Curriculum: Integrating Text Structure and Content Area Instruction. In D. S. McNamara (Ed.), *Reading Comprehension Strategies: Theories, Interventions, and Technologies*. Erlbaum.

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

Wolfe, M. B. W. (2005) Memory for narrative and expository text: Independent influences of semantic associations and text organization. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*, 2, 359-364.

Zwaan, R.A. (1993). Aspects of literary comprehension. Amsterdam: John Benjamins.

**Tables**

Table 1: Accuracy and misclassifications for Narrative, History, and Science texts, and "Don't Know"(DK) classifications.

| | Accuracy | | | Correct | | | Misclassification | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Recall | Precision | F1 | Narrative | History | Science | Narrative | History | Science | DK |
| Rater 1 | .824 | .840 | .832 | .914 | .829 | .729 | .081 | .052 | .023 | .019 |
| Rater 2 | .824 | .892 | .856 | .871 | .857 | .743 | .062 | .038 | .000 | .076 |
| Rater 3 | .810 | .817 | .813 | .886 | .757 | .786 | .081 | .076 | .024 | .029 |
| Mean | .819 | .850 | .834 | .890 | .814 | .752 | .075 | .055 | .016 | .041 |

Table 2: Accuracy and misclassifications of expert raters by domain for Narrative, History, and Science texts, and unclassified "Don't Know" (DK) texts.

| Decisions | | | Accuracy | | | Misclassifications | | | |
|---|---|---|---|---|---|---|---|---|---|
| Domain | Selected | Correct | Recall | Precision | F1 | Narrative | History | Science | DK |
| Narrative | 234 | 187 | 0.890 | 0.799 | 0.842 | / | 10 | 3 | 10 |
| History | 206 | 171 | 0.814 | 0.830 | 0.822 | 25 | / | 7 | 7 |
| Science | 168 | 158 | 0.752 | 0.940 | 0.836 | 22 | 25 | / | 5 |

Table 3: The six sentences identified by all raters as narratives.

| Example | Domain | Sentence |
|---|---|---|
| 1 | History | We cannot[1] sell the lives of men and[3] animals[2], said one Blackfoot chief in the 1800s, "therefore we cannot sell this land." |
| 2 | History | I[1] had vainly[3] flattered[2] myself that without very much bloodshed it might be done. |
| 3 | History | Much to[1] my surprise[2], I[3] had forgotten my glasses in prison, so I used my wife's. |
| 4 | Science | Taking no joy[1] in life, looking forward[3] to nothing, wanting to withdraw from people and activities[2]. |
| 5 | Science | This, he thought[1], would[2] demonstrate[3] that emotions can be mechanically induced (Cohen, 1979). |
| 6 | Science | Watson[1,] 3 went even[2] further and suggested that at the human level, deep emotions are also just the result of association and learning. |

Note: The superscript number indicates the point at which the genre selection was made

Table 4: The three longest, misclassified sentences.

| Domain | Classification | Sentence |
|---|---|---|
| Narrative | Don’t Know | Friends in the barrio explained that the director was called a principal, and that it was a lady and not a man. |
| History | Narrative | The governor presided over an advisory council, usually appointed by the governor, and a local assembly elected by landowning white males. |
| History | Don’t Know | We blow the whistle that's heard round the world, and all peoples stop to heed and welcome it. |

Table 5: Accuracy of genre evaluation

| Genre | Accuracy | Mean | SD |
|---|---|---|---|
| Narrative | Recall | 0.86 | 0.09 |
| | Precision | 0.71 | 0.12 |
| | F1 | 0.77 | 0.09 |
| History | Recall | 0.71 | 0.14 |
| | Precision | 0.76 | 0.09 |
| | F1 | 0.72 | 0.11 |
| Science | Recall | 0.67 | 0.12 |
| | Precision | 0.88 | 0.09 |
| | F1 | 0.75 | 0.11 |

Table 6: Frequencies of number of words used in correct genre assessments.

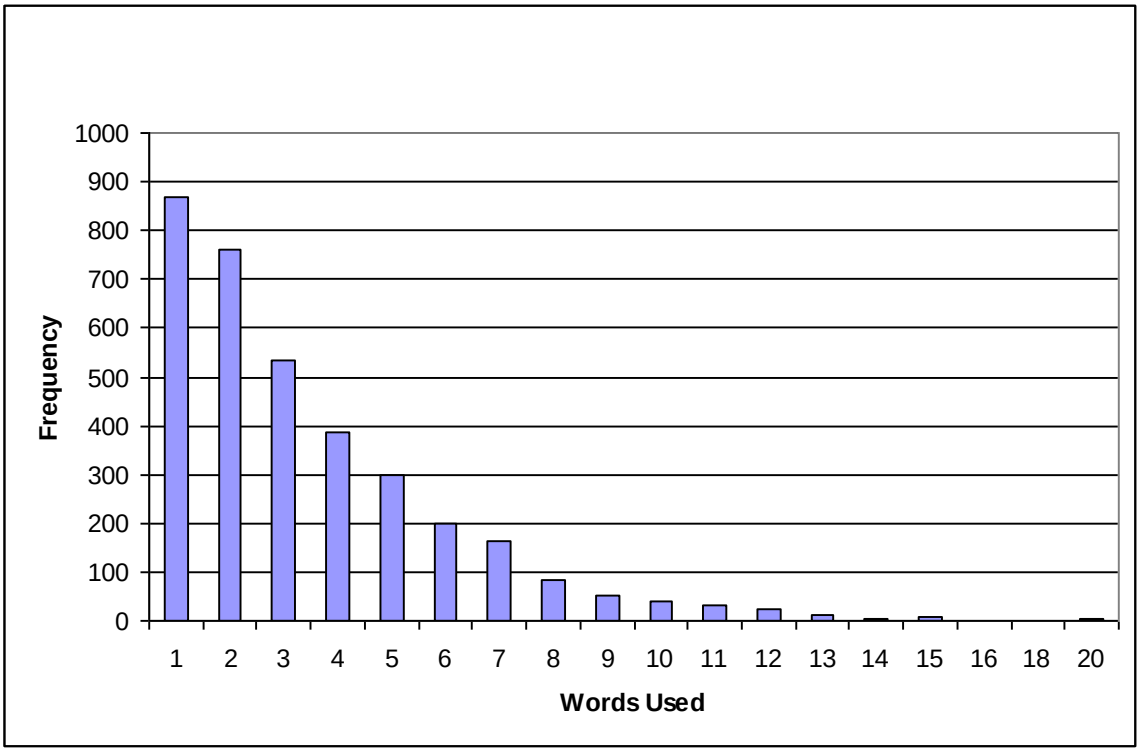

Table 7: Five most predictive variables in distinguishing low/high skill readers

| | Low skill | | High Skill | | | | |
|---|---|---|---|---|---|---|---|
| Dependent Variable | Mean | SD | Mean | SD | F | P | $\eta^2$ |
| Narrative precision | 0.66 | 0.12 | 0.79 | 0.08 | 7.55 | 0.01 | 0.27 |
| Time: 3rd word (History) | 1.01 | 0.29 | 0.72 | 0.21 | 6.72 | 0.02 | 0.25 |
| Science Recall | 0.62 | 0.13 | 0.74 | 0.06 | 6.52 | 0.02 | 0.25 |
| Science F1 | 0.71 | 0.11 | 0.81 | 0.07 | 5.87 | 0.02 | 0.23 |
| Time: 3rd word (Narrative) | 0.96 | 0.29 | 0.70 | 0.16 | 5.75 | 0.03 | 0.22 |

Table 8: Most significant genre predictor variables for “3 word”, “5 word”, and “whole sentence.”

| Words | Dependent Variable | Mean Narrative | Mean History | Mean Science | F | $\eta^2$ |
|---|---|---|---|---|---|---|
| 3 | Past tenses | 177.3 (168.12) | 68.18 (136.01) | 13.33 (65.98) | 20.01 | 0.23 |
| | Pronoun/noun phrases | 184.04 (167.93) | 51.14 (119.51 | 38.33 (105.42) | 17.29 | 0.20 |
| | Syllables | 3.70 (.86) | 4.89 (1.46) | 4.94 (1.46) | 13.21 | 0.16 |
| 5 | Reading Grade | 1.49 (2.12) | 6.40 (3.8) | 4.72 (3.57) | 29.44 | 0.30 |
| | Past tense verbs | 150.00 (115.61) | 60.31 (91.43) | 9.52 (43.10) | 29.31 | 0.30 |
| | Pronoun/noun phrases | 141.71 (125.77) | 32.54 (96.70) | 27.78 (69.03) | 19.51 | 0.22 |
| Whole | Reading Ease | 83.06 (15.41) | 48.98 (21.78) | 51.63 (21.78) | 22.43 | 0.40 |
| | CELEX frequency | 2.81 (0.25) | 2.34 (0.33) | 2.45 (0.31) | 15.91 | 0.32 |
| | Past tense verbs | 66.67 (66.83) | 62.87 (48.06) | 5.85 (19.83) | 10.68 | 0.24 |

Note: All F-values are significant at $p<.001$; SD appear in parentheses

Table 9: “Three word” recall, precision, and F1 results for computational model (test set; all data) compared to participant’s performance.

| | Narrative | | | History | | | Science | | |
|---|---|---|---|---|---|---|---|---|---|
| | Recall | Precision | F1 | Recall | Precision | F1 | Recall | Precision | F1 |
| Test set | 0.61 | 0.74 | 0.67 | 0.23 | 0.33 | 0.27 | 0.60 | 0.38 | 0.46 |
| All data | 0.67 | 0.73 | 0.70 | 0.46 | 0.52 | 0.49 | 0.71 | 0.59 | 0.65 |
| Participants | 0.85 | 0.71 | 0.77 | 0.71 | 0.76 | 0.72 | 0.68 | 0.87 | 0.76 |

Table 10: “Five word” recall, precision, and F1 results for computational model (test set; all data) compared to participant’s performance.

| | Narrative | | | History | | | Science | | |
|---|---|---|---|---|---|---|---|---|---|
| | Recall | Precision | F1 | Recall | Precision | F1 | Recall | Precision | F1 |
| Test set | 0.47 | 0.56 | 0.51 | 0.55 | 0.39 | 0.46 | 0.61 | 0.71 | 0.66 |
| All data | 0.72 | 0.71 | 0.72 | 0.66 | 0.53 | 0.59 | 0.60 | 0.73 | 0.66 |
| Participants | 0.85 | 0.71 | 0.77 | 0.71 | 0.76 | 0.72 | 0.68 | 0.87 | 0.76 |

Table 11: “Whole sentence” recall, precision, and F1 results for computational model (test set; all data) compared to participant’s performance.

| | Narrative | | | History | | | Science | | |
|---|---|---|---|---|---|---|---|---|---|
| | Recall | Precision | F1 | Recall | Precision | F1 | Recall | Precision | F1 |
| Test set | 0.70 | 0.56 | 0.62 | 0.62 | 0.67 | 0.46 | 0.60 | 0.68 | 0.64 |
| All data | 0.75 | 0.70 | 0.73 | 0.64 | 0.67 | 0.66 | 0.75 | 0.77 | 0.76 |
| Participants | 0.85 | 0.71 | 0.77 | 0.71 | 0.76 | 0.72 | 0.68 | 0.87 | 0.76 |

**Appendix.** Sample of the sentences used in the study

| Index | Topic sentence value (1-6) | Sentence 1 (1)/ Sentence 3 (0) | Narrative (0); History (1); Science (2) | No. of words | Sentence |
|---|---|---|---|---|---|
| 1 | 2.67 | 0 | 1 | 15 | Because of the fragmented nature of Mayan society, the different cities frequently went to war. |
| 2 | 1.33 | 0 | 1 | 11 | They moved slowly, not as invading hordes but as small communities. |
| 23 | 2.00 | 1 | 0 | 18 | When the time was up, Mr.Dooley asked us to put down our pencils and pass our tests in. |
| 52 | 2.00 | 0 | 2 | 14 | However, more material is ultimately moved by the greater number of slow mass movements. |
| 53 | 2.33 | 0 | 2 | 20 | Likewise, it's easier to express the concentration of a solution as the number of moles of material dissolved in it. |

All sentences used in this study can be downloaded at http://tinyurl.com/55ex2x