TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 3 (2011) 371–398

ISSN: 1756-8757 print / 1756-8765 online
DOI: 10.1111/j.1756-8765.2010.01081.x

Computational Analyses of Multilevel Discourse Comprehension

Arthur C. Graesser, Danielle S. McNamara

Department of Psychology, University of Memphis

Received 20 January 2009; received in revised form 17 August 2009; accepted 28 November 2009

Abstract

The proposed multilevel framework of discourse comprehension includes the surface code, the textbase, the situation model, the genre and rhetorical structure, and the pragmatic communication level. We describe these five levels when comprehension succeeds and also when there are communication misalignments and comprehension breakdowns. A computer tool has been developed, called Coh-Metrix, that scales discourse (oral or print) on dozens of measures associated with the first four discourse levels. The measurement of these levels with an automated tool helps researchers track and better understand multilevel discourse comprehension. Two sets of analyses illustrate the utility of Coh-Metrix in discourse theory and educational practice. First, Coh-Metrix was used to measure the cohesion of the text base and situation model, as well as potential extraneous variables, in a sample of published studies that manipulated text cohesion. This analysis helped us better understand what was precisely manipulated in these studies and the implications for discourse comprehension mechanisms. Second, Coh-Metrix analyses are reported for samples of narrative and science texts in order to advance the argument that traditional text difficulty measures are limited because they fail to accommodate most of the levels of the multilevel discourse comprehension framework.

Keywords: Discourse processes; Text comprehension; Coherence; Cohesion; Semantics; Computational linguistics

1. Introduction

The purpose of this issue of *topics* is to provide an overview of current methods that extract meaning from text. The approach in this paper is distinctive in its attempt to accommodate multiple levels of discourse, whereas most of the other contributions in this issue

Correspondence should be sent to Arthur C. Graesser, Psychology Department, 202 Psychology Building, University of Memphis, Memphis, TN 38152-3230. E-mail: a-graesser@memphis.edu

focus on particular levels of semantic or conceptual processing. Contemporary theories of comprehension have identified the representations, structures, strategies, and processes at various levels of discourse, ranging from the surface code (comprised of the words and syntax) to the deeper meaning and pragmatic intentions of discourse. An objective analysis of multilevel discourse comprehension should benefit from computational methods that dissect and measure each of the levels, as opposed to relying on human intuitions for scoring and annotating texts. We have developed a computer tool called Coh-Metrix that analyzes texts on most of the discourse levels (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara, Louwerse, & Graesser, 2008). Our goal is to use the Coh-Metrix tool to acquire a deeper understanding of language-discourse constraints and the associated psychological mechanisms underlying multilevel discourse comprehension.

It is important to clarify what we mean by discourse. Our definition of discourse includes both oral conversation and printed text. The utterances in oral conversation and the sentences in printed text are composed by the speaker/writer with the intention of communicating interesting and informative messages to the listener/reader. Therefore, ''considerate'' discourse is likely to be coherent, understandable to the community of discourse participants, and relevant to the situational goals. There are times when discourse communication breaks down, however. Communication breakdowns occur when the writer and reader (or speaker and listener) encounter substantial gulfs in language, common ground, prior knowledge, or discourse skills. Minor misalignments may also occur and capture one's attention, as in the case of a mispronounced word, a rare word in a text, an ungrammatical sentence, or a sentence that does not fit into the thread of discourse. A model of discourse comprehension should handle instances when there are communication breakdowns in addition to successful comprehension.

Multiple levels of comprehension have been identified and explored by numerous discourse researchers over the years (Clark, 1996; van Dijk & Kintsch, 1983; Graesser, 1981; Graesser, Millis, & Zwaan, 1997; Kintsch, 1998; Kintsch, Welsh, Schmalhofer, & Zimny, 1990; McNamara & Magliano, 2008; Pickering & Garrod, 2004; Schmalhofer & Glavanov, 1986; Snow, 2002). The taxonomy we adopt in this article is the one presented by Graesser et al. (1997) who included five levels: the *surface code*, the explicit *textbase*, the *situation model* (sometimes called the mental model), the discourse *genre and rhetorical structure* (the type of discourse and its composition), and the *pragmatic communication* level (between speaker and listener, or writer and reader). The first three levels are equivalent to the levels proposed by other researchers (Kintsch et al., 1990; Schmalhofer & Glavanov, 1986; Zwaan, 1994) who used a recognition memory paradigm to validate the distinctions among the surface code, textbase, and situation model.

Table 1 elaborates on these five levels by identifying the codes, constituents, and content associated with each level. It is beyond the scope of this article to crisply define each level and provide an exhaustive specification of the theoretical entities. Moreover, it is sometimes debatable which level to assign a particular component of language or discourse; this is quite expected because of interactions between levels. Table 1 merely illustrates the various levels and provides examples of what each level may contain. Moreover, the table depicts the levels of discourse as compositional components that are constructed as a *result* of

Table 1
Levels of discourse

(1) Surface code
Word composition (graphemes, phonemes, syllables, morphemes, lemmas, tense, aspect)
Words (lexical items)
Part of speech categories (noun, verb, adjective, adverb, determiner, connective)
Syntactic composition (noun-phrase, verb-phrase, prepositional phrases, clause)
Linguistic style and dialect
(2) Textbase
Explicit propositions
Referents linked to referring expressions
Connectives that explicitly link clauses
Constituents in the discourse focus versus linguistic presuppositions
(3) Situation model
Agents, objects, and abstract entities
Dimensions of temporality, spatiality, causality, intentionality
Inferences that bridge and elaborate ideas
Given versus new information
Images and mental simulations of events
Mental models of the situation
(4) Genre and rhetorical structure
Discourse category (narrative, persuasive, expository, descriptive)
Rhetorical composition (plot structure, claim + evidence, problem + solution, etc.)
Epistemological status of propositions and clauses (claim, evidence, warrant, hypothesis)
Speech act categories (assertion, question, command, promise, indirect request, greeting, expressive evaluation)
Theme, moral, or point of discourse
(5) Pragmatic communication
Goals of speaker/writer and listener/reader
Attitudes (humor, sarcasm, eulogy, deprecation)
Requests for clarification and backchannel feedback (spoken only)

comprehension. It is important not to lose sight of the fact that this *compositional* viewpoint is incomplete. Each level and compositional component also has an affiliated *knowledge* and *process* viewpoint. That is, for any given compositional entity C, the comprehender needs to have had the prerequisite knowledge about C through prior experiences and training. The comprehender also needs to be able to process C by identifying its occurrence in the discourse and by executing the cognitive processes, procedures, and strategies that are relevant to C. The processing of C becomes automatized after extensive experience, often to the point of being executed unconsciously. Under ideal circumstances, the comprehender is sufficiently proficient that the composition, knowledge, and processes are intact for a wide universe of discourse experiences.

This article does not propose a particular cognitive model that specifies precisely how the representations are constructed within each level and between levels. Discourse researchers who have modeled some of these levels have typically used a hybrid between production system and connectionist architectures, as in the case of Collaborative Action-based Production System (CAPS) Reader model (Just & Carpenter, 1987, 1992), the

Construction-Integration model (Kintsch, 1998), the constructivist model (Graesser, Singer, & Trabasso, 1994), and the landscape model (Van den Broek, Virtue, Everson, Tzeng, & Sung, 2002). We assume that such hybrid architectures can be successfully implemented once we consider all five levels. One important step in achieving a complete cognitive processing model is to understand the constraints of each level, ideally to the point of having automated algorithms. This is the step we focus on in the present article.

The remainder of this article has three sections. The first section discusses some of the mechanisms that operate when readers/listeners experience comprehension difficulties, breakdowns, or communication misalignments. Such difficulties, breakdowns, and misalignments in comprehension are particularly informative because they have a dramatic impact on attention, reading time, memory, reasoning, behavior, and other manifestations of cognition. The second section describes a computer tool, called Coh-Metrix (Graesser et al., 2004; McNamara et al., 2008), that scales discourse on dozens of measures associated with the first four of the five discourse levels. Coh-Metrix can be applied to any text, so measures can be compared for texts in different genre categories and different experimental conditions. The third section presents some examples of how Coh-Metrix can be used to investigate mechanisms of multilevel discourse comprehension.

2. Breakdowns, misalignments, and complexity in multilevel comprehension

Comprehension can misfire at any of the five levels depicted in Table 1. The cause of the misfire may be attributed to either deficits in the reader (i.e., lack of knowledge or processing skill) or the discourse (e.g., incoherent text, unintelligible speech). The consequence of a misfire can range from a complete breakdown in comprehension to a modest irregularity that captures the comprehender's attention. The comprehender may do nothing, which runs the risk of a comprehension failure. Alternatively, the comprehender may attempt to compensate for the misfire by using information from other levels of discourse, from prior knowledge, or from external sources (e.g., other people or technologies). For example, the scenarios below illustrate misfires at various discourse levels, along with the resulting consequences.

Scenario 1: An immigrant arrives in the United States and does not understand the English language at all. Phoneme deficits prevent him from understanding any of the conversations, so his attention settles on the sounds of words. A complete breakdown at discourse level 1 also blocks the deeper levels of 2–5 (see Table 1).

Scenario 2: A 4-year-old child is learning to read but has trouble recognizing some of the words. She has trouble reading aloud a textbook on dinosaurs because she stumbles on the rare words. However, she does a pretty good job reading aloud Snow White because she has heard similar stories dozens of times. World knowledge compensates for a lexical deficit at level 1.

Scenario 3: Two homebuyers read a legal document that has lengthy sentences with embedded clauses and numerous Boolean operators (*and, or, not, if*). They have only a vague idea what the document explicitly states because of complex syntax and a dense

textbase (i.e., deficits at levels 1 and 2). However, the man and wife sign the contract because they understand the purpose of the document and trust the real estate agent. Thus, levels 4 and 5 circumvent the need to understand levels 1–3 completely.

Scenario 4: A father and son read the directions to assemble a new entertainment center. They argue on how to connect the TV to a DVD and on which washer to pair with a particular screw. The father and son have no problem understanding the words and textbase in the directions (levels 1 and 2) and no problem understanding the genre and purpose of the document (levels 4 and 5), but they do have a deficit at the situation model level (level 3).

Scenario 5: A college student majoring in biology asks his roommate to proofread a term paper. The roommate complains that the logical flow was problematic for him even though he knows the jargon. The biology major revises the text by adding connectives (e.g., *because, so, therefore, before*) and other coherence-based signaling devices to improve the cohesion. The roommate applauds the changes because he found the text much more comprehensible. In this case an augmentation in the language at levels 1 and 2 compensated for a deficit at level 3.

Scenario 6: A professional novelist had always loved reading Russian novels, but recalls his college days when he had received a C in Russian literature. Years later the novelist realizes that he had had a great appreciation for the settings and plots of those Russian novels but never understood the deep meaning because he had missed intentions and attitudes of the authors. In this case, discourse levels 1–4 are intact in college, but the novelist did not make it to level 5.

These scenarios illustrate how deficits in one or more discourse levels can have substantial repercussions on processes at other levels. Researchers who pursue a multilevel discourse comprehension model need to understand the processing mechanisms *within* levels and *between* levels. Our hope is that a computer tool such as Coh-Metrix will assist researchers in better understanding multilevel discourse comprehension.

Available psychological research supports a number of generalizations about the processing order, constraints, interaction, coordination, and compensatory mechanisms of the different levels of discourse comprehension. Some of these generalizations are briefly described below.

2.1. Bottom-up dependencies of meaning

In a strictly bottom-up model of reading, the ordering on depth is assumed to be 1 $\rightarrow$ 2 $\rightarrow$ 3 $\rightarrow$ 4 $\rightarrow$ 5. Although most researchers endorse an interactive model of reading rather than a strictly bottom-up model (McClelland, 1986; Rayner & Pollatsek, 1994; Rumelhart, 1977; Van den Broek, Rapp, & Kendeou, 2005), the prevailing assumption is that the lower levels constrain the higher levels more than vice versa. The asymmetry in dependencies between levels perhaps follows a cascade model in which at least a partial analysis of level N is needed prior to the initiation of levels N + 1 and higher. The encoding of words during reading is robustly influenced by the bottom-up constraints of the letters and syllables (Gough, 1972; Rayner, 1998; Rayner & Pollatsek, 1994). The quality of a person's lexicon has a large impact on the fidelity and speed of interpreting sentences and

generating inferences at levels 2, 3, and higher (Perfetti, 2007; Stanovich, 1986). Similarly, it is important for readers to establish an interpretation of the textbase before they can productively move on to the construction of the situation model and higher levels (McNamara, Graesser, & Louwerse, in press). A partial-to-full analysis of levels 1–3 is presumably needed to adequately construct the rhetorical structure.

From the other direction, there is some question about the extent to which top-down processes influence lower order processes. It is well documented that top-down processing influences the speed and construction of word meanings (Hess, Foss, & Carroll, 1995; McClelland, 1986; Rayner & Pollatsek, 1994), but top-down influences on the construction of the textbase and situation model do not have a rich empirical base. Zwaan (1994) reported that the encoding of the surface code, textbase, and situation model had different profiles when college students were told they were reading a newspaper article versus literature. As predicted, the literature instructions enhanced the surface code, whereas the newspaper instructions enhanced the situation model. However, these effects are confined to texts that are sufficiently ambiguous or malleable that they can fall under the umbrella of two genres. Such texts are not representative of normal texts in which the correct genre can be identified by linguistic features within a couple of sentences (McCarthy, Myers, Briner, Graesser, & McNamara, 2009).

There are two implications of the principle that there are asymmetrical, bottom-up dependencies on meaning. One implication is that misfires at a particular level propagate problems to higher levels, but not necessarily to the lower levels. If the reader fails to construct a textbase, for example, the reader will not be able to readily construct an adequate situation model, genre, and pragmatic communication, even though level 1 is intact. Scenario 1 illustrates this generalization. A second implication of this principle is that comprehenders routinely achieve the deepest level of comprehension that is supported by the discourse constraints, their knowledge, and their processing skill (Graesser et al., 1994; Hess et al., 1995). For example, the immigrant in scenario 1 had to settle for level 1 and went no higher, whereas the novelist in scenario 6 made it through level 4 but never achieved level 5 while he was in college.

2.2. Novel information requires more processing effort than familiar and automatized components

Novelty of information is a foundational cognitive dimension that attracts attention and effort and that is salient in memory (Mandler, 1988; Tulving & Kroll, 1995). The relevant question in this context is what components and levels have the most novel information. It could be argued that lower frequency words, the textbase level, and the situation model tend to have the highest density of novel information. In contrast, most aspects of levels 1, 4, and 5 tend to have components that are frequently experienced and therefore overlearned and automatized.

There is evidence from reading time studies that more processing time is allocated to rare words than high-frequency words (Carver, 1990; Just & Carpenter, 1987; Rayner, 1998) as

well as to new information expressed in the textbase and situation model than to old information already mentioned (Haberlandt & Graesser, 1985; Haviland & Clark, 1974; Kintsch & Van Dijk, 1978). In contrast, levels of genre, rhetorical structure, and author characteristics are normally familiar structures that are invisible to the comprehender unless there are irregularities or breakdowns.

The above predictions on novelty are compatible with the data on word reading times reported by Haberlandt and Graesser (1985). They collected self-paced word reading times using a moving window method on 3,278 words in 12 passages. Half of the passages were narrative and half informational text on science, social studies, and other topics that would normally appear in an encyclopedia. Orthogonal to the narrative-informational split, half were rated as being on familiar topics/ideas and half unfamiliar. Multiple regression analyses on mean reading times for individual words were performed as a function of the levels, including a layout variable that reflected the layout of the words on the screen. The predictors listed below were the only measures that had significant semi-partial correlations that statistically removed the contributions of the remaining predictors; normalized beta weights are in parentheses.

> Layout: beginning of line (0.08), end of line (0.09), beginning of screen (0.07), end of screen (0.04)
> Surface code (level 1): number of syllables (0.30), logarithm of word frequency (−0.17)
> Textbase and situation model (levels 2 and 3): beginning of sentence (0.12), end of sentence (0.39), beginning of clause (0.04), end of clause (0.07), number of new content words in sentence (0.11), sentence imagery (0.04), type-token ratio for content words (0.04), content word overlap with all previous sentences (−0.05), content word overlap with previous adjacent sentence (−0.05)
> Genre and rhetorical composition (level 4): narrativity (−0.12), serial position of sentence in text (−0.12)

It should be noted that the two surface code variables that ended up being significant directly reflected the novelty of the words. That is, words with lower frequency in the English language are comparatively novel and longer words (many syllables) tend to be less frequent according to Zipf's (1949) law. These 17 significant predictor variables accounted for 54% of the variance of mean word reading times. Of these four groups of variables, the layout variables alone accounted for 5% of the variance, the surface code by itself (which is essentially the novelty of words) accounted for 30%, the textbase and situation model together accounted for 31% of the variance, and the discourse level 13% of the variance. When these four groups are entered in a stepwise fashion assuming bottom–up processing, the trend is layout (5%) $\rightarrow$ surface code (an increment of 22%) $\rightarrow$ textbase and situation model (an increment of 24%) $\rightarrow$ genre and rhetorical composition (an increment of 3%). These results support the claim that the processing of lower frequency words, the textbase, and the situation model explain most of the reading time variance and also capture the most novel information.

2.3. Attention, consciousness, and effort gravitate to misfires

Misfires at any level of analysis are likely to draw cognitive resources. Reading time studies have shown that extra processing time is allocated to pronouns that have unresolved or ambiguous referents (Gernsbacher, 1990; Rayner, 1998), to sentences that have breaks in textbase cohesion (Gernsbacher, 1990), to sentences that have coherence breaks in the situation model on the dimensions of temporality, spatiality, causality, and intentionality (Magliano & Radvansky, 2001; Zwaan, Magliano, & Graesser, 1995; Zwaan & Radvansky, 1998), and to sentences that contradict ideas already established in the evolving situation model (O'Brien, Rizzella, Albrecht, & Halleran, 1998). Attention drifts toward sources of cognitive disequilibrium, such as obstacles, anomalies, discrepancies, and contradictions (Graesser, Lu, Olde, Cooper-Pye, & Whitten, 2005).

2.4. Misfires may be repaired or circumvented by world knowledge, information at other discourse levels, or external sources

The scenarios illustrate some compensatory mechanisms that repair or circumvent the misfires. World knowledge fills in the lexical deficits in scenario 2. The syntax and textbase deficits in scenario 3 are circumvented by the information in discourse levels 4 and 5; in this case the meaning of the discourse is never understood, but the couple has enough information about levels 4 and 5 to know that deep understanding is unnecessary. The gaps and misalignments in the situation model of scenario 4 are rectified by extended conversations between father and son and by active problem solving. Coherence gaps in the textbase and situation model of scenario 5 are rectified by augmenting the discourse at levels 1 and 2 with connectives and other discourse cohesion markers. Inserting these connectives and markers are known to improve comprehension for readers with low domain knowledge on topics in the text (Britton & Gulgoz, 1991; McNamara, 2001; McNamara & Kintsch, 1996; McNamara, Kintsch, Songer, & Kintsch, 1996; O'Reilly & McNamara, 2007).

A rather counterintuitive finding reported by McNamara and her colleagues is that the connectives and discourse markers are frequently ineffective and sometimes even counterproductive for readers with high domain knowledge and reading skill. The liability of connectives and discourse markers is that they sometimes lower processing effort, inferences, and elaborations for high knowledge skilled readers. This reverse-cohesion effect underscores the challenges in assigning texts to readers for those who wish to optimize reading experiences in literacy programs.

The multilevel comprehension framework outlined in this section has provided a plausible sketch of the complexities of constructing meaning on different levels during discourse comprehension. There are multiple levels of meaning that mutually, but asymmetrically, constrain each other. The components at each level are successfully built if the text is considerate and the reader has prerequisite background knowledge and reading skills. However, Murphy's law (''Anything that can go wrong, will go wrong'') is routinely applicable, so there are periodic misfires that range from minor misalignments and comprehension

difficulties to comprehension breakdowns. The misfires are magnets of attention that sometimes trigger compensatory mechanisms that repair or circumvent the problems.

3. Coh-Metrix: A computer tool for analyzing language and discourse at multiple levels

Coh-Metrix was developed to analyze and measure text on levels 1 through 4 (Graesser et al., 2004; McNamara, Louwerse, McCarthy, & Graesser, in press; McNamara et al., 2008). The original purpose of the Coh-Metrix project was to concentrate on the cohesion of the textbase and situation model because those levels needed a more precise specification and much of our research concentrated on discourse coherence. We soon acknowledged the importance of considering cohesion and language processing at all of the levels under the rubric of the multilevel comprehension framework. Our theoretical vision behind Coh-Metrix was to use the tool to (a) assess the overall cohesion and language difficulty of discourse on multiple levels, (b) investigate the constraints of discourse within levels and between levels, and (c) test models of multilevel discourse comprehension. There were also some practical goals in our vision: (a) to enhance standard text difficulty measures by providing scores on various cohesion and language characteristics and (b) to determine the appropriateness of a text for a reader with a particular profile of cognitive characteristics.

A distinction is made between cohesion and coherence (Graesser, McNamara, & Louwerse, 2003). *Cohesion* consists of characteristics of the explicit text that play some role in helping the reader mentally connect ideas in the text. *Coherence* is a cognitive representation that reflects the interaction between linguistic/discourse characteristics and world knowledge. When we put the spotlight on the text as an object of investigation, coherence can be defined as characteristics of the text (i.e., aspects of cohesion) that are likely to contribute to the coherence of the mental representation. Coh-Metrix provides indices of such cohesion characteristics.

Coh-Metrix is available in both a public version for free on the Web (http://cohmetrix.memphis.edu, version 2.0) and an internal version (versions 2.1 and 3.0). The public version has over 60 measures of discourse on levels 1–4 (see Table 1), whereas the internal research version has nearly a thousand measures that are at various stages of testing. The Coh-Metrix tool is very easy to use. The researcher enters text and then the system produces a long list of measures on the text. There is a help system that defines the measures and that provides various forms of contextual support.

Coh-Metrix was designed to move beyond standard readability formulas, such as Flesch-Kincaid Grade Level (Klare, 1974–1975), that rely on word length and sentence length. Formula 1 shows the Flesch-Kincaid Grade Level metric. *Words* refers to mean number of words per sentence and *syllables* refers to mean number of syllables per word.

$$\text{Grade Level} = 0.39^{*}\text{Words} + 11.8^{*}\text{Syllables} - 15.59 \quad (1)$$

The lengths of words and sentences no doubt have important repercussions on psychological processing, but certainly there is more to reading difficulty than these two parameters. Coh-Metrix provides more diverse indices of language and also deeper measures.

The Coh-Metrix measures (i.e., indices, metrics) cover many of the components in Table 1 and more. Some measures refer to characteristics of individual words. Much can be discovered from computer facilities that link words to psychological dimensions, as in the case of *WordNet* (Fellbaum, 1998; Miller, Beckwith, Fellbaum, Gross, & Miller, 1990) and *Linguistic Inquiry Word Count* (Pennebaker, Booth, & Francis, 2007). However, the majority of the Coh-Metrix indices include deeper or more processing-intensive algorithms that analyze syntax, referential cohesion, semantic cohesion, dimensions of the situation model, and rhetorical composition. This section describes measures associated with Levels 1–4 in Table 1.

3.1. Surface code

3.1.1. Word measures

Coh-Metrix measures words on dozens of characteristics that were extracted from established psycholinguistic and corpus analyses. The MRC Psycholinguistic Database (Coltheart, 1981), for example, is a collection of human ratings of several thousands of words along several psychological dimensions: meaningfulness, concreteness, imagability, age of acquisition, and familiarity. Coh-Metrix computes scores for word frequency, ambiguity, abstractness, and parts of speech, as documented below. There are several other measures of words that are not addressed in this article.

3.1.1.1. Word frequency: The primary word frequency counts in Coh-Metrix come from *CELEX*, the database from the Dutch Centre for Lexical Information (Baayen, Piepenbrock, & Gulikers, 1995) that analyzed 17.9 million words. Word frequency is the frequency (per million words) of a word appearing in published documents in the real world. The logarithm of word frequency is often computed because reading times are linearly related to the logarithm of word frequency, not raw word frequencies (Haberlandt & Graesser, 1985; Just & Carpenter, 1987).

3.1.1.2. Polysemy and hypernymy: Coh-Metrix computes the ambiguity and abstractness of content words (e.g., nouns, main verbs, adjectives) by calculating the values of polysemy and hypernymy with WordNet (Fellbaum, 1998; Miller et al., 1990). Polysemy refers to the number of senses that a word has; ambiguous words have more senses. Hypernymy refers to the number of levels deep a word appears in a conceptual, taxonomic hierarchy. For example, *table* (as a concrete object) has seven hypernym levels: seat → furniture → furnishings → instrumentality → artifact → object → entity. A low score means the word tends to be comparatively superordinate in the hierarchy and is therefore more abstract. Mean scores are computed over the set of content words in a document, with an important contrast between nouns and main verbs.

3.1.1.3. Parts of speech: Coh-Metrix provides the part of speech (POS) for every word contained in a text. There are over 50 POS tags derived from the Penn Treebank (Marcus, Santorini, & Marcinkiewicz, 1993). The tags include content words (e.g., *nouns*, *verbs*, *adjectives*, *adverbs*) and function words (e.g., *prepositions*, *determiners*, *pronouns*). Coh-Metrix incorporates a natural language processing tool, the Brill (1995) POS tagger, for assigning POS tags to each word. This assignment allows for an *incidence score* of POS categories, calculated as the relative frequency of a particular category per 1,000 words.

3.1.2. Syntax

Coh-Metrix analyzes sentence syntax with the assistance of a syntactic parser developed by Charniak (2000). The parser assigns part-of-speech categories to words and syntactic tree structures to sentences. Our evaluations of several parsers showed better performance of Charniak's parser than other major parsers when comparing the assigned structures to judgments of human experts (Hempelmann, Rus, Graesser, & McNamara, 2006). The root of the tree, or highest level, divides the sentence into intermediate branches that specify nodes that include noun phrase (NP), verb phrase (VB), prepositional phrase (PP), and embedded sentence constituents. The tree terminates at leaf nodes, or words of the sentence that are labeled for their part of speech (POS) by the Brill (1995) POS tagger. Syntactic complexity is assumed to increase with a greater degree of embedded phrases, dense syntactic structures, and load on working memory. Coh-Metrix has several indices of syntactic complexity, but only two of them are reported in this article.

3.1.2.1. Modifiers per noun phrase: The mean number of modifiers per noun-phrase is an index of the complexity of referencing expressions. For example, *big bad wolf* is a noun-phrase with two modifiers of the head noun *wolf*.

3.1.2.2. Words before main verb of main clause: The number of words before the main verb of the main clause is an index of syntactic complexity because it places a burden on the working memory of the comprehender (Graesser, Cai, Louwerse, & Daniel, 2006; Just & Carpenter, 1987, 1992). Sentences with preposed clauses and left-embedded syntax require comprehenders to keep many words in working memory before getting to the meaning of the main clause.

3.2. Textbase

The *textbase* contains (a) explicit propositions in the text plus (b) referential links and a small set of inferences that connect the explicit propositions (van Dijk & Kintsch, 1983; Kintsch & Van Dijk, 1978). The propositions are in a stripped-down form that removes surface code features captured by determiners, quantifiers, tense, aspect, and auxiliary verbs. Suppose, for illustration, that the following excerpt about a corporation was read in a newspaper:

> When the board met on Friday, they discovered they were bankrupt. They needed to take some action, so they fired the president.

The first sentence would have the following propositions.
PROP 1: meet (board, TIME = Friday)
PROP 2: discover (board, PROP 3)
PROP 3: bankrupt (corporation)
PROP 4: when (PROP 1, PROP 2)

When the second sentence is comprehended, a referential bridging inference is needed to specify that *they* refers to *board*. This anaphor *they* provides referential cohesion between the two sentences in the textbase. Otherwise there would be no connection between the two sentences in the textbase. However, the textbase does not contain deeper links of coherence that would be provided by the situation model, such as the board fired the president because the president caused the bankruptcy by virtue of being incompetent and that a new president would make the corporation more solvent.

3.2.1. Coreference

Coreference is an important linguistic method of connecting propositions, clauses, and sentences in the textbase (Britton & Gulgoz, 1991; Halliday & Hasan, 1976; Kintsch & Van Dijk, 1978; McNamara et al., 1996). Referential cohesion occurs when a noun, pronoun, or noun-phrase refers to another constituent in the text. For example, in the sentence *When the intestines absorb the nutrients, the absorption is facilitated by some forms of bacteria,* the word *absorption* in the second clause refers to the event (or alternatively the verb *absorb*) in the first clause. There is a referential cohesion gap when the words in a sentence do not connect to other sentences in the text.

Coh-Metrix tracks five major types of lexical coreference: *common noun overlap*, *pronoun overlap*, *argument overlap*, *stem overlap*, and *content word overlap*. Common noun overlap is the proportion of all sentence pairs that share one or more common nouns, whereas pronoun overlap is the proportion of sentence pairs that share one or more pronoun. Argument overlap is the proportion of all sentence pairs that share common nouns or pronouns (e.g., *table/table, he/he,* or *table/tables*). Stem overlap is the proportion of sentence pairs in which a noun (or pronoun) in one sentence has the same semantic morpheme (called a lemma) in common with any word in any grammatical category in the other sentence (e.g., the noun *photograph* and the verb *photographed*). The fifth co-reference index, content word overlap, is the proportion of content words that are the same between pairs of sentences.

There are different variants of the five measures coreference. Some indices consider only pairs of *adjacent* sentences, whereas others consider *all possible pairs* of sentences in a paragraph. When all possible pairs of sentences are considered, there is the distinction between weighted and unweighted metrics that are sensitive to the distance between sentences.

3.2.2. Pronoun anaphors

Coh-Metrix treats pronouns carefully because pronouns are known to create problems in comprehension when readers have trouble linking the pronouns to referents. Coh-Metrix computes the incidence scores for personal pronouns (*I, you, we*) and the proportion of noun-phrases that are filled with any pronoun (including *it, these, that*).

Anaphors are pronouns that refer to previous nouns and constituents in the text. For any sentences s1 and a later sentence s2, if there exists a pronoun in s2 that refers to a pronoun or noun in s1, then the two sentences are considered *anaphorically overlapped*. There are measures of anaphor overlap in Coh-Metrix that approximate binding the correct referent to a pronoun, but the pronoun resolution mechanism is not perfect. A pronoun is scored as having been filled with a referent corresponding to a previous constituent if there is any prior noun that agrees with the pronoun in number and gender and that satisfies some syntactic constraints (Lappin & Leass, 1994). One measure computes the proportion of adjacent sentence pairs in the text that are anaphorically overlapped. The second is the proportion of sentence pairs within a 10-sentence window in the text that are anaphor overlapped.

3.2.3. Connectives and discourse markers that link clauses and sentences

There are word classes that have the special function of connecting clauses and other constituents in the textbase (Halliday & Hasan, 1976; Louwerse, 2001; Sanders & Noordman, 2000). The categories of connectives in Coh-Metrix include additive (*also, moreover*), temporal (*and then, after, during*), causal *(because, so*), and logical operators (*therefore, if, and, or*). The additive, temporal, and causal connectives are subdivided into those that are positive (*also, because*) versus negative (*but, however*). A higher incidence of these connectives increases cohesion in the textbase and also the situation model. The *incidence* of each word class is computed as the number of occurrences per 1,000 words.

3.2.4. Lexical diversity

Indices of lexical diversity are presumably related to both text difficulty and textbase cohesion. Greater lexical diversity adds more difficulty because each unique word needs to be encoded and integrated into the discourse context. Lexical diversity provides a global measure of the cohesiveness of the text: The lower the lexical diversity of the text, the greater the repetition of the terms in the text. The most well-known lexical diversity index is *type-token ratio* (TTR, Templin, 1957), but there are a number of analogous measures (McCarthy & Jarvis, 2007). TTR is a ratio value: The number of unique words in a text (i.e., types) is divided by the overall number of words (i.e., tokens) in the text.

3.3. Situation model

An important level of text comprehension consists of constructing a *situation model* (or mental model), which is the referential content or microworld of what a text is about (Graesser et al., 1994, 1997; Kintsch, 1998; Zwaan & Radvansky, 1998).

3.3.1. Dimensions of the situation model

Text comprehension researchers have investigated five dimensions of the situational model (Zwaan & Radvansky, 1998; Zwaan et al., 1995): causation, intentionality, time, space, and protagonists. A break in cohesion or coherence occurs when there is a discontinuity on one or more of these situation model dimensions. Whenever such discontinuities occur, it is important to have connectives, transitional phrases, adverbs, or other signaling

devices that convey to the readers that there is a discontinuity; we refer to these different forms of signaling as *particles*. Cohesion is facilitated by particles that clarify and stitch together the actions, goals, events, and states in the text.

Coh-Metrix analyzes the situation model dimension on causation, intentionality, space, and time, but not protagonists. There are many measures of the situation model, far too many to address in this article. For causal and intentional cohesion, Coh-Metrix computes the ratio of cohesion particles to the incidence of relevant referential content (i.e., main verbs that signal state changes, events, actions, and processes, as opposed to states). The ratio metric is essentially a conditionalized incidence of cohesion particles: Given the occurrence of relevant content (such as clauses with events or actions, but not states), Coh-Metrix computes the density of particles that stitch together the clauses. For example, the referential content for intentional information includes intentional actions performed by agents (as in stories, scripts, and common procedures); in contrast, the intentional cohesion particles include infinitives and intentional connectives (*in order to, so that, by means of*). Similarly, the referential content for causation information includes various classes of events that are identified by change-of-state verbs and other relevant classes of verbs in WordNet (Fellbaum, 1998). The causal particles are the causal connectives and other word classes that denote causal connections between constituents. In the case of temporal cohesion, Coh-Metrix computes the uniformity of the sequence of main verbs with respect to tense and aspect. More details are provided below for causality, intentionality, temporality, and spatiality.

3.3.1.1. Causality and intentionality: The distinction between causality and intentionality is based on the event-indexing model (Zwaan & Radvansky, 1998; Zwaan et al., 1995). Intentionality refers to the actions of animate agents as part of plans in pursuit of goals. Narrative texts are replete with such intentionality because they are stories about people with plans that follow a plot. In contrast, the causal dimension refers to mechanisms in the material world or psychological world that are not driven by goals of people. A text about scientific processes and mechanisms is a prototypical example of the causal dimension. Some researchers consider it important to distinguish between intentional and causal dimensions because they are fundamentally different categories of knowledge (Graesser & Hemphill, 1991; Keil, 1981) and may partly explain why science is more difficult to comprehend than stories. Other researchers believe that the distinction is unimportant or murky, so they choose to combine the causal and intentional dimensions into an overarching causal category. The distinction is made in Coh-Metrix, but researchers can decide for themselves whether to separate or combine them. In this article, we have combined these two dimensions into a single causality dimension.

3.3.1.2. Temporality: Assessing temporality in text is important because of its ubiquitous presence in organizing language and discourse. Time is represented through inflected tense morphemes (e.g., *-ed, is, has*) in every sentence of the English language. The temporal dimension also depicts unique internal event timeframes, such as an event that is complete (i.e., *telic*) or ongoing (i.e., *atelic*), by incorporating a diverse tense-aspect system (Ter Meulen, 1995). The occurrence of events at a point in time can be established by a large

repertoire of adverbial cues, such as *before*, *after*, *then*. These temporal features provide several different indices of the temporal cohesion of a text (Duran, McCarthy, Graesser, & McNamara, 2007).

Coh-Metrix temporal indices function through a repetition score that tracks the consistency of tense (e.g., *past* and *present*) and aspect (*perfective* and *progressive*) across a passage of text. The repetition scores decrease as shifts in tense and aspect are encountered. A low score indicates that the representation of time in the text is disjointed, thus having a possible negative influence on the construction of a mental representation. When such temporal shifts occur, the readers would encounter difficulties without explicit particles that signal such shifts in time, such as the temporal adverbial (*later on*), temporal connective (*before*), or prepositional phrases with temporal nouns (*on the previous day*). A low particle-to-shift ratio is a symptom of problematic temporal cohesion.

3.3.1.3. Spatiality: The Coh-Metrix spatial cohesion algorithm was designed to identify the spatial content of text. Herskovits (1998) proposed that there are two kinds of spatial information: *location information* and *motion information*. Herskovits also provided a list of particles that capture these two aspects of spatiality. For example, *beside, upon, here,* and *there* indicate location spatiality, whereas *into* and *through* indicate motion spatiality. Herskovits' theory was extended by assuming that motion spatiality is represented by motion verbs and that location spatiality is represented by location nouns. Classifications for both motion verbs (*move*, *go*, and *run*) and location nouns (*place, region*) were found in WordNet (Fellbaum, 1998). The Coh-Metrix algorithm computed location spatiality by counting the proportion of words that are location nouns, whereas the motion spatiality is the proportion of words that are motion verbs. In addition to estimating the amount of spatial information in a text (spatial *density*), we also measure the spatial *cohesion.* Our algorithm captures cohesion by computing the ratio of the density of location and motion particles (i.e., instances per 1,000 words) and the amount of spatial information in the text. The ratio score incorporates the joint influence of location and motion information.

3.3.2. Latent semantic analysis

In addition to the coreference variables discussed earlier, Coh-Metrix assesses conceptual overlap between sentences by a statistical model of word meaning: Latent Semantic Analysis (LSA; Landauer & Dumais, 1997; Landauer, McNamara, Dennis, & Kintsch, 2007; Millis et al., 2004). It should be noted that LSA is used to extract meaning from text in several of the articles featured in this issue. In our case, LSA is an important method of computing similarity because it considers implicit knowledge. LSA is a mathematical, statistical technique for representing world knowledge, based on a large corpus of texts. The central intuition is that the meaning of a word is captured by the company of other words that surround it in naturalistic documents. Two words have similarity in meaning to the extent that they share similar surrounding words. For example, the word *hammer* will be highly associated with words of the same functional context, such as *screwdriver*, *tool*, and *construction*. LSA uses a statistical technique called singular value decomposition to condense a very large corpus of texts to 100–500 statistical dimensions (Landauer et al., 2007). The conceptual

similarity between any two text excerpts (e.g., word, clause, sentence, text) is computed as the geometric cosine between the values and weighted dimensions of the two text excerpts. The value of the cosine typically varies from 0 to 1. LSA-based cohesion was measured in several ways in Coh-Metrix, such as LSA similarity between adjacent sentences, LSA similarity between all possible pairs of sentences in a paragraph, and LSA similarity between adjacent paragraphs.

The statistical representation of words in LSA depends on the corpus of texts on which they are trained. The users of Coh-Metrix have the option of declaring which corpus to use, but the corpus that is routinely used and serves as the default is the TASA (Touchstone Applied Science Associates) corpus of academic textbooks. TASA is a corpus of over 10 million words that cover a broad range of topics.

3.3.3. Given versus new information

Coh-Metrix supplies a LSA-based measure of given versus new information in text. Given information is recoverable either anaphorically or situationally from the preceding discourse, whereas new information is not recoverable (Haviland & Clark, 1974; Prince, 1981). The Coh-Metrix measure for given/new was based on a variant of LSA and statistically segregated new versus old information. The LSA-based statistical method is called *span* (Hempelmann et al., 2005; Hu et al., 2003; VanLehn et al., 2007). Rather than simply adding vectors, span constructs a hyperplane out of all previous vectors from sentences in the text. The comparison vector (in this case the current sentence in the text) is projected onto the hyperplane. The projection of the sentence vector on the hyperplane is considered to be the component of the vector that is shared with the previous text, or given (G). The component of the vector that is perpendicular to the hyperplane is considered new (N). To calculate the newness of the information, a proportion score is computed as: Span(new information) = N/(N + G). Hempelmann et al. (2005) reported that the span method has a high correlation with the theoretical analyses of give/new proposed by Prince (1981).

3.4. Genre and rhetorical composition

3.4.1. Genre

Coh-Metrix attempts to distinguish texts in three genres: *narrative*, *social studies*, and *science*. A reader's comprehension of a text can be facilitated by correctly identifying the textual characteristics that signal its genre (Biber, 1988). There is some evidence that training struggling readers to recognize genre and other aspects of global text structure helps them improve comprehension (Meyer & Wijekumar, 2007; Williams, 2007). Skilled readers activate particular expectations and strategies depending on the genre that is identified.

The genre indices in Coh-Metrix focus on the three domains most typical of high-school reading exercises: narrative, social studies, and science. The indices are derived from discriminant analyses conducted to identify the features that diagnostically predict the genre to which a text belongs (McCarthy et al., 2009). The algorithm produces three values, one for each genre. Another index is under development that assesses the extent to which a text is in a single genre as opposed to a blend of different genres.

3.4.2. Topic sentencehood

Researchers spanning the fields of composition, linguistics, and psychology frequently claim that topic sentences help readers better remember text and facilitate comprehension (Kieras, 1978). Topic sentences theoretically have a number of intrinsic features. They are a *claim* or assertion about the main theme or topic of the paragraph and are elaborated by other sentences in the paragraph. They ideally occur in the *paragraph initial* position. Topic sentences are more likely to appear in informational texts than narrative texts. In spite of this theoretical analysis, the empirical research has revealed that topic sentences appear in only about 50% of paragraphs (Popken, 1991). Coh-Metrix provides a number of indices of topic sentencehood that are reported in McCarthy et al. (in press). Some indices analyze the intrinsic characteristics of sentences, whereas other indices are relative in the sense that there are comparisons between sentences in a paragraph.

4. Example uses of Coh-Metrix to investigate multilevel discourse comprehension

Coh-Metrix has been used in dozens of projects that investigate characteristics of discourse, comprehension, memory, and learning (McNamara et al., 2008, in press). These studies have validated the Coh-Metrix measures by comparing the computer output to language and discourse annotations by experts, to texts scaled on cohesion, to psychological data (e.g., ratings, reading times, memory, test performance), and to samples of texts that serve as gold standards. Coh-Metrix has uncovered differences among a wide range of discourse categories in levels 4 and 5, such as differences between (a) spoken and written samples of English (Louwerse, McCarthy, McNamara, & Graesser, 2004), (b) physics content in textbooks, texts prepared by researchers, and conversational discourse in tutorial dialogue (Graesser, Jeon, Yang, & Cai, 2007), (c) articles written by different authors (McCarthy, Lewis, Dufty, & McNamara, 2006), (d) sections in typical science texts, such as *introductions*, *methods*, *results*, and *discussions* (McCarthy, Briner, Rus, & McNamara, 2007), and (e) texts that were *adopted* (or authentic) versus *adapted* (or simplified) for second language learning (Crossley, Louwerse, McCarthy, & McNamara, 2007).

It is beyond the scope of this article to review the large body of research in support of Coh-Metrix or to present a systematic analysis of a new corpus of texts. Instead, we will focus on a theoretical issue and a practical issue. Regarding theory, we show how Coh-Metrix can help discourse researchers better understand cohesion in the multilevel discourse comprehension framework. As discussed earlier, the textbase and situation model have a large impact on reading time and memory because these are the levels that contain the most novel information (along with lower frequency words) and these are the levels that most strongly impact discourse coherence. Regarding a practical issue, we address the problems of scaling texts on difficulty and of matching texts to different populations of readers. The texts assigned to a reader allegedly should not be too hard or too easy, but at the zone of proximal development (Wolfe et al., 1998). We make the case that Coh-Metrix has the foundation to go beyond standard text difficulty formulae in this endeavor. We do this by

presenting some output from Coh-Metrix on particular discourse samples and texts in the narrative and science genre that are judiciously selected to make our argument.

4.1. Cohesion at the textbase and situation model levels

As discussed earlier, it is well documented that cohesion at the textbase and situation model levels has a robust impact on reading time, comprehension, inference generation, and memory. Breaks in discourse cohesion substantially increase reading time (Zwaan & Radvansky, 1998; Zwaan et al., 1995), invite inferences to fill in or explain the coherence gaps (Graesser et al., 1994; McKoon & Ratcliff, 1992; O'Brien et al., 1998), and negatively influence memory for the material (Britton & Gulgoz, 1991; McNamara & Kintsch, 1996; McNamara et al., 1996; O'Reilly & McNamara, 2007). However, it is not the case that more cohesion is always better. The reverse-cohesion effect documented by McNamara and her colleagues (McNamara, 2001; McNamara & Kintsch, 1996; McNamara et al., 1996; O'Reilly & McNamara, 2007) has often shown that higher cohesion texts result in a less coherent mental representation of the text for those readers who have high domain knowledge. Whereas low-knowledge readers more consistently gain from increases in text cohesion, high-knowledge readers either show no effects of cohesion or show a reversed effect, whereby they benefit from the cohesion gaps. There are different explanations for this reverse-cohesion effect. The explanation proffered by McNamara and colleagues is that the readers with high knowledge are lulled into a sense that they have good comprehension of the high-cohesion texts so they expend less effort generating inferences; in contrast, the cohesion gaps force them to generate inferences and self-explanations of the material.

The complex interactions among characteristics of the text, the reader, as well as the level of cognitive representation (i.e., textbase vs. situation model) underscore the importance of systematically scaling the discourse. There is a large literature on cohesion manipulation studies in which researchers prepare low-versus high-cohesion texts and investigate the impact of the cohesion manipulation on psychological measures (for reviews, see Britton, Gulgoz, & Glynn, 1993; McNamara et al., in press). Most of the studies increased cohesion by either improving referential cohesion or inserting connectives and other discourse markers. Referential cohesion is potentially compromised when there are pronouns and when the same discourse entity is referred to with different noun-phrases. Referential cohesion is enhanced when, for example, the pronouns are filled with head nouns (e.g., *it* $\rightarrow$ *evaporation*) and a particular discourse entity is referred to with the same head noun (*rapid evaporation … conversion to vapor* $\rightarrow$ *rapid evaporation … evaporation*). Connectives and various forms of discourse markers (e.g., *because, also, therefore, however*) specify how clauses and sentences are linked at either the textbase- or situation-model level.

If Coh-Metrix produces *sensitive* measures of cohesion, it should produce values on the cohesion measures that directly reflect the text manipulations in these cohesion manipulation studies. If a measure of cohesion is accurately *discriminating*, it should influence the intended measure, but not other aspects of cohesion or other levels of discourse analysis. A neat and tidy measure would be ideal, but it is widely acknowledged that there are potential unintended consequences and extraneous variables associated with any text manipulation.

For example, increasing cohesion tends to increase sentence length and thereby may impose a load on working memory (McNamara et al., in press; Millis, Graesser, & Haberlandt, 1993; Ozuru, Dempsey, & McNamara, 2009). Side effects at the surface code level may also include changes in word frequency and syntactic complexity. Coh-Metrix can assess these possible side effects, or what we call *fallout* variables.

McNamara et al. (in press) analyzed the texts in 19 experiments that manipulated text cohesion (high vs. low) through referring expressions, connectives, and discourse markers. Results showed that the Coh-Metrix indices of cohesion (individually and combined) significantly distinguished the high versus low cohesion versions of these texts. Compared with the low-cohesion texts, the texts with high cohesion had significantly higher referential cohesion (noun, argument, stem), more causal connectives, higher causal cohesion, and higher LSA overlap scores between sentences within paragraphs. The effect sizes (*d*, in standard deviation units) for these cohesion measures varied from 0.51 to 1.08 ($M = 0.82$). The results also showed that potential confounding variables had very small or nonsignificant effects. The effect sizes for all of the potential fallout variables were: words per sentence (0.59), word frequency (0.58), number of sentences (0.25), Flesch-Kincaid grade level (0.20), and word concreteness (0.16). These effect sizes for the fallout variables tended to be lower than the cohesion variables. When a discriminant analysis was conducted with these predictor variables to classify high- versus low-cohesion texts, the cohesion variables were significant but the fallout variables were not. It is also informative to point out that the directions of the tends for the fallout variables did not make sense with respect to impact on reading time, memory, and other psychological measures. The high-cohesion texts had slightly more words per sentence, lower word frequency, more sentences per text, a higher grade level, and more abstract words. If anything, these trends in the fallout variables should increase reading time and decrease comprehension and memory. This is opposite to the predicted facilitation of cohesion on psychological processing (ignoring the reverse-cohesion effect).

4.2. Challenges to standard text difficulty measures: The need for measures of multilevel discourse comprehension

Texts in school systems are normally measured with readability formulae that assess text difficulty. The purpose of these difficulty measures is to make sure that students receive age appropriate texts in their curriculum. That is, the texts should not be too hard or too easy, but just right—at the zone of proximal development (Wolfe et al., 1998). Widely adopted measures of text difficulty are the Flesch-Kincaid Grade Level (Klare, 1974–1975), Degrees of Reading Power (DRP; Koslin, Zeno, & Koslin, 1987), and Lexile scores (Stenner, 2006). Despite the seeming diversity of these difficulty formulae, these measures are all based on, or highly correlated with, two variables: the frequency or familiarity of the words, and the length of the sentences. Word length has a strong negative correlation with word frequency (Haberlandt & Graesser, 1985; Zipf, 1949), so number of letters or syllables provides an excellent proxy for word frequency or familiarity. Thus, many measures are based simply on the length of the words and sentences.

In this section, we present an argument that standard text difficulty metrics are limited by the fact that they do not consider important levels of a multilevel discourse comprehension framework. We make the argument with two ways. First, we take a sample of texts at a particular grade level (namely seventh grade) and show that there are a large number of theoretically expected differences between science and narrative texts at multiple levels of language and discourse. These differences, which are exposed by Coh-Metrix, are not captured by the standard text difficulty metrics. Second, we identify some very successful books in science and literature that are not aligned with the projections of text readability; however, Coh-Metrix provides some clues on what aspects of language and discourse might predict their success. Additional research is needed, however, to assess the generality of our argument with respect to a larger age span and sample of texts.

Texts in the narrative and science genre are interesting to compare for both theoretical and practical reasons. World knowledge tends to be very high for narrative discourse and low for informational texts (including science). Narrative texts are more likely to convey life experiences, person-oriented dialogue, and familiar language in the oral tradition than are informational expository texts (Bruner, 1986; Graesser, 1981; Rubin, 1995). The purpose of informational texts is to inform the reader of new knowledge, so by definition they generally use more unfamiliar words and demand more knowledge that fewer people possess. Narrative has foundation in oral discourse, so it typically has simpler words, syntax, and information pacing in order to accommodate face-to-face synchronous communication is a specific context (Clark, 1996; Tannen, 1982). In contrast, informational text (including science) has a language aligned with print media that can accommodate higher information density because it allows rereading and reflection at the comprehender's own pace. Narrative texts are read nearly twice as fast as informational texts but remembered nearly twice as well (Graesser, 1981; Haberlandt & Graesser, 1985). The contrast between narrative and informational text is fundamental in education and discourse research (Biber, 1988; Louwerse et al., 2004; VanderVeen et al., 2007).

Coh-Metrix output was compared for texts in narrative versus science genres. The first set of analyses consisted of a normative set of texts from the TASA corpus. The TASA corpus has been frequently analyzed by discourse researchers who have investigated contemporary techniques in computational linguistics (Landauer & Dumais, 1997; McNamara, Louwerse, & Graesser 2008). TASA is a corpus of 12 million words in over 30,000 documents that cover a broad range of topics that high school students would have experienced. TASA classifies the texts into the science versus narrative categories, as well as other genres/registers. Therefore, there is an independent, objective, operational definition for classifying the texts into these two categories. We randomly selected 100 science texts and 100 narrative texts from the large TASA corpus. The sampled texts were approximately at the seventh-grade level according to the Flesch-Kincaid Grade Level.

The first four columns of Table 2 present the means and standard deviations collected from the TASA sample. The science and narrative text samples are both at the seventh grade level according to the Flesch-Kincaid Grade Level metric (which correlates with DRP at $r > .90$). Therefore, the framework behind the text difficulty metric offers no a priori prediction that there should be any differences between these genres. In contrast, researchers who

investigate differences between narrative and science texts would anticipate many differences in the surface code, textbase, and situation model. Given the sample size of 100 for the TASA texts, one can derive the 95% confidence intervals for assessing whether a particular score is outside of the range of the mean, given an alpha level of .05. The general formula would be Mean ± [1.96 * *SD*/SQRT(100)]. For example, the mean number of negations for narrative is 9.6 and the standard deviation (*SD*) is 7.2. The 95% confidence interval would be 9.6 ± 1.4. That is, scores between 8.2 and 11.0 are not significantly different than the mean of 9.6. Given that science texts have a mean of 6.3 negations, there are significantly more negations in narrative text than science texts.

The results in Table 2 support the need to consider a large number of measures associated with the multilevel discourse framework rather than a simple metric of text difficulty.

Table 2
Coh-Metrix analysis of printed texts

	TASA							
	Science		Narrative		Hewitt	Einstein	Way Things	Harry
	M	*SD*	*M*	*SD*	Physics	Dreams	Work	Potter
Basic statistics								
Number of texts	100		100		8	10	1	1
Text length	276		290		5,967	667	67	63
Words per sentence	12.7		18.0		18.2	15.1	16.8	31.5
Flesch-Kincaid grade level	7		7		9	7	7	12
Surface code								
Logarithm of frequency of content words	2.24	(0.15)	2.27	(0.15)	2.15	2.27	1.11	1.94
Noun concreteness in hierarchy (hypernym)	4.86	(0.50)	5.28	(0.51)	4.89	4.98	6.07	5.22
Verb concreteness in hierarchy (hypernym)	1.42	(0.19)	1.49	(0.19)	1.44	1.48	1.69	1.63
All connectives	69.7	(19.4)	79.0	(19.4)	69.3	66.8	104.5	63.5
Logical connectives	30.8	(12.2)	31.0	(11.8)	38.0	29.8	29.8	31.7
Negations	6.3	(6.4)	9.5	(7.2)	7.9	14.1	0	0
Personal pronouns	43.4	(26.9)	95.2	(33.9)	43.4	26.9	95.2	33.9
Pronoun ratio per noun-phrase	0.15	(0.10)	0.34	(0.12)	0.10	0.24	0.06	0.44
Modifiers per noun phrase	0.91	(0.18)	0.80	(0.14)	0.93	0.81	1.11	1.28
Words before main verb of main clause	3.82	(1.12)	3.93	(1.52)	5.21	4.20	6.00	5.50
Textbase								
Adjacent argument overlap	0.63	(0.17)	0.49	(0.21)	0.66	0.42	0.67	1.00
Adjacent stem overlap	0.62	(0.18)	0.24	(0.18)	0.64	0.26	0.67	1.00
Content word overlap	0.15	(0.06)	0.09	(0.04)	—	0.09	0.23	0.11
Type-token ratio	0.66	(0.09)	0.82	(0.06)	0.36	0.72	0.71	0.97
Situation model dimensions								
LSA sentence overlap—adjacent	0.43	(0.11)	0.28	(0.09)	0.36	0.12	0.51	0.04
LSA sentence overlap—all	0.32	(0.11)	0.25	(0.09)	—	0.09	0.51	0.04
Causal cohesion	1.21	(1.15)	3.05	(2.42)	0.33	2.14	0.50	0.33
Temporal cohesion	0.84	(0.09)	0.85	(0.09)	0.87	0.86	1.00	1.00

TASA, Touchstone Applied Science Associates.

Table 2 lists 18 variables measured by Coh-Metrix and 15 are statistically significant. The only nonsignificant measures are number of logical connectives, number of words before the main verb of the main clause, and temporal cohesion. Compared to the science texts, the narrative texts have significantly more frequent words, more concrete nouns and verbs, more connectives (notably the additive connective *and* in our follow-up analyses), more negations, more personal pronouns, more pronouns per noun-phrases, fewer modifiers per noun-phrase, lower cohesion in all textbase measures, lower LSA overlap scores, and higher causal cohesion. Narrative appears to favor word and sentence processing, but there is lower semantic cohesion between sentences. It suffices to say that Flesch-Kincaid text difficulty measure hardly goes the distance in differentiating narrative and science texts.

We now turn to some outstanding texts that are highly respected in educational communities or have high sales. Half are narrative and half in the science genres. Graesser et al. (2007) analyzed chapters in a popular introductory physics text in its eighth edition, entitled *Conceptual Physics* (Hewitt, 1998). Graesser, Jeon, Cai, and McNamara (2008) analyzed 10 chapters in a best-selling novelette in the narrative genre that was written by a physicist, entitled *Einstein's Dreams* (Lightman, 1993). The texts in both studies were written by physicists and had very large sales, so the writers came from similar intellectual communities.

Hewitt's *Conceptual Physics* textbook differs from the TASA gold standard science texts on a few Coh-Metrix dimensions in addition to the fact that the Hewitt texts are at a ninth grade level and the TASA norms are grade 7. Of the 16 comparisons in Table 2, the physics texts have significantly less frequent content words, more logical connectives, more negations, fewer pronouns per noun-phrase, more words before the main verb of the main clause, a lower type-token ratio, lower LSA overlap in adjacent sentences, and lower causal cohesion. Eight of 16 measures are significantly different; 7 of the 8 suggested more difficult comprehension for the Hewitt physics text. These results suggest that this physics textbook is too difficult for seventh graders, an outcome that squares away with the Flesch-Kincaid grade level. So here we see a compatibility between Flesch-Kincaid text difficulty metrics and Coh-Metrix measures.

Next consider the *Einstein's Dreams* text, which is in the narrative genre and at the seventh grade level. The Einstein's Dream texts differ from the TASA gold standard narratives on several dimensions when inspecting the 18 measures in Table 2. Given the confidence intervals of the TASA narrative measures (the Mean ± approximately 1⁄5 of the standard deviation), the Einstein's Dream texts have significantly fewer concrete nouns, fewer connectives, more negations, fewer personal pronouns, fewer pronouns per noun-phrase, lower adjacent argument overlap, lower type token ratio (lexical diversity), and lower LSA overlap scores for adjacent sentences and all possible pairs of sentences. There were significant differences for 9 of 18 measures, with 7 of 9 measures suggesting the Einstein's Dream texts were more difficult to comprehend than the narrative texts in the gold standard TASA sample; the only two measures signaling easier processing in Einstein's Dream were fewer pronouns per noun-phrase and a lower type-token ratio. This profile of Coh-Metrix data suggests that *Einstein's Dreams* texts may be too difficult to assign to seventh graders unless the teacher wants to stretch the literacy levels of the students. However, the text difficulty metrics suggest these texts would be perfectly fine for seventh graders.

Analyses of a couple of specific texts further illustrate the value of using Coh-Metrix to examine the multiple levels of discourse that underlie text difficulty. Consider a text snippet from an illustrated text on cylinder locks in *The Way Things Work* written by David Macaulay (1988). This is a popular book that describes how mechanical and electrical artifacts work, with texts, pictures, labels, and arrows. The text below is a snippet from an illustrated text on the cylinder lock.

> When the door is closed, the spring presses the bolt into the door frame. Inserting the key raises the pins and frees the cylinder. When the key is turned, the cylinder rotates, making the cam draw back the bolt against the spring. When the key is released, the spring pushes back the bolt, rotating the cylinder to its initial position and enabling the key to be withdrawn.

This snippet of text is at the seventh grade reading level so we examined how its profile of Coh-Metrix values compared with the confidence intervals in the TASA science sample. The snippet would be considered difficult when inspecting its comparatively rare words, higher lexical diversity, and complex syntax (see Table 2). However, comprehension is facilitated by virtue of the absence of negations, more concrete words, and higher cohesion scores on most measures of the textbase and situation model. These latter measures, interestingly, are not publically available in computerized text analysis programs (other than Coh-Metrix). In contrast, there are many commercial computer sites that analyze word frequency, lexical diversity, and syntax. These programs would present a misleading conclusion on the difficulty level of cylinder lock snippet because discourse cohesion, word concreteness, and negation density would have been ignored. This example illustrates the importance of analyzing the full spectrum of multilevel discourse comprehension.

Consider next a text snippet from *Harry Potter and the Sorcerer's Stone*, by J.K. Rowling (1998). This is the top-selling novel written in the last 40 years—read by children and adults.

> The escape of the Brazilian boa constrictor earned Harry his longest-ever punishment. By the time he was allowed out of his cupboard again, the summer holidays had started and Dudley had already broken his new video camera, crashed his remote control airplane, and, first time out on his racing bike, knocked down old Mrs. Figg as she crossed Privet Drive on her crutches.

This snippet would be considered too difficult for all of the children in elementary and middle school who avidly read it with tremendously high motivation. When comparisons are made to the narrative TASA norms, we see from the profile of measures (in the right column in Table 2) that it is written at the 12th grade reading level, has comparatively low frequency words, fewer connectives, more noun-phrases filled with pronouns (but fewer personal pronouns), more complex syntax, higher lexical diversity, and lower cohesion for most measures at the situation model. This example illustrates that children can read narratives quite beyond their reading grade level and beyond the language and discourse norms

of TASA. The example also complicates the challenge of assigning texts to readers in a fashion that optimizes text comprehension.

This subsection illustrates how texts can be scaled on multiple levels of discourse, how text samples vary on difficulty, and how challenging it might be to match texts to readers. We argue that multiple levels of language and discourse need to be considered in a comprehensive computational model of extracting meaning from text and that a simple text difficulty formula will not go the distance in scaling texts in the educational enterprise.

Acknowledgments

This research was supported by the National Science Foundation (ITR 0325428, ALT-0834847, BCS 0904909, DRK-12-0918409), the Institute of Education Sciences (R305G020018, R305H050169, R305B070349, R305A080589, R305A080594), and the U.S. Department of Defense Counterintelligence Field Activity (H9C104-07-0014). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, IES, and DoD. Our gratitude goes to Zhiqiang Cai for his software development of Coh-Metrix, to Phil McCarthy, Max Louwerse, and Moongee Jeon for their testing of Coh-Metrix on text samples, and to David Dufty and Karl Haberlandt for supplying and analyzing the reading time data.

References

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database (Release 2) {CD-ROM}*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania {Distributor}.

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

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, *21*(4), 543–565.

Britton, B. K., & Gulgoz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology*, *83*, 329–345.

Britton, B., Gulgoz, S., & Glynn, S. (1993). Impact of good and poor writing on learners: Research and theory. In B. K. Britton, A. Woodward, & M. R. Binkley (Eds.), *Learning from textbooks: Theory and practice* (pp. 1–46). Hillsdale, NJ: Erlbaum.

Bruner, J. (1986). *Actual minds, possible worlds*. Cambridge, MA: Harvard University Press.

Carver, R. P. (1990). *Reading rate: A review of research and theory*. New York: Academic Press.

Charniak, E. (2000). A maximum-entropy-inspired parser. In J. Wiebe (Ed.), *Proceedings of the First Conference on North American Chapter of the Association for Computational Linguistics* (pp. 132–139). San Francisco: Morgan Kaufmann Publishers.

Clark, H. H. (1996). *Using language*. Cambridge, England: Cambridge University Press.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, *33A*, 497–505.

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

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

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

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database [CD-ROM]*. Cambridge, MA: MIT Press.

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Erlbaum.

Gough, P. B. (1972). One second of reading. In J. F. Kavanaugh & J. G. Mattingly (Eds.), *Language by ear and by eye* (pp. 331–358). Cambridge, MA: MIT Press.

Graesser, A. C. (1981). *Prose comprehension beyond the word*. New York: Springer-Verlag.

Graesser, A. C., Cai, Z., Louwerse, M., & Daniel, F. (2006). Question Understanding Aid (QUAID): A web facility that helps survey methodologists improve the comprehensibility of questions. *Public Opinion Quarterly*, *70*, 3–22.

Graesser, A. C., & Hemphill, D. (1991). Question answering in the context of scientific mechanisms. *Journal of Memory and Language*, *30*, 186–209.

Graesser, A.C., Jeon, M., Cai, Z., & McNamara, D.S. (2008). Automatic analyses of language, discourse, and situation models. In J. Auracher & W. Van Peer (Eds.), *New beginnings in literary studies* (pp. 72–88. Cambridge, England: Cambridge Scholars Publishing.

Graesser, A. C., Jeon, M., Yang, Y., & Cai, Z. (2007). Discourse cohesion in text and tutorial dialogue. *Information Design Journal*, *15*, 199–213.

Graesser, A. C., Lu, S., Olde, B. A., Cooper-Pye, E., & Whitten, S. (2005). Question asking and eye tracking during cognitive disequilibrium: Comprehending illustrated texts on devices when the devices break down. *Memory and Cognition*, *33*, 1235–1247.

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

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

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology*, *48*, 163–189.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*, 371–395.

Haberlandt, K. F., & Graesser, A. C. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General*, *114*, 357–374.

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

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

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

Hempelmann, C. F., Rus, V., Graesser, A. C., & McNamara, D. D. (2006). Evaluating the state-of-the-art treebank-style parsers for Coh-Metrix and other learning technology environments. *Natural Language Engineering*, *12*, 131–144.

Herskovits, A. (1998). Schematization. In P. Olivier & K. P. Gapp (Eds.), *Representation and processing of spatial expressions* (pp. 149–162). Mahwah, NJ: Erlbaum.

Hess, D. J., Foss, D. A., & Carroll, P. (1995). Effects of global and local context on lexical processing during language comprehension. *Journal of Experimental Psychology: General*, *124*, 62–82.

Hewitt, P. G. (1998). *Conceptual physics*. Menlo Park, CA: Addison-Wesley.

Hu, X., Cai, Z., Louwerse, M., Olney, A., Penumatsa, P., Graesser, A. C., & the Tutoring Research Group. (2003). A revised algorithm for latent semantic analysis. In G. Hlob & T. Walsh (Eds.), *Proceedings of the 2003 International Joint Conference on Artificial Intelligence* (pp. 1489–1491. San Francisco: Morgan Kaufmann.

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Boston: Allyn & Bacon.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122–149.

Keil, F. C. (1981). Constints on knowledge and cognitive development. *Psychological Review*, *88*, 197–227.

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. (1998). *Comprehension: A paradigm for cognition*. Cambridge, England: Cambridge University Press.

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

Kintsch, W., Welsh, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language*, *29*, 133–159.

Klare, G. R. (1974–1975). Assessing readability. *Reading Research Quarterly*, *10*, 62–102.

Koslin, B. I., Zeno, S., & Koslin, S. (1987). *The DRP: An effective measure in reading*. New York: College Entrance Examination Board.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.) (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Erlbaum.

Lappin, S., & Leass, H. J. (1994). An algorithm for pronominal coreference resolution. *Computational Linguistics*, *20*, 535–561.

Lightman, A. P. (1993). *Einstein's Dreams*. New York: Random House, Inc.

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

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). Mahwah, NJ: Erlbaum.

Macaulay, D. (1988). *The way things work*. Boston: Houghton Mifflin.

Magliano, J. P., & Radvansky, G. A. (2001). Goal coordination in narrative comprehension. *Psychonomic Bulletin and Review*, *8*, 372–376.

Mandler, J. M. (1988). How to build a baby: On the development of an accessible representational system. *Cognitive Development*, *3*, 113–136.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*, 313–330.

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, UK: Springer-Verlag.

McCarthy, P. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing*, *24*, 4.

McCarthy, P. M., Lewis, G. A., Dufty, D. F., & McNamara, D. S. (2006). Analyzing writing styles with Coh-Metrix. In G. C. J. Sutcliffe and R. G. Goebel (Eds.), *Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS)* (pp. 764–770). Menlo Park, CA: AAAI Press.

McCarthy, P. M., Myers, J. C., Briner, S. W., Graesser, A. C., & McNamara, D. S. (2009). Are three words all we need? A psychological and computational study of genre recognition. *Journal for Computational Linguistics and Language Technology*, *1*, 23–57.

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

McClelland, J. L. (1986). The programmable blackboard model of reading. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing*, Vol. 2. (pp. 122–169). Cambridge, MA: MIT Press.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review*, *99*, 440–466.
McNamara, D. S. (2001). Reading both high and low coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology*, *55*, 51–62.
McNamara, D. S., Graesser, A., & Louwerse, M. (in press). Sources of text difficulty: Across the ages and genres. In J. P. Sabatini & E. Albro (Eds.), *Assessing reading in the 21st century: Aligning and applying advances in the reading and measurement sciences*. Lanham, MD: Rowman & Littlefield Education.
McNamara, D. S., & Kintsch, W. (1996). Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*, 247–287.
McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*, 1–43.
McNamara, D. S., Louwerse, M. M., & Graesser, A. C. (2008). *Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension*. Final report on Institute of Education Science grant (R305G020018). University of Memphis, Memphis, TN.
McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (in press). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*.
McNamara, D. S., & Magliano, J. (2008). Towards a comprehensive model of comprehension. In B. Ross (Ed.), *The psychology of learning and motivation*, Vol. 51, (pp. 297–384). New York: Elsevier Science.
Meyer, B. J. F., & Wijekumar, K. (2007). Web-based tutoring of the structure strategy: Theoretical background, design, and findings. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 347–375). Mahwah, NJ: Erlbaum.
Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *Journal of Lexiography*, *3*, 235–244.
Millis, K., Graesser, A. C., & Haberlandt, K. (1993). The impact of connectives on memory for expository texts. *Applied Cognitive Psychology*, *7*, 317–340.
Millis, K. K., Kim, H. J., Todaro, S., Magliano, J., Wiemer-Hastings, K., & McNamara, D. S. (2004). Identifying reading strategies using latent semantic analysis: Comparing semantic benchmarks. *Behavior Research Methods, Instruments, & Computers*, *36*, 213–221.
O'Brien, E. J., Rizzella, M. L., Albrecht, J. E., & Halleran, J. G. (1998). Updating a situation model: A memory-based text processing view. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *24*, 1200–1210.
O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes*, *43*, 121–152.
Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction*, *19*, 228–242.
Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic inquiry and word count*. Austin, TX: LIWC.net (http://www.liwc.net).
Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific studies of Reading*, *11*, 357–383.
Pickering, M., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169–226.
Popken, R. (1991). A study of topic sentence use in technical writing. *The Technical Writing Teacher*, *18*, 49–58.
Prince, E. F. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Radical pragmatics* (pp. 223–255). New York: Academic Press.
Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372–422.
Rayner, K., & Pollatsek, A. (1994). *The psychology of reading*. Mahwah, NJ: Erlbaum.
Rowling, J. K. (1998). *Harry Potter and the sorcerer's stone*. New York: Scholastic.
Rubin, D. C. (1995). *Memory in oral traditions: The cognitive psychology of epic, ballads, and counting-out rhymes*. New York: Oxford University Press.

Rumelhart, D. E. (1977). Toward an interactive model of reading. In S. Dornie (Ed.), *Attention and performance* (pp. 573–603). Hillsdale, NJ: Erlbaum.

Sanders, T. J. M., & Noordman, L. G. M. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes*, *29*, 37–60.

Schmalhofer, F., & Glavanov, D. (1986). Three components of understanding a programmer's manual: Verbatim, propositional, and situational representations. *Journal of Memory and Language*, *25*, 279–294.

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND Corporation.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*, 360–407.

Stenner, A. J. (2006). *Measuring reading comprehension with the Lexile framework*. Durham, NC: Metametrics, Inc. Presented at the California Comparability Symposium, October 1996. Available at http://www.lexile.com/DesktopDefault.aspx?view=re. Accessed January 30, 2006.

Tannen, D. (1982). The oral/literate continuum in discourse. In D. Tannen (Ed.), *Spoken and written language* (pp. 1–16). Norwood, NJ: Ablex.

Templin, M. (1957). *Certain language skills in children: Their development and interrelationships*. Minneapolis, MN: The University of Minnesota Press.

Ter Meulen, A. G. B. (1995). *Representing time in natural language: The dynamic interpretation of tense and aspect*. Cambridge, MA: MIT Press.

Tulving, E., & Kroll, N. (1995). Novelty assessment in the brain: Long-term memory encoding. *Psychonomic Bulletin & Review*, *2*, 387–390.

Van den Broek, P., Rapp, D. N., & Kendeou, P. (2005). Integrating memory-based and constructionist processes in accounts of reading comprehension. *Discourse Processes*, *39*, 299–316.

Van den Broek, P., Virtue, S., Everson, M. G., Tzeng, Y., & Sung, Y. (2002). Comprehension and memory of science texts: Inferential processes and the construction of a mental representation. In J. Otero, J. Leon, & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 131–154). Mahwah, NJ: Erlbaum.

VanderVeen, A., Huff, K., Gierl, M., McNamara, D. S., Louwerse, M., & Graesser, A. C. (2007). Developing and validating instructionally relevant reading competency profiles measured by the critical reading sections of the SAT. In D. S. McNamara (Ed.), *Theories of text comprehension: The importance of reading strategies to theoretical foundations of reading comprehension* (pp. 137–172). Mahwah, NJ: Erlbaum.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, *31*, 3–62.

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* (pp. 199–219). Mahwah, NJ: Erlbaum.

Wolfe, M. B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Learning from text: Matching readers and text by latent semantic analysis. *Discourse Processes*, *25*, 309–336.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Reading, MA: Addison-Wesley.

Zwaan, R. A. (1994). Effect of genre expectations on text comprehension. *Journal of Experimental Psychology: Learning, Memory, Cognition*, *20*, 920–933.

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 386–397.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*, 162–185.