

# Coh-Metrix: Analysis of text on cohesion and language

ARTHUR C. GRAESSER, DANIELLE S. McNAMARA, MAX M. LOUWERSE, and ZHIQIANG CAI  
*University of Memphis, Memphis, Tennessee*

Advances in computational linguistics and discourse processing have made it possible to automate many language- and text-processing mechanisms. We have developed a computer tool called Coh-Metrix, which analyzes texts on over 200 measures of cohesion, language, and readability. Its modules use lexicons, part-of-speech classifiers, syntactic parsers, templates, corpora, latent semantic analysis, and other components that are widely used in computational linguistics. After the user enters an English text, Coh-Metrix returns measures requested by the user. In addition, a facility allows the user to store the results of these analyses in data files (such as Text, Excel, and SPSS). Standard text readability formulas scale texts on difficulty by relying on word length and sentence length, whereas Coh-Metrix is sensitive to cohesion relations, world knowledge, and language and discourse characteristics.

The vision of having a computer understand natural language has persisted for nearly half a century, but it has been challenged by the computational difficulty of simulating many of the processing components. The standard wisdom has been that it is particularly more difficult to implement the deeper and more global levels of comprehension than the shallow and more local levels. The deeper and global levels require semantic interpretation, the construction of mental models, text cohesion, pragmatics, rhetorical composition, and world knowledge (Lehnert & Ringle, 1982; Schank & Riesbeck, 1981). The computer is more reliable in managing the more shallow and local components, such as accessing words from electronic dictionaries, identifying misspelled words, and splitting up words into component syllables, basic meaning units (morphemes), or sound units (phonemes).

Recent landmark progress across several disciplines has made it possible to explore computational measures of language and text comprehension that go beyond these surface components. These disciplines include computa-

tional linguistics (Allen, 1995; Jurafsky & Martin, 2000; Moore & Wiemer-Hastings, 2003), corpus linguistics (Biber, Conrad, & Reppen, 1998; Marcus, Santorini, & Marcinkiewicz, 1993), information extraction (DARPA, 1995; Lehnert, 1997; Pennebaker & Francis, 1999), information retrieval (Belew, 2002; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Graesser, Burger, et al., 2003; Robertson, 2001; Voorhees, 2001), and discourse processing (Graesser, Gernsbacher, & Goldman, 2003; Kintsch, 1998). As a consequence of these advances in scientific research and language technologies, we can go some distance in automating many of the deeper and global levels of text and language analysis.

One such level of language analysis that presents particular computational challenges is called *coherence* or *cohesion* (Graesser, McNamara, & Louwerse, 2003; Louwerse & Mitchell, 2003; McNamara, E. Kintsch, Songer, & W. Kintsch, 1996). For the purpose of clarity, we make a distinction between cohesion and coherence. Specifically, cohesion is a characteristic of the text, whereas coherence is a characteristic of the reader's mental representation of the text content. Cohesion is an objective property of the explicit language and text. There are explicit features, words, phrases, or sentences that guide the reader in interpreting the substantive ideas in the text, in connecting ideas with other ideas, and in connecting ideas to higher level global units (e.g., topics and themes). These cohesive devices cue the reader on how to form a coherent representation. The coherence relations are constructed in the mind of the reader and depend on the skills and knowledge that the reader brings to the situation. If the reader has adequate world knowledge about the subject matter and if there are adequate linguistic and discourse cues, then the reader is likely to form a coherent mental representation of the text. A reader perceives a text to be coherent to the extent that the ideas conveyed in the text hang together in a meaningful and organized manner. Thus, coherence is an

---

The research was supported by Institute for Education Sciences Grant IES R3056020018-02 and National Science Foundation Grant SES 9977969. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the IES or the NSF. We thank Xiangen Hu, who provided invaluable advice and help at all levels of this project. We thank Zhijun Lu, who created the Data Viewer facility for Coh-Metrix. We are grateful to David Dufty for helping to write the Help files on Coh-Metrix and for extensive testing of the system. Finally, we thank the following students and postdoctoral fellows, who provided help in testing this system and various other aspects of the project: Rachel Best, Corina Castellano, Kyle Dempsey, Randy Floyd, Phil McCarthy, Tenaha O'Reilly, Yasuhiro Ozuru, Margie Petrowski, Srinivas Pillarisetti, Mack Reese, Mike Rowe, Jayme Sayroo, Kim Sumara, and Fang Yang. Correspondence concerning this manuscript should be sent A. C. Graesser or D. S. McNamara, Department of Psychology, 202 Psychology Building, University of Memphis, Memphis, TN 38152-3230 (e-mails: a-graesser@memphis.edu; d.mcnamara@mail.psy.memphis.edu).

achievement that is a product of psychological representations and processes. Simply put, coherence is a psychological construct, whereas cohesion is a textual construct (Graesser, McNamara, & Louwerse, 2003; Louwerse, 2002; Louwerse & Graesser, in press).

Our focus on coherence and cohesion has multiple inspirations in the arena of science. First, the *coherence assumption* was one of the central theoretical constructs in the constructivist theory of discourse comprehension proposed by Graesser, Singer, and Trabasso (1994). According to this assumption, readers routinely attempt to construct coherent meanings and connections among text constituents unless the text is very poorly composed and they give up trying. Given the significance of the coherence assumption in the theory, it would be important to dissect and possibly automate coherence (and cohesion). Second, McNamara and colleagues have discovered some intriguing interactions between cohesion and world knowledge when students construct and subsequently use mental models underlying science texts (McNamara, 2001; McNamara et al., 1996; McNamara & W. Kintsch, 1996). Readers with less prior knowledge about the science domain are helped by texts with better cohesion, whereas readers with greater science knowledge can benefit from cohesion gaps. Cohesion gaps require the reader to make inferences using either world knowledge or previous textual information. When inferences are generated, the reader makes more connections between ideas in the text and knowledge. This process results in a more coherent mental representation. Hence, cohesion gaps can be beneficial for high-knowledge readers because their knowledge affords successful inference making. These results highlight the importance of pinning down linguistic and discourse features of cohesion and of better understanding the properties of world knowledge. Third, we have had an ongoing interest in directly investigating the processing of conjunctions, connectives, discourse markers, and other classes of linguistic elements that connect constituents in sentences, text, and oral discourse (Graesser, Burger, et al., 2003; Louwerse, 2002; Louwerse & Mitchell, 2003).

Our interest in cohesion and coherence also has a practical side. Readability formulas (Klare, 1974–1975) have had a major influence on the textbook industry and on the selection of texts in Grades K–12 and college. Readability formulas have widespread use even though they rely exclusively on word length and sentence length, two very simple and shallow metrics. Readability formulas ignore dozens of language and discourse components that are theoretically expected to influence comprehension difficulty. Texts are no doubt more difficult to read when they contain longer words and lengthier sentences. Longer words tend to be less frequent in the language, as we know from Zipf's (1949) law, and infrequent words take more time to access and interpret during reading (Just & Carpenter, 1980). Longer sentences tend to place more demands on working memory and are therefore more difficult (Graesser, Karnavat, et al., 2001). We do not deny that the word- and sentence-length parameters in these readability formulas

have some approximate degree of validity. However, these two-parameter multiple regression equations will not go the distance in explaining text difficulty. Even worse, they will end up being misused. Textbook writers are known to shorten sentences in basal readers for the purpose of downsizing the grade levels of their texts. The unfortunate liability of shortening sentences is that the texts end up having lower cohesion and coherence.

These theoretical advances and practical needs have led us to develop a Web-based software tool called *Coh-Metrix* (for additional information, visit [coh-metrix.memphis.edu](http://coh-metrix.memphis.edu)). Coh-Metrix analyzes texts on over 50 types of cohesion relations and over 200 measures of language, text, and readability. Unlike standard readability formulas, Coh-Metrix is sensitive to a broad profile of language and cohesion characteristics. There are modules that use lexicons, part-of-speech categorizers, syntactic parsers, templates, corpora, statistical representations of world knowledge, and other components that are widely used in computational linguistics. One important contribution of Coh-Metrix is that all of these modules are located in one central Web facility. Some of these modules incorporate or expand the modules that were developed in a Web facility that analyzes questions on surveys (called *Question Understanding Aid*, or *QUAID*; Graesser, K. Wiemer-Hastings, Kreuz, P. Wiemer-Hastings, & Marquis, 2000) and in a computer tutor that helps students learn about subject matter by holding conversations in natural language (called *AutoTutor*; Graesser, Person, Harter, & the Tutoring Research Group, 2001; Graesser, VanLehn, Rose, Jordan, & Harter, 2001).

## HOW TO USE COH-METRIX

Coh-Metrix is very easy to use. After accessing the Web site and reading the description of the tool, the facility is ready for the user to enter the text. As is illustrated in Figure 1, the user enters identifying information about the text and then enters the text in a text window. In the example, the text has the following fields of identifying information:

Title: The Needs of Plants  
 Source: Research  
 User Code: GraesserTest  
 Genre: Science  
 LSA Space: College level (This will be described later)

The text is typically entered by a cut-and-paste facility from a Text file. After this identifier information and the text are entered, the user clicks the Submit button located at the bottom center of the screen. At that point, values of a set of 13 *primary measures* are listed; these will be defined later in this article. The user is free to click on a link button to view values of over 200 other measures of language, text, and readability. Definitions of the primary measures can be obtained by clicking on a particular labeled measure. The Help links column at the right has a list of words and measures, the definitions of which can be viewed by clicking on the labels.

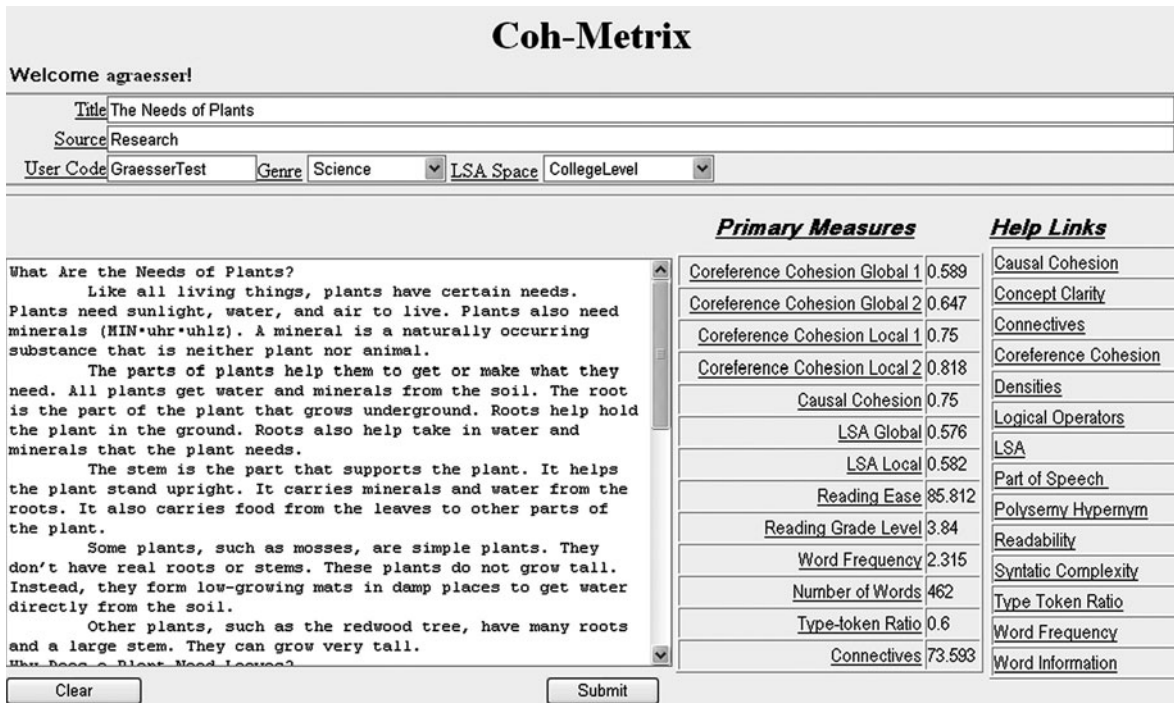


Figure 1. Screen shot of Coh-Metrix with text input and measures of language and cohesion.

The Coh-Metrix DataViewer facility is shown in Figure 2. This facility allows the user to specify what measures are desired for data analyses. As is shown in Figure 2, there are several categories of measures that correspond to the labels in Figure 1. The user selects the measures by clicking on the toggle buttons (including a *selectall* option within each category). These measures will be printed in a column in the output file that is created. The user specifies the Output File Format, which includes Excel, Text, SPSS Output and None. The user then clicks on *Go* at the bottom, and the file is automatically created.

Currently, the Web facility is for internal use only. Those who wish to use the facility can contact us through the Web site and inquire about the process of using it. The use of Coh-Metrix is not automatically available to the public because of copyright restrictions on some of the measures, because the tool is currently used for experimental purposes, and because the system could slow down substantially if a large number of researchers used the tool simultaneously.

In the remainder of this article, we will list and describe the identifier information, the measures listed under the Help links in Figure 1, and the primary measures. We will not specify the details of all of the measures because of space limitations. The measures are classified according to the labels listed in the Help links. A small subset of these measures has been selected as the primary measures because they are most pertinent to our research on coherence and cohesion. The values of these primary measures are listed directly to the right of the text after it is analyzed. The values of the measures are listed and organized according

to the categories under Help links; the user can inspect these values by clicking on a link button.

The categories of measures are listed alphabetically under the Help links. However, instead of describing these categories in alphabetical order, we will start out with measures of words, then move to measures of sentences, and then to measures that connect sentences of the text as a whole. That is, we will adopt a bottom-up, local-to-global scheme in describing these components.

### IDENTIFIER INFORMATION AND MEASURES SUPPLIED BY COH-METRIX

#### Identifier Information

**Title.** This is the title of the text to be analyzed. It is important to use a title that best reflects the content of the text so that it can be identified and remembered easily in the future. The title of the text is used together with the user code to identify the results of the Coh-Metrix analyses.

**Source.** It is sometimes useful to remember where the text came from.

**User code.** The user code identifies the researcher.

**Genre.** Genre is a major category of texts. For example, Brooks and Warren (1972) have categorized texts into four major categories: narrative, expository, description, and persuasion. However, there are many different category schemes and taxonomies of genres. The current version of Coh-Metrix has only two major genre categories plus an *other* option. *Informational* text covers material that is scientific, factual, or informational, as in the case of

**DataViewer For CohMetrix Tools**

**Step 3: Select Output Variables:** [help](#)

BasicInfo	ArticleInfo		Syntactic	Relation/Structure
<input checked="" type="checkbox"/> SelectAll <input type="checkbox"/> Title <input type="checkbox"/> Source <input type="checkbox"/> Genre <input type="checkbox"/> SubGenre <input type="checkbox"/> L Grade <input type="checkbox"/> LSASpace	<b>Levels</b> <input type="checkbox"/> SelectAll <input type="checkbox"/> Word <input type="checkbox"/> Sentence <input type="checkbox"/> Paragraph	<b>WInformation</b> <input type="checkbox"/> SelectAll <input type="checkbox"/> Familiarity <input type="checkbox"/> Concrete <input type="checkbox"/> Imagability <input type="checkbox"/> ColorMean <input type="checkbox"/> PaivioMean <input type="checkbox"/> AgeAcq	<b>ConceptClarity</b> <input type="checkbox"/> SelectAll <input type="checkbox"/> CLARABSN <input type="checkbox"/> CLARAMB <input type="checkbox"/> CLARVAG	<b>Readability</b> <input type="checkbox"/> SelectAll <input type="checkbox"/> READNP <input type="checkbox"/> READNS <input type="checkbox"/> READNW <input type="checkbox"/> READAPL <input type="checkbox"/> READASL <input type="checkbox"/> READASW <input type="checkbox"/> READFRE <input type="checkbox"/> READFKGL
<b>Primary Measures</b> <input checked="" type="checkbox"/> Select All <input type="checkbox"/> Confidence Cohesion Global 1 <input type="checkbox"/> Confidence Cohesion Global 2 <input type="checkbox"/> Confidence Cohesion Local 1 <input type="checkbox"/> Confidence Cohesion Local 2 <input type="checkbox"/> Casual Cohesion <input type="checkbox"/> LSA Global <input type="checkbox"/> LSA Local <input type="checkbox"/> Reading Ease <input type="checkbox"/> Reading Grade Level <input type="checkbox"/> Word Frequency <input type="checkbox"/> Number of Words <input type="checkbox"/> Type-token Ratio <input type="checkbox"/> Connectives	<b>Frequency</b> <input type="checkbox"/> SelectAll <input type="checkbox"/> FRQK <input type="checkbox"/> FRQT <input type="checkbox"/> FRQB <input type="checkbox"/> FRQC <input type="checkbox"/> FRQW <input type="checkbox"/> FRQS	<b>Poly&amp;Hyper</b> <input type="checkbox"/> SelectAll <input type="checkbox"/> Polysemy <input type="checkbox"/> Hyphenym	<b>Connectives</b> <input type="checkbox"/> SelectAll <input type="checkbox"/> CONP <input type="checkbox"/> CONN <input type="checkbox"/> CONALL	<b>LSA</b> <input type="checkbox"/> SelectAll <input type="checkbox"/> LSAss <input type="checkbox"/> LSAsp <input type="checkbox"/> LSAct <input type="checkbox"/> LSApp <input type="checkbox"/> LSAtf
	<b>POS</b> <input type="checkbox"/> SelectAll <input type="checkbox"/> Nouns <input type="checkbox"/> Verbs <input type="checkbox"/> Adjective <input type="checkbox"/> Adverb <input type="checkbox"/> Pronoun <input type="checkbox"/> OtherWTypes		<b>SynComplexity</b> <input type="checkbox"/> SelectAll <input type="checkbox"/> SYNNP <input type="checkbox"/> SYNHs <input type="checkbox"/> SYNHw <input type="checkbox"/> SYNLOGIC	<b>LogicalOperators</b> <input type="checkbox"/> SelectAll <input type="checkbox"/> DENANDi <input type="checkbox"/> DENIFI <input type="checkbox"/> DENORI <input type="checkbox"/> DENCONDi <input type="checkbox"/> DENNEGi <input type="checkbox"/> DENLOGi
				<b>CorefCohesion</b> <input type="checkbox"/> SelectAll <input type="checkbox"/> CREFUW <input type="checkbox"/> CREFW
				<b>CausalCohesion</b> <input type="checkbox"/> SelectAll <input type="checkbox"/> CAUSV <input type="checkbox"/> CAUSP <input type="checkbox"/> CAUSC
				<b>Density</b> <input type="checkbox"/> SelectAll <input type="checkbox"/> DENSN <input type="checkbox"/> DENSPR
				<b>TypeToken</b> <input type="checkbox"/> SelectAll <input type="checkbox"/> TYPTOKc <input type="checkbox"/> TYPTOKa <input type="checkbox"/> TYPTOKo
Display Format: <input checked="" type="radio"/> Portrait <input type="radio"/> Landscape			OutPut File Format: <input checked="" type="radio"/> None <input type="radio"/> Excel <input type="radio"/> Text <a href="#">SPSS Output</a>	
<input type="button" value="Go .."/>				

Figure 2. Coh-Metrix data viewer.

a textbook or an encyclopedia article. *Narrative* text covers material that describes fictional or actual events that occur, as in the case of a story or a news article about a recent event.

**LSA space.** Latent semantic analysis (LSA) is a statistical representation of word and text meaning (Foltz, 1996; Landauer & Dumais, 1997). A world knowledge “space” needs to be declared for some of the analyses that are performed using LSA. A general default LSA space is assigned for users who do not declare a particular LSA space. More details about LSA and its uses are discussed later in this article.

### Word Information

The words in the text have particular characteristics that have been measured in previous research in corpus linguistics and psycholinguistics. In particular, the MRC Psycholinguistics Database (Coltheart, 1981) contains 150,837 words and provides information about 26 different linguistic properties of these words. Some of the linguistic properties are absent for particular words. For example, researchers have not yet collected imagery ratings on all words, but there are imagery ratings on 9,240 of the words in the database. The word information in Coh-

Metrix includes the six MRC properties of words listed below, with values ranging from 100 to 700.

*Familiarity:* How frequently a word appears in print.

*Concreteness:* How concrete or nonabstract a word is, on the basis of human ratings.

*Imageability:* How easy it is to construct a mental image of the word in one’s mind, according to human ratings.

*Colorado meaningfulness:* These are the meaningfulness ratings from a corpus developed by Toglia and Battig (1978), multiplied by 100.

*Paivio meaningfulness:* This is the rated meaningfulness of the word, based on the norms of Paivio, Yuille, and Madigan (1968) and Gilhooly and Logie (1980), multiplied by 100 to produce a range from 100 to 700.

*Age of acquisition:* This is the score of the age-of-acquisition norms (Gilhooly & Logie, 1980) multiplied by 100 to produce a range from 100 to 700. Age of acquisition captures the fact that some words appear in children’s language earlier than others.

Coh-Metrix computes a number of measures for each of the six properties listed above. There is a mean of the words over the entire text, a mean of the paragraph word averages, and a mean of the sentence word averages. In addition, maximum (or minimum) values and their means are com-

puted for words per sentence (and per paragraph). These details about sampling of observations and numerical calculations are specified on the Web site but will not be covered in this article (for this set of measures or for those below).

### Word Frequency

Word frequency refers to metrics of how frequently particular words occur in the English language. Most of the frequency measures are based on corpora of printed texts, as opposed to spoken discourse. A corpus is a collection of printed documents or discourse excerpts. Word frequency is an important measure because frequent words are normally read more quickly and understood better than infrequent words. Researchers have investigated the impact of frequency on word processing at great depth. One finding is that word processing time tends to decrease linearly with the *logarithm* of word frequency rather than with raw word frequency (Haberlandt & Graesser, 1985; Just & Carpenter, 1980). This is the case because some words (such as *the* and *is*) have extremely high frequencies, with minimal incremental facilitation in reading time over words that are common but not nearly as frequent. The logarithmic transformation makes the distribution of word frequencies better fit a normal distribution and have a linear fit with reading times.

The measures of word frequency in Coh-Metrix are based on four corpus-based standards. First, the primary frequency counts come from *CELEX*, the database from the Dutch Centre for Lexical Information (Baayen, Piepenbrock, & Gulikers, 1995). It consists of frequencies taken from the early 1991 version of the 17.9-million-word COBUILD corpus. About 1 million of these are tokens of spoken English; the remainder are from written corpora. Written sources include newspapers and books. Spoken sources include the BBC World Service and taped telephone conversations. Second is the written frequency count from the Kučera–Francis norms (Francis & Kučera, 1982). Third is the written frequency count from the norms of Thorndike and Lorge (1944). Fourth is the frequency count of spoken English analyzed by Brown (1984). Both raw and logarithm values are computed for each of these. Also, separate measures are computed for content words (nouns, lexical verbs, adjectives, or adverbs), function words (such as prepositions or determiners), and all words. The primary measure of word frequency consists of the mean logarithm of word frequencies for content words. Instead of using only one of these four standards, Coh-Metrix provides data for all four, both to increase scope (as in the case of *CELEX*) and to provide commonly used standards (as in Francis & Kučera, 1982).

### Part of Speech

Researchers sometimes want to know how often a particular part of speech (POS) occurs in the text. For example, there is an important contrast between content words (e.g., nouns, lexical verbs, adjectives, and adverbs) and function words (e.g., prepositions, determiners, and pro-

nouns). There are over 50 POSs in Coh-Metrix. The POS categories are adopted from the Penn Treebank (Marcus et al., 1993) and the Brill (1995) POS tagger. These POS categories are segregated into content and function words. When a word can be assigned to more than one POS category, the most likely category is assigned on the basis of its syntactic context, using the Brill POS tagger. Moreover, the syntactic context can assign the most likely POS category for words it does not know.

An *incidence score* is computed for each POS category and for different sets of POS categories. An incidence score is defined as the number of occurrences of a particular category per 1,000 words. The incidence score for content words, for example, is an important measure because it is a quick index of how much substantive content there is in the text.

### Density Scores

Density scores measure the incidence, ratio, or proportion of particular word classes or constituents in the text. The higher the number, the more classes or word constituents there are in the text. An incidence score is the number of word classes or constituents per 1,000 words, as has already been mentioned. A ratio is the number of instances of Category A divided by the number of instances of Category B. A proportion is  $(A \& B)/B$ , where  $(A \& B)$  is the number of observations that are in both A and B. It is appropriate to think in terms of incidence for some metrics, in terms of ratios for others, and in terms of proportions for yet others. An incidence score is appropriate when the researcher needs a count of the number of categorical units in a text. Ratios and proportions are used when an incidence score needs to be compared with the quantities of some other unit. Ratios can exceed 1.0, whereas proportions vary from 0 to 1.

The density of pronouns is one important metric of potential comprehension difficulty. Texts are more difficult to comprehend when there is a higher density of pronouns, all else being equal. Pronoun density consists of the proportion of noun phrases (NPs, as defined by a syntactic parser, which will be described later) that are captured by pronouns (as defined by the Brill POS tagger). The incidence of nouns, pronouns, and NPs is computed first, followed by a proportion score that varies from 0 to 1. Scores approaching 1 indicate that nearly all of the NPs are captured by pronouns. As the density of pronouns increases, comprehension is expected to be more difficult. There are different ways of defining nouns, pronouns, and NPs. These variants are defined in the Coh-Metrix Web site, but not in this article.

### Logical Operators

Logical operators include variants of *or*, *and*, *not*, and *if-then*. The scope of these operators can apply to either phrases (such as NPs or verb phrases [VPs]), clauses, or sentences. If a text has a high density of logical operators, the text is analytically dense and places a high demand on

working memory. An incidence score is computed for each type of logical operator and for the entire set of logical operators.

### Connectives

Connectives are extremely important words for assessments of cohesion. Therefore, the density of connectives and different subcategories of connectives receive special focus. On one dimension, there are connectives associated with particular classes of cohesion, as identified by Halliday and Hasan (1976), Louwerse (2002), and Graesser, McNamara, and Louwerse (2003): These are (1) clarifying connectives, such as *in other words* and *that is*; (2) additive connectives, such as *also* and *moreover*; (3) temporal connectives, such as *after*, *before*, and *when*; and (4) causal connectives, such as *because*, *so*, and *consequently*. On another dimension, there is a contrast between positive and negative connectives. For example, adversative additive connectives (e.g., *however*, *in contrast*) and adversative causal connectives (e.g., *although*) are negative.

### Type:Token Ratio

Each unique word in a text is a word *type*. Each instance of a particular word is a *token*. For example, if the word *dog* appears in the text seven times, its type value is 1, whereas its token value is 7. The type:token ratio is the number of unique words divided by the number of tokens of the words. When the type:token ratio is 1, each word occurs only once in the text; comprehension should be comparatively difficult because many unique words need to be encoded and integrated with the discourse context. A low type:token ratio indicates that words are repeated many times in the text, which should generally increase the ease and speed of text processing. Type:token ratios are computed for content words but not for function words. Content words are also segregated into those that are nouns versus those that are content words other than nouns.

### Polysemy and Hypernym

A word is *ambiguous* when it has multiple senses. For example, the word *bank* has at least two senses: a place to store money and the land next to a body of water. A word is *abstract* when it has few distinctive features and few attributes that can be pictured in the mind. One way of measuring the ambiguity of a word is by the *polysemy* values in WordNet (Fellbaum, 1998; Miller, Beckwith, Fellbaum, Gross, & Miller, 1990), whereas a plausible index of abstractness is measured by the *hypernym* values in WordNet. WordNet is an online lexicon whose design is inspired by current psycholinguistic theories of human lexical representations. English nouns, verbs, adjectives, and adverbs are organized into semantic fields of underlying lexical concepts. Some sets of words are functionally synonymous because they have the same meaning or very similar meanings. There are also relations between synonym sets. Polysemy is measured as the number of senses of a word. A word with more senses will generally be more am-

biguous and slow to process, particularly for less skilled (Gernsbacher & Faust, 1991) and less knowledgeable (McNamara & McDaniel, 2004) readers. The hypernym count is defined as the number of levels in a conceptual taxonomic hierarchy that is above (i.e., superordinate to) a word. For example, *chair* (in the sense of *seat*) has seven hypernym levels: *seat* → *furniture* → *furnishings* → *instrumentality* → *artifact* → *object* → *entity*. Words having many hypernym levels tend to be more concrete, whereas those with few hypernym levels tend to be more abstract. Mean values of polysemy and hypernym are computed for the words in the text that have entries in the WordNet lexicon.

### Concept Clarity

This module is currently under development. The goal is to identify the clarity of content words by computing a composite of multiple factors that measure ambiguity, vagueness, and abstractness.

### Syntactic Complexity

Syntactic complexity involves a number of metrics that assess how difficult it is to analyze the syntactic composition of sentences. Sentences with difficult syntactic composition are structurally dense, are syntactically ambiguous, have many embedded constituents, or are ungrammatical. The syntactic analyses are based on the ApplePie parser (Sekine & Grishman, 1995) and the Brill (1995) POS tagger. A syntactic parser assigns a tree structure to every sentence in the text. The syntactic parse is the most likely tree structure, given the underlying grammar and the most likely structure when syntactic ambiguities arise. The terminal nodes in the syntactic tree structures are the words and their associated POS tags. There are several levels of intermediate nodes, such as NP, VP, prepositional phrase (PP), and embedded sentence constituents.

Syntactic complexity is measured by Coh-Metrix in three major ways. First, NP density, which consists of the mean number of modifiers per NP, is measured. A modifier is an optional element describing the property of the head of a phrase. Examples of these are adjectives, which modify heads of NPs, and adverbs, which modify heads of VPs. For example, the NP *the lovely little girl* is an NP that has three modifiers and a head. A second metric is the mean number of high-level constituents per word. Higher level constituents are sentence and embedded sentence constituents. Sentences with difficult syntactic compositions have a higher ratio of high-level constituents per word. A third measure of syntactic complexity consists of the incidence of word classes that signal logical or analytical difficulty (such as *and*, *or*, *if-then*, conditionals, and negations).

### Readability

As was discussed earlier, the traditional method of assessing text difficulty consists of readability formulas. More than 40 readability formulas have been developed over the years (Klare, 1974–1975). The most common

formulas are the *Flesch Reading Ease* score and the *Flesch–Kincaid Grade Level*. These two readability formulas are listed as primary measures in Coh-Metrix, because in our future research we will be routinely comparing other measures of Coh-Metrix against this popular standard.

The output of the Flesch Reading Ease formula is a number from 0 to 100, with a higher score indicating easier reading. The average document has a Flesch Reading Ease score between 6 and 70. Formula 1 specifies how this score is computed.

$$\text{Flesch Reading Ease} = 206.835 - 1.015 \times \text{ASL} - 84.6 \times \text{ASW}, \quad (1)$$

where ASL refers to the average sentence length, computed as the ratio of the number of words in the text divided by the number of sentences, and ASW refers to the average number of syllables per word, computed as the ratio of the number of syllables divided by the number of words. The more common Flesch–Kincaid Grade Level formula converts the Reading Ease score to a U.S. grade-school level. The higher the number, the harder it is to read the text. Formula 2 specifies how this score is computed.

$$\text{Flesch–Kincaid Grade Level} = .39 \times \text{ASL} + 11.8 \times \text{ASW} - 15.59. \quad (2)$$

In general, a text should have more than 200 words before the Flesch Reading Ease and Flesch–Kincaid Grade Level scores can successfully be applied.

### Co-Reference Cohesion

Co-reference occurs when a noun, pronoun, or NP refers to another constituent in the text. It has been extensively investigated in the fields of text linguistics and discourse processes. One form of co-reference that has been studied extensively in discourse psychology is *argument overlap* (Kintsch & Van Dijk, 1978). This occurs when a noun, pronoun, or NP in one sentence is a co-referent of a noun, pronoun, or NP in another sentence. The word *argument* is used in a special sense in this context, as a contrast between arguments and predicates in propositional representations (Kintsch & Van Dijk, 1978). In this early work, two sentences were regarded as being linked by co-reference if they shared a common argument (i.e., an overlapping noun, pronoun, or NP). However, the early theory was eventually expanded to allow referential overlap between a {noun|pronoun|NP} and a referential proposition that contains a similar *morphological stem*. For example, consider the two sentences:

*When water is heated, it boils and eventually evaporates. When the heat is reduced, it turns back into a liquid form.*

*Heat* in the second sentence refers to the proposition *water is heated*; note that *heat* and *heated* share the same morphological stem *heat*, even though one is a noun and the other is a verb.

Coh-Metrix currently considers three forms of co-reference between sentences. For any two sentences  $s_1$  and  $s_2$ , if there exists a common noun, then the two sentences have *noun overlap*. If there are two nouns (one from  $s_1$  and the other from  $s_2$ ) sharing a common stem, then the two sentences have *argument overlap*. If a noun from  $s_i$  has a stem that is shared by any category of word in  $s_j$ , then the two sentences have *stem overlap*.

**Matrices, subdiagonals, and adjacencies.** It is important to understand how referential cohesion is calculated. Specifically, we make use of matrices to depict overlap between text segments. Suppose that the text has  $n$  sentences, designated as  $s_1, s_2, \dots, s_n$ . Two sentences,  $s_i$  and  $s_j$ , may or may not be related by co-reference. The status of the co-referential relation between the two sentences is designated as  $R_{ij}$ . The value of  $R_{ij}$  is 1 if the two sentences have at least one co-referential relation; otherwise, it is 0. When all  $n$  sentences are considered, there is an  $n \times n$  symmetrical matrix, with all values on the diagonal being 1 because all sentences co-refer to themselves. The  $n \times n$  co-reference cohesion matrix is designated as **R**. We define the  $k$ th subdiagonal of the matrix by the entries  $\{R_{i,i+k} | i = 1, 2, \dots, n-k\}$ . Adjacent sentences in the text have the subdiagonal of  $k = 1$ . In the Appendix, Table A1 shows a matrix **R** for an example sentence. The subdiagonal of the matrix with  $k = 1$  is perfectly symmetrical, with (1,0,0) for the upper right adjacency and (0,1,0) for the lower left adjacency. The subdiagonal of the matrix with  $k = 2$  has the values of (0,1). The subdiagonal of the matrix with  $k = 3$  has the value of 1.

**Weighted distance between sentences.** It is quite plausible that the overlap of sentences near each other in the text will be particularly important for enhancing the coherence or readability of texts. When multiple subdiagonals are considered in assessments of co-reference, we have the option of giving larger weights to closer sentences. The distance between two sentences is weighted through the reciprocal of the distance. That is, when the distance between two sentences in the text is 1, 2, 3, ...,  $k$ , the corresponding weights are 1, 1/2, 1/3, ..., 1/ $k$ , respectively. The cohesion matrix with distance-weighted co-reference is shown in Table A2 of the Appendix.

**Co-reference cohesion local.** A simple measure of co-referential text cohesion is the proportion of adjacent sentence pairs ( $k = 1$ ) in the text that share a common noun argument. This measure is designated as *co-reference cohesion local-1* and is one of the primary measures in Coh-Metrix. Formula 3 specifies how this is computed.

$$\text{Co-reference cohesion local-1} = \frac{\sum_{i=1}^{n-1} R_{i,i+1}}{n-1}. \quad (3)$$

The metric for another primary measure, called *co-reference cohesion local-2*, is the same as in Formula 3 except that stem overlap, instead of argument overlap, determines whether adjacent sentences in the text have a co-referential cohesion relation.

Sentence pairs within some threshold distance from each other may be evaluated with respect to co-reference. For example, if there were five sentences (1, 2, 3, 4, and 5) in the text and we accepted a distance of 2 ( $k = 1$ ), we would evaluate whether the following pairs of sentences were linked by a co-reference: 1–2, 1–3, 2–3, 2–4, 3–4, 3–5, and 4–5. The sentence pairs 1–4, 1–5, and 2–5 would not be considered because the distance between them is 3 or more. Coh-Matrix computes co-reference cohesion metrics for distances of 2, 3, and higher.

**Co-reference cohesion global.** This measure includes all possible pairs of sentences when co-referential cohesion is computed. The metric is the proportion of pairs that have a co-referential connection, as is specified in Formula 4.

$$\text{Co-reference cohesion global} = \frac{\sum_{i=1}^n \sum_{j=i}^n R_{ij} |i < j|}{n \times \frac{n-1}{2}}. \quad (4)$$

The *global-1* primary measure uses argument overlap for its operational definition of co-reference, whereas the *global-2* measure uses stem overlap as its operational definition. Coh-Matrix has additional metrics for co-referential cohesion that weight sentence pairs by their distance between each other, as is illustrated in Table A2.

### Causal Cohesion

Causal cohesion reflects the extent to which sentences are related by causal cohesion relations. Causal cohesion relations are appropriate only when the text refers to events and actions that are related causally, as in the case of science texts with causal mechanisms and stories with an action plot (Graesser et al., 1994; Trabasso & van den Broek, 1985; van den Broek, Virtue, Everson, Tzeng, & Sung, 2002; Zwaan & Radvansky, 1998). Causality is not relevant, for example, in texts that describe static scenes and texts that convey abstract logical arguments.

Coh-Matrix must first estimate how much of the text refers to events and actions that may be part of causal content. This is accomplished by counting the number of main verbs that are causal, on the basis of WordNet (Fellbaum, 1998; Miller et al., 1990). The WordNet lexicon contains a large number of semantic characteristics of words, including verb causality. A verb is considered causal if the action or event it represents causes something to happen. For example, the action of the verb *kill* causes some animate being to die. The higher the incidence of causal verbs in a text, the more the text is assumed to convey causal content.

Having causal verbs in a text does not ensure that the reader can connect these events and actions with causal relations. According to Coh-Matrix, causal cohesion relations are signaled by causal particles. Some causal particles are conjunctions, transitional adverbs, and other forms of connectives, such as *since*, *so that*, *because*, *the cause of*, and *as a consequence*. These particles are used to indicate some causal relationship between clauses that refer to events and actions. Other causal particles consist of a

small number of verbs that explicitly assert that there is a causal relationship between constituents, without specifying the nature of the causal content (e.g., *cause*, *enable*, and *make*). The total list of causal particles comes either from this short list of verbs or from the causal conjunctions, transitional adverbs, and causal connectives.

The current metric of *causal cohesion*, which is a primary measure, is simply a ratio of causal particles (P) to causal verbs (V). The denominator is incremented by the value of 1 to handle the rare case in which there are zero causal verbs in a text. It should be noted that this causal cohesion metric is unstable when the text is very short or when there are very few causal verbs in the text. As in the case of the readability formulas, these measures are more stable when there is a sufficient volume of content.

### LSA Information

World knowledge has recently been statistically represented in the form of higher dimensional spaces that accommodate the constraints of a large corpus of texts. Notable examples of these approaches are the Hyperspace Analog to Language (Burgess, Livesay, & Lund, 1998) and LSA (Foltz, 1996; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). LSA was adopted in Coh-Matrix as a measure of semantic cohesion and coherence.

LSA uses a statistical method called *singular value decomposition* (SVD) to reduce a large word  $\times$  document co-occurrence matrix to approximately 100–500 functional dimensions. The word  $\times$  document co-occurrence matrix is simply a record of the number of times word  $W_i$  occurs in document  $D_j$ . A document may be defined as a sentence, paragraph, or section of an article. Each word, sentence, or text ends up being a weighted vector on the  $K$  dimensions. The *match* (i.e., similarity in meaning, conceptual relatedness) between two unordered bags of words (single words, sentences, or texts) is computed as a geometric cosine between the two vectors, with values ranging from  $-1$  to  $1$ . From the present standpoint, LSA was used to compute the similarity between two sentences or that between the entire text and a sentence.

Text cohesion (and sometimes coherence) is assumed to increase as a function of higher cosine scores between text constituents. The *LSA global measure*, one of the primary measures of Coh-Matrix, is simply the mean cosine value of all possible pairs of sentences. The *LSA local measure* is the mean cosine value between adjacent pairs of sentences in the text. These computations are the same as Formulas 3 and 4, except that the LSA cosines are the values for sentence pairs instead of values for co-referential cohesion.

Coh-Matrix has several other methods of computing LSA cohesion. Some of these are listed below.

*Sentence to paragraph:* This measures how similar each sentence is to its paragraph.

*Sentence to text:* This measures how similar a sentence is to the text.

*Paragraph to paragraph:* This measures how similar a paragraph is to the other paragraphs in the text.



*Paragraph to text:* This measures how similar a paragraph is to the entire text.

*Family resemblance score:* This score is used to measure the similarity between sentences and core sentences in the text. A core sentence is defined by a high family resemblance score, a metric that computes how similar a sentence is to all other sentences in the text.

These and other metrics of LSA cohesion can be found on the Web site.

## CLOSING COMMENTS

Coh-Metrix 1.0 is our first version of a tool that analyzes texts on multiple levels of language, discourse, cohesion, and world knowledge. It makes use of existing modules from computational linguistics and other fields that automatically extract information from text. One important contribution from Coh-Metrix 1.0 is that it will allow researchers to collect a great deal of information about bodies of text with little effort. Moreover, in our attempts to provide measures of text cohesion and text difficulty, we are providing the research community with measures heretofore unavailable. Indeed, we are confident that having these measures so readily available will revolutionize text and discourse research. It will not only allow investigators to improve their empirical research (by having more information during the development of their experimental corpora), but it will also allow researchers to investigate existing corpora, unearthing new and exciting understandings about language processing.

Our next steps in this research activity are twofold. First, we will evaluate the validity of these measures on text corpora and data from psychology experiments. The measures will be modified and tuned to fit the constraints of the empirical findings. Second, we will explore more sophisticated algorithms and language discourse patterns that capture vestiges of semantic interpretation, mental models, discourse coherence, rhetorical structures, and pragmatics. We anticipate that this is merely the beginning of a large-scale enterprise that could become an alternative to common readability scores and help publishing houses, educators, and students in the selection of appropriate textbooks.

## REFERENCES

- ALLEN, J. (1995). *Natural language understanding*. Redwood City, CA: Benjamin/Cummings.
- BAAYEN, R. H., PIEPENBROCK, R., & GULIKERS, L. (1995). *The CELEX lexical database* (CD-ROM). Philadelphia: University of Pennsylvania, Linguistic Data Consortium.
- BELEW, R. K. (2002). Finding out about: A cognitive perspective on search engine technology and the WWW. *Information Retrieval*, *5*, 269-278.
- BIBER, D., CONRAD, S., & REPPEN, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: 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*, 543-566.
- BROOKS, C., & WARREN, R. P. (1972). *Modern rhetoric*. New York: Harcourt Brace Jovanovich.
- BROWN, G. D. A. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. *Behavior Research Methods, Instruments, & Computers*, *16*, 502-532.
- BURGESS, C., LIVESAY, K., & LUND, K. (1998). Explorations in context space: Words, sentences, and discourse. *Discourse Processes*, *25*, 211-257.
- COLTHEART, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33A*, 497-505.
- DARPA (1995). *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. San Francisco: Morgan Kaufman.
- DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K., & HARSHMAN, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391-407.
- FELLBAUM, C. (ED.) (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.
- FOLTZ, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, & Computers*, *28*, 197-202.
- FRANCIS, W. N., & KUČERA, H. (1982). *Frequency analysis of English usage*. Boston: Houghton-Mifflin.
- GERNSBACHER, M. A., & FAUST, M. (1991). The mechanism of suppression: A component of general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *17*, 245-262.
- GILHOOLY, K. J., & LOGIE, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, *12*, 395-427.
- GRAESSER, A. C., GERNSBACHER, M. A., & GOLDMAN, S. R. (2003). Introduction to the *Handbook of discourse processes*. In A. C. Graesser, M. A. Gernsbacher, and S. R. Goldman (Eds.), *Handbook of discourse processes* (pp. 1-24). Mahwah, NJ: Erlbaum.
- GRAESSER, A. C., BURGER, J., CARROL, J., CORBETT, A., FERRO, L., GORDON, D., GREIFF, W., HARABAGIU, S., HOWELL, K., KELLY, H., LITMAN, D., LOUWERSE, M., MOORE, A., PELL, A., PRANGE, J., VOORHEES, E., & WARD, W. (2003). *Question generation and answering systems: R&D for technology-enabled learning systems. Research roadmap for the Federation of American Sciences*. Unpublished manuscript.
- GRAESSER, A. C., KARNAVAT, A. B., DANIEL, F. K., COOPER, E., WHITTEN, S. N., & LOUWERSE, M. (2001). A computer tool to improve questionnaire design. In *Statistical Policy Working Paper 33, Federal Committee on Statistical Methodology* (pp. 36-48). Washington, DC: Bureau of Labor Statistics.
- GRAESSER, A. C., MCNAMARA, D. S., & LOUWERSE, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text? In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 82-98). New York: Guilford.
- GRAESSER, A. C., PERSON, N., HARTER, D., & THE TUTORING RESEARCH GROUP (2001). Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education*, *12*, 257-279.
- GRAESSER, A. C., SINGER, M., & TRABASSO, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*, 371-395.
- GRAESSER, A. C., VANLEHN, K., ROSE, C. P., JORDAN, P. W., & HARTER, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine*, *22*(4), 39-52.
- GRAESSER, A. C., WIEMER-HASTINGS, K., KREUZ, R., WIEMER-HASTINGS, P., & MARQUIS, K. (2000). QUAID: A questionnaire evaluation aid for survey methodologists. *Behavior Research Methods, Instruments, & Computers*, *32*, 254-262.
- HABERLANDT, K., & 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., & HASAN, R. (1976). *Cohesion in English*. London: Longman.
- JURAFSKY, D., & MARTIN, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice-Hall.
- JUST, M. A. & CARPENTER, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*, 329-354.
- KINTSCH, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

- KINTSCH, W., & VAN DIJK, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, **85**, 363-394.
- KLARE, G. R. (1974-1975). Assessing readability. *Reading Research Quarterly*, **10**, 62-102.
- LANDAUER, T. K., & DUMAIS, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, **104**, 211-240.
- LANDAUER, T. K., FOLTZ, P. W., & LAHAM, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, **25**, 259-284.
- LEHNERT, W. G. (1997). Information extraction: What have we learned? *Discourse Processes*, **23**, 441-470.
- LEHNERT, W. G., & RINGLE, M. H. (Eds.) (1982). *Strategies for natural language processing*. Hillsdale, NJ: Erlbaum.
- LOUWERSE, M. M. (2002). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics*, **12**, 291-315.
- LOUWERSE, M. M., & GRAESSER, A. C. (in press). Coherence in discourse. In P. Strazny (Ed.), *Encyclopedia of linguistics*. Chicago: Fitzroy Dearborn.
- LOUWERSE, M. M., & MITCHELL, H. H. (2003). Toward a taxonomy of a set of discourse markers in dialog: A theoretical and computational linguistic account. *Discourse Processes*, **35**, 199-239.
- MARCUS, M., SANTORINI, B., & MARCINKIEWICZ, M. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, **19**, 313-330.
- 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., KINTSCH, E., SONGER, N. B., & KINTSCH, W. (1996). Are good texts always better? Text coherence, background knowledge, and levels of understanding in learning from text. *Cognition & Instruction*, **14**, 1-43.
- McNAMARA, D. S., & KINTSCH, W. (1996). Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes*, **22**, 247-287.
- McNAMARA, D. S., & MCDANIEL, M. (2004). Suppressing irrelevant information: Knowledge activation or inhibition? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **30**, 465-482.
- MILLER, G. A., BECKWITH, R., FELLBAUM, C., GROSS, D., & MILLER, K. (1990). *Five papers on WordNet* (Tech. Rep. No. 43). Princeton, NJ: Princeton University, Cognitive Science Laboratory.
- MOORE, J. D., & WIEMER-HASTINGS, P. (2003). Discourse in computational linguistics and artificial intelligence. In A. C. Graesser, M. A. Gernsbacher, and S. R. Goldman (Eds.), *Handbook of discourse processes* (pp. 439-486). Mahwah, NJ: Erlbaum.
- PAIVIO, A., YUILLE, J. C., & MADIGAN, S. A. (1968). Concreteness, imagery and meaningfulness values for 925 words. *Journal of Experimental Psychology Monograph Supplements*, **76**(3, Part 2).
- PENNEBAKER, J. W., & FRANCIS, M. E. (1999). *Linguistic inquiry and word count (LIWC)*. Mahwah, NJ: Erlbaum.
- ROBERTSON, S. (2001). *Evaluation in information retrieval: Lectures on information retrieval*. New York: Springer-Verlag.
- SCHANK, R., & RIESBECK, C. K. (Eds.) (1981). *Inside computer understanding*. Hillsdale, NJ: Erlbaum.
- SEKINE, S., & GRISHMAN, R. (1995). A corpus-based probabilistic grammar with only two nonterminals. In *Fourth International Workshop on Parsing Technologies* (pp. 260-270). Prague: Karlovy Vary.
- THORNDIKE, E. L., & LORGE, I. (1944). *The teacher's word book of 30,000 words*. New York: Teachers College.
- TOGLIA, M. P., & BATTIG, W. R. (1978). *Handbook of semantic word norms*. Hillsdale, NJ: Erlbaum.
- TRABASSO, T., & VAN DEN BROEK, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory & Language*, **24**, 612-630.
- 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.
- VOORHEES, E. (2001). The TREC Question Answering Track. *Natural Language Engineering*, **7**, 361-378.
- ZIPF, G. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley.
- ZWAAN, R. A., & RADVANSKY, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, **123**, 162-185.

## APPENDIX

### Co-reference Cohesion Matrices for an Example Text

Example Text, With Sentences Numbered

- (1) When we heat water, it will boil and eventually evaporate.  
 (2) When the heat is reduced, evaporation turns back into its liquid water form. (3) This process can be summarized as changing matter. (4) Heating an object can basically change its matter.

**Table A1**  
 $n \times n$  Symmetrical Matrix

	S1	S2	S3	S4
S1	1	1	0	1
S2	1	1	0	1
S3	0	0	1	0
S4	1	1	0	1

**Table A2**  
 Matrix With Values Weighted by the Distance  
 Between Sentences

	S1	S2	S3	S4
S1	1	1	0	0.33
S2	1	1	0	0.50
S3	0	0	1	0
S4	.33	.50	0	1

Note—When the distances between two sentences in the text are 1, 2, 3, ...,  $k$ , the corresponding weights are 1, 1/2, 1/3, ..., 1/ $k$ , respectively.