# Introduction

This book describes Coh-Metrix, a computational tool that provides a wide range of language and discourse measures. It is a linguistic workbench that researchers, teachers, and students of many different disciplines can use to obtain information about their texts on numerous levels of language. This book consists of two parts. The first section focuses on the theoretical motivations and perspectives that led to the development of Coh-Metrix. Part I describes its technological foundations, the measures it provides, and empirical work that has been conducted using Coh-Metrix. We see Part I as being invaluable to researchers who wish to situate their Coh-Metrix work within the theoretical and empirical fields of discourse processing, psycholinguistics, text design, and related fields.

Part II shifts to the practical and pedagogical arena, describing how to use Coh-Metrix and how to analyze, interpret, and describe Coh-Metrix results. This section is written for computational novices and students who wish to not only use Coh-Metrix (or similar computational tools), but also describe the resulting studies and their outcomes.

Coh-Metrix was developed, refined, and tested between 2002 and 2011 at the University of Memphis. The initial funding for the Coh-Metrix project was awarded in 2002 (R305G020018) from the Office of Educational Research and Improvement (OERI), which became the Institute for Education Sciences (IES) the following year. Our initial discussions that led to the Coh-Metrix grant proposal revolved around establishing common ground between an interdisciplinary collection of researchers with very different backgrounds. One fundamental issue that called for a common understanding was whether we all believed that *cohesion* was observable in text, or alternatively whether it could only be measured with respect to the reader. We all agreed, fortunately, that cohesion could be measured in a text. We finally agreed to use the term *cohesion* when referring to observable aspects of the text, and *coherence* when

referring to the consequences of cohesion in the mind of the reader (see Chapter 5). This definition of terms was crucial to our moving forward. Since that time, we have been working on developing, refining, and playing with Coh-Metrix.

Coh-Metrix has quickly and effectively moved well beyond its original goals of developing measures of cohesion to better *match text to readers*. It is arguably the broadest and most sophisticated automated textual assessment tool currently available on the Web. Coh-Metrix empowers anyone with an interest in text to pursue a wide array of previously unanswerable research questions. Coh-Metrix automatically provides numerous measures of evaluation at the levels of the text, the paragraph, the sentence, and the word. Coh-Metrix uses lexicons, part-of-speech classifiers, syntactic parsers, semantic analyzers, Latent Semantic Analysis (a statistical representation of world knowledge based on corpus analyses), and several other components that are widely used in computational linguistics. For example, the MRC (Medical Research Council) Psycholinguistic Database (Coltheart, 1981) is used for psycholinguistic information about words. WordNet has linguistic and semantic features of words, as well as semantic relations between words (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990). Latent Semantic Analysis computes the semantic similarities between words, sentences, and paragraphs (Landauer & Dumais, 1997; Landauer, McNamara, Dennis, & Kintsch, 2007). And, syntax is analyzed by syntactic parsers (e.g., Charniak, 2000).

This book describes a plethora of studies that have been conducted since Coh-Metrix was first launched in 2003. Our research labs have collectively published well over a hundred studies that have used Coh-Metrix to analyze texts in print and oral discourse. Among those publications are studies that have validated the use of Coh-Metrix to assess the cohesion of text (e.g., McNamara, Louwerse, McCarthy, & Graesser, 2011). Collectively, these studies have used Coh-Metrix to distinguish a wide range of texts. For example, Louwerse, McCarthy, McNamara, and Graesser (2004) identified significant differences between spoken and written samples of English. Graesser, Jeon, Yang, and Cai (2007) identified differences between physics context that occurred in textbooks, texts prepared by researchers, and conversational discourse in tutorial dialogue. Lightman, McCarthy, Dufty, and McNamara (2007a) distinguished the *beginnings*, *middles*, and *ends* of chapters in a corpus of history and science textbooks for high school. Crossley, Louwerse, McCarthy, and McNamara's (2007) investigations of second language learner texts revealed a wide variety of structural and lexical differences between texts that were *adopted* (or authentic) versus *adapted* (or simplified) for second language learning purposes. These few studies only begin to represent the

extensive body of research that has evolved since Coh-Metrix was launched to discourse processing researchers and scholars in other fields.

The Coh-Metrix facility and the associated theoretical framework would never have been accomplished without an interdisciplinary team of researchers. The relevant major fields have included psychology, computer science, linguistics, and education but it is the more specialized hybrid fields that have provided the more useful, targeted contributions: discourse processing, psycholinguistics, reading, computational linguistics, corpus linguistics, cognitive science, artificial intelligence, information retrieval, and composition.

Some of us brand ourselves as *computational discourse scientists*. We use the term *discourse* as a general umbrella term for analyses of language, texts, communication, and social interaction through various communication channels. Our work is *computational* in two ways. First, we precisely specify the algorithms or symbolic procedures that identify text categories, units, or patterns at the various levels of a multilevel theoretical framework. Second, we attempt to program the computer to implement these algorithms and procedures. Many computer implementations are successful, but there are no guarantees. Coh-Metrix includes only the successful automated algorithms and procedures. And finally, we are *scientists* because we embrace scientific methods in all stages of our research. That is, we sample texts in a systematic manner when we empirically test well-formulated claims about text characteristics. We perform statistical analyses that assess the generality of our claims regarding targeted text categories. We collect data from human participants to test claims and predictions about the impact of text characteristics on comprehension and other psychological processes.

We are hopeful that Coh-Metrix will be useful to scholars in both the sciences and humanities and to all sectors of the public. Coh-Metrix opens the door to a new paradigm of research that coordinates studies of language, discourse, corpus analysis, computational linguistics, education, and cognitive science (Graesser, McNamara, & Rus, 2007). We hope that this book will be of use to a wide range of readers, including researchers, educators, writers, publishers, and students. Our vision is broad. There is the student in a literature course who analyzes differences between various works by Shakespeare, and the student in an educational psychology course who compares textbooks written for elementary versus middle school courses. There are the students who want to know about the nature of their own writing and whether it improves over time. There is the book publisher who wants to know whether a text in biology is written coherently compared with other books on the market. There are the school superintendents who want to evaluate all of the books being used in their school system. There is the

attorney who wants to know the difficulty of the Miranda Rights when defending a client who has a modest understanding of the English language.

The uses and applications of Coh-Metrix are endless. Enjoy!

## RECOMMENDED SUPPLEMENTARY READINGS

An introduction to Coh-Metrix is provided in a number of publications (Graesser & McNamara, 2011; Graesser, McNamara, & Kulikowich, 2011; Graesser, McNamara, Louwerse, & Cai, 2004; McNamara & Graesser, 2012; McNamara, Louwerse, & Graesser, 2010). The Coh-Metrix research group has published well over 50 articles in journals, books, and conference proceedings. Many of these articles can be accessed on the Coh-Metrix website (www.cohmetrix.com) and many can be accessed from Danielle McNamara's lab website (soletlab.com). Most importantly, the cohmetrix.com site also provides access to Coh-Metrix 3.0, the focus of this book.

A book edited by McCarthy and Boonthum-Denecke (2012) provides many examples of research efforts in computational discourse science. This interdisciplinary field is closely aligned with a number of other hybrid fields that investigate language and discourse, including discourse processing (Graesser, Goldman, & Gernsbacher, 2003; Sanford & Emmott, 2012), psycholinguistics (Spivey, Joanisse, & McRae, 2010), reading (Kamil, Pearson, Moje, & Afflerbach, 2011; McNamara, 2007), computational linguistics (Jurafsky & Martin, 2008), corpus linguistics (Biber, Conrad, & Reppen, 1998), and cognitive science (Kintsch, 1998; Landauer, McNamara, Dennis, & Kintsch, 2007).

We have adopted a multilevel theoretical framework for analyzing text difficulty with Coh-Metrix (Graesser & McNamara, 2011). An alternative perspective assigns a text to a single dimension of text difficulty, as in the case of Lexiles (Stenner, 2006). Another alternative positions a text in a multiple dimensional space, as in the case of analyses by Biber (1988).

Multilevel theoretical frameworks have been proposed that include the levels of words, syntax, textbase, situation model, and genre/rhetorical structure (Graesser & McNamara, 2011; Graesser, Millis, & Zwaan, 1997; Kintsch, 1998; Pickering & Garrod, 2004). More detailed theoretical and empirical discussions of these levels are provided for words (Pennebaker et al., 2007; Perfetti, 2007), syntax (Charniak, 2000; Rus et al., 2006), textbase (McNamara et al., 2010; van Dijk & Kintsch, 1983), situation model (Graesser, Singer, & Trabasso, 1994; Zwaan & Radvansky, 1998), and genre/rhetorical structure (Biber, 1988). The book edited by McCarthy and Boonthum-Denecke (2012) reports computational measures and psychological evidence for these five levels and other aspects of language, discourse, and text.

PART I

# COH-METRIX: THEORETICAL, TECHNOLOGICAL, AND EMPIRICAL FOUNDATIONS

1

# What Is Text and Why Analyze It?

Some texts are easy to read. Others are difficult. That is perfectly obvious. The challenge lies in devising an objective means to measure texts on *how* difficult they are to read. That is one of the puzzles that motivated our development of Coh-Metrix and ultimately the writing of this book. How do we scale texts on comprehension difficulty? Or on the flip side: *easability*?

It is often quite clear when texts are difficult or easy. Consider the two texts below and cast your vote on which is difficult and which is easy.

**Lady Chatterley's Lover**
He spread the blankets, putting one at the side for a coverlet. She took off her hat, and shook her hair. He sat down, taking off his shoes and gaiters, and undoing his cord breeches. "Lie down then!" he said, when he stood in his shirt. She obeyed in silence, and he lay beside her, and pulled the blanket over them both.

**A Mortgage**
The assignment, sale, or transfer of the servicing of the mortgage loan does not affect any term or condition of the mortgage instrument, other than terms directly related to the servicing of your loan. Except in limited circumstances, the law requires your present servicer send you this notice within 15 days before this effective date or at closing.

We do not need to conduct a survey to discover how most English speakers will vote. The *Chatterley* text by D. H. Lawrence is clearly easier than the mortgage text. The question is *why*?

Some obvious hypotheses fail to discriminate these two excerpts on comprehension difficulty. Both passages have pronouns that require inferences to understand what they refer to. And, both texts have low-frequency words in the English language. Readers will be challenged by *coverlet*, *gaiters*, and *cord breeches*, just as they will be challenged by words such as *mortgage*, *instrument*, and *present servicer*. The core topics underlying these two texts are both

important. Sex and romance are on par with money and domestic security, although it could be argued that sex and romance are considerably more interesting. Both texts require a sociocultural context for a complete understanding, be it knowledge of romance or of finance. Moreover, a deep understanding of the D. H. Lawrence story requires knowledge of the status of women in the early 20th century (i.e., not great), when it was written. The differences in comprehension difficulty for these two texts are indeed much more complex and subtle than is readily apparent from the text alone.

This book will unveil the many ways that texts vary in comprehension difficulty. What we sometimes call *comprehension easability* is aligned with reading ease or readability, the other end of the continuum being text difficulty or text complexity. Our theoretical approach is to analyze texts on many levels of language, meaning, and discourse (Graesser & McNamara, 2011). A computer program called Coh-Metrix (and Coh-Metrix-TEA) performs these analyses automatically for many of the levels that researchers have identified over the years (Graesser, McNamara, & Kulikowich, 2011; Graesser, McNamara, Louwerse, & Cai, 2004; McNamara & Graesser, 2012; McNamara, Graesser, & Louwerse, 2012; McNamara, Louwerse, McCarthy, & Graesser, 2010). The Coh-Metrix output on these many levels provides the foundation for scaling texts on difficulty (versus easability).

## WHAT TEXT?

Our emphasis in this book is on printed texts, although the texts may derive from virtually any source and be composed for any English language community. For example, they may be newspaper articles, entries in encyclopedias, science texts in schools, legal documents, advertisements, short stories, or theatrical scripts – the list goes on. The Coh-Metrix program holds up quite well for most of the texts that we have analyzed. The majority of our analyses have been on naturalistic texts, but we have also analyzed well-controlled texts that discourse researchers have prepared or manipulated for psychology experiments (McNamara et al., 2010). Our goal is to accommodate virtually any text in the English language that people write with the intention of communicating messages to readers.

Our theoretical framework and the Coh-Metrix program can also be used to analyze transcripts of naturalistic oral discourse. We have analyzed conversations in tutoring sessions, chat rooms, e-mail exchanges, and various forms of informal conversation. Transcribed texts of conversations are replete with speech disfluencies (*um*, *ah*, *er*), ungrammatical utterances, interruptions, overlapping speech, slang, and semantically vague expressions (Clark,

1996). These deviations from well-formed, edited, neat and tidy text have a major impact on some of the Coh-Metrix measures, but many of the measures are minimally disturbed. It is also possible to analyze students' written responses, explanations, and essays that are similarly replete with untidy language and discourse (Crossley & McNamara, 2011; Louwerse, McCarthy, McNamara, & Graesser, 2004; McNamara, Raine, et al., 2012; Renner, McCarthy, Boonthum-Denecke, & McNamara, 2012).

While Coh-Metrix analyses of more naturalistic discourse (e.g., dialogues) have been highly successful, it remains important to acknowledge that some classes of printed texts will stress the boundaries of Coh-Metrix. Current versions of Coh-Metrix are not well equipped to handle mathematical expressions, pictures, diagrams, and other forms of nonverbal media. Coh-Metrix can be applied to poetry (Lightman, McCarthy, Dufty, & McNamara, 2007b), but measures at some levels (such as syntax) will be compromised and Coh-Metrix will not do justice to metaphorical expressions (Graesser, Dowell, & Moldovan, 2011). Likewise, many aspects of the quality of writing, such as rhetorical and pragmatic aspects of language, are not fully captured by Coh-Metrix alone (McNamara, Crossley, & Roscoe, 2013). These challenges are on deck for future research endeavors.

## WHY SHOULD WE SCALE TEXTS ON DIFFICULTY?

Skeptics ask why we bother scaling texts on difficulty. What problems will this solve? Text is qualitative verbal material, so what's the point in assigning numbers to the morass of qualitative symbolic codes? Wouldn't it be better to have a group of experts describe particular texts on qualitative attributes and to scrap the mission of assigning numbers to texts?

Our response to the skeptics is that the assignment of Coh-Metrix values to texts is quite important and eminently humane. Consider the following applications of Coh-Metrix and the practical implications for quality of life.

***Assigning texts to students in school***. Ideally, the texts assigned to students should be within an optimal zone of comprehension difficulty. The optimal zone is a matter of debate and is likely to depend on the characteristics of the student (Graesser et al., 2011) as well as the teacher's pedagogical goals. Some students are best served by texts at an intermediate level of difficulty for them: Not too easy, not too difficult, but just right. If the texts are too easy, the students are not challenged and they may become bored. If the texts are too difficult, the students are overwhelmed, become discouraged, and tune out. Some students are eager to read texts considerably above their comfort level and others need to build self-confidence in reading by receiving texts that are

easy for them to read. The assignment of texts can also be tailored to particular deficits that a student has at particular levels of language or discourse. A student who is reading quite well but has trouble understanding the global meaning of stories should be receiving different texts than students who are having trouble with syntax or those who experience challenges with vocabulary. Many claim that text assignment should be adapted to the student's profile of reading skills and proficiencies, and moreover, that student motivation and learning improve when this happens (Connor, Morrison, Fishman, Schatschneider, & Underwood, 2007).

***Quality of public documents***. The comprehension difficulty of many public documents is too high for a large percentage of the population. The earlier mortgage text illustrates the problem. Legal documents, medical documents, and employment agreements are also excellent examples of challenging texts that are difficult to understand for most of the public. Similarly, questionnaires and surveys administered to the public, such as tax forms and census surveys, have a high percentage of questions that pose comprehension difficulties to a significant portion of the public (Conrad & Schober, 2007; Graesser, Cai, Louwerse, & Daniel, 2006). The reliability and validity of data collected from these surveys is compromised when the questions have difficult words, ambiguous meaning, complex syntax, or content that excessively burden cognitive resources. Individuals and society suffer the consequences.

***Drug prescriptions and medical procedures***. It is obviously important to take the proper dosage of drugs, to be mindful of side effects, and to understand medical procedures. Failure to do so may be a matter of life or death. Unfortunately, the complexity of medical information is too high for most of the public to comprehend, particularly when there is a large amount of jargon, incoherent descriptions of procedures, and complex models of health and biological mechanisms (Day, 2006). Interestingly, the advertisements tend to be much easier to read than the warnings. Consider the following warning on a nonprescription drug:

> Do not use if you are now taking a prescription monoamine oxidase inhibitor (MAOI) (certain drugs for depression, psychiatric, or emotional conditions, or Parkinson's disease), or for 2 weeks after stopping the MAOI drug.

These examples illustrate the value of analyzing texts on difficulty and including quantitative scales in this process. We would argue that public documents and medical instructions need to be within a reasonable zone of text difficulty. The education of students hinges on the assignment of texts, tests, and other materials that are within the students' proficiency zones at