Chapter 11
Coh-Metrix:
An Automated Tool for Theoretical and Applied Natural Language Processing

Danielle S. McNamara
Arizona State University, USA

Arthur C. Graesser
The University of Memphis, USA

ABSTRACT

Coh-Metrix provides indices for the characteristics of texts on multiple levels of analysis, including word characteristics, sentence characteristics, and the discourse relationships between ideas in text. Coh-Metrix was developed to provide a wide range of indices within one tool. This chapter describes Coh-Metrix and studies that have been conducted validating the Coh-Metrix indices. Coh-Metrix can be used to better understand differences between texts and to explore the extent to which linguistic and discourse features successfully distinguish between text types. Coh-Metrix can also be used to develop and improve natural language processing approaches. We also describe the Coh-Metrix Text Easability Component Scores, which provide a picture of text ease (and hence potential challenges). The Text Easability components provided by Coh-Metrix go beyond traditional readability measures by providing metrics of text characteristics on multiple levels of language and discourse.

INTRODUCTION

Coh-Metrix is an automated tool that provides linguistic indices for text and discourse (Graesser, McNamara, Louwerse, & Cai, 2004). Coh-Metrix was developed to meet three practical needs. First, at the time that the Coh-Metrix research project began in 2002, there were no readily available tools that provided an array of indices on words or texts. For example, if a researcher needed the word frequency values for words or sentences in a text, one tool might be available (though challenging to find). But another tool would have to be used for measures of word concreteness,

DOI: 10.4018/978-1-60960-741-8.ch011

familiarity, imagery, syntactic complexity, and so on. In other words, there existed no linguistic workbench capable of providing a wide array of measures on language and discourse. Second, traditional measures of text difficulty, referred to as *readability*, were outdated given the maturation of our understanding of text and discourse (Clark, 1996; Graesser, Gernsbacher, & Goldman, 2003; Kintsch, 1998). There was a growing recognition of a number of factors contributing to text difficulty that are not considered within traditional measures of text readability. Third, there existed no automated measures of text cohesion. Whereas recognition of the importance of cohesion had flourished in the 80s and 90s (Gernsbacher, 1990; Goldman, Graesser, & Van den Broek, 1999; Louwerse, 2001; McNamara & Kintsch, 1996; Sanders & Noordman, 2000), there were no objective, implemented measures of cohesion available. Thus, with the overarching goal of providing more informative measures of text complexity, particularly considering text cohesion, we embarked in 2002 on a mission to develop Coh-Metrix (initially funded by an Institute of Education Sciences). This chapter describes some motivating factors that led to Coh-Metrix, an overview of the measures provided by Coh-Metrix, some of the many NLP studies that have been completed over the last eight years, and the ultimate outcome of our endeavors: Coh-Metrix *Text Complexity* Components.

READABILITY VS. COHESION: WHY COH-METRIX WAS DEVELOPED

Readability measures are the most common approach to estimating the difficulty of a text and hundreds have been developed over the past century. Readability formulas became popular in the 1950s and by the 1980s over 200 readability algorithms had been developed, with over a 1000 supporting studies (Chall & Dale, 1995; Dubay, 2004). The most well known readability measures include Flesch-Kincaid Grade Level (Klare, 1974-5), Degrees of Reading Power (DRP; Koslin, Zeno, & Koslin, 1987), and Lexile scores (Stenner, 2006). Measures of readability are highly correlated because they are based on the same constructs: the difficulty of the individual words and the complexity of the separate sentences in the text. However, the way in which these constructs are operationalized and the underlying statistical assumptions vary somewhat across readability measures. The Flesch-Kincaid Grade Level metric is based on the length of words (i.e., number of letters or syllables) and length of sentences (i.e., number of words). DRP and Lexile scores relate these characteristics of the texts to readers' performance on cloze tasks. In a cloze task, the reader reads a text with some words left blank; the reader is asked to fill in the words by generating them or by selecting a word from a set of options (usually the latter). Using this methodology, the appropriateness of a text for a particular reader can be calculated based on the characteristics of the texts and the reader's performance on cloze tasks. A particular text would be predicated to be at the reader's level of proficiency if the reader can perform the cloze task at a threshold of performance (75%) for texts with similar characteristics (i.e., with the same word and sentence level difficulties). A text can be defined as too easy if performance is higher than 75% and too difficult to the extent it is lower than 75%.

Readability measures based on word and sentence characteristics (i.e., usually length) have validity as indices of text difficulty. When words contain more letters or syllables, they tend to be less frequently used in a language. Readers need to have greater exposure to language and text in order to encounter less frequent words and to know what they mean. Clearly, a requisite to comprehension is knowing the meaning of the words in a text. In turn, to the extent that a sentence contains more words, there is a greater likelihood that the sentence is more complex syntactically. Readers who have had less exposure to language