

Language and Discourse Analysis with Coh-Metrix: Applications from Educational Material to Learning Environments at Scale

Nia M. M. Dowell

University of Memphis and Institute for Intelligent Systems, USA
ndowell@memphis.edu

Arthur C. Graesser

University of Memphis and Institute for Intelligent Systems, USA

Zhiqiang Cai

University of Memphis and Institute for Intelligent Systems, USA

ABSTRACT: The goal of this article is to preserve and distribute the information presented at the LASI (2014) workshop on Coh-Metrix, a theoretically grounded, computational linguistics facility that analyzes texts on multiple levels of language and discourse. The workshop focused on the utility of Coh-Metrix in discourse theory and educational practice. We discuss some of the motivating factors that led to the development of Coh-Metrix, situated within the context of multilevel theoretical frameworks of discourse comprehension and learning. A review of published studies will highlight the applications of Coh-Metrix, ranging from the scaling and selection of educational material to learning environments at scale. The examples illustrate the relationship between discourse and cognitive, affective, and social processes. We walk through the methodological guidelines that should be followed when analyzing texts using Coh-Metrix. Finally, we conclude the paper with a general discussion of the future directions for Coh-Metrix including methodological and practical implications for the learning analytics (LA) and educational data mining (EDM) communities.

Keywords: Coh-Metrix, computational linguistics, learning analytics, ITSs, MOOCs, scaling of educational material

1 INTRODUCTION

Capturing the defining characteristics of language and discourse has enormous practical and theoretical value in education. The significant applications of computational linguistic analyses fall under two broad categories: 1) detecting and monitoring learning experiences, and 2) scaling and assessing educational texts. Language, discourse, and communication have been regarded as a gold mine that can offer powerful insights into learners' cognitive, affective, motivational, and social processes among other learning-related phenomena. Consequently, automated text analysis has garnered considerable attention among learning analytics (LA) and educational data mining (EDM) researchers attempting to

(2016). Language and discourse analysis with Coh-Metrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

improve emerging environments, such as intelligent tutoring systems (ITS), computer-mediated learning (CML), and massive open online courses (MOOCs). Furthermore, text complexity (or ease) is a central component that influences successful comprehension. As such, the selection and scaling of texts on complexity is a major priority for teachers, principals, and superintendents. Our purpose in this article is to provide information about a computational linguistic analysis facility, and the procedures by which it can be applied to educational data. More specifically, we hope this will preserve and distribute the information provided in the Learning Analytics Summer Institute (LASI, 2014) workshop on Coh-Metrix, a theoretically grounded, computational linguistics facility that analyzes texts on multiple levels of language and discourse (Graesser et al., 2014; Graesser, McNamara, Louwerse, & Cai, 2004; McNamara, Graesser, McCarthy, & Cai, 2014).

The subsequent sections of the paper are organized as follows. In section two, we discuss some of the motivating factors that led to Coh-Metrix situated within the context of multilevel theoretical frameworks of discourse comprehension and learning. Section three provides an overview of the types of measures and discourse dimensions provided by Coh-Metrix¹ and Coh-Metrix-Text Easability Assessor (TEA)² tools. Then, in section four, we review published studies to highlight the wide-ranging applications of Coh-Metrix, and illustrate the relationship between discourse and cognitive, affective, and social processes. In section five, we walk through the pedagogical guidelines that should be followed when analyzing texts using Coh-Metrix. Finally, we conclude the paper with a general discussion of future directions for Coh-Metrix, including methodological and practical implications for the EDM and LA communities.

2 MOTIVATION & THEORETICAL FRAMEWORK

Cognitive scientists have spent over half a decade studying how the human mind actively constructs meaning from discourse, which includes both oral communication and printed text. This endeavor stimulated a large body of research, in the '80s and '90s, highlighting the importance of cohesion and coherence (Gernsbacher, 1990; Goldman, Graesser, & van den Broek, 1999; Louwerse, 2001; McNamara & Kintsch, 1996; Sanders & Noordman, 2000). Cohesion is defined as characteristics of the explicit text that play some role in helping the reader mentally connect ideas in the text. Coherence is defined as a cognitive representation that reflects the interaction between linguistic/discourse characteristics and world knowledge. The prominent view from cognitive models in discourse psychology assumes that cognitive mechanisms within the reader/listener dynamically interact with the discourse during comprehension, and it is these processes that collaboratively generate cognitive representations (i.e., meaning) (Graesser, Singer, & Trabasso, 1994; Kintsch, 1998). This is supported by many theoretical frameworks, including the construction–integration model (Kintsch, 1998; Singer & Kintsch, 2001), the constructionist theory (Graesser et al., 1994; Singer, Graesser, & Trabasso, 1994), the structure building framework (Gernsbacher, 1997), the event-indexing model (Zwaan, Langston, & Graesser, 1995; Zwaan

¹ <http://cohmetrix.com>

² <http://tea.cohmetrix.com>

(2016). Language and discourse analysis with Coh-Metrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

& Radvansky, 1998), memory-based resonance models (Lorch, 1998), and the landscape model (van den Broek, Everson, Virtue, Sung, & Tzeng, 2002). These theoretical advances and practical needs stimulated the development of Coh-Metrix in 2002³ (Graesser et al., 2004; McNamara et al., 2014). However, Coh-Metrix quickly expanded beyond its initial goal of affording objective measures of cohesion. Now, 14 years later, Coh-Metrix has transformed into arguably the most comprehensive automated linguistics tool available to the public. The measures provided by Coh-Metrix reflect the advanced multilevel theoretical view of language, communication, and comprehension (Clark, 1996; Graesser, Gernsbacher, & Goldman, 2003; Graesser & McNamara, 2011; Kintsch, 1998).

This multiple-level view of discourse is the foundation of Coh-Metrix as well as many psychological theories of comprehension and learning. These theoretical frameworks identify representations, structures, strategies, and processes at different levels of conversation and printed text (Clark, 1996; Graesser, Millis, & Zwaan, 1997; Kintsch, 1998; McNamara & Magliano, 2009; Pickering & Garrod, 2004; Snow, 2002; van Dijk & Kintsch, 1983). A central tenet across these discourse frameworks is that, in both communication and text comprehension, misalignments, complications, and breakdowns can occur at different levels. These can be a product of deficits in the reader/listener (i.e., lack of knowledge or skill) or the discourse (e.g., incoherent text, unintelligible speech). In the learning context, such obstacles have important consequences. Indeed, numerous studies have highlighted a detrimental impact on students' attention, reading time, memory, logic, and other manifestations of cognition that influence subsequent behaviour and comprehension (e.g., Graesser, Lu, Olde, Cooper-Pye, & Whitten, 2005; Millis, King, & Kim, 2000; Zwaan & Radvansky, 1998).

Coh-Metrix automatically analyzes discourse on five of the six levels commonly proposed: *words*, *syntax*, the explicit *textbase*, the *situation model* (or mental model), the discourse *genre and rhetorical structure* (Graesser & McNamara, 2011; Kintsch, 1998; Snow, 2002). Words and syntax are exactly what their names imply, and together constitute what is called the surface code. The textbase consists of the explicit ideas (or propositions) in the discourse, and so it is the meaning rather than the surface code of wording and syntax. The situation model is the subject matter content or narrative world described by, but not necessarily explicitly stated within the discourse; this includes any inferences readers/listeners generate. It is important to note that inferences are critical factors of situation models. Inferences allow students to make connections between different elements, which facilitate the construction of a coherent memory of what the discourse is about. The discourse genre and rhetorical structure is the type of discourse and its structural composition (e.g., narration, exposition, and persuasion). We will elaborate on these levels later, and more information is available in previous journal publications and the Coh-Metrix book (Graesser et al., 2014; Graesser, McNamara, & Kulikowich, 2011; Graesser & McNamara, 2011; McNamara et al., 2014). The multilevel framework summarized in this section provides a sketch of the complexities involved in constructing meaning on different levels during communication and text comprehension that motivated the development of Coh-Metrix.

³ <http://cohmetrix.com>

3 COH-METRIX MEASURES OF LANGUAGE AND DISCOURSE

During the last fourteen years, the Coh-Metrix research group has gathered and evaluated hundreds of measures of language and discourse. This section gives an overview of the measures provided by Coh-Metrix. The emphasis of Coh-Metrix is on language and discourse characteristics closely related to deeper levels of cognition. As such, some linguistic features fall outside the bounds of this goal. For instance, Coh-Metrix is not well suited for capturing the basic reading components, such as the alphabet, letter-sound correspondences, lexical decoding, morphological awareness, and reading fluency (words read per minute). Similarly, some tools (e.g., Linguistic Inquiry Word Count) classify words into psychological categories based on the ratings of human experts (Pennebaker, Booth, & Francis, 2007). This dictionary based word-counting approach is quite useful for assessing very specific psychological word categories. Coh-Metrix includes some content focused word measures, but the primary emphasis is beyond the word into sentence interpretations, inferences, and more global discourse structures.

There are different versions of Coh-Metrix, so the number and specific measure offered depends on the version and the type of tool. We have an internal version of Coh-Metrix that preforms batch analyses, and is quite useful for larger volumes of texts. We offer a text analysis service to help researchers with larger corpora batch analyze many texts or analyze texts that exceed the limit of the online tools, which is limited to 15,000 characters per text. Additionally, two free public versions of Coh-Metrix, which differ in terms of complexity, are available on the web.⁴ The web version of Coh-Metrix 3.0 currently provides 108 measures. However, 108 linguistic features can be a bit overwhelming, especially for novice Coh-Metrix users. There were requests to make a more teacher friendly version (Elfenbein, 2011). As such, we made an effort to reduce the large number of measures provided by Coh-Metrix into a more manageable set. This was accomplished in a study that assessed 53 Coh-Metrix measures for 37,520 texts in the TASA (Touchstone Applied Science Association) corpus, which represents what typical high school students have read throughout their lifetime (Graesser et al., 2011). A principal components analysis was conducted on the corpus, yielding eight components that explained a striking 67.3% of the variability among texts; the top five components explained over 50% of the variance. Most importantly, the components aligned with the language-discourse levels previously proposed in multilevel theoretical frameworks of cognition and comprehension (Graesser & McNamara, 2011; Kintsch, 1998; Snow, 2002). The main five linguistic dimensions are currently being used to analyze texts in K–12 for the Common Core literacy standards (CCSSONGA, 2010) and states throughout the U.S. The Common Core Standards provide clear and consistent learning goals to help prepare students for college, career, and life.⁵ The standards clearly demonstrate what students are expected to learn at each grade level, so that every parent and teacher can understand and support their learning. The Coh-Metrix TEA tool illustrates these dimensions quite well for new users, allowing educators to enter a short passage (of fewer than 1000 words) and quickly receive a readability profile of the text. The interface is quite user-friendly (copy

⁴ The regular version (<http://www.cohmetrix.com>) and Coh-Metrix-TEA (Text Easability Assessor) (<http://tea.cohmetrix.com>).

⁵ <http://www.corestandards.org/>

(2016). Language and discourse analysis with Coh-Metrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

paste and analyze) and provides immediately interpretable results through an informative visual illustration and short explanation. The Coh-Metrix TEA is ideal for classroom use.

The potential applied contributions of the Coh-Metrix software are extensive. Since 2006, the public web version of the tool has attracted more than 10,000 registered users. Coh-Metrix has been applied in a variety of contexts, both within and beyond education, including assessing text readability (Gates Foundation), emotion detection, deception detection, terrorist/authoritarian leaders' speeches, writing styles, second language writing proficiency, and psychological disorders. It is beyond the scope of this article to specify precisely how each measure is computed. Such information is available from the Help system of Coh-Metrix and the referenced Coh-Metrix publications. However, we do briefly discuss some of the major computational measures.

Descriptive. Coh-Metrix provides descriptive indices such as the number and length of words, sentences, and paragraphs. These indices help the user to check the Coh-Metrix output (e.g., to make sure that the numbers make sense) and to interpret patterns of data.

Words. It is important to analyze words on multiple characteristics that have relevance to the construction of meaning. Coh-Metrix evaluates words on abstractness, parts of speech, familiarity, age of acquisition, and many other psychological features.

Lexical Diversity. Coh-Metrix provides three measures of lexical diversity. The most commonly used measure of lexical diversity is *type–token ratio* (McCarthy & Jarvis, 2007). Type–token ratio is the number of unique words in a text (i.e., types) divided by the overall number of words (i.e., tokens). Type–token ratio influences the cohesion of text. For instance, when the number of word types is equal to the total number of words (tokens), then all of the words are unique. In this situation, when lexical diversity is at a maximum, the text is either very low in cohesion or perhaps the text is very short.

Syntax. Coh-Metrix can scale texts on a variety of syntactic dimensions. Models of syntax ascribe words to part-of-speech categories (e.g., nouns, verbs, adjectives, conjunctions), group words into phrases (noun phrases, verb phrases, prepositional phrases, clauses), and assign syntactic tree structures to sentences (Jurafsky & Martin, 2009). Oral discourse typically has simpler syntactic structures with few if any embedded clauses, and active rather than passive voice (Tannen, 1982). Conversely, sentences in print, like academic articles, frequently have a complex, embedded syntax that creates demands on an individual's working memory. The following sentence illustrates this: "Due to hormone-induced shifts in the body's internal Circadian Clock or severely impacted schedules, adolescents stay up exceedingly late during the school week, compared to the weekend, accumulating a sleep debt which exacts a substantial physical and mental toll."

This sentence has several forms of complex syntax. First, it contains dense noun phrases with many modifiers. Second, it places a high number of words (i.e., 15) before the main verb (i.e., "stay") of the main clause, thus taxing the reader's working memory. Third, it requires the reader to keep track of

(2016). Language and discourse analysis with Coh-Metrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

many combinations of meaning with logic-based words such as “and,” “or,” and “not.” Other syntactic measures captured by Coh-Metrix include frequency of passive voice, which is more difficult to process than active voice (Snow, 2002); and syntactic similarity, or similarity in syntactic structure between pairs of sentences in a paragraph, which facilitates reading speed and comprehension.

Co-Referential Cohesion. Coh-Metrix tracks various forms of word co-reference: LSA, content word overlap, noun overlap, argument overlap, and stem overlap. These measures vary in terms of locality. That is, some measures reflect only reflect overlap between adjacent sentences in the text (local), whereas others compute co-reference of all possible pairs of sentences in a paragraph (global).

Latent Semantic Analysis. Coh-Metrix measures Latent Semantic Analysis cohesion (LSA; Landauer, McNamara, Dennis, & Kintsch, 2013). LSA provides measures of semantic overlap between sentences or between paragraphs. LSA considers meaning overlap between explicit words and words that are implicitly similar or related in meaning. For instance, *home* in one sentence will have a relatively high degree of semantic overlap with *house*, *cook*, and *table* in another sentence. LSA utilizes a statistical technique called singular value decomposition to condense a large corpus of thousands of texts to 100–500 statistical dimensions. The conceptual similarity between any two text excerpts (e.g., word, clause, sentence, text) is calculated as the geometric cosine between the values and weighted dimensions of the two text excerpts. The value of the cosine normally varies from 0 to 1.

Connectives. Connectives also represent an important category because they play a non-trivial role in establishing situation model cohesion (or Deep Cohesion). Coh-Metrix delivers a relative frequency (index) score (occurrence per 1000 words) for all connectives as well as different types of connectives. Indices are provided on five broad categories of connectives: causal (*because*, *so*), additive (*and*, *moreover*), temporal (*first*, *until*), logical (*and*, *or*), and adversative/contrastive (*although*, *whereas*) which Coh-Metrix classifies based on prior research (Halliday & Hasan, 1976; Louwerse, 2001). Additionally, Coh-Metrix differentiates between positive connectives (*also*, *moreover*) and negative connectives (*however*, *but*).

Situation Model. Scholars in cognitive science and discourse processing use the expression *situation model* to refer to the level of conceptual representation for a text that goes beyond the explicit words and sentences (Graesser & McNamara, 2011; Graesser et al., 1994; Kintsch, 1998; van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998). The situational model is the subject matter content that the text is describing. In narrative text, this includes the characters, objects, spatial settings, actions, events, processes, plans, thoughts and emotions of characters, and other details about the story. In informational text, the situation model corresponds to the substantive subject matter (i.e., domain knowledge, topics) that the text describes. For example, the lead sentence in a recent *Economist* (2014) article stated, “The rise of online instruction will upend the economics of higher education.” This sentence would potentially activate the following background knowledge: (a) causal networks of the events, processes, and enabling states that explain the rise of online instruction, (b) properties of online instruction (and likely activation of MOOCs) and higher education, (c) the mechanisms of upending the

(2016). Language and discourse analysis with Coh-Metrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

economics of higher education, and (d) goal-oriented actions of online instructors. At least some world knowledge about the economics of traditional higher education and online instruction is needed to comprehend the example sentence. The situation model includes inferences activated by the explicit text and encoded in the meaning representation (Goldman, Braasch, Wiley, Graesser, & Brodowinska, 2012; Graesser et al., 1994; Kintsch, 1998; D. S. McNamara & Kintsch, 1996; Wiley et al., 2009). Zwaan and Radvansky (1998) proposed five dimensions of the situational model that apply to the thread of deep comprehension: causation, intentionality (goals), time, space, and people. A break in text cohesion occurs when there is a discontinuity on one or more of these situation model dimensions. Such cohesion breaks result in an increase in reading time and generation of inferences (Rapp, van den Broek, McMaster, Kendeou, & Espin, 2007; Zwaan & Radvansky, 1998). When such discontinuities arise, it is important to have connectives, transitional phrases, adverbs, or other signalling devices that convey to the readers that there is a discontinuity. Coh-Metrix provides multiple measures of causal, temporal, and intention cohesion to capture the breath of situation model cohesion.

Coh-Metrix Principal Components

- **Narrativity.** Narrative text tells a story, with characters, events, places, and things familiar to the reader. Narrative is closely affiliated with everyday oral conversation. This robust component is highly affiliated with word familiarity, world knowledge, and oral language. Informational expository texts on less familiar topics would lie at the opposite end of the continuum.
- **Deep Cohesion.** This dimension reflects the degree to which the text contains causal, intentional, and temporal connectives and conceptual links. These connectives help the reader to form a more coherent and deeper understanding of the causal events, processes, and actions in the text.
- **Referential Cohesion.** This component includes Coh-Metrix indices that assess referential cohesion. High-cohesion text contains words and ideas that overlap across sentences and the entire text, forming explicit threads that connect the text for the reader. Low cohesion text is typically more difficult to process because there are fewer threads that tie the ideas together for the reader.
- **Syntactic Simplicity.** This component reflects the degree to which the sentences in the text contain fewer words and use simpler, familiar syntactic structures, which are less challenging to process. At the opposite end of the continuum are texts that contain sentences with more words, embedded constituents, unfamiliar syntactic structures, noun-phrases with many modifiers, and many words before the main verb of the main clause (i.e., left-embedded syntax that is taxing on working memory).
- **Word Concreteness.** Texts that contain content words that are concrete, meaningful, and evoke mental images are easier to process and understand. Abstract words represent concepts that are difficult to represent visually. Texts that contain more abstract words are more challenging to understand.

Recently, Graesser and colleagues (2014) also defined a composite *formality* score that increases with low narrativity, syntactic complexity, word abstractness, and high cohesion. The formality metric was derived from the first five principal components listed above: [(referential cohesion + deep causal

(2016). Language and discourse analysis with Coh-Metrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

cohesion – narrativity – syntactic simplicity – word concreteness)/5]. This formality metric has a high correlation with unidimensional metrics of text difficulty as well as other psychological measures that reflect processing difficulty.

4 TOUR OF COH-METRIX APPLICATIONS IN LEARNING SCIENCES

There have been over 100 published studies validating Coh-Metrix indices (McNamara, Louwse, McCarthy, & Graesser, 2010; McNamara et al., 2014). When the keyword “Coh-Metrix” is entered into Google Scholar, it returns 1,300 results. Obviously, some of these will be redundant, but it does give a rough sense of the growing interest in Coh-Metrix. In this section, we review a broad set of published studies to illustrate the wide-ranging applications of Coh-Metrix in the learning sciences. The most popular applications of Coh-Metrix from the perspective of learning analytics fall under two broad categories: detecting and monitoring of cognitive, affective, motivational, and social processes and the scaling and assessment of educational texts.

4.1 Creation and Evaluation

Researchers who analyze language and discourse processing often compare particular text segments with control or comparison text segments that differ on some particular text feature. A rigorous comparison requires the researcher to rule out extraneous text features. However, the potential for uncontrolled variability in experimental texts is daunting, and quite difficult to address without objective measures. Coh-Metrix can be used to quickly detect any unintended linguistic differences between control and experimental texts (Dodell-Feder, Koster-Hale, Bedny, & Saxe, 2011; McNamara et al., 2010). Early research in psycholinguistics was limited to sentences, or shorter passages, because of the increasing complexity associated with tracking sources of linguistic variability in longer texts. As one would imagine, the issue is amplified in naturalistic texts, such as newspaper articles or textbooks. One of the practical benefits of Coh-Metrix is providing computational measures that can track various aspects of language and discourse more effortlessly and reliably. Indeed, systematic investigations of language have provided a number of exciting insights and challenges for both theory and practice.

As one example, common sense would predict that high-cohesion texts yield better comprehension than low-cohesion texts. However, researchers in discourse processing have discovered that the relationship between cohesion and comprehension is less straightforward than one might intuitively suspect. McNamara and her colleagues have documented that complex interactions occur between text cohesion and the readers’ prior knowledge (O’Reilly & McNamara, 2007). A considerable amount of research has documented the benefits of increasing cohesion for readers with low knowledge (McNamara, Kintsch, Songer, & Kintsch, 1996; McNamara & Kintsch, 1996; O’Reilly & McNamara, 2007). These studies substantiate that all types of cohesion can help these readers. The low knowledge readers are simply not equipped with enough background knowledge to generate the inferences needed to connect constituents in low cohesion texts. Their lack of prior knowledge makes it impossible to bridge

(2016). Language and discourse analysis with Coh-Metrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

the cohesion gaps without explicit text cues for cohesion, such as connectives and overlap of words in noun-phrases among sentences.

Interestingly, the story is quite different when it comes to students with adequate background knowledge. Across several studies, students with more background knowledge either do not benefit from cohesion, or actually profit from a lack of cohesion in the text (McNamara et al., 1996; McNamara & Kintsch, 1996; O’Reilly & McNamara, 2007). This phenomenon has been referred to as the *expertise reversal effect*. Subsequent studies identified the main explanations for this phenomenon. Essentially, the high knowledge readers in McNamara et al. (1996) were able to gain from low cohesion text because it forced them to generate inferences, and that active construction of inferences resulted in deeper comprehension and enhanced understanding of the situation model.

This intriguing finding would never have been revealed using traditional measures of text readability, namely Flesch-Kincaid Grade Level or Reading Ease (Klare, 1974) and Lexile scores (Stenner, 2006). These formulas provide an indication of text readability based on the word and sentence lengths found in the text. Thus, readability measures often predict a decrease in ease when cohesion is increased because adding cohesion often results in increasing the length of the sentences through connectives and adding more unfamiliar or longer words. This is indicative of one of the many deficiencies of unidimensional readability formulas. As Graesser and colleagues (2011) have pointed out, their simplicity and association with grade level is attractive, but they lack the ability to capture the more global levels of discourse meaning, cohesion, and differences in text genre (e.g., narrative versus informational texts). Additionally, unidimensional measures are not useful for identifying specific deficit areas in a text or providing students with personalized support on particular reading problems (Rapp et al., 2007).

Recently, Coh-Metrix has been recruited by the Common Core Standards to aid in the scaling and selection of text using the multilevel analysis approach. The multilevel framework can be used to guide the selection of texts according to particular pedagogical goals. For instance, supporters of Vygotsky’s (1978) zone of proximal development would agree that educational material should not be too difficult or too easy for students, but should occupy an intermediate zone of difficulty. That is, sometimes learners benefit from challenging material. In this context, a teacher might want texts that aggressively push the envelope of what they can handle and provide scaffolding support to help them through the text comprehension. At the other end of the continuum, students occasionally need a self-confidence boost, and thus would benefit from easier material that they can readily comprehend. Our vision is that perhaps it is best for students to receive a “diet” balanced across the difficulty dimension, with a bias toward the intermediate zone. As suggested by Graesser et al. (2011), texts can be recommended or assigned by teachers based on this multifaceted profile of text characteristics. Consider the types or combinations of texts that might be assigned depending on certain pedagogical goals:

- **Challenging texts with associated explanations.** Some assigned texts are considerably beyond students’ ability level. In such cases, students need comments by a teacher, tutor, group, or

(2016). Language and discourse analysis with Coh-Metrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

computer that explains technical vocabulary and points of difficulty. Students are greatly stretched by exposure to difficult content, strategies, and associated explanations.

- **Texts at the zone of proximal development.** Some assigned texts are slightly above the difficulty level that students can handle. These texts gently push the envelope — they are not too easy or too difficult, but just right.
- **Easy texts to build self-efficacy.** Easy texts are assigned to build reading fluency and self-efficacy. Struggling readers can lose self-confidence, self-efficacy, and motivation when beset with a high density of texts that they can barely handle, if at all.
- **A balanced diet of texts at varying difficulty.** Texts may be assigned according to a distribution of alternatives 1, 2, and 3 above, mostly in the zone of proximal development. The balanced diet benefits from exposure to challenging texts, texts that gradually push the envelope, and texts that build self-efficacy. This approach also includes texts in different genres.
- **Texts tailored to develop particular reading components.** Texts may be assigned adaptively in a manner sensitive to the student’s complex profile of reading components. The texts attempt to rectify particular reading deficits or to advance particular reading skills.

While discourse researchers have explored these five approaches, it is beyond the scope of this article to comment on which approach best serves particular populations of readers. Instead, the point of listing these is to offer examples that highlight the landscape of possibilities.

4.2 Prediction & Detection

Advances in educational technologies and a desire for increased access to learning are enabling the development of pedagogical environments at scale, such as intelligent tutoring systems (ITSs), computer-mediated collaborative learning (CMCL) environments, and massive open online courses (MOOCs). The insulated nature of the computer-mediated platforms allows valuable learning dynamics to be detailed at unprecedented resolution and scale. As such, the digital traces left by learners are regarded as a goldmine that can offer powerful insights into the learning process, resulting in the advancement of educational sciences and substantially improved learning environments. Regarding analytical approaches, there has been extensive knowledge gleaned from manual content analyses of learners’ discourse during educational interactions, but these methods are no longer a viable option with the increasing volume of educational data. Consequently, researchers have been incorporating automated linguistic analyses that range from shallow level word counts to deeper level discourse analysis approaches. Both levels of linguistic analysis are informative. In this section, we review some of the recent applications of Coh-Metrix in these emerging learning environments.

Affect-sensitive learning environments have practical and theoretical interest. The growing interest in this has surfaced as a result of research showing that cognition and emotion are inextricably linked (Baker, D’Mello, Rodrigo, & Graesser, 2010; Dalgleish & Power, 1999; D’Mello, Lehman, Pekrun, & Graesser, 2014; Lehman, D’Mello, & Person, 2010). From a practical view, affect detection is a cornerstone of affect-aware interfaces that aspire to automatically detect and intelligently respond to

(2016). Language and discourse analysis with Coh-Metrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

students' emotions. The ITSs community has leveraged recent advances in affective computing to detect the learners' affective states (Calvo & D'Mello, 2010; D'Mello & Kory, 2012; D'Mello, Dowell, & Graesser, 2013). Though the use of physiological and bodily sensors represent feasible options to detect affect in the lab and classroom settings, they are not viable options for scaled learning environments (Calvo & D'Mello, 2010). There are several advantages to utilizing textual features as an independent channel for affect detection. First, textual features are abundant and inexpensive to collect in ITSs that support natural language dialogues. Second, textual features derived from tutorial dialogues are contextually constrained in a fashion that provides cues regarding the social dynamics of the student and tutor.

Recently, Coh-Metrix has been used to explore intelligent tutoring systems dialogues. This research has revealed that the language and discourse features of students and tutors are particularly good diagnostics of the learners' affective states (D'Mello, Dowell, & Graesser, 2009; D'Mello & Graesser, 2012; D'Mello & Graesser, 2010). In fact, when all of the learning-centred emotions are considered, language/discourse features predict learner emotional states as well as facial expressions and better than body posture. For instance, D'Mello, Dowell, and Graesser (2009) explored the possibility of predicting learners' affective states (boredom, flow/engagement, confusion, and frustration) by monitoring variations in the cohesiveness of tutorial dialogues during interactions with AutoTutor, an intelligent tutoring system with conversational dialogues. Multiple measures of cohesion (e.g., pronouns, connectives, semantic overlap, causal cohesion, co-reference) were automatically computed using Coh-Metrix. Cohesion measures in multiple regression models predicted the proportional occurrence of each affective state, yielding medium to large effect sizes. Specifically, the findings indicated the incidence of negations, pronoun referential cohesion, causal cohesion, and co-reference cohesion were the most diagnostic predictors of the affective states. We subsequently used Coh-Metrix to explore more socio-affective constructs. Specifically, Coh-Metrix was used to detect learners' socio-affective attitudes towards fellow students in computer-mediated collaborative environments, which may have long-term consequences for their motivation and continued use of such systems (Cade, Dowell, Graesser, Tausczik, & Pennebaker, 2014). These and other findings illustrate the utility of automated text analysis in emerging learning environments. Similar to ITSs, we could aim to create affect-sensitive MOOC environments that could intelligently support learners with pedagogical and motivational strategies.

Coh-Metrix has also been used to explore learners' cognitive processes in the context of collaborative learning. Dowell and others explored the possibility of using discourse features to predict student and group performance during collaborative learning interactions (Dowell, Cade, Tausczik, Pennebaker, & Graesser, 2014). They investigated the linguistic patterns of group chats, within an online collaborative learning exercise, on five discourse dimensions using Coh-Metrix. The results indicated that students who engaged in deeper cohesive integration and generated more complicated syntactic structures performed significantly better. Interestingly, the overall group level results indicated collaborative groups who engaged in deeper cohesive and expository style interactions and performed significantly better. Although students do not directly express the nature of knowledge construction and cognitive

(2016). Language and discourse analysis with Coh-Metrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

processes at a meta level, these states can be automatically tracked by analyzing language and discourse. Another interesting finding is regarding the topic of granularity in collaborative learning analyses. Dowell and others' research shows that it takes an analysis of both the student level and group level discourse to acquire a comprehensive understanding of the linguistic properties that influence knowledge acquisition during collaborative group interactions. These findings stimulate an interesting discussion because, until recently, most research on groups has concentrated on the individual people in the group as the cognitive agents (Stahl, 2009). This traditional granularity uses the individual as the unit of analysis both to understand behavioural characteristics of individuals working within groups and to measure performance or knowledge-building outcomes of the individuals' in-group contexts. However, the present findings support the claims of many in the computer supported collaborative learning (CSCL) community to also consider group levels of granularity in discourse tracking (Graesser, Jeon, Yan, & Cai, 2007).

In the context of MOOCs, Social Network Analysis (SNA) is increasingly used to explore learning-related phenomena (Gašević, Kovanović, Joksimović, & Siemens, 2014). Automated linguistic analysis of student interactions within computer-mediated learning environments can complement SNA techniques by adding rich contextual information to the structural patterns of learner interactions. Coh-Metrix has recently been involved in pioneering research exploring the potential methodological and theoretical advantages of combining SNA and computational linguistic analyses (Dowell et al., 2015; Joksimović et al., 2015). Joksimović and colleagues' (2015) research used Coh-Metrix to analyze learners' forum posts in a distributed (Twitter, blogs, and Facebook) MOOC. Social Network Analysis was used to determine students' social centrality. Linear mixed-effect modelling was used to reveal the linguistic profiles associated with more centrality located learners. Overall, the results indicated that learners in the MOOC connected more easily to individuals who use a more informal narrative style, but still maintain a deeper cohesive structure in their communication. However, this linguistic profile cannot be immediately interpreted as beneficial for learning. Dowell et al. (2015) used a similar methodological design, but also included a measure of student performance in the MOOC. Specifically, they explored the extent to which characteristics of discourse diagnostically reveal learners' performance and social position in a MOOC. Their results for performance mirrored the pattern observed for learning in the computer-mediated collaborative learning study discussed earlier (Dowell et al., 2014). Specifically, students who performed significantly better engaged in more expository style discourse, with surface and deep level cohesive integration, abstract language, and simple syntactic structures. However, linguistic profiles of the centrally positioned learners differed from the high performers. Learners with a more significant and central position in their social network engaged using a more narrative style discourse with less overlap between words and ideas, simpler syntactic structures, and abstract words. These results are similar to those observed by Joksimović and colleagues (2015). Interestingly, their findings highlight a misalignment between the linguistic features associated with improved performance and more centrally located network positions. In other words, high performers and those with central positions in the network are not necessarily the same individuals. Additional research is needed to track the far-reaching implications of these two different profiles of individuals. Nevertheless, the results pose

(2016). Language and discourse analysis with Coh-Matrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

some provocative theoretical and practical implications for transferring analytic approaches to scaled environments.

Language, discourse, and communication are at the foundation of emerging, computer-mediated learning environments. As such, they are regarded as a goldmine that can offer powerful insights into the learning process. Computational linguistics tools, like Coh-Matrix, can be particularly useful for exploring learning-related phenomena in scaled learning environments because it is domain-independent, unobtrusive, inexpensive, computationally powerful, and theoretically grounded in learning sciences. The studies reviewed in this section highlight some of the recent work showing the advantages of using Coh-Matrix to identify pedagogically valuable discourse features that can be applied in collaborative learning, intelligent tutoring systems (ITS), computer-mediated collaborative learning (CMCL), and MOOC environments.

5 METHODOLOGICAL GUIDELINES FOR USING COH-MATRIX

At this point, our assumption is that the reader has general knowledge of Coh-Matrix (see McNamara et al., 2014). That is, we assume that the reader knows *what Coh-Matrix is, has a general grasp of the theoretical foundations, is familiar with some of the measures, and has a good understanding of the scope of applications*. In this section, we describe some of the technical details, methodological guidelines, and best practices to follow when conducting Coh-Matrix analyses. This is written for novice users to guide them in understanding the steps in choosing a corpus, important pre-processing best practices, use of Coh-Matrix online tools, and the nature of the resulting data. More detailed information is provided in the Coh-Matrix book (McNamara et al., 2014), including how to write up a research paper using a tool like Coh-Matrix.

5.1 The Corpus, Pre-Processing, and Best Practices for Text Analytics

Whether the project starts with a research question or a theory, researchers must consider the *corpus*, and continue considering the corpus during most of the research process. A corpus is a collection of texts. For example, corpora may be newspaper articles, entries in encyclopaedias, science texts in schools, legal documents, ITS and MOOC transcripts, advertisements, short stories, theatrical scripts — the list goes on. The texts are of enormous importance because they are the empirical manifestations of the hypothesis the researcher is testing. The Coh-Matrix 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 in 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. Building a corpus is no simple matter and many criteria have to be considered (e.g., what kinds of texts should be in it, how large does it have to be, etc.). Careful considerations of these and other questions are just as important as forming the research question, the hypotheses, and the theory.

(2016). Language and discourse analysis with Coh-Metrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

There are different ways to classify corpora. For instance, a corpus can be complete or a sample. In many cases, the researcher might not have access to the complete corpus or the complete corpus is so large it is not feasible to collect and analyze in full. In these instances, the discourse corpus needs to be sampled systematically and scientifically, rather than haphazardly or with bias. Ideally, the science of selecting a discourse corpus should be on par with the science of selecting participants in experimental studies. That is, a corpus should be randomly sampled, representative of the population, and a large enough sample. Other examples of incomplete corpora would be the texts used in the MOOC and collaborative learning studies discussed earlier. These might be seen as complete because all transcripts were analyzed from those courses. However, they are not complete because we did not analyze all MOOC interaction in the history of MOOCs. In this context, it is important to use statistical methods that help address the variance associated with individual learners or courses, so that the results are more representative of the full population of discourse samples. These examples show that corpora are rarely complete in the strictest sense. Instead, the researcher will have an incomplete corpus and the best practice for addressing potential limitations will be determined by the type of incomplete corpus. All of these issues need to be carefully considered so that the corpus and subsequent findings can be justified as representative.

The format of the corpus is another important criterion to keep in mind when selecting texts. Specifically, Coh-Metrix can only analyze that which is computationally analyzable. More simply, there is no slot in Coh-Metrix through which we can deposit hand-written texts, painted texts, CDs of talks, movie cassettes, or any example of sign language or brail. Although making such remarks might seem obvious, it is nevertheless important to consider these limitations of Coh-Metrix because 1) many people ask us, 2) future developments in Coh-Metrix need to consider these aspects because they are, after-all, language too, and 3) if the researcher's texts are in any of these forms then they will have to be changed to .txt documents, a process that might be extraordinarily long and painful.

Once the researcher has a representative and balanced corpus, in a computationally analyzable format for Coh-Metrix, the next phase will be the pre-processing. In this phase of the project, other characteristics of the texts need to be considered. Whether the corpora are collected by the researcher, designed by professionals, or borrowed from other studies, few of them are ever clean. The best way to think about a clean corpus is imagining it as close to human readable form as possible. In other words, a clean text looks exactly like it would appear if the writer had just finished typing it, had it checked for typos and errors by a large group of copy editors, printed if off, and then handed it to the researcher.

So when are corpora ever *dirty*? Many professional corpora are annotated for such features as parts of speech, intonation, and even the actions of the speaker (e.g., “applause”). In other cases, such as student essays, odd line breaks may have occurred, and bizarre spelling is ubiquitous. Similarly, corpora that have been passed around from computer to computer tend to “grow” various oddities such as the odd Spanish letter, or a string of mathematical symbols. Particularly in cases where researchers have converted documents that include pictures into text files, the pictures in the document disappear, often leaving the captions lurking oddly in the middle of the text. Each of these *dirty* has the potential to

(2016). Language and discourse analysis with Coh-Matrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

seriously undermine the validity of Coh-Matrix analyses. The biggest issue with these dirties is that they are never consistent. In other words, where they have been found to be consistent, we have designed algorithms to correct for them. As such, the researcher is ultimately responsible for making sure that the corpus is sufficiently clean. An appropriate phrase here is, “Garbage in, garbage out” (GIGO).

In the LASI and other workshops on Coh-Matrix, we have discussed some useful tools and approaches for dealing with these issues. For example, general regular expressions and programs like Textcrawler⁶ can be very powerful tools for batch cleaning texts. Many students and participants have inquired about what should be removed from the text, and what can be left in (e.g., headers, typos, spelling mistakes, pronunciation guides etc.). There are two golden “best practices” to help guide students and researchers in these decisions:

1. If there is no good reason to take it out, the researcher should leave it in
2. What the researcher does to one text, should be done to all

Best practice 1 states that the default condition of the text is exactly the way the researcher found it. Therefore, all changes made to it after that should be documented and reported for future replications. Most commonly, researchers decide to remove annotations and picture captions. The logic behind this decision is that they make the text unreadable, and consequentially any Coh-Matrix results are likely to be seriously flawed. A different motivation might be reported for removing the picture captions. Here our strong argument would be that they are not part of the continuous text that the writer intended. Additionally, their insertion into the document renders the sentence meaningless, and the corresponding evaluations will be misleading. Best practice 2 is extremely important. It means a researcher should never pick and choose which texts to modify. If something is removed from one text (e.g., a day, month, and year that happens to be at the end of a text) then one must confirm that none of the other texts also have that pattern (and if they do, they must all be removed, or all kept). Similarly, the same consistency should be used for spelling corrections and typos. Finally, it is important to understand that having a few dirties across the corpus is not considered unusual. As a general rule, the corpus needs to be at least 95% clean. That is, about 95% of the texts should have no problems at all, and at least 95% of each text should be thoroughly correct. When researchers have very large corpora, reading through all of them is not feasible. Note, that in this context assessing a random sample of the text (e.g., 10%) is generally considered sufficient.

5.2 Coh-Matrix Tools, Data and Illustrative Example

Analyzing texts with the free Coh-Matrix tools online is the easiest part of the process. Both of the Coh-Matrix websites are set up to be quick and user friendly for students, teachers, and researchers. Figures 1 and 2 show the main pages that users will find when they visit the Coh-Matrix-Text Easability Assessor (TEA) and the Coh-Matrix 3.0 websites, respectively. To set up an account on the Coh-Matrix TEA site,

⁶ <http://textcrawler.en.softonic.com>

(2016). Language and discourse analysis with Coh-Metrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

simply click on “New user click here” and fill in the requested information. Similarly, for the Coh-Metrix 3.0 website, click on the “Web tool” button and set up an account. Users are encouraged to explore the other links on the sites, which provide tons of useful information, including a full list of relevant Coh-Metrix references, detailed descriptions of the 109 indices (click “Documentation” on the Coh-Metrix 3.0 website), information on our new text analysis service, and links to a new Chinese version of Coh-Metrix.

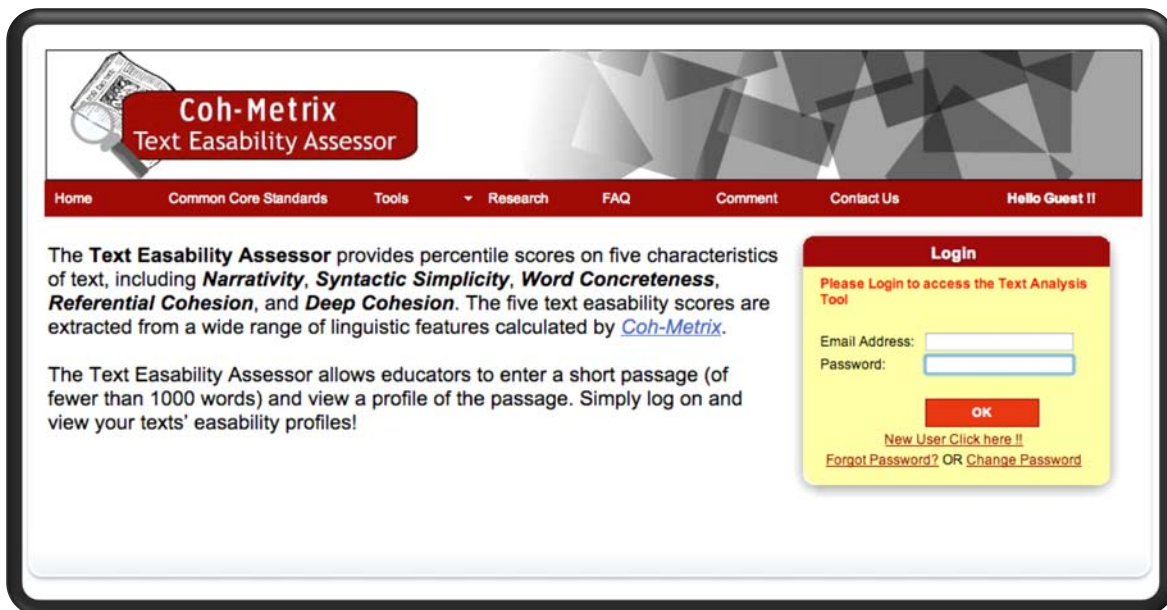


Figure 1. Coh-Metrix TEA tool main page (<http://tea.cohmetrix.com>).

On both websites, the text is entered using simple cut-and-paste from a text file. Note that both tools are limited to about 15,000 characters. Researchers and students who wish to use the internal batch facility can contact us through the website and inquire about the process of using it. The Coh-Metrix Data Viewer facility is shown in Figure 3. This facility allows the user to inspect the sentence segmentation of the texts prior to analysis. This is one of the ways users can make sure the texts are “clean.”

(2016). Language and discourse analysis with Coh-Metrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

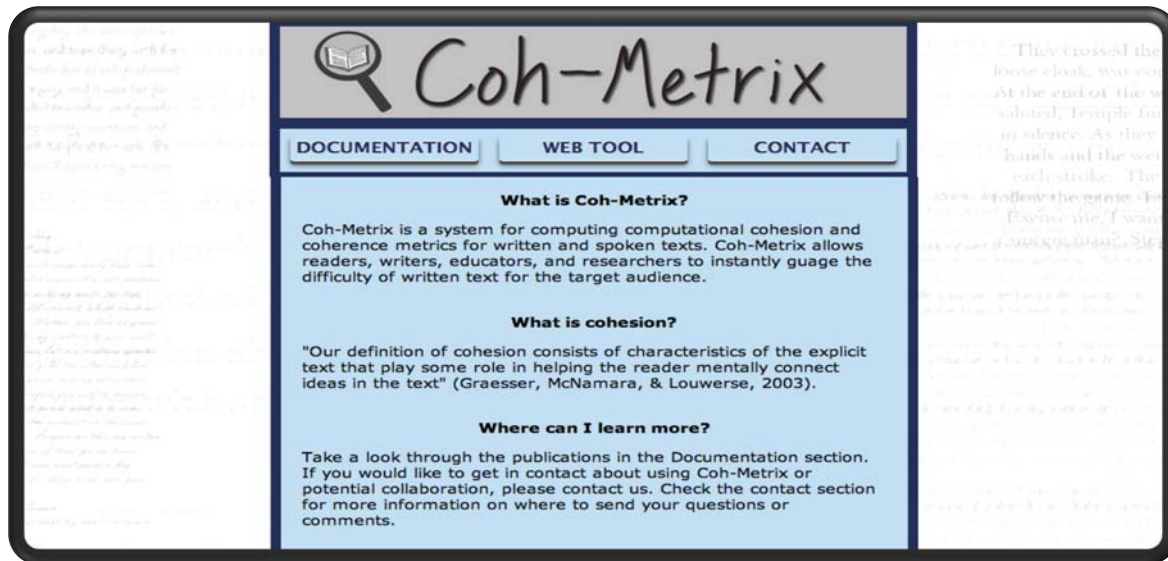


Figure 2. Coh-Metrix 3.0 main page (<http://cohmetrix.com>)

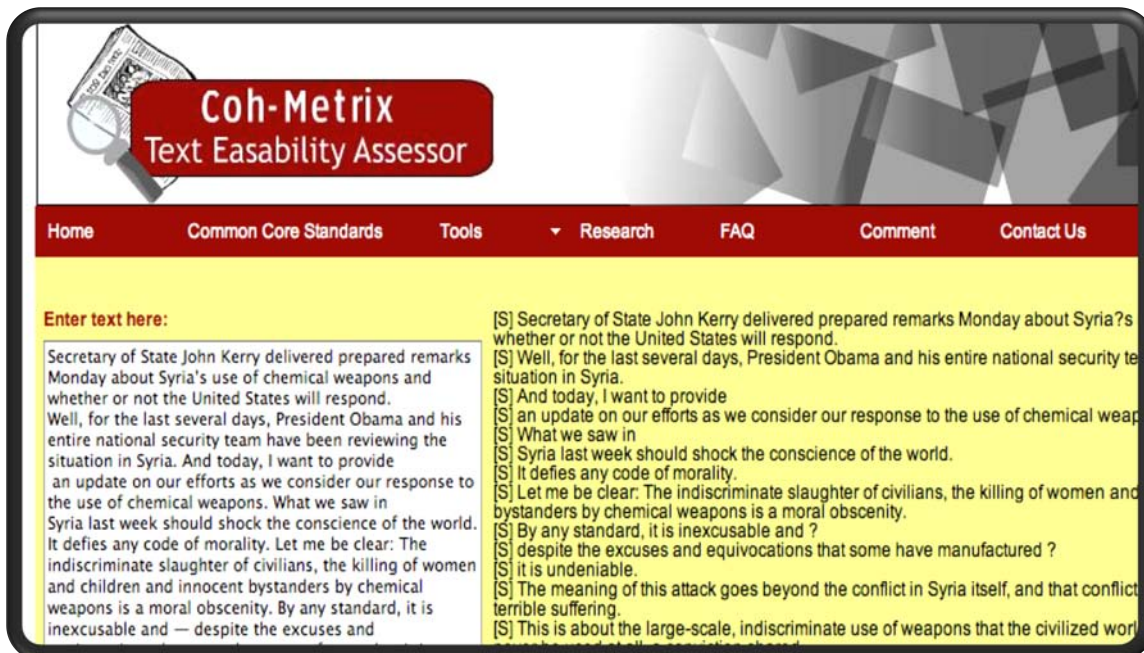


Figure 3. Coh-Metrix TEA pre-process facility

Once a clean text is entered into either or both of the Coh-Metrix websites, simply hit the analyze button to receive the linguistic profile of the text. However, as we mentioned earlier, the websites provide different types of output. The results from Coh-Metrix 3.0, which can be downloaded in an analysis ready format (.csv), include the full scope of measures.

(2016). Language and discourse analysis with Coh-Matrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

An example of Coh-Matrix use is provided in Figures 4 and 5 to illustrate, through a concrete case study, some of the main features and strengths of the tool and framework. In this case study, we are interested in exploring the discourse characteristics of MOOC participants’ forum posts. Suppose we are exploring similar research questions as the Dowell, Skrypnik, Joksimović, and colleagues (2015) studies reviewed earlier. Specifically, we want to see if the posts from centrally located MOOC participants exhibit different linguistic profiles than the posts of more peripherally located participants. The Coh-Matrix TEA analysis results for the centrally and peripherally located MOOC participants are presented in Figures 4 and 5, respectively.

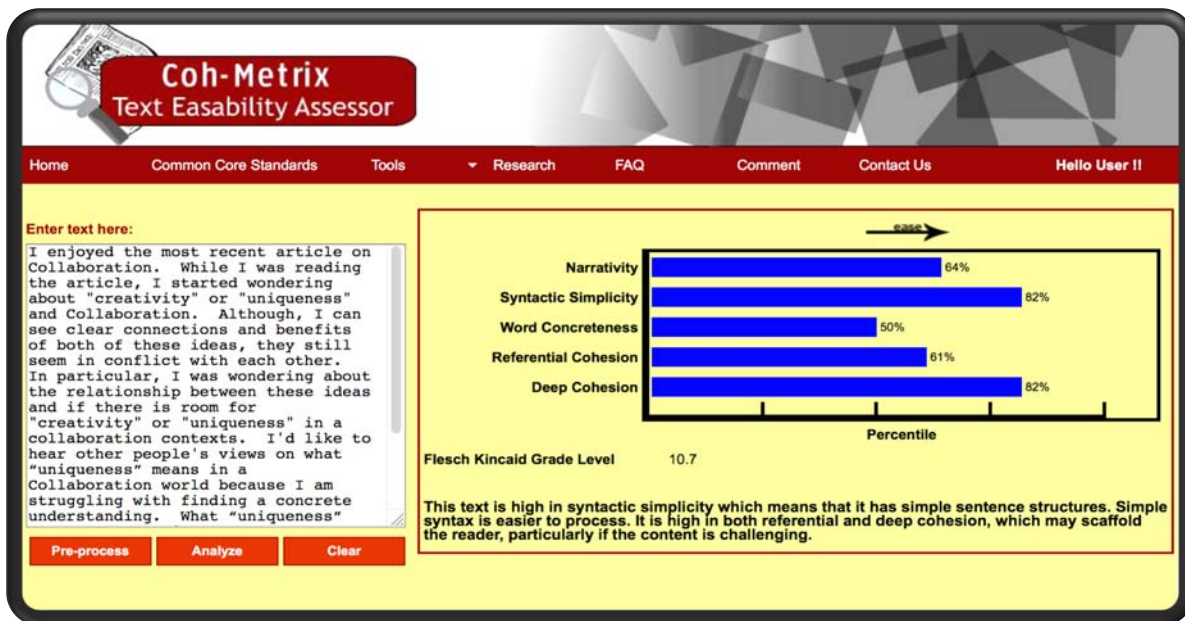


Figure 4. Coh-Matrix TEA example analysis of a centrally located MOOC participant’s forum posts

The results presented in Figure 4 suggest that the participant who attained a more prominent social centrality position used more conversational style discourse overall. Specifically, the centrally located MOOC participant engaged using a more narrative style of discourse with high overlap between words and ideas (referential cohesion), deep level cohesive integration, concrete language, and simple syntactic structures. The example results, presented in Figure 5, for the posts from the peripherally located MOOC participant reveal a very different linguistic profile. Here we see this participant engaged in a more expository style of discourse (less narrative), with little cohesive overlap between words and ideas (low referential cohesion) and a deep level of cohesive integration, abstract language, and complex syntactic structures. Ideally, this sample Coh-Matrix analysis has illustrated some of the main features and strengths of the tool and framework. If this were a real study, the next step would be to interpret and ground these observed findings in the theoretical frameworks of relevant learning sciences, discourse, and social interaction.

(2016). Language and discourse analysis with Coh-Matrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

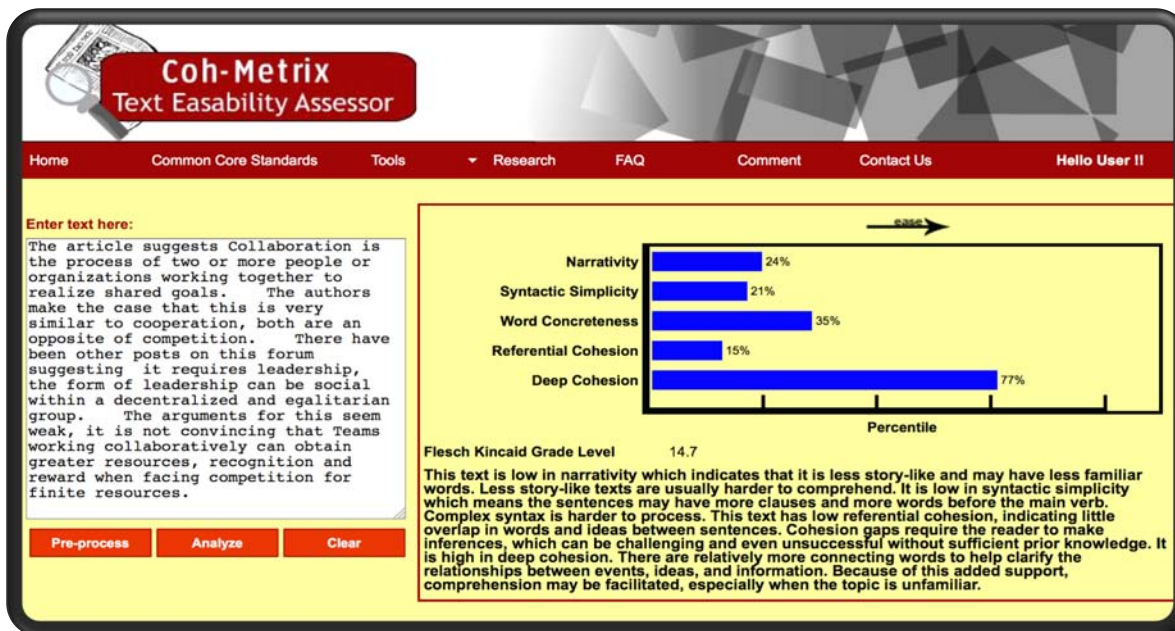


Figure 5. Coh-Matrix TEA example analysis of a peripherally located MOOC participant’s forum posts.

6 CONCLUSIONS & FUTURE DIRECTIONS

We hope this will preserve and distribute the information provided in the Learning Analytics Summer Institute (LASI, 2014) workshop on Coh-Matrix. The workshop focused on the utility of Coh-Matrix in discourse theory and educational practice. In this article, we have reviewed most of the important information presented in the workshop. Unfortunately, an article is not a good substitution for the hands-on experience gained by the participants of the Coh-Matrix workshop. In light of that, we extend a standing offer to provide one-on-one tutorials via Skype or other platforms. Any students or researchers who need additional assistance, or would like to use the text analysis service, may contact the authors at the Institute for Intelligent Systems (IIS).⁷ Our contact information is available on the IIS and Coh-Matrix websites.

We have received an increase in requests for Coh-Matrix analyses from the learning analytics and educational data mining communities. This has stimulated a new Coh-Matrix project, the goal of which is to expand the architecture drastically for a scalable web-based Coh-Matrix text analysis service. The existing Coh-Matrix software has great potential both as a research tool, and as a basis for numerous commercial services. The end Coh-Matrix product will be more flexible and extensible so that researchers can easily apply the base functionality to different services. In our view, an interdisciplinary approach that combines psychological theories of discourse comprehension with computational linguistics methodologies holds the potential for enabling substantially improved learning environments.

⁷ <http://www.memphis.edu/iis/>

(2016). Language and discourse analysis with Coh-Matrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

7 ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation under Grant BCS 0904909, DRK-12-0918409; the Institute of Education Sciences (Grants R305G020018 and R305A080589); and Minerva Initiative, Air Force Office of Scientific Research, Department of Defense (FA9550-14-0308).

REFERENCES

- Baker, R. S. J. d., D’Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners’ cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human–Computer Studies*, 68(4), 223–241. <http://dx.doi.org/10.1016/j.ijhcs.2009.12.003>
- Cade, W. L., Dowell, N. M., Graesser, A. C., Tausczik, Y. R., & Pennebaker, J. W. (2014). Modeling student socioaffective responses to group interactions in a collaborative online chat environment. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 399–400). Berlin: Springer.
- Calvo, R. A., & D’Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1), 18–37. <http://dx.doi.org/10.1109/T-AFFC.2010.1>
- CCSSONGA (Council of Chief State School Officers and National Governors Association). (2010). *Common core state standards for English language arts and literacy in history/social studies, science, and technical subjects*. (Common Core State Standards Initiative). Washington, DC. <http://www.corestandards.org/>
- Clark, H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.
- Dalgleish, T., & Power, M. (Eds.). (1999). *Handbook of cognition and emotion* (1st ed.). Chichester, UK: Wiley.
- D’Mello, S., Dowell, N., & Graesser, A. C. (2009). Cohesion relationships in tutorial dialogue as predictors of affective states. *Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling* (pp. 9–16). Amsterdam, Netherlands: IOS Press
- D’Mello, S. K., Dowell, N., & Graesser, A. (2013). Unimodal and multimodal human perception of naturalistic non-basic affective states during human–computer interactions. *IEEE Transactions on Affective Computing*, 4(4), 452–465. <http://dx.doi.org/10.1109/T-AFFC.2013.19>
- D’Mello, S. K., & Graesser, A. C. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction*, 20(2), 147–187. <http://dx.doi.org/10.1007/s11257-010-9074-4>
- D’Mello, S., & Graesser, A. C. (2012). Language and discourse are powerful signals of student emotions during tutoring. *IEEE Transactions on Learning Technologies*, 5(4), 304–317. <http://dx.doi.org/10.1109/TLT.2012.10>
- D’Mello, S., & Kory, J. (2012). Consistent but modest: A meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. *Proceedings of the 14th ACM International*

(2016). Language and discourse analysis with Coh-Metrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

- Conference on Multimodal Interaction, 31–38. New York: ACM. <http://dx.doi.org/10.1145/2388676.2388686>
- D’Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, 29, 153–170. <http://dx.doi.org/10.1016/j.learninstruc.2012.05.003>
- Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). fMRI item analysis in a theory of mind task. *NeuroImage*, 55(2), 705–712. <http://dx.doi.org/10.1016/j.neuroimage.2010.12.040>
- Dowell, N. M., Cade, W. L., Tausczik, Y. R., Pennebaker, J. W., & Graesser, A. C. (2014). What works: Creating adaptive and intelligent systems for collaborative learning support. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Proceedings of the 12th International Conference on Intelligent Tutoring Systems* (pp. 124–133). Berlin: Springer. http://dx.doi.org/10.1007/978-3-319-07221-0_15
- Dowell, N. M., Skrypnik, O., Joksimović, S., Graesser, A. C., Dawson, S., Gašević, S., ... Kovanović, V. (2015). Modeling learners’ social centrality and performance through language and discourse. In C. Romero & M. Pechenizkiy (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining* (pp. 250–257). International Educational Data Mining Society.
- Elfenbein, A. (2011). Research in text and the uses of Coh-Metrix. *Educational Researcher*, 40(5), 246–248. <http://dx.doi.org/10.3102/0013189X11414181>
- Gašević, D., Kovanović, V., Joksimović, S., & Siemens, G. (2014). Where is research on massive open online courses headed? A data analysis of the MOOC Research Initiative. *The International Review of Research in Open and Distributed Learning*, 15(5). <http://dx.doi.org/10.19173/irrodl.v15i5.1954>
- Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hove, UK: Psychology Press.
- Gernsbacher, M. A. (1997). Two decades of structure building. *Discourse Processes*, 23(3), 265–304. <http://dx.doi.org/10.1080/01638539709544994>
- Goldman, S. R., Braasch, J. L. G., Wiley, J., Graesser, A. C., & Brodowinska, K. (2012). Comprehending and learning from internet sources: Processing patterns of better and poorer learners. *Reading Research Quarterly*, 47(4), 356–381. <http://dx.doi.org/10.1002/RRQ.027>
- Goldman, S. R., Graesser, A. C., & van den Broek, P. W. (1999). *Narrative comprehension, causality, and coherence: Essays in honor of Tom Trabasso*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Graesser, A. C., Gernsbacher, M. A., & Goldman, S. (Eds.). (2003). *Handbook of discourse processes*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Graesser, A. C., Jeon, M., Yan, Y., & Cai, Z. (2007). Discourse cohesion in text and tutorial dialogue. *Information Design Journal*, 15(3), 199–213. <http://dx.doi.org/10.1075/idj.15.3.02gra>
- 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 & Cognition*, 33(7), 1235–1247. <http://dx.doi.org/10.3758/BF03193225>
- Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, 3(2), 371–398. <http://dx.doi.org/10.1111/j.1756-8765.2010.01081.x>

(2016). Language and discourse analysis with Coh-Metrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

- Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, 115(2), 210–229. <http://dx.doi.org/10.1086/678293>
- Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5), 223–234. <http://dx.doi.org/10.3102/0013189X11413260>
- Graesser, A. C., McNamara, D. S., Louwse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society*, 36(2), 193–202. <http://dx.doi.org/10.3758/BF03195564>
- Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology*, 48(1), 163–189. <http://dx.doi.org/10.1146/annurev.psych.48.1.163>
- Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101(3), 371–95. <http://dx.doi.org/10.1037/0033-295X.101.3.371>
- Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Routledge.
- Joksimović, S., Dowell, N. M., Skrypnik, O., Kovanović, V., Gašević, D., Dawson, S., & Graesser, A. C. (2015). How do you connect? Analysis of social capital accumulation in connectivist MOOCs. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge (LAK '15)*, 66–68. <http://dx.doi.org/10.1145/2723576.2723604>
- Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Pearson Prentice-Hall.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.
- Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly*, 10, 62–102.
- Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2013). *Handbook of latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.
- LASI. (2014). Learning Analytics Summer Institute (LASI), 30 June–2 July 2014, Cambridge, MA. <https://solaresearch.org/events/lasi/lasi2014/>
- Lehman, B., D’Mello, S., & Person, N. (2010). The intricate dance between cognition and emotion during expert tutoring. In V. Alevan, J. Kay, & J. Mostow (Eds.), *Intelligent Tutoring Systems* (pp. 1–10). Berlin/Heidelberg: Springer. http://dx.doi.org/10.1007/978-3-642-13437-1_1
- Lorch, R. F. (1998). Memory-based text processing: Assumptions and issues. *Discourse Processes*, 26(2–3), 213–221. <http://dx.doi.org/10.1080/01638539809545045>
- Louwse, M. (2001). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics*, 12, 291–315. <http://doi.org/10.1515/cogl.2002.005>
- Economist*. (2014, February 8). Massive open online forces: The rise of online instruction will upend the economics of higher education. Retrieved from <http://www.economist.com/news/finance-and-economics/21595901-rise-online-instruction-will-upend-economics-higher-education-massive>

(2016). Language and discourse analysis with Coh-Matrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

- McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4), 459–488. <http://dx.doi.org/10.1177/0265532207080767>
- McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Matrix*. Cambridge, MA: Cambridge University Press.
- McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, 22(3), 247–288. <http://dx.doi.org/10.1080/01638539609544975>
- McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, 14(1), 1–43. http://dx.doi.org/10.1207/s1532690xci1401_1
- McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Matrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47(4), 292–330. <http://dx.doi.org/10.1080/01638530902959943>
- McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 51, pp. 297–384). Salt Lake City, UT: Academic Press. [http://dx.doi.org/10.1016/S0079-7421\(09\)51009-2](http://dx.doi.org/10.1016/S0079-7421(09)51009-2)
- Millis, K. K., King, A., & Kim, H.-J. J. (2000). Updating situation models from descriptive texts: A test of the situational operator model. *Discourse Processes*, 30(3), 201–236. http://dx.doi.org/10.1207/S15326950dp3003_1
- 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(2), 121–152. <http://dx.doi.org/10.1080/01638530709336895>
- Pennebaker, J. W., Booth, R., & Francis, M. (2007). *LIWC2007: Linguistic inquiry and word count*. Austin, TX: liwc.net
- Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *The Behavioral and Brain Sciences*, 27(2), 169–190; discussion 190–226. <https://dx.doi.org/10.1017/S0140525X04000056>
- Rapp, D. N., van den Broek, P., McMaster, K. L., Kendeou, P., & Espin, C. A. (2007). Higher-order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific Studies of Reading*, 11(4), 289–312. <http://doi.org/10.1080/10888430701530417>
- Sanders, T. J. M., & Noordman, L. G. M. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes*, 29(1), 37–60. http://dx.doi.org/10.1207/S15326950dp2901_3
- Singer, M., Graesser, A. C., & Trabasso, T. (1994). Minimal or global inference during reading. *Journal of Memory and Language*, 33(4), 421–441. <http://dx.doi.org/10.1006/jmla.1994.1020>
- Singer, M., & Kintsch, W. (2001). Text retrieval: A theoretical exploration. *Discourse Processes*, 31(1), 27–59. http://dx.doi.org/10.1207/S15326950dp3101_2
- Snow, C. E. (2002). *Reading for understanding: Toward a research and development program in reading comprehension*. Santa Monica, CA: Rand Corporation.

(2016). Language and discourse analysis with Coh-Metrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), 72–95. <http://dx.doi.org/10.18608/jla.2016.33.5>

- Stahl, G. (2009). From individual representations to group cognition. In G. Stahl (Ed.), *Studying Virtual Math Teams* (pp. 57–73). New York: Springer. http://dx.doi.org/10.1007/978-1-4419-0228-3_5
- Stenner, A. J. (2006, October). *Measuring reading comprehension with the Lexile framework*. Paper presented at the California Comparability Symposium, Durham, NC.
- Tannen, D. (1982). *Spoken and written language: Exploring orality and literacy*. New York: ABLEX Publishing Corporation.
- van den Broek, P., Everson, M., Virtue, S., Sung, Y.-T., & Tzeng, Y. (2002). Comprehension and memory of science texts: Inferential processes and the construction of a mental representation. In O. Leon & A. C. Graesser (Eds.), *The psychology of science text comprehension*. Mahwah, NJ: Lawrence Erlbaum Associates.
- van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.
- Vygotsky, L. S. (1978). *Mind in Society*. Cambridge, MA: Harvard University Press.
- Wiley, J., Goldman, S. R., Graesser, A. C., Sanchez, C. A., Ash, I. K., & Hemmerich, J. A. (2009). Source evaluation, comprehension, and learning in internet science inquiry tasks. *American Educational Research Journal*, 46(4), 1060–1106. <http://dx.doi.org/10.3102/0002831209333183>
- Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, 6(5), 292–297. <http://dx.doi.org/10.1111/j.1467-9280.1995.tb00513.x>
- Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2), 162–185. <http://dx.doi.org/10.1037/0033-2909.123.2.162>