

Arthur C. Graesser, Moongee Jeon, Yan Yan & Zhiqiang Cai

Discourse cohesion in text and tutorial dialogue

Keywords: cohesion, discourse types, readability, software tools, automatic analysis

Discourse cohesion is presumably an important facilitator of comprehension when individuals read texts and hold conversations. This study investigated components of cohesion and language in different types of discourse about Newtonian physics: *A textbook, textoids* written by experimental psychologists, *naturalistic tutorial dialogue* between expert human tutors and college students, and *AutoTutor tutorial dialogue* between a computer tutor and students (AutoTutor is an animated pedagogical agent that helps students learn about physics by holding conversations in natural language). We analyzed the four types of discourse with Coh-Metrix, a software tool that measures discourse on different components of cohesion, language, and readability. The cohesion indices included co-reference, syntactic and semantic similarity, causal cohesion, incidence of cohesion signals (e.g., connectives, logical operators), and many other measures. Cohesion data were quite similar for the two forms of discourse in expository monologue (textbooks and textoids) and for the two types of tutorial dialogue (i.e., students interacting with human tutors and AutoTutor), but very different between the discourse of expository monologue and tutorial dialogue. Coh-Metrix was also able to detect subtle differences in the language and discourse of AutoTutor versus human tutoring.

There has been a dramatic increase in computer analyses of large text corpora during the last decade. This can partly be explained by revolutionary advances in computational linguistics (Jurafsky & Martin, 2000; Walker et al., 2003), discourse processes (Pickering & Garrod, 2004; Graesser et al., 2003), the representation of world knowledge (Lenat, 1995; Landauer et al., 2007), and corpus analyses (Biber et al., 1998). Because thousands of texts can be quickly accessed and analyzed on thousands of measures in a short amount of time, data mining is emerging as a standard methodology in a broad spectrum of fields.

Researchers at the University of Memphis have recently developed a system called Coh-Metrix (<http://cohmetrix.memphis.edu>, Graesser et al., 2004), a computational tool that produces measures of the linguistic and discourse characteristics of text (both printed texts and transcripts of oral discourse). The values on the Coh-Metrix measures can be used to investigate the cohesion of the explicit text and the coherence of the mental representation of the text. Our definition of *cohesion* consists of linguistic characteristics of the explicit text that play some role in connecting ideas in the text. *Coherence* includes characteristics of the text (i.e., aspects of cohesion) that are likely to contribute to the coherence of mental representations.

Researchers at the University of Memphis have also developed an intelligent tutoring system called AutoTutor (Graesser et al., 2005; Graesser, Lu et al., 2004). AutoTutor is a learning environment that tutors students by holding a conversation in natural language. AutoTutor tutors students in Newtonian qualitative physics, computer literacy, critical thinking, and other topics that exhibit explanations and verbal reasoning. The dialogue of AutoTutor is sufficiently stable that it holds its own in conversing with students for hours. It is also designed to mimic the discourse patterns of human tutors (Graesser, Person, & Magliano, 1995).

The purpose of the present study was to use the Coh-Metrix tool to analyze the components of cohesion and language in different types of discourse about Newtonian physics. We analyzed a sample of chapters in a *textbook*, *textoids* written by experimental psychologists, *naturalistic tutorial dialogues* between expert human tutors and college students, and *AutoTutor tutorial dialogues* between a computer tutor and students. One strong virtue of this study was our attempt to achieve information equivalence with respect to the content covered in the four corpora. We did this by filtering content that covered the same set of core constructs in physics, namely Newtonian laws of force and motion. Given this control over information equivalence, we investigated how the four types of discourse differ with respect to language and cohesion. We might expect the two types of expository monologues (textbook and textoids) to be different from the two types of interactive dialogues (human tutors and AutoTutor). We might also expect differences between the two monologue types or between the two dialogue types. For example, textbook and textoids presumably differ from each other because textbooks are written by professional writers, whereas textoids are generated by experimental psychologists to satisfy methodological constraints. What is less clear is the nature of these

differences. If the Coh-Metrix tool is valid and useful, it should detect subtle and explainable differences in the four types of discourse. The present study investigated whether this is indeed the case.

Coh-Metrix

There are approximately 60 indices in the Coh-Metrix version (v. 2.0) that is available to the public. After the user of Coh-Metrix enters a text into the Web site, it prints out measures of the text on indices that span different levels of discourse and language. Coh-Metrix was designed to move beyond standard readability formulas, such as Flesch-Kincaid Grade Level (Klare, 1974–1975). Such formulas rely exclusively on word length and sentence length. For example, in the Flesch-Kincaid Grade Level index (Figure 1) *words* refers to mean number of words per sentence and *syllables* refers to mean number of syllables per word.

$$(1) \text{ Grade Level} = .39 * \text{Words} + 11.8 * \text{Syllables} - 15.59$$

Sentence length and word length do in fact robustly predict reading time (Haberlandt & Graesser, 1985), but certainly there is more to reading difficulty than word and sentence length. There must also be important deeper measures of language and cohesion. Coh-Metrix aims to provide these deeper measures.

The Coh-Metrix *indices* (i.e., measures, metrics) cover multiple levels of language and discourse. Some indices refer to characteristics of individual words, as has been achieved in many other computer facilities, such as WordNet (Fellbaum, 1998) and Linguistic Inquiry Word Count (Pennebaker & Francis, 1999). However, the majority of the Coh-Metrix indices include deeper or more processing-intensive algorithms that analyze syntax, referential cohesion, semantic cohesion, and dimensions of the situation model. Coh-Metrix is the only computer facility available to the public for free

that analyzes language and discourse on a broad set of components at multiple levels.

A snapshot of the landscape of indices is provided in this section. Researchers at the University of Memphis have over 600 indices in their internal computer system, of which 60 are available on the public Web site (<http://cohmetrix.memphis.edu/>). This article focuses exclusively on the set of publicly available measures. Researchers at the University of Memphis have also evaluated the accuracy of the Coh-Metrix indices in over 60 published studies, which can be accessed at the public Web site. However, it is beyond the scope of this study to review the research in these assessments.

Word measures. Coh-Metrix measures words on a large number of characteristics, most of which will not be defined in this article (see the help system on the Web site <http://cohmetrix.memphis.edu/>). There are measures of word frequency in the English language, which is based on the CELEX lexicon (Baayen et al., 1993) and other similar lexicons. Coh-Metrix also distinguishes between content words (e.g., noun, main verb, adjective) and function words (e.g., prepositions, articles), based on standard part-of-speech categories that are accepted in the computational linguistics community.

Several word indices are directly relevant to cohesion, coherence, and comprehension difficulty. In particular, there are word classes that have the special function of connecting clauses and other constituents in the text (Halliday & Hasan, 1976; Louwerse, 2002; Sanders & Noordman, 2000). The categories of connectives in Coh-Metrix include additive (*also, moreover*), temporal (*and then, after, during*), causal (*because, so*), and logical operators (*therefore, if, and, or*). The additive, temporal, and causal connectives are subdivided into those that are positive (*also, because*) or negative (*but, however*). The word indices include negations (*not, n't*) that span different levels of constituent structure and various conditional expressions (*if, given*). Negations, conditional

expressions, and negative connectives are predicted to be affiliated with complex conceptualizations and rhetorical structures, such as counterfactuals, hypothetical worlds, multiple perspectives, qualifications, hedges, and argumentation. A higher incidence of these words should therefore predict text difficulty.

The *incidence* of each word class is computed as the number of occurrences per 1000 words. An incidence score is necessary for comparing texts of different sizes. A text with higher cohesion would have a higher incidence of word classes that connect constituents.

Syntax. Coh-Metrix analyzes sentence syntax with the assistance of a syntactic parser developed by Charniak (2000). The parser assigns part-of-speech categories to words and syntactic tree structures to sentences. Our evaluations of several parsers showed better performance of Charniak's parser than other major parsers when comparing the assigned structures to judgments of human experts (Hemphill et al., 2006). Coh-Metrix has several indices of syntactic complexity, two of which (the mean number of modifiers per noun-phrase, and the number of words before the main verb of the main clause) are reported in this article. The mean number of modifiers per noun-phrase is an index of the complexity of referencing expressions. For example, *very large accelerating objects* is a complex noun-phrase with 3 modifiers of the head noun *objects*. The number of words before the main verb of the main clause is an index of syntactic complexity because it places a burden on the working memory of the comprehender (Graesser, Cai, Louwerse, & Daniel, 2006).

Referential and semantic cohesion. Referential cohesion occurs when a noun, pronoun, or noun phrase refers to another constituent in the text. For example, in the sentence *As the earth orbits the sun, it exerts a force*, the word *it* refers to the word *earth* by virtue of a syntactic rule of pronoun assignment. A referring expression (E) is the noun, pronoun, or noun-phrase that refers to

another constituent (C). C is designated as the referent of E. In the example sentence, the word *it* is the referring expression E, whereas the referent C is the word *earth*. One form of co-reference that has been extensively studied is argument overlap (Kintsch & van Dijk, 1978). This occurs when a noun, pronoun, or noun-phrase in one sentence is a co-referent of a noun, pronoun, or noun-phrase in another sentence. The word “argument” is used in a special sense in this context, namely it is a contrast with predicates in propositional representations. The argument overlap index of Coh-Metrix currently considers exact matches of arguments between two sentences. The value of this metric, which varies from 0 to 1, is the proportion of adjacent sentence pairs that share a common argument in the form of an exact match.

Another form of co-reference is stem overlap, where a noun in one sentence has a similar morphological root (i.e., lemma) as a content word in another sentence. For example, consider the two sentences *As the earth orbits the sun, it exerts a force. The orbit is not perfectly round. Orbits* and *orbit* have common stems, so there is stem overlap, even though one is a main verb and the other a noun. The value of this metric is the proportion of adjacent sentence pairs that have a stem overlap.

Yet another form of co-reference is anaphoric pronominal co-reference. A pronoun (*he, hers, it*) in one sentence refers to a referent in another sentence. A pronoun can present a coherence problem when the comprehender does not know the referent of the pronoun. Pronouns often require a conversational or social context to resolve their referents, as opposed to their referring to other text constituents. Coh-Metrix computes the referents of pronouns on the basis of syntactic rules, semantic fit, and discourse pragmatics by some existing algorithms in computational linguistics (see Jurafsky & Martin, 2000; Lappin & Leass, 1994). The value of this metric is the proportion of adjacent sentence pairs in which the second sentence has a

pronoun that can be successfully linked to a constituent in the previous sentence by executing the pronoun assignment mechanisms.

In addition to referential cohesion indices, Coh-Metrix has indices that assess the extent to which the content of sentences, turns, or paragraphs is similar semantically or conceptually. Cohesion and coherence are predicted to increase as a function of similarity. Latent Semantic Analysis (LSA) is the primary method of computing similarity because it considers implicit knowledge. LSA is a mathematical, statistical technique for representing world knowledge, based on a large corpus of texts. The central intuition is that the meaning of a word is captured by the company of other words that surround it in naturalistic documents; two words have similar meanings to the extent that they share similar surrounding words. LSA uses a statistical technique called singular value decomposition to condense a very large corpus of texts to 100-500 statistical dimensions (Landauer et al., 2007). The conceptual similarity between any two text excerpts (e.g., word, clause, sentence, text) is computed as the geometric cosine between the values and weighted dimensions of the two text excerpts. The value of the cosine varies from 0 to 1. LSA-based cohesion was measured in two ways relevant to the present study: (1) LSA similarity between adjacent sentences and (2) LSA similarity between adjacent paragraphs.

Lexical diversity provides a simple, but less computationally expensive, approach to computing semantic cohesion of a text. The lexical diversity metric in Coh-Metrix is the type-token ratio score. This is the number of unique words in a text (i.e., types) divided by the overall number of words (i.e., tokens) in the text. A low value means there is a large amount of redundancy in the content words of a text. Cohesion and coherence should increase inversely with type-token ratio.

Situation model dimensions. Many aspects of a text

can contribute to the *situation model* (or mental model), which is the referential content or microworld of what a text is about (Graesser et al., 1994; Kintsch, 1998). Text comprehension researchers have investigated at least five dimensions of the situational model (Zwaan & Radvansky, 1998): causation, intentionality, time, space, and protagonists. A break in cohesion or coherence occurs when there is a discontinuity on one or more of these situation model dimensions. Whenever such discontinuities occur, it is important to have connectives, transitional phrases, adverbs, or other signaling devices that convey to the readers that there is a discontinuity; we refer to these different forms of signaling as *particles*. Cohesion is facilitated by particles that clarify and stitch together the actions, goals, events, and states conveyed in the text.

Coh-Matrix 2.0 analyzes the situation model dimension on causation, intentionality, space, and time, but not protagonists. There are many measures of the situation model, far too many to address in this article. The present study concentrated on three indices that measure cohesion on the dimensions of causality, intentionality, and temporality. For causal and intentional cohesion, Coh-Matrix computes the ratio of cohesion particles to the incidence of relevant referential content (i.e., main verbs that signal state changes, events, actions, and processes, as opposed to states). The ratio metric is essentially a conditionalized incidence of cohesion particles: Given the occurrence of relevant content (such as clauses with events or actions, but not states), what is the density of particles that stitch together the clauses? For example, the referential content for intentional information includes intentional actions performed by agents (as in stories, scripts, and common procedures); in contrast, the intentional cohesion particles would include infinitives and intentional connectives (*in order to, so that, by means of*). Similarly, the referential content for causation information includes various classes of events that are identified by change-of-state verbs and

other relevant classes of verbs in WordNet (Fellbaum, 1998); the causal particles are the causal connectives and other word classes that denote causal connections between constituents. In the case of temporal cohesion, Coh-Matrix computes the uniformity of the sequence of main verbs with respect to tense and aspect. The Coh-Matrix help facility is available at the Web site for more details.

AutoTutor

Student conversations with AutoTutor were one of the four types of discourse analyzed by Coh-Matrix. AutoTutor is a pedagogical agent that helps students learn by holding a conversation in natural language (Graesser et al., 2005; Graesser, Lu et al., 2004). The learning gains of AutoTutor have been assessed in the areas of computer literacy (Graesser, Lu et al., 2004) and Newtonian physics (VanLehn et al., 2007). AutoTutor increases learning by approximately one letter grade when compared to reading textbooks for an equivalent amount of time.

AutoTutor's dialogues are organized around difficult questions and problems that require reasoning and explanations in the answers. The example below is one of the challenging questions on Newtonian physics.

Physics question: If a lightweight car and a massive truck have a head-on collision, upon which vehicle is the impact force greater? Which vehicle undergoes the greater change in its motion, and why?

Such questions require the learner to construct approximately 3–7 sentences in an ideal answer and to exhibit reasoning in natural language. The dialogue for one challenging question typically requires 50–200 conversational turns between AutoTutor and the student. The deep learning expected to occur during this process is distributed over many turns.

The structure of the dialogue in both AutoTutor and human tutoring (Chi et al., 2001, Graesser et al., 1995;

Van Lehn et al., 2007) can be segregated into three levels or aspects: (1) expectation and misconception-tailored dialogue, (2) a five-step dialogue frame, and (3) composition of a conversational turn. These three levels can be automated and produce respectable tutorial dialogue.

Expectation and misconception tailored dialogue. This is the primary pedagogical method of scaffolding good student answers. Both AutoTutor and human tutors typically have a list of expectations (anticipated good answers) and a list of anticipated *misconceptions* associated with each main question. For example, expectation E and misconception M are relevant to the example physics problems.

- E: The magnitudes of the forces exerted by A and B on each other are equal.
- M: A lighter or smaller object exerts no force on a heavier or larger object.

AutoTutor guides the student in articulating the expectations through a number of dialogue moves: generic *pumps* (what else?) to get the student to do the talking, *hints*, and *prompts* for the student to fill in missing words. Hints and prompts are carefully selected by AutoTutor to produce content in the answers that fill in missing content words, phrases, and propositions. For example, a hint to get the student to articulate expectation E might be “What about the forces exerted by the vehicles on each other?”; this hint would ideally elicit the answer “The magnitudes of the forces are equal.” A prompt to get the student to say “equal” would be “What are the magnitudes of the forces of the two vehicles on each other?” AutoTutor adaptively selects those hints and prompts that fill missing constituents and thereby achieves pattern completion. For those students who cannot fill in the content of an expectation after multiple conversational turns, AutoTutor steps in as a last resort and simply expresses the expectation as an *assertion*. AutoTutor ends up generating a high proportion of

pumps and hints for articulate students with high knowledge, but a high proportion of prompts and assertions for low verbal, low knowledge students. The list of expectations is eventually covered after the multi-turn dialogue and the main question is scored as answered.

AutoTutor adapts to the learners in other ways than scaffolding them to articulate expectations. AutoTutor corrects misconceptions that periodically arise in the students’ talk. When the students articulate a misconception, AutoTutor acknowledges the error and corrects it. AutoTutor gives feedback to the students on their contributions in most conversational turns. AutoTutor gives short feedback on the quality of student contributions: positive (very good, bravo), negative (not quite, almost), or neutral (uh huh, okay). AutoTutor attempts to answer the students’ questions when they are asked. The answers to the questions are retrieved from glossaries or from paragraphs in textbooks via intelligent information retrieval.

Five-step dialogue frame. This dialogue frame is prevalent in human tutoring (Graesser et al., 1995; VanLehn et al., 2007) and is also implemented in AutoTutor. The five steps of the dialogue frame are: (1) Tutor asks main question, (2) student gives initial answer, (3) tutor gives short feedback on the quality of the student’s answer in #2, (4) tutor and student collaboratively interact via expectation and misconception tailored dialogue, and (5) tutor verifies that the student understands (e.g., *Do you understand?*)

Managing one conversational turn. Each turn of AutoTutor in the conversational dialogue has three information slots (constituents). The first slot of most turns is short feedback on the quality of the student’s last turn (i.e., positive, negative, or neutral). The second slot advances the coverage of the ideal answer with either pumps, hints, prompts for specific words, assertions, corrections of misconceptions, or answers to student questions. The third slot is a cue to the student for the floor to shift from

AutoTutor as the speaker to the student. For example, AutoTutor ends each turn with a question or a gesture to cue the learner to do the talking. Discourse markers (e.g., *and also, okay, well*) connect the utterances of these three slots of information within a turn.

The three levels of AutoTutor go a long way in simulating a human tutor. AutoTutor can keep the dialogue on track because it is always comparing what the student says to anticipated input (i.e., the expectations and misconceptions in the curriculum script). Pattern matching operations and pattern completion mechanisms drive the comparison. These matching and completion operations are based on latent semantic analysis (Landauer et al., 2007) and symbolic interpretation algorithms (Rus et al., 2006) that are beyond the scope of this article to address. AutoTutor cannot interpret student contributions that have no matches to content in the curriculum script. For example, AutoTutor cannot explore topic changes and tangents as students introduce them. However, available studies of naturalistic tutoring (Chi et al., 2001; Graesser et al., 1995) reveal that (a) human tutors rarely tolerate true mixed-initiative dialogue with student topic-changes that steer the conversation off course and (b) most students rarely change topics, ask questions, and spontaneously grab the conversational floor. Instead, it is the tutor that drives the dialogue and leads the dance.

Using Coh-Metrix to analyze four types of text on Newtonian physics

Coh-Metrix was used to analyze the language and discourse of four types of discourse on Newtonian physics: *Textbook chapters*, *textoids*, *AutoTutor tutorial dialogue*, and *naturalistic tutorial dialogue*. The textbook corpus was the first 8 chapters from Hewitt's 1998 textbook on *Conceptual Physics*. The textoid corpus was 12 physics passages prepared by van den Broek and his

colleagues for research on the cognitive processes that occur during science comprehension (Kendeou & van den Broek, in press). The AutoTutor corpus was dialogue transcripts from a published experiment conducted on 10 physics problems with 22 students (Experiment 1 of VanLehn et al., 2007); there were 213 conversations total because seven of them were incomplete. The human tutoring corpus included dialogue transcripts on the same 10 physics problems but with a different group of 16 students, also in Experiment 1 of VanLehn et al. (2007). The human tutors held conversations with the students through computer mediated conversation. That is, the student and tutor were in different rooms and interacted on computers. The five human tutors had Ph.D.s in physics and were highly trained in pedagogy. The students in both tutoring corpora were college students enrolled in a physics course.

The content of the four corpora were very similar in the sense that they covered Newtonian laws of force and motion. The goal was to achieve information equivalence in domain knowledge among the four types of discourse so that differences in language and discourse could be attributed to the types of texts (i.e., genres or registers).

One way of viewing our selection of the four text types is to cross two dimensions. One dimension distinguishes expository *monologues* that are designed to be read (textbook and textoids) from conversational *dialogues* (tutoring with humans and AutoTutor). The language of the former is expected to be more compact, literate, and structurally dense than the language of dialogue that has an affinity to the oral tradition (Biber, 1988; Tannen, 1982). Orthogonal to this monologue-dialogue dimension is a second dimension that contrasts natural and artificial discourse. Textbooks and human tutoring are natural, ecologically valid discourse samples, whereas textoids and AutoTutor interactions have some modicum of artificiality. The researchers attempted to make the textoids and AutoTutor interactions well-

structured and coherent, of course, but in truth they are constrained by a research agenda or computational algorithms. It is an empirical question how close the artificial discourse samples are to the naturalistic samples.

We used Coh-Metrix 2.0 to analyze the four corpora. Each conversational turn was treated as a paragraph in the analyses of the two tutoring corpora. Therefore, a turn in a dialogue is analogous to a paragraph in an expository monologue.

Means and standard deviations of the Coh-Metrix indices

Table 1 presents the means and standard deviations of the Coh-Metrix indices, segregated by the four corpora. Table 2 presents a follow-up analysis that segregates tutor turns from student turns within the two tutoring corpora. In order to assess whether the means significantly differ from each other, one can compute 95% confidence intervals around each mean. The general formula is $\text{Mean} \pm [1.96 * \text{SD} / \text{SQRT}(N)]$. For example, the mean number of negations per 1000 words in AutoTutor dialogue is 11.5 and the standard deviation (SD) is 6.2. The 95% confidence interval would be $11.5 \pm [1.96 * 6.2 / \text{SQRT}(213)]$. That is, scores between 10.6 and 12.4 are not significantly different from the mean of 11.5. It follows that the mean scores for negations in textbooks (7.9) and textoids (6.8) are clearly outside of the range for AutoTutor, whereas the mean score for human tutoring (10.6) is within the range for AutoTutor. The 95% confidence interval for the textbooks is 7.9 ± 1.2 , or 6.7 to 9.1; the negations in the textoids are within this range, but not the AutoTutor dialogues. The 95% confidence interval for textoids is 2.3 to 11.3, so the textbooks are within this range, whereas the AutoTutor dialogues are outside of the range.

T-tests can also be mathematically derived from the means and standard deviations in these tables. The set

of *t*-tests would support the following comparisons on the incidence of negations: AutoTutor = human tutoring > textbook = textoids. Our discussion of the data below do not explicitly report inferential statistics because the large number of statistical tests would be cumbersome. However, our claims are supported by statistical tests at $\alpha = .05$ without adjustments for alpha inflation from multiple tests.

Simple measures of texts

The top cluster of indices in Tables 1 and 2 present simple measures of texts. The expository monologues (textbook and textoids) had a higher Flesch-Kincaid grade level than the tutoring dialogues (AutoTutor and human tutoring). The higher grade level can be attributed to longer sentences in the monologues because there were small differences in syllables per word (recall that grade level is based on sentence length and word length). The monologues also had more sentences in the paragraphs than the tutoring sessions had sentences in the conversational turns. The flow of information in tutoring clearly has smaller packages of information (sentences, paragraphs) distributed over more turns (i.e., paragraphs) compared with the expository monologues that are designed to be read. Stated differently, tutoring is more *fragmented* and *distributed* than the discourse designed for print.

A more detailed analysis of the tutoring can be derived from the data in Table 2. The AutoTutor dialogue was more fragmented and distributed than the human tutoring. AutoTutor had comparatively more turns, fewer sentences per turn, and fewer words per sentence. For both types of tutoring, the contributions of the students were much shorter than the tutor. Most of the student turns were only one sentence with 8 or 9 words. It was the tutor who did most of the talking in both AutoTutor and human tutoring. Educational research-

Table 1. Means and standard deviations for the measures of Coh-Metrix by physics corpora

	Textbook	Textoids	AutoTutor tutoring	Human tutoring
SIMPLE MEASURES OF TEXTS				
Number of texts	8	12	213	160
Total number of words in the text	5967 (2333)	177 (15)	913 (369)	406 (215)
Total number of sentences in the text	329 (136)	14.4 (2.5)	104.6 (49.7)	36.9 (21.2)
Total number of paragraphs/turns in the text	76.6 (27.7)	3.9 (1.08)	46.5 (22)	15.5 (11.8)
Average words per sentence	18.2 (.86)	12.5 (1.97)	9.3 (1.73)	11.5 (2.83)
Average sentences per paragraph/turn	4.3 (.58)	4 (1.54)	2.3 (.17)	2.7 (.84)
Average syllables per word	1.51 (.03)	1.45 (.13)	1.45 (.06)	1.43 (.08)
Flesch-Kincaid Grade level (0-12)	9.3 (.46)	6.4 (1.55)	5.2 (.8)	5.8 (1.53)
WORD LEVEL				
Logarithm of frequency of content words	2.15 (.05)	2.3 (.14)	2.27 (.11)	2.31 (.16)
Incidence score of all connectives	69.3 (5.9)	71.1 (20.4)	59.9 (13.4)	71.5 (52.8)
Incidence of positive causal connectives	9.4 (2.2)	12.4 (11.8)	11.9 (5.5)	18.4 (16.7)
Incidence of negative causal connectives	1.35 (.52)	2.72 (4.35)	.53 (.8)	.36 (1.04)
Incidence of positive additive connectives	22 (2.8)	29.5 (11.1)	22.5 (8.7)	22.7 (21.2)
Incidence of negative additive connectives	10.7 (1.8)	6 (5.4)	3.9 (2.6)	6.1 (5.6)
Incidence of positive temporal connectives	10.4 (2.9)	10.6 (8.9)	14.6 (6)	12.3 (11.8)
Incidence of negative temporal connectives	.29 (.29)	.48 (1.66)	.1 (.38)	.56 (1.77)
Incidence of all logical operators (and +if+or+cond+neg)	38 (2.7)	32.6 (17.6)	34.6 (9.9)	36.9 (27.6)
Incidence of conditionals in the text	5.19 (.83)	1.33 (2.41)	4.82 (3.03)	5.38 (4.26)
Incidence of negations in the text	7.9 (1.7)	6.8 (7.9)	11.5 (6.2)	10.6 (8)
SYNTAX				
Words before main verb of main clause in sentences	5.21 (.35)	3.56 (1.11)	2.52 (.64)	2.76 (1.18)
Average number of modifiers per noun phrase	.93 (.07)	.75 (.17)	.87 (.07)	.86 (.13)
REFERENTIAL AND SEMANTIC COHESION				
Argument overlap of adjacent Sentences	.66 (.03)	.53 (.26)	.24 (.07)	.35 (.12)
Stem overlap of adjacent sentences	.64 (.03)	.49 (.25)	.23 (.07)	.3 (.11)
LSA cosine of adjacent sentence to sentence	.36 (.02)	.28 (.12)	.19 (.08)	.21 (.11)
LSA cosine of paragraph/turn to paragraph/turn	.48 (.07)	.45 (.18)	.32 (.09)	.26 (.16)
Anaphor pronominal coreference of adjacent sentences	.26 (.07)	.29 (.21)	.09 (.04)	.21 (.1)
Type-token ratio of all content words	.36 (.03)	.73 (.09)	.37 (.09)	.57 (.1)
SITUATION MODEL DIMENSIONS				
Causal cohesion: Causal particles divided by causal verbs	.33 (.08)	.26 (.24)	.26 (.12)	.4 (.22)
Intentional cohesion: Intentional particles / intentional actions	1.49 (.2)	1.04 (1.62)	.58 (.29)	.94 (.8)
Temporal cohesion: Tense and aspect repetition scores	.87 (.02)	.82 (.08)	.9 (.03)	.84 (.07)

Table 2. Means and standard deviations for students and tutor turns

	AutoTutor (Tutor Turns)	AutoTutor (Student Turns)	Human Tutoring (Tutor Turns)	Human Tutoring (Student Turns)
SIMPLE MEASURES OF TEXTS				
Number of texts	213	213	160	160
Total number of words in the text	725 (329)	190 (91)	309 (153)	102 (92)
Total number of sentences in the text	77.7 (38.8)	26.9 (11.7)	26.7 (14.1)	10.2 (7.9)
Total number of paragraphs/turns in the text	23.4 (11)	23.1 (11)	7.9 (5.9)	7.5 (5.9)
Average words per sentence	9.8 (1.77)	7.6 (3.04)	11.9 (3.04)	9.4 (5.13)
Average sentences per paragraph/turn	3.3 (.31)	1.2 (.2)	3.9 (1.38)	1.4 (.56)
Average syllables per word	1.44 (.06)	1.51 (.11)	1.44 (.07)	1.41 (.17)
Flesch-Kincaid Grade level (0-12)	5.2 (.84)	5.2 (1.48)	6.0 (1.51)	4.9 (2.84)
WORD LEVEL				
Logarithm of frequency of content words	2.29 (.11)	2.2 (.17)	2.3 (.13)	2.36 (.3)
Incidence score of all connectives	59.4 (14.2)	59 (21.9)	64.1 (19.7)	64.3 (57.2)
Incidence of positive causal connectives	10.1 (5.6)	17.6 (11.4)	15.4 (8)	19.3 (25.3)
Incidence of negative causal connectives	.61 (.99)	.19 (.95)	.29 (1.03)	.45 (1.86)
Incidence of positive additive connectives	22.9 (9.4)	19.8 (14.4)	20.5 (10.6)	18.1 (21.7)
Incidence of negative additive connectives	3.9 (2.4)	4 (6.5)	5.4 (5.1)	7.2 (10.9)
Incidence of positive temporal connectives	15.6 (6.7)	10.6 (10)	11.2 (6.8)	11 (18)
Incidence of negative temporal connectives	.0 (.0)	.43 (1.87)	.29 (1)	1.05 (4.75)
Incidence of all logical operators (and +if+or+cond+neg)	32.8 (11.2)	40.4 (16.6)	31.1 (12.2)	44.6 (39.8)
Incidence of conditionals in the text	5.16 (3.01)	3.27 (5.91)	5.36 (4.55)	3.76 (7.28)
Incidence of negations in the text	9.8 (6.9)	18.4 (12.5)	7.5 (5.8)	24.5 (35.7)
SYNTAX				
Words before main verb of main clause in sentences	2.72 (.72)	1.92 (1.23)	2.91 (1.43)	2.06 (1.91)
Average number of modifiers per noun phrase	.87 (.07)	.85 (.18)	.86 (.14)	.68 (.32)
REFERENTIAL AND SEMANTIC COHESION				
Argument overlap of adjacent Sentences	.24 (.07)	.24 (.15)	.37 (.14)	.28 (.25)
Stem overlap of adjacent sentences	.21 (.06)	.25 (.16)	.31 (.14)	.23 (.25)
LSA cosine of adjacent sentence to sentence	.15 (.08)	.23 (.09)	.19 (.09)	.23 (.15)
LSA cosine of paragraph/turn to paragraph/turn	.44 (.11)	.20 (.08)	.34 (.16)	.25 (.17)
Anaphor pronominal co-reference of adjacent sentences	.09 (.05)	.10 (.08)	.23 (.12)	.17 (.19)
Type-token ratio of all content words	.42 (.12)	.49 (.12)	.62 (.1)	.76 (.16)
SITUATION MODEL DIMENSIONS				
Causal cohesion: Causal particles divided by causal verbs	.21 (.11)	.47 (.7)	.36 (.23)	.39 (.41)
Intentional cohesion: Intentional particles / intentional actions	.53 (.27)	.84 (1.25)	.82 (.74)	.78 (1.12)
Temporal cohesion: Tense and aspect repetition scores	.88 (.04)	.95 (.05)	.81 (.09)	.9 (.11)

ers encourage active learning on the part of the student, with attempts to get the student to do the talk and action. However, this is a challenge even in one-on-one tutoring.

Word-level indices

The words in the four types of discourse did not appreciably vary in word frequency but differences did emerge in connectives, conditionals, and negations. The overall incidence of connectives was lower in AutoTutor than the other three genres, which were approximately the same. The distribution of subcategories of connectives differed among the discourse types. However, it is difficult to discern any simple picture from the data.

Differences that emerged between AutoTutor and human tutoring appear in Table 2. The students learning from AutoTutor had fewer negative connectives, negations, logical operators, and conditional expressions, but approximately the same number of positive causal, additive, and temporal connectives. Although this suggests that the human tutor extracted more complex analytical content from the students than did AutoTutor, the distribution of word categories was very similar for the tutor turns in AutoTutor and the human tutor. This supports the claim that the automated tutor did a reasonable job simulating the human tutors, at least from the perspective of the distribution of word categories.

Syntax

The syntactic composition of sentences systematically differed among the four types of discourse. The expository monologues had more complex syntax than the tutoring dialogues when we examined the mean number of words before the main verb of the main clause, which reflects a greater load on working memory. The textbook clearly had the highest score on this syntactic index. Within the tutoring discourse, the student contributions

were not different on this index for AutoTutor versus human tutors; the tutor contributions were also not significantly different for AutoTutor versus human tutors. The other measure of syntactic complexity, namely the mean number of modifiers per noun-phrase, was not remarkably different among the discourse types.

Referential and semantic cohesion

Referential and semantic cohesion was consistently higher for the expository monologues than the tutoring dialogues when we examined argument overlap, stem overlap, and LSA scores. Therefore, in addition to tutoring being more fragmented and distributed, the discourse also has lower cohesion on these referential and semantic indices. When the two types of expository monologues were compared, cohesion was higher for the textbooks than the textoids. Unfortunately, it is difficult to interpret the anaphor pronominal reference index because the measure confounds the incidence of pronouns and the likelihood that the pronoun referent can be resolved; additional analyses will need to be conducted to differentiate these two components. Nevertheless, the scores were higher for expository monologues than the tutoring dialogues.

The type-token index of lexical diversity showed extremely high scores for the textoids, followed by human tutoring, and the lowest for the textbooks and AutoTutor. Therefore, there is more redundancy in the content words of the textbook and AutoTutor. The textoids are very packed with information. Apparently experimental psychologists fill their stimulus texts with a large amount of new information, much more than the professional writers of textbooks.

In-depth analyses of the tutoring sessions uncovered a few informative results. The student contributions in AutoTutor had lower argument overlap, LSA-turn similarity, anaphor pronominal reference, and lexi-

cal diversity, whereas stem overlap and LSA-sentence similarity scores were the same. The tutor contributions in AutoTutor had lower argument overlap, stem overlap, LSA-sentence similarity, anaphor pronominal reference, and lexical diversity than in human tutoring, but higher LSA-turn similarity scores. It is interesting that students' scores in referential and semantic cohesion tended to match that of the tutors. The low lexical diversity in AutoTutor can be explained by the fact that the content covered in AutoTutor was more narrow and constrained by the information in the curriculum scripts. Human tutors tended to cover topics beyond the realm of the curriculum.

Situation model dimensions

The three dimensions of the situation model we measured were causal, intentional, and temporal cohesion. The expository monologues had higher intentional cohesion than the tutoring dialogues, but such differences did not occur for causal and temporal cohesion. In-depth analyses of the tutoring discourse revealed two differences. First, the student contributions had higher cohesion on all three dimensions of the situation model for AutoTutor than for human tutoring sessions. Second, the tutor contributions of AutoTutor had lower causal and intentional cohesion, but higher temporal cohesion. There was no style matching of students to tutor discourse when situation model dimensions were analyzed. This result is rather different from our analyses of referential and semantic cohesion.

Discussion

The Coh-Metrix analyses have uncovered a large number of differences among the four types of physics discourse. The rich set of findings supports two general claims. First, the categories of text differ quite a bit on aspects of

language and discourse. Second, Coh-Metrix is sufficiently sensitive to uncover many of these differences.

We can offer several conclusions from our data analyses conducted on the four text categories. Consider first the contrast between the two types of expository monologues (which are designed to be read) and the two types of tutoring dialogues. Compared to the tutoring discourse, the two expository monologues tended to be less fragmented, less distributed, have more complex sentence syntax, have higher referential and semantic cohesion, and higher intentional cohesion. Some of these differences are compatible with the reported differences between print and oral language that were identified in the early 1980s (see Tannen, 1982). However, we are uncertain about the precise cause of the differences (see Biber, 1988). Is it the difference between the structure of monologue and dialogue? Is it the amount of interactivity between speech participants? Is it the difference between carefully composed printed discourse that is designed to be read versus speech acts composed on-the-fly in socially constrained conversations? Is it the grain size of messages that get composed, i.e., lengthy documents versus single turns? Answers to such questions await future research.

There were notable differences between the textbook and the textoids written by an experimental psychologist. It is important to be tentative, however, because our sample was limited to a single textbook and a single laboratory investigating textoids. Nevertheless, based on this limited sample, the textoids had shorter sentences, a lower grade level, less complex syntax, lower referential and semantic cohesion, a much higher type-token ratio, and lower cohesion on the three situation model dimensions (causality, intentionality, and temporality). Such differences suggest that there is some value in behavioral scientists using Coh-Metrix to analyze their stimulus materials. For example, discourse psychologists and education researchers are encouraged to analyze their

text stimuli to see whether the language and discourse characteristics are aligned with naturalistic texts on similar topics.

Coh-Metrix uncovered subtle differences in the discourse characteristics of AutoTutor and human tutoring. When we consider tutor contributions (as opposed to student contributions), AutoTutor and the human tutors had quite similar profiles of language and discourse characteristics. We would of course want this if AutoTutor were expected to simulate a human tutor. But some differences did emerge. AutoTutor had many more turns, fewer sentences per turn, fewer words per sentence, fewer connectives in some categories, lower sentence-to-sentence referential and semantic cohesion, higher turn-to-turn LSA cohesion, lower lexical diversity (type-token ratio), lower causal and intentional cohesion, and higher temporal cohesion. Therefore, AutoTutor was more fragmented, more distributed and had less cohesion than the human tutors. However, the curriculum script of AutoTutor was more narrow, so the lexical diversity was lower (more redundant) and the turn-to-turn LSA similarity scores were higher. Such differences had repercussions on the turns of students. The students of AutoTutor had fewer sentences per turn, shorter sentences, fewer negations, fewer connectives in some categories, lower lexical diversity, and somewhat higher cohesion on the three dimensions of the situation model.

We are currently developing a number of other measures of cohesion, but have not yet analyzed these sufficiently to release them to the public. For example, we are in the process of developing indices on genre uniformity and on contrasts between given and new information. Biber (1988) conducted a factor analysis on a corpus of texts on the basis of 67 features of language and discourse. We have automated nearly all of these features so Coh-Metrix can compute the extent to which a text fits different genres (such as narrative, science,

vs. history texts). However, Coh-Metrix incorporates many other characteristics of text and other algorithms that Biber never considered when he performed his analysis of genre nearly 20 years ago. Another analysis contrasts *new* information from *given* (old) information. The traditional approach segregates constituents that are introduced for the first time in the text from references to previous text constituents (Prince, 1981). Previous analytical treatments of the given-new distinction have been compositional and symbolic, whereas we are exploring LSA-based algorithms that quantitatively segregate the amount of new versus old information in a sentence automatically as sentences are received incrementally in a text.

This article provides only a glimpse of the analyses that can be conducted on texts with Coh-Metrix. There is an open frontier of questions to explore now that Coh-Metrix and other computer systems can automatically and quickly analyze texts in large corpora. We invite our colleagues to use Coh-Metrix and explore some new frontiers.

Acknowledgements

The research on AutoTutor and Coh-Metrix was supported by the National Science Foundation (SBR 9720314, REC 0106965, REC 0126265, IIS 0416128, ITR 0325428, REESE 0633918), the Institute of Education Sciences (R305H050169, R3056020018-02), and the Department of Defense Multidisciplinary University Research Initiative (MURI) administered by the Office of Naval Research under grant N00014-00-1-0600. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, IES, DoD, or ONR.

References

- Baayen, R. H., Piepenbrock, R., & Rijn, H. van (Eds.) (1993). *The CELEX Lexical Database* (CD-ROM). University of Pennsylvania, Philadelphia (PA): Linguistic Data Consortium.
- Biber, D. (1988). *Variations across speech and writing*. Cambridge, MA: Cambridge University Press.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.
- Charniak, E. (2000). A maximum-entropy-inspired parser. *Proceedings of the First Conference on North American Chapter of the Association For Computational Linguistics* (pp. 132-139). San Francisco, CA: Morgan Kaufmann Publishers.
- Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, 25, 471-533.
- Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.
- Graesser, A. C., Cai, Z., Louwerse, M., & Daniel, F. (2006). Question Understanding Aid (QUAID): A web facility that helps survey methodologists improve the comprehensibility of questions. *Public Opinion Quarterly*, 70, 3-22.
- Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions in Education*, 48, 612-618.
- Graesser, A. C., Gernsbacher, M. A., & Goldman, S. (Eds.). (2003). *Handbook of discourse processes*. Mahwah, NJ: Erlbaum.
- Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H., Ventura, M., Olney, A., & Louwerse, M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers*, 36, 180-193.
- Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, 36, 193-202.
- Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, 9, 359-387.
- Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101, 371-395.
- Haberlandt, K., & Graesser, A. C. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General*, 114, 357-374.
- Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.
- Hempelmann, C. F., Rus, V., Graesser, A. C., & McNamara, D. D. (2006). Evaluating the state-of-the-art Treebank-style parsers for Coh-Metrix and other learning technology environments. *Natural Language Engineering*, 12, 131-144.
- Hewitt, P. G. (1998). *Conceptual Physics*. Menlo Park, CA: Addison-Wesley.
- Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice-Hall.
- Kendeou, P., & Broek, P. van den (In Press). The role of readers' misconceptions on comprehension of scientific text. *Journal of Educational Psychology*.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.
- Kintsch, W., & Dijk, T. A. van (1978). Toward a model of text comprehension and production. *Psychological Review*, 85, 363-394.
- Klare, G. R. (1974-1975). Assessing readability. *Reading Research Quarterly*, 10, 62-102.
- Landauer, T., McNamara, D., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Erlbaum.
- Lappin, S., & Leass, H. J. (1994). An algorithm for pronominal coreference resolution. *Computational Linguistics*, 20, 535-561.
- Lenat, D. B. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38, 33-38.
- Louwerse, M. M. (2002). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics*, 291-315.
- Pennebaker, J. W., & Francis, M. E. (1999). *Linguistic inquiry and word count (LIWC)*. Mahwah, NJ: Erlbaum.
- Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27, 169-190.
- Prince, E. (1981). Towards a taxonomy of given-new information. In P. Cole (Ed.), *Radical Pragmatics* (pp. 223-255). New York: Academic Press.
- Rus, V., McCarthy, P. M., & Graesser, A. C. (2006). Analysis of a text entailment. In A. Gelbukh (Ed.), *Lecture notes in computer science:*

Computational linguistics in intelligent text processing: 7th international conference (pp. 287-298). New York: Springer Verlag.

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

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

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

Walker, M., Whittaker, S., Stent, A., Maloor, P., Moore, J., Johnson, M., & Vasireddy, G. (2003). Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science*, 28, 811-840.

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

About the authors

Art Graesser is a professor in the Department of Psychology and co-director of the Institute for Intelligent Systems at the University of Memphis. He received his Ph.D. in psychology from the University of California at San Diego. Graesser has worked in several areas of cognitive science, artificial intelligence, and discourse processing, including text comprehension, inference generation, conversation, question asking and answering, tutoring, and advanced learning environments. He is currently associate editor of the journal *Discourse Processes* (former senior editor) and was senior editor of the 2003 *Handbook of Discourse Processes*. He has designed, developed, and tested cutting edge software in learning, language, and discourse technologies, including AutoTutor, Coh-Metrix, HURA Advisor, SEEK Web Tutor, Question Understanding Aid (QUAID), QUEST, and Point&Query.

Moongee Jeon is a doctoral student in the Department of Psychology and the Institute for Intelligent Systems at the University of Memphis. His research interests are in discourse processing, text cohesion and coherence, computational linguistics, learning technologies, eye tracking, and tutoring. He has developed interactive simulation and dialogue facilities for AutoTutor, and designed and tested SEEK Web Tutor and Coh-Metrix.

Yan Yan's main area of interest includes Text Mining, Computational Linguistics, and Computational Geometry. She received her Bachelor degree in Computational Mathematics from Harbin Institute of Technology, China in 1999, Master degree in Computational Geometry from Fudan University, China in 2002, and Master degree in Computer Science from University of Memphis in 2006.

Zhiqiang Cai is now a computing specialist of the Institute for Intelligent Systems, the University of Memphis. He has a M. Sc. degree in computational mathematics received in 1985 from Huazhong University of Science and Technology, P. R. China. He is the chief software developer of Coh-Metrix and other systems.

Contact

Art Graesser
Psychology Department
202 Psychology Building
University of Memphis
Memphis, TN, 38152-3230
901-678-2146
901-678-2579 (fax)
e-mail: a-graesser@memphis.edu

Copyright of Information Design Journal & Document Design is the property of John Benjamins Publishing Co. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.