Running head: AUTOMATED TEXT ANALYSIS

Methods of Automated Text Analysis

Arthur C. Graesser, Danielle S. McNamara, and Max M. Louwerse

University of Memphis

Send correspondence to:

Art Graesser
Psychology Department
202 Psychology Building
University of Memphis
Memphis, TN, 38152-3230
901-678-2742
901-678-2579 (fax)
a-graesser@memphis.edu

Chapter in the Handbook of Reading Research, Volume IV, edited by Michael L. Kamil, P. David Pearson, Elizabeth B. Moje, and Peter Afflerbach (in press). Routledge/Erlbaum.

This chapter describes methods of analyzing the structures, functions, and representations of text. The lens is on the text, but we will selectively identify salient implications for cognitive processes and educational practice. The primary emphasis is also on automated methods of text analysis. That is, a computer system receives the text, performs processes that implement computational algorithms, and produces a text analysis on various levels of language and discourse structure. However, sometimes there are theoretical components of text that cannot be handled by computers so it is necessary to have human experts annotate or structure the text systematically. Human annotation schemes are identified in cases when it is beyond the scope of computer technologies to perform such text processing mechanisms.

A systematic analysis of the text is no doubt important to any comprehensive theory of reading and any application designed to improve reading in school systems. Reports on reading comprehension research acknowledged that there is a complex interaction among characteristics of the reader, the tasks that the reader is to perform, the socio-cultural context, and the properties text itself (National Reading Panel, 2000; McNamara, 2007; Snow, 2002). Investigations of such interactions are facilitated by precise measures of the properties of the text at various levels of analysis: words, sentences, paragraphs, entire texts. Readers may excel at some levels, but have deficits at others, so it is necessary to measure and record the various levels. Interventions to improve reading are expected to be precise on what characteristics of the text are being targeted in the intervention. From the standpoint of assessment, developers of high stake tests need to be specific on the characteristics of the text in addition to the cognitive processes that are being measured and aligned to reading standards.

This is a unique point in history because there is widespread access to computer tools that analyze texts at many levels of language and discourse. This increased use of automated text analysis tools can be attributed to landmark advances in such fields as computational linguistics (Jurafsky & Martin, 2000), discourse processes (Pickering & Garrod, 2004; Graesser, Gernsbacher, & Goldman, 2003), cognitive science (Lenat, 1995; Landauer, McNamara, Dennis, & Kintsch, 2007), and corpus linguistics (Biber, Conrad, & Reppen, 1998). Thousands of texts can be quickly accessed and analyzed on thousands of measures in a short amount of time.

The chapter is into divided four sections. The first section provides a brief historical background on theoretical approaches to analyzing texts between 1970 and 2000. The second section covers some current theoretical and methodological trends, most of which are interdisciplinary in scope. The third section describes how the text analysis systems are scored in assessments of accuracy and reliability. The fourth section identifies text analysis tools that assist researchers and practitioners. These tools span a large range of text units (words, clauses, propositions, sentences, paragraphs, lengthy documents) and levels of representation and structure (syntax, semantics, mental models, text cohesion, genre). Throughout the chapter we will point out some of the ways that automated text analyses have been put into practice, as in the case of intelligent tutoring systems that coach reading skills, feedback on student writing, and the selection of texts to match reader profiles.

Historical Background from 1970 to 2000

Reading researchers have always explored methods of analyzing the structures, functions, and representations of text (Ausubel, 1968; Gibson & Levin, 1975; Goldman

& Rakestraw, 2000; Williams, 2007), but there were some dramatic breakthroughs in the 1970s when the field of reading became more interdisciplinary. The fields of text linguistics, artificial intelligence, psychology, education, and sociology were particularly influential. Below are some of the landmark contributions that launched the 1970s.

(1) **Text linguistics**. Structural grammars that had originally been applied to phonology and sentence syntax were applied to meaning units in text and connected discourse (Van Dijk, 1972). In some analyses, texts were decomposed into basic units of meaning called propositions, which refer to events, actions, goals, and states that are organized in a hierarchical structure. Each proposition contains a predicate (e.g., main verb, adjective, connective) and one or more arguments (e.g., nouns, embedded propositions) that play a thematic role, such as agent, patient, object, time, or location. Below is an example sentence and its propositional meaning representation.

When the board met on Friday, they discovered they were bankrupt.

PROP 1: meet (AGENT=board, TIME = Friday)

PROP 2: discover (PATIENT=board, PROP 3)

PROP 3: bankrupt (OBJECT: corporation)

PROP 4: when (EVENT=PROP 1, EVENT=PROP 2)

The arguments are placed within the parentheses and have role labels, whereas the predicates are outside of the parentheses. The propositions, clauses, or other similar conceptual units are connected by principles of cohesion, many of which were identified by Halliday and Hasan (1976). Referential cohesion occurs when a noun, pronoun, or noun-phrase refers to another constituent in the text. For example, if the above example sentence were followed by “The meeting lasted several hours,” the noun-phase the meeting refers to PROP-1. Cohesion between propositions or clauses are often signaled by various forms of discourse markers, such as connectives (e.g., because, in order to, so that), adverbs (therefore, afterwards), and transitional phrases (on the other hand).

At a higher level of rhetorical structure, the propositions or other analogous conceptual units are organized into rhetorical structures that are affiliated with particular text genres. The rhetorical structure specifies the global organization of discourse, such as setting+plot+moral, problem+solution, compare-contrast, claim+evidence, question+answer, and argue+counter-argue (Meyer, 1975). Formal text grammars specify the elements and composition of these rhetorical patterns explicitly and precisely. For example, story grammars (Rumelhart, 1975; van Dijk, 1972) decompose simple stories in the oral tradition into structures that are captured by rewrite rules, such as:

Story → Setting+Plot+Moral

Setting → Location+Time+Characters

Plot → Complication+Resolution

Complication → Episode*

The specific goals of these compositional analyses in text linguistics were to segment texts into units, to assign the units to theoretical categories, and to organize the units into structures (typically hierarchical structures). Such a detailed decomposition was viewed as essential to any rigorous analysis of meaning, content, and discourse. Human experts were needed to segment, annotate, and structure these text representations because the theoretical distinctions were too complex or subtle for naïve coders to understand. However, it is hardly the case that the experts agreed on these structured representations when social scientists collected inter-judge agreement in the experts' judgments of segmentations, annotation, and structure. Inter-judge agreement was typically found to be significantly above chance, but modest, unless the experts had substantial training and feedback on their analyses to the point that they had similar analytical mindsets. Given that experts yielded imperfect agreement, automated computer analyses were similarly limited. At present, there are no computer programs that can translate texts into these structured text representations automatically, although there have been some attempts with modest success to automate propositional analyses (see section on text analysis tools).

(2) **Artificial intelligence (AI)**. Computer models in the 1970s attempted to interpret texts (Woods, 1977), generate inferences (Rieger, 1978; Schank, 1972), and comprehend simple, scripted narratives (Schank & Abelson, 1977). Most AI researchers were convinced that syntactic parsers and formal semantics would not go the distance in achieving natural language understanding because it is necessary to have world knowledge about people, objects, situations, and other dimensions of everyday experience. AI researchers identified packages of the generic world knowledge, such as person stereotypes, spatial frames, scripted activities, and schemas (Schank & Abelson, 1977). For example, scripts are generic representations of everyday activities (e.g., eating at a restaurant, washing clothes, playing baseball) that have actors with goals and roles, sequences of actions that are typically enacted to achieve these goals, spatial environments with objects and props, and so on. These scripts and other generic knowledge packages are activated during comprehension through pattern recognition processes and subsequently guide the course of comprehension by monitoring attention, generating inferences, formulating expectations, and interpreting explicit text.

AI researchers quickly learned that it was extremely difficult to program computers to comprehend text even when the systems were fortified with many different classes of world knowledge (Lehnert & Ringle, 1982). Modest successes were achieved when the texts were organized around a central script (Schank & Riesbeck, 1981) or when the computer achieved ahallow rather than deep comprehension (Lehnert, 1997). Shallow comprehension is sufficient to answer questions such as who, what, when, and where, which elicit a single word that is likely to be mentioned in a text. Deep comprehensions requires substantial inferences and answer questions such as why, how, what if, and so what. AI research in natural language comprehension eventually became transformed into a new field called computational linguistics. This new field systematically evaluated the accuracy of computer programs that processed language or discourse at particular levels.

(3) **Psychology and education**. Researchers in psychology and education empirically tested some of the theories in text linguistics and AI, as well as models of their own. They did this by collecting reading times, recall for text units, summarization protocols, answers to questions, ratings on various dimensions, and other data. Some researchers sampled naturalistic texts in their investigations whereas others prepared experimenter-constructed textoids that controlled for extraneous variables.

The results of the psychological research were quite illuminating on a number of fronts. For example, the number of propositions in a text predicted reading times (after controlling for the number of words), whereas recall for the text could be predicted by structural composition (Frederiksen, 1975; Haberlandt & Graesser, 1985; Kintsch, 1974). The distinction between given (old) and new information in a text predicted reading times for sentences and activation of inferences (Haviland & Clark, 1974). The structures generated by story grammars predicted recall and summaries of narrative text (Mandler & Johnson, 1977; Rumelhart, 1975). Recall memory for expository text also was systematically influenced by the text's rhetorical organization (Meyers, 1975). Memory for a vague or ambiguous text dramatically improved when there was a world knowledge schema that clarified and organized the sentences in the text (Bransford & Johnson, 1974). Indeed, reading times, memory, inferences, and other psychological processes were all facilitated by the scripts postulated by the AI researchers and the world knowledge schemas postulated by the social scientists (Anderson, Spiro, & Montague, 1977; Bower, Black, & Turner, 1979; Graesser, Gordon, & Sawyer, 1979; Spilich, Vesonder, Chiesi, & Voss, 1979).

The complexity of the psychological models grew as researchers collected more data in these psychological studies. Comprehension came to be viewed as a transaction between an author and a reader through the medium of a text, as opposed to being a bottom-up extraction of language codes and meaning (Rosenblatt, 1978). Interactive models of reading assumed there was a mixture of top-down and bottom-up processes among the multiple levels of language and discourse during the process of comprehension (Rumelhart & Ortony, 1977). These models were very different from the strictly bottom-up models (Gough, 1972).

(4) **Sociology, philosophy, and communication**. Researchers in these fields emphasized the social, pragmatic, and communication processes that underlie text and discourse. Speech acts were basic units of conversation that could be assigned to categories such as question, command, assertion, request, promise, and expressive

evaluation (Searle, 1969). These same categories were believed to play a functional role in printed text just as they do in oral conversation. Rommetveit (1974), for example, proposed that a printed text is a structured, pragmatic, social interaction between an author and reader. Patterns of turn-taking in conversation were explored by Sacks, Schegloff, and Jefferson (1974), whereas Grice (1975) identified the conversational postulates that underlie smooth conversation, including the cooperation principle and the maxims of quality, quantity, relation, and manner.

Some AI researchers attempted to capture the pragmatic foundations of speech acts in discourse in structured symbolic representations (Cohen & Perrault, 1979). Unfortunately, these pragmatic mechanisms have never been successfully automated on computers, except for very narrow applications (airline reservations, verifying train schedules; Allen, 1995) in which there are few alternative actions, goals, and people.

By the end of the 1970s a new multidisciplinary field of discourse processes was launched on the foundations of the above research traditions. The first publication of the journal Discourse Processes appeared in 1978, founded by Roy Freedle, a research scientist at Educational Testing Service in the fields of psychology and education. In 1990, the Society for Text and Discourse (ST&D) was founded as a society dedicated to investigating text and discourse processing through the lenses of the above disciplines. Discourse processing researchers continued to apply systematic methods of text analysis by human experts and computers. They continued to investigate the psychological processes that underlie the comprehension and production of discourse. However, the field changed in some fundamental ways that are addressed in the next section.

Current Theoretical and Methodological Trends

The theoretical metaphors in text analysis have shifted in recent years. There has been a shift from deep, detailed, structured representations of a small sample of texts to comparatively shallow, approximate, statistical representations of large text corpora. In the 1970s, text analysts were absorbed in identifying idea nodes in a complex web of knowledge structures. Today's researchers are inspecting high-dimensional semantic spaces that serve as statistical representations of document corpora as large as an encyclopedia or Wikipedia. This is the era when computers can analyze millions of words in thousands of documents in minutes or even seconds. Researchers are exploring computational models that specify how statistical patterns of words in documents map onto theoretical components of form and meaning. This section describes some new trends.

Corpus-based Statistical Representations

As we discussed in the previous section, researchers in psychology and education concluded that world knowledge has a large impact on reading and comprehension. Among the various classes of world knowledge are scripted activities, spatial frames, stereotypes about people, taxonomic hierarchies of plants and animals, the functioning of devices and artifacts, and so on. A psychological or computational model has to get a handle on how to represent the vast repertoire of world knowledge in the cognitive system. However, world knowledge has traditionally been an insurmountable challenge to text analysts because it is boundlessly large and semantically unruly. Decades of research in the cognitive sciences has converged on the view that the most forms of world knowledge, other than pure mathematics, are open-ended, imprecise, ill-defined, incomplete, and often vague.

Fortunately, however, the new statistical approach to representing world knowledge and the meaning of texts has provided an approximate solution to the problem of world knowledge. Two notable examples of statistical, corpus-based approaches to analyzing text meaning and world knowledge are Latent Semantic Analysis (Kintsch, 1998; Landauer & Dumais, 1997; Landauer, McNamara, Dennis, & Kintsch, 2007) and the Linguistic Inquiry Word Count (LIWC, Pennebaker, Booth, & Francis, 2007).

Latent Semantic Analysis (LSA) uses a statistical method called “singular value decomposition” (SVD) to reduce a large Word-by-Document co-occurrence matrix to approximately 100-500 functional dimensions. The Word-by-Document co-occurrence matrix is simply a record of the number of times word W_i occurs in document D_j. A document may be defined as a sentence, paragraph, or section of an article. Each word, sentence, paragraph, or entire document ends up being a weighted vector on the K dimensions. One important use of LSA is it provides a match score between any two texts (A and B) on the extent to which the texts are similar in meaning, relevance, or conceptual relatedness. The match score between two texts A and B (where a text can be either a single word, a sentence, or a larger text excerpt) is computed as a geometric cosine between the two vectors, with values ranging from approximately 0 to 1.

LSA-based technology is currently being used in a large number of learning technologies and applications in education (see Landauer et al., 2007). The Intelligent Essay Assessor grades essays as reliably as experts in English composition (Landauer, Laham, & Foltz, 2003); a similar achievement on essay grading in computational linguistics has been achieved by the E-rater at Educational Testing Service (Burstein, 2003). LSA is used in Summary Street to give feedback on students' summaries of texts that they read (E. Kintsch, Caccamise, Franzke, Johnson, & Dooley, 2007), in iSTART to give feedback on students' self-explanations on the text as they read (McNamara, Levinstein, & Boonthum, 2004), in RSAT to infer reading strategies from a reader's think aloud protocols (Millis et al., 2004), and in AutoTutor to guide tutorial dialogue as students work on problems by holding a conversation in natural language (Graesser, Lu et al., 2004; VanLehn, Graesser et al., 2007). LSA-based technologies are currently being assimilated in the textbook industry for retrieving documents and for giving students feedback on their writing. Because students get immediate feedback on their writing, they spend substantially more time revising their essays than they normally would if they had to wait days or weeks to receive feedback from their instructor. The accuracy of LSA technologies are often quite impressive and have been steadily improving, but are not perfect (either are humans, of course). The impact of these technologies on reading and writing quality will undoubtedly receive more attention in the future.

Pennebaker's LIWC accesses 70+ dictionaries that search for specific words or word stems in any individual or group of text files (Pennebaker, Booth, & Francis, 2007). Each dictionary in LIWC was constructed with the help of groups of judges who evaluated the degree to which each word is conceptually related to the broader category of which it is part. Most of the categories are psychologically-oriented (e.g., negative or positive emotions, intentions, traits, causality), whereas others are more standard linguistic categories (e.g., 1st person singular pronouns, prepositions, connectives). LIWC has been used to analyze a vast diversity of texts and has uncovered relations between these text analyses and psychological variables. For example, the narratives written by victims of traumatic events can predict how well they cope with the trauma and the number of times they visit a medical doctor (Pennebaker & Chung, 2007). Analyses of hundreds of thousands of text files reveal that people are remarkably consistent in their function word use across context and over time. For example, most everyone uses far more pronouns in informal settings than formal ones (Pennebaker & King, 1999) and are quite consistent in their writing over the course of their careers (Pennebaker & Stone, 2003). The distribution of function words in a person's writing is diagnostic of different social and personality characteristics, such as leadership, loneliness, deception, and inclinations toward suicide. LIWC has been developed in English, Arabic, Dutch, German, Hungarian, Italian, Korean, Norwegian, and Spanish at the time of this writing.

It is important to acknowledge that LSA and LIWC do not address the order of words in the text so they are hardly perfect representations of meaning. Word order is known to be an important aspect of comprehension. For example, the sequence of words are you here? conveys a question, whereas here you are is an assertion and you are here is possibly a command. John loves Mary has a different meaning than Mary loves John. Advances in computational linguistics (Allen, 1995; Jurafsky & Martin, 2000) have provided syntactic parsers and semantic modules that are more structured and sensitive to word order. Some of these systems are reported later in the section on text analysis tools. One brut-force approach to word order is an n-gram analysis that examines sequences of words of length n. For example, an n-gram analysis of length 3 considers word triplets. Landauer (2007) reported that essay grading is nearly perfect, i.e., equivalent to experts in English composition, with an algorithm that combines LSA and n-gram analyses.

It is quite apparent that computers can glean much from a text by merely inspecting the distribution of words, their co-occurrence with other words in a text, and their ordering. These statistical analyses of the new millennium are an order of magnitude more informative than the computer analyses that measure text difficulty with readability formulae. Most readability formulas rely exclusively on word length and sentence length, with occasional consideration of word frequency in the language. For example, Flesch-Kincaid Grade Level (Klare, 1974-5) considers only the number of words in a sentence and the number of syllables in a word in the readability formula. Sentence length and word length do in fact robustly predict reading time (Haberlandt & Graesser, 1985; Just & Carpenter, 1987; Rayner, 1998), but certainly there is more to text difficulty than word and sentence length. These other measures of text difficulty are addressed later when we discuss the Coh-Metrix system (Graesser, McNamara, Louwerse, & Cai, 2004).

Psychological models of text comprehension

Discourse psychologists have developed and tested a number of models that attempt to capture how humans comprehend text. Among these are the Collaborative Action-based Production System (CAPS) Reader model (Just & Carpenter, 1992), the Construction-Integration model (Kintsch, 1998), the constructivist model (Graesser, Singer, & Trabasso, 1994), the structure-building framework (Gernsbacher, 1990), the event indexing model (Zwaan & Radvansky, 1998), the landscape model (Van den Broek, Virtue, Everson, Tzeng, & Sung, 2002), and embodiment models (Glenberg, 1997; deVega, Glenberg, & Graesser, 2008). Many of these are complex processing models that combine symbolic representations and statistical representations that satisfy constraints at multiple levels of language and discourse. This subsection considers the CAPS/Reader and Construction-Integration (CI) model because these are the closest candidates to building a completely automated computer system that comprehends naturalistic text.

Just and Carpenter's (1992) CAPS/Reader model directs comprehension with a large set of production rules. The production rules play a variety of roles in the cognitive system, such as (a) scanning the words in the explicit text, (b) governing the operations of working memory, (c) changing activation values of information in working memory and long-term memory, and (d) performing other cognitive or behavioral actions. The production rules have an "IF <state>, THEN <action>" form and are probabilistic, with activation values and thresholds. For example, if the contents of working memory has some state S that is activated to a degree that meets or exceeds some threshold T, then action A is executed by spreading activation to one or more other information units in working memory, long-term memory, or response output. All of the production rules are evaluated in parallel within in each cycle of the production system, and multiple rules may get activated within each cycle. The researcher can therefore trace the activation of information units (i.e., word or proposition nodes) in the text, working memory, and long-term memory as a function of the cycles of production rules that get activated. Just and Carpenter have reported that these profiles of nodal activation can predict patterns of reading times for individual words, eye tracking behavior, and memory for text constituents. However, one drawback to the CAPS Reader model is that the researcher needs to formulate all of the production rules ahead of time.

Kintsch's (1998) CI model directs comprehension with a connectionist network. As text is read, sentence by sentence (or alternatively, clauses by clause), a set of word concepts and proposition nodes are activated (constructed). Some nodes match constituents in the explicit text whereas others are activated inferentially by world knowledge. The activation of each node fluctuates systematically during the course of comprehension, sentence by sentence. When any given sentence S (or clause) is comprehended, the set of activated nodes include (a) N explicit and inference nodes affiliated with S and (b) M nodes that are held over in working memory from the previous sentence by virtue of meeting some threshold of activation. The resulting N+M nodes are fully connected to each other in a weight space. The set of weights in the resulting (N+M) by (N+M) connectivity matrix specifies the extent to which each node activates or inhibits the activation of each of the other N+M nodes. The values of the weights in the connectivity matrix are theoretically motivated by multiple levels of language and discourse. For example, if two word nodes (A and B) are closely related in a syntactic parse, they would have a high positive weight; if two propositions contradict each other, they would have a high negative weight.

The dynamic process of comprehending sentence S has a two stage process, namely construction and integration. During construction, the N+M nodes are activated and then the connectivity matrix operates on these initial node activations in multiple activation cycles until there is a settling of the node activations to a new final stable activation profile. At that point, integration of the nodes has been achieved. Sentences that are more difficult to comprehend would require more cycles to settle. These dynamic processes have testable implications for psychological data. Reading times should be correlated with the number of cycles during integration. Recall of a node should be correlated with the number of cycles of activation. Inferences should be activated to the extent that they are activated and survive the integration phases. Kintsch (1998) summarizes substantial empirical evidence that support these and other predictions from the CI model.

The original CI model was not completely automated because researchers had to supply the world knowledge and the connectivity matrix that captures the language and discourse constraints. However, now that there has been substantial progress in the field of computational linguistics, it is possible to generate the weights in a principled fashion computationally by a computer. There are syntactic parsers (Charniak, 2000; Linn, 1998) that can assign syntactic tree structures to sentences automatically and these can be used to generate the syntactic connectivity matrix. Kintsch (1998) has used LSA to automatically activate concepts (near neighbors) from long-term memory that are associated with explicit words and to generate weights that connect the N+M nodes. One current technical limitation is that there is no reliable mechanism for translating language to propositions, an important functional unit in the CI model. However, there has been progress on this front, as will be discussed in the section on text analysis tools.

Annotation schemes in linguistics

The linguistics literature does not have a uniform method for representing text, but propositions are not normally the functional discourse units. Linguists have assumed, for example, that the functional discourse segments are clauses (Givón, 1983), conversational turns (Sacks, et al., 1974), sentences (Polanyi, 1988), prosodic units (Grosz & Hirschberg, 1992) or intentional units (Grosz and Sidner, 1986). Such differences in fundamental discourse units end up having an impact on the resulting taxonomies of discourse categories and coherence relations.

Two of the most frequently mentioned annotation schemes for discourse relations were developed by Hobbs (1985) and Mann and Thompson (1988). Hobbs (1985) integrates a theory of coherence relations within a larger knowledge-based theory of discourse interpretation. According to Hobbs, readers attempt to establish text coherence as they read. Coherence relations guide the readers' text building strategies but inferences are often needed to establish the coherence. Hobbs (1985) identified 9 coherence relations, such as occasion, evaluation, background, explanation, contrast, parallel, and elaboration.

Mann and Thompson's (1988) Rhetorical Structure Theory (RST) is similar to Hobbs but there are more relations and parts of it have been computationally implemented (Marcu, 2000). RST specifies the relations among text spans, regardless of whether or not they are marked by linguistic devices. RST assumes that relations in the text are between text spans, which are usually but necessarily identical to clauses. The text spans have variable size, ranging from two clauses to multi-sentence segments. RST proposes that a set of rhetorical relations tend to dominate in most texts, but the door is open for additional rhetorical relations that the writer needs. Mann and Thompson (1988) identified 23 rhetorical relations, including circumstance, solutionhood, elaboration, background, purpose, and non-volitional result. Thus, RST analysis starts by dividing the text into functional units that are called text spans. Two text spans form a nucleus and a satellite (Mann & Thompson, 1988); the nucleus is the part that is more essential to the writer's purpose than the satellite. Rhetorical relations are then composed between two non-overlapping text spans and form schemas. These schemas are rearranged into larger schema applications. The result of the analysis is a rhetorical structure tree, which is a hierarchical system of schema applications.

Scoring of Text Analysis Systems

A systematic scoring method is needed to assess the reliability of humans or computers in analyzing texts. There are two fundamental questions that guide these scoring systems: How similar are the text analyses of two or more humans? How similar are the analyses of humans and computers? Scoring procedures are needed to assess the segmentation of texts into theoretical units (such as propositions, clauses, sections, or text spans), the assignment of text units to theoretical categories (such as speech act categories, cohesion relations, or rhetorical categories), the structural relations between text units (such as relational links, connectives, or superordinate/subordinate relations), and the ratings of texts on quantitative dimensions (such as importance, quality, difficulty, or cohesion). We refer to these four procedures as segmentation, classification, linking, and rating, respectively.

Some of these scoring decisions are categorical (or qualitative) whereas others are continuous (or quantitative). An example categorical classification task would involve assigning sentences to one of several speech act categories (e.g., assertion, command, question, request, etc.). A Cohen's kappa score is normally used to assess the similarity of humans in their classification categories, or the similarity of humans and computers. The kappa scores vary from 0 to 1 and statistically adjust for the base rate likelihood that observations are in the various categories. It should be noted that percent agreement in the decisions of 2 or more parties (humans or computers) is inappropriate because of base rate problems and possibilities of inflating the score via highly skewed frequency distributions. An example quantitative rating task is the grading of essays on quality, with values ranging from 0 to 1. A Cronbach's alpha score is normally used to assess the similarity of ratings of humans or the similarity of humans and computers, although correlation coefficients (Pearson, Spearman) can also serve the sample purpose. These measures of agreement are quite familiar to researchers in education, psychology, and most other fields.

Researchers in computational linguistics use different quantitative methods of scoring agreement between human and computer decisions (Jurafsky & Martin, 2008). Human experts normally serve as the gold standard when measuring the performance of computers in making a decision or in assigning a text unit to category X. Recall and precision scores are routinely reported in the field of computational linguistics, as defined below.

Recall score is the proportion of computer decisions that agree with human decisions: $p(X_{computer} \mid X_{expert})$.

Precision score is the proportion of human decisions that agree with computer decisions: $p(X_{expert} \mid X_{computer})$.

An F-measure is a combined overall score (varying from 0 to 1) that takes both recall and precision into account. An alternative to these recall, precision, and F-measure scores is to perform signal detection analyses (Green & Swets, 1966), as defined below.

Hit rate = recall score

Miss rate = (1.0 - hit rate)

False alarm rate = $p(X_{computer} \mid \text{not } X_{expert})$

Correct rejection rate = (1.0 - false alarm rate)

An overall *d'* score is a measure in theoretical standard deviation units of the computer's discriminating the occurrence of X versus not-X, when using the human expert as the gold standard. The *d'* score is highest when the hit rate is 1 and the false alarm rate is 0.

The field of computational linguistics has benefited from some large-scale initiatives, funded by Department of Defense, that have systematically measured the performance of different systems developed by the computational linguists. These systems perform a variety of different useful functions, such as (a) accessing relevant documents from large document corpora (called information retrieval) and (b) extracting lexical, syntactic, semantic, or discourse information from text (called information extraction, or automated content extraction). The performance of information and content extraction systems has been assessed and reported in the Message Understanding Conferences (MUC, DARPA, 1995) and the Document Understanding Conferences (DUC), sponsored by the Association of Computational Linguistics. The National Institute of Standards and Technology (NIST) is a neutral party that selects the benchmark tasks, performance measures, and scheduling of the assessments. This assures that the performance of dozens of different systems can be evaluated and compared with fairness and objectivity. The different systems are compared quantitatively on various capabilities, although it is explicitly emphasized that the goal is to mark progress in the field as a whole rather than to hold competitions.

Comparisons of performance between a computer program and human experts (or between experts) are not expected to be perfect, but the question arises as to what level of performance is considered impressive, modest, or disappointing. How high should we expect the scores to be when examining kappa, alpha, recall, precision, or F-measures? There is no iron-clad, defensible answer to this question, but our labs regard kappa and F-measure scores of .70 or higher to be impressive, .30 to .69 to be modest, and .29 or lower to be disappointing. However, comparisons between computer and humans should be evaluated relative to the scores between two human judges. If two experts have a kappa score of only .40, then a kappa score between computer and expert of .36 would be impressive. We often consider the ratio of [kappa(computer, expert) / kappa (expert1, expert2)] to be the most relevant metric of the performance of the computer system (see Graesser, Cai, Louwerse, & Daniel, 2006), as long as the agreement between experts is modest or higher.

Text Analysis Tools

This section identifies computer tools that can be used to analyze texts on different levels of language and discourse. We focus here on completely automated text analysis systems, as opposed to the large array of tools in which the human and computer collaboratively annotate the text. We will start with the most conventional simple systems and end with the most complex systems that analyze text at global levels.

Conventional measures of text difficulty

Readability formulae. Readability formulae (Klare, 1974-5) have had a major influence on the textbook industry because they are routinely used as a standard for the selection of texts in K12 and college. For example, if the students are in the sixth grade, a textbook at the 3rd grade level would be considered too easy and a textbook at the 9th grade level would be considered too difficult for the students. A text that is scaled at the

5th through 7th grade level would be considered closer to the sixth graders' zone of proximal development.

Readability formulas have widespread use even though they rely exclusively on word length, sentence length, and sometimes word frequency. For example, the output of the Flesch Kincaid Reading Grade Level is specified in formulae 1.

$$\text{Flesch Reading Grade Level} = (.39 \text{ x } ASL + 11.8 \text{ x } ASW - 15.59 \qquad (1)$$

ASL refers to the average sentence length in the text whereas ASW is the average number of syllables per word. This simple metric is easy to score but ignores dozens of language and discourse components that are theoretically expected to influence comprehension difficulty, as will be described later. Texts tend to take more time to read when there are longer words and lengthier sentences (Haberlandt & Graesser, 1985; Just & Carpenter, 1987, 1992). Longer words tend to be less frequent in the language, as we know from Zipf's law (1949), and longer sentences tend to place more demands on working memory (Just & Carpenter, 1992). Nevertheless, there are a host of other variables that are also expected to influence reading difficulty.

There is also a potential risk of these readability formulas when they are mechanically applied to alter texts. It would not be wise to superficially shorten words and sentences in a text in order to have it fall under the rubric of an earlier grade level. For example, one could shorten words by substituting pronouns (it, this, that, he, she, we) for nouns, noun-phrases, or clauses. But that would functionally increase difficulty whenever the reader incorrectly binds the pronoun to a referent. One could shorten a sentence by chopping it up into shorter sentences. But that would functionally increase text difficulty whenever the reader has trouble conceptually relating the shorter sentences.

The textbook industry has on occasion shorted texts in this mechanical way and thereby made texts less coherent rather than easier to read.

Lexile and DRP. Measures such as Lexile (Stenner, 1996) and Degrees of Reading Power (DRP, B.I. Koslin, Zeno, & S. Koslin, 1987) capitalize on the predictive power of word and sentence length as automated signatures of word familiarity and sentence complexity. The precise formulae that compute reading difficulty of texts are not publically released, but the Lexile and DRP scores for texts are highly correlated with Flesch Kincaid Readibility scores (r > .90 in our analyses). Moreover, the metrics of text difficulty also take into consideration the students' comprehension performance at different ages. Students' comprehension ability can be measured using cloze tests, for example, in which the student fills in a missing word in a sentence, usually by choosing one of four possibilities. A cloze score of .75 means that the correct word is selected 75% of the time; this .75 benchmark may be adopted as an adequate threshold for comprehension in the Lexile analysis. The Lexile score measures both text difficulty and student ability in terms of the same unit, namely lexiles. A student is expected to comprehend 75% of a text, if the text has a Lexile score of 800 lexiles and the student's ability is estimated at 800 lexiles A students' expected comprehension of a text is a function of the difference between the difficulty of the text and the student's comprehension ability. One practical use of these measures is that they match readers to texts by providing automated text difficulty and comprehension ability indices, with the assumption that students can be encouraged to read texts that they understand at a specified level.

Some advantages of these types of reading programs are that they encourage students to read, the students can see their progression from level to level, and the predictive power of these types of formulae is relatively good. However, their predictive power largely stems from their predicting the same shallow level of comprehension as the level of difficulty measured in the text. Once again, we believe it is important to develop automated indices of text difficulty and comprehension that go beyond surface understanding and surface text characteristics by measuring deeper, more conceptual levels of comprehension.

Word-level measures

Hundreds of lexicons are available that analyze words on different dimensions of form, syntax, meaning, and psychological attributes. In this section we identify many of the word-based measures. Most of these measures can be accessed through Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara, Lowerse, & Graesser, 2007; http://cohmetrix.memphis.edu/cohmetrixpr/index.html), a facility on the Web that analyzes texts on characteristics of language and discourse.

Word frequency. Word frequency estimates the frequency of a word appearing in published documents in the real world, based on a designated corpus of texts with millions of words. One impressive estimate is CELEX, the database from the Dutch Centre for Lexical Information (Baayen, Piepenbrock, & Gulikers, 1995), based on a corpus of 17.9 million words. About 1 million of the word tokens are collected from spoken English (news transcripts and telephone conversations) whereas the remainder come from written corpora (newspapers and books). Some other well-known word frequency norms, all based on smaller corpora, are those of Francis and Kucera (1982) and Brown (1984).

Psychological dimensions of words. We have already described Pennebaker's Linguistic Inquiry Word Count (LIWC, Pennebaker & Francis, 1999), which includes psychological indices of words on such dimensions as positive and negative emotions, causality, and personality traits. Researchers had to rate or classify words in the lexicon on each of these psychological indices. Another important lexicon is the MRC Psycholinguistic Database (Coltheart, 1981), a collection of human ratings of 150,837 words along four psychological dimensions: meaningfulness, concreteness, imagability, familiarity, and age of acquiring the word. These ratings are based on work by Paivio, Yuille and Madigan, (1968), Toglia and Battig (1978), and Gilhooly and Logie (1980).

WordNet. WordNet® (Fellbaum, 1998; Miller, Beckwith, Fellbaum, Gross, & K. Miller, 1990) is an online lexicon whose design was inspired by cognitive science and psycholinguistics. English nouns, verbs, adjectives and adverbs are organized into semantic fields of underlying lexical concepts. For example, some words are functionally synonymous because they have the same or a very similar meaning. Polysemy is the number of senses of a word; for example, bank has one sense that is affiliated with money and another that is affiliated with rivers. A word's hypernym count is defined as the number of levels in a conceptual taxonomic hierarchy that is above (superordinate to) a word; table (as a concrete object) has 7 hypernym levels: seat -> furniture -> furnishings -> instrumentality -> artifact -> object -> entity. A word having many hypernym levels tends to be more concrete, whereas few hypernym levels is diagnostic of abstractness.

Parts of speech. There are over 70 part-of-speech (POS) tags derived from the Penn Treebank (Marcus, Santorini, & Marcinkiewicz, 1993). The tags include content words (e.g., nouns, verbs, adjectives, adverbs) and function words (e.g., prepositions, determiners, pronouns). Brill (1995) developed a POS tagger that automatically assigns a POS tag to each word. The Brill tagger assigns POS tags even to words that are not stored in its lexicon, based on the syntactic context of the other words in the sentence. Content words are more important than function words when it comes to identifying the background world knowledge for a text. However, the function words are also important for syntactic processing and are affiliated with the psychological profile of the writer (Pennebaker & Stone, 2003). Pronouns are particularly diagnostic of the writers' psychological states according to research by Pennebaker (Pennebaker & Chunk, 2007).

Syntax

Computational linguists have developed a large number of syntactic parsers that automatically assign syntactic tree structures to sentences (Jurafsy & Martin, 2008). Two popular contemporary parsers are Apple Pie (Sekine & Grishman, 1995) and the Charniak parser (2000). These parsers capture surface phrase-structure composition rather than deep structures, logical forms, or propositional representations. The Charniak (2000) parser generates a parse tree from an underlying formal grammar which can be induced from a corpus of texts via machine learning algorithms. The root of the tree, or highest level, divides the sentence into intermediate branches that specify nodes that include noun phrase (NP), verb phrase (VB), prepositional phrase (PP), and embedded sentence constituents. The tree terminates at leaf nodes, or words of the sentence that are labeled for their part of speech. Hempelman, Rus, Graesser, and McNamara (2006) evaluated the accuracy and speed of generating the parse trees for a number of syntactic parsers and concluded that the Charniak parser fared the best.

The syntactic complexity of sentences can be evaluated by two tools on the Web, namely Coh-Metrix (Graesser et al., 2004) and QUAID (Question Understanding Aid, Graesser et al., 2006). Syntactic complexity is assumed to increase with the degree to which sentences have embedded phrases, nodes that directly dominate many subnodes, and high working memory loads. One index is noun phrase density, calculated by taking the mean number of modifiers (e.g., adjectives) per noun phrase. A second index computes the extent to which there are embedded constituents by calculating the mean number of high-level constituents per word. A third index computes the number of words in the sentence that appear before the main verb of the main clause; as this number increases, the comprehender is expected to hold more words in working memory. Coh-Metrix also provides a metric of the extent to which sentences in the text have different syntactic structures, i.e., a form of syntactic diversity.

Propositions

Researchers in AI and computational linguistics have not been able to develop a computer program that can automatically and reliably translate sentences into a propositional representation or logical form, even in large-scale evaluations that aspire to such a goal (Rus, 2004; Rus, McCarthy, & Graesser, 2007). The assignment of noun-phrases to thematic roles (e.g., agent, recipient, object, location) is also well below 80% correct in the available computer systems (DARPA, 1995). However, progress will hopefully be made in future years through two avenues. First, there is a corpus of annotated propositional representations in PropBank (Palmer, Kingsbury, & Gildea, 2005), so researchers can work on developing and refining their algorithms for automatic proposition extraction. Second, a tool called AutoProp (Briner, McCarthy, & McNamara, 2007) has been designed to "propositionalize" texts that have already been reduced to clauses. This is a promising tool that might eventually achieve adequate performance.

Coreference and cohesion

Coh-Metrix (Graesser et al., 2004) was explicitly designed to analyze text cohesion and coherence by incorporating recent advances in computational linguistics. McNamara, Louwerse, and Graesser (2007) have reviewed over 40 studies that evaluated the accuracy of Coh-Metrix and also how it was tested in psychological experiments on text comprehension and memory. This subsection covers measures of Coh-Metrix, but it is beyond the scope of this chapter to review the research that has assessed its validity.

Coreference**.** Coreference is perhaps the most frequent definition of cohesion among researchers in discourse processing and linguistics (Britton & Gulgoz, 1991; Halliday & Hasan, 1976; Kintsch & van Dijk, 1978; McNamara, Kintsch, Songer, & Kintsch, 1996). As discussed earlier in the chapter, referential cohesion occurs when a noun, pronoun, or noun-phrase refers to another constituent in the text. There is a referential cohesion gap when the words in a sentence do not connect to words in surrounding text or sentences. Coh-Metrix tracks five major types of lexical coreference by computing overlap in nouns, pronouns, arguments, stems, and content words. Noun overlap is the proportion of all sentence pairs that share one or more common nouns, whereas pronoun overlap is the proportion of sentence pairs that share one or more pronoun. Argument overlap is the proportion of sentence pairs that share nouns or pronouns (e.g. table/table, he/he). Stem overlap is the proportion of sentence pairs in which a noun (or pronoun) in one sentence has the same semantic morpheme (called a lemma) in common with any word in any grammatical category in the other sentence (e.g. the noun photograph and the verb photographed). The fifth coreference index, content word overlap, is the proportion of content words that are the same between pairs of sentences. Some of these measures consider only pairs of adjacent sentences, whereas others consider all possible pairs of sentences in a paragraph.

Connectives**.** Connectives help increase the cohesion of a text by explicitly linking ideas at the clausal and sentential level (Britton & Gulgoz, 1991; Halliday & Hasan, 1976; Louwerse & Mitchell, 2003; McNamara et al., 1996; Sanders & Noordman, 2000). Most of the connectives in Coh-Metrix are defined according to the subcategories of cohesion identified by Halliday and Hasan (1976) and Louwerse (2001). These include connectives that correspond to additive cohesion (e.g., also, moreover, however, but), temporal cohesion (e.g., after, before, until), and causal/intentional cohesion (e.g., because, so, in order to). Logical operators (e.g., variants of or, and, not, and if–then) are also cohesive links that influence the analytical complexity of a text. The measures of connectives are computed as a relative frequency score, the number of instances of a category per 1000 words.

Cohesion of the situation model. An important level of text comprehension consists of constructing a 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 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-Metrix computes the ratio of cohesion particles to the incidence of relevant referential content; given the occurrence of relevant content (such as clauses with events or actions), what is density of particles that stitch together the clauses. In the case of temporality, Coh-Metrix computes a repetition score that tracks the consistency of tense (e.g., past and present) and aspect (perfective and progressive) across a passage of text. The repetition scores decrease as shifts in tense and aspect are encountered. A low particle-to-shift ratio is a symptom of problematic temporal cohesion.

Latent Semantic Analysis (LSA)**.** Coh-Metrix assesses conceptual overlap between sentences by LSA (Landauer et al., 2007), the corpus-based statistical representation that considers implicit knowledge. LSA-based cohesion was measured in several ways in Coh-Metrix, such as LSA similarity between adjacent sentences, LSA similarity between all possible pairs of sentences in a paragraph, and LSA similarity between adjacent paragraphs. The Coh-Metrix research team has also developed a tool with an LSA-based measure that automatically computes the relative amount of given versus new information that each sentence has in a text and then computes the average newness among all sentences in the text (Chafe, 1976; Halliday, 1967; Haviland & Clark, 1974; Prince, 1981). Hempelmann et al. (2005) reported that the span method has a high correlation with the theoretical analyses of give/new developed by Prince (1981), as well as other linguists who have analyzed the given-new distinction in discourse.

Genre

There are many types of discourse, or what some researchers call genre (category in French), conversational registers, or simply discourse categories (the expression we adopt here). There are prototypical discourse categories in the American culture, such as folktales, scientific journal articles, and news editorials. These three examples would funnel into more superordinate classes that might be labeled as narrative, exposition, and argumentation, respectively. Some texts will be hybrids, of course.

Biber (1988) conducted a very ambitious investigation of discourse categories . Biber used 23 spoken and written categories from the Lancaster-Oslo-Bergen (LOB) corpus and the London-Lund corpus, and computed the frequency of 67 linguistic features in these categories. The normalized frequencies of these features in each of the discourse categories were then entered in a factor analysis, from which six factors emerged. These factors can be seen as dimensions on which discourse categories can be placed. Biber's analysis showed that no single dimension comprised a difference between speech and writing, but there were the following relations or features among the texts: (1) Involved versus informational production, (2) narrative versus non-narrative, (3) explicit versus situation dependent reference, (4) overt expression of persuasion, (5) abstract versus non-abstract information, and (6) on-line informational elaboration. For example, categories such as romantic fiction, mystery fiction and science fiction were positioned high on the second dimension (narrative). In contrast, categories such as academic prose, official documents, hobbies, and broadcasts scored low (non-narrative).

Biber’s (1988) study and the multi-feature multidimensional approach have become a standard in corpus linguistics (McEnery, 2003) and have led to various extensions (Conrad & Biber, 2001), as well as to assessments of the validity, stability, and meaningfulness of the approach (Lee, 2004). An automated version of Biber’s system is available in his laboratory whereas the 67 features of language and discourse are automated in Coh-Metrix. Coh-Metrix also has other algorithms that significantly differentiate science, narrative, versus history texts (McCarthy, Myers, Briner, Graesser, & McNamara, in press).

Closing Comments

As we close the first decade of the new millennium we are confident that automated text analyses will continue to progress and lead to useful new applications. Few of our colleagues would have placed their bets 20 years ago on computer systems that would grade student essays as well as experts in composition, that would train students to read at deeper levels, or that would tutor students on science topics in natural language. However, there are now systems that achieve these practical goals, as we have pointed out in this chapter. Moreover, they are now being scaled up to the point of being used in school systems, the textbook industry, and eLearning. Of course, the use of these technologies throughout the world is at the early phase of adoption and it will take awhile before they are fully evaluated with respect to their impact on reading proficiency.

This chapter will close with two examples that should have a profound impact on reading researchers in the future. The first is that the textbook industry will eventually use these tools to improve the quality of textbooks. Existing textbooks in science and other academic topics are frequently not well written because they fail to consider the world knowledge of the reader and they have gaps in text cohesion. As discussed earlier, writers in the textbook industry run the risk of shortening words and sentences in order to minimize the text difficulty scores of the texts that are targeted for grades 1-4. The problem with shortening words by substituting pronouns for long content words is that it is sometimes difficult to ground pronouns in appropriate referents, so comprehension and coherence suffers. The problem with shortening the sentences is that there is a potential penalty in lowering the cohesion among the ideas expressed in sentences. The unfortunate consequence of these mechanical alterations is that students end up with an incoherent reading experience. Writers of textbooks of the future are expected to take a closer look at the automated text analysis tools as they prepare materials for school systems. Moreover, the text analysis systems need to consider the full range of levels of language and discourse – not merely word and sentence length.

The second example addresses the selection of texts for the learner. How can we optimize the assignment of the next text for the student to read? We presumably would not want to assign a text that is too easy or too difficult, but rather to assign a text that is at the reader’s zone of proximal development. The computer can play an important role in suggesting the ideal text for the reader at the right time and place. It would be important for the text to be relevant to the immediate curriculum and also to match the reader’s profile of world knowledge, comprehension skills, and interests. Technologies are currently available to make such assignments of texts to readers. However, more research and development is needed to refine these technologies and advance the science of reading comprehension mechanisms. Once again, however, the text analysis systems need to consider the full range of levels of language and discourse.

Author Notes

The research on was supported by the National Science Foundation (SBR 9720314, SES 9977969, REC 0106965, REC 0126265, ITR 0325428, REESE 0633918), the Institute of Education Sciences (R305G020018**,** R305H050169, R305B070349, R305A080589), and the Office of Naval Research (N00014-00-1-0917). 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, or ONR. Requests for reprints should be sent to Art Graesser, Department of Psychology, 202 Psychology Building, University of Memphis, Memphis, TN 38152-3230, a-graesser@memphis.edu.

References

Allen, J. (1995). Natural language understanding. Redwood City, CA: Benjamin/Cummings.

Anderson, R.C., Spiro, R.J., & Montague, W.E. (1977) (Eds.). Schooling and the acquisition of knowledge. Hillsdale, NJ: Erlbaum.

Ausubel, D. (1968). Educational psychology: A cognitive view. New York: Holt, Rinehart, and Winston.

Baayen, R. H., Piepenbrock, R., & Gulikers. L. (1995). The CELEX lexical database (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Bransford, J.D. & Johnson, M.K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. Journal of Verbal Learning and Verbal Behavior, 11, 717-726.

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.

Bower, G.H., Black, J.B., and Turner, T.J. (1979). Scripts in memory for text. Cognitive Psychology, 11, 177-220.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics, 21, 543-566.

Briner, S.W., McCarthy, P.M., McNamara, D.S. (2006). Automating text propositionalization: An assessment of AutoProp. In R. Sun & N. Miyake (Eds.), Proceedings of the 28th Annual Conference of the Cognitive Science Society (pp. 2449). Austin, TX: Cognitive Science Society.

Britton, B. K., & Gulgoz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. Journal of Educational Psychology, 83, 329-345.

Brown, G.D.A. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. Behavioral Research Methods Instrumentation and Computers, 16, 502-532

Burstein, J. (2003). The e-rater® scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), Automated essay scoring: A cross-disciplinary perspective. Hillsdale, NJ: Erlbaum.

Chafe, W.. (1976). Givenness, contrastiveness, definiteness, subjects, and topics. In C. Li (Ed.), Subject and Topic (pp. 25–76). New York: Academic Press.

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. Proceedings of the 14th National Conference on Artificial Intelligence, Menlo Park, CA: AAAI Press/MIT 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.

Clark, H.H. (1996). Using language. Cambridge: Cambridge University Press.

Cohen, P.R., & Perrault, C.R. (1979). Elements of a plan-based theory of speech acts. Cognitive Science, 3, 177-212.

Conrad, S. & Biber, D. (2001). Variation in English: Multi-dimensional studies. Harlow: Longman

Coulthard, M. (1981). The MRC Psycholinguistic Database. Quarterly Journal of Experimental Psychology, 33A, 497-505.

DARPA (1995). Proceedings of the Sixth Message Understanding Conference (MUC-6). San Francisco: Morgan Kaufman Publishers.

De Vega, M., Glenberg, A. M., & Graesser, A. C. (Eds.)(2008). Symbols and embodiment: Debates on meaning and cognition. Oxford, UK: Oxford University Press.

Fellbaum, C. (1998) (Ed.) WordNet: An electronic lexical database. Cambridge, MA: MIT Press.

Francis, W.N., & Kucera, N. (1982). Frequency analysis of English usage. Houghton-Mifflin.

Frederiksen, C.H. (1975). Representing logical and semantic structure of knowledge acquired from discourse. Cognitive Psychology, 7, 371-458.

Gernsbacher, M.A. (1990). Language comprehension as structure building. Hillsdale: Lawrence Erlbaum.

Gibson, E.J., & Levin, H. (1975). The psychology of reading. Cambridge, MA: MIT Press.

Gilhooly, K. J., & Logie, R. H. (1980). Age of acquisition, imagery, familiarity and ambiguity measures for 1944 words. Behavioral Research Methods and Instrumentation. 12**,** 395–427.

Givón, T. (1983). Topic continuity in discourse: An introduction. In T. Givón (Ed.), Topic continuity in discourse: Quantified cross-linguistic studies (pp. 347-63). Amsterdam: John Benjamins.

Glenberg, A.M. (1997). What memory is for. Behavioral and Brain Sciences, 20, 1-19.

Goldman, S.R., & Rakestraw, J.A. (2000). Structural aspects of constructing meaning from text. In M.L. Kamil, P.B. Mosenthal, P.D. Pearson, and R. Barr (Eds.), Handbook of Reading Research,Vol. III. Mahwah, NJ: Erlbaum.

Gough, P.B. (1972). One second of reading. In J.F. Kavanaugh and J.G. Mattingly Eds.), Language by ear and by eye (pp. 331-358). 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., Gernsbacher, M.A., & Goldman, S.R. (Eds.)(2003). Handbook of discourse processes. Mahwah, NJ: Erlbaum.

Graesser, A. C., Gordon, S. E., & Sawyer, J. D. (1979). Recognition memory for typical and atypical actions in scripted activities: Tests of a script pointer plus tag hypothesis. Journal of Verbal Learning and Verbal Behavior, 18, 319-322.

Graesser, A.C., Lu, S., Jackson, G.T., Mitchell, H., Ventura, M., Olney, A., & Louwerse, M.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. Behavioral Research Methods, Instruments, and Computers, 36, 193-202.

Graesser, A.C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. Psychological Review, 101, 371-95.

Green, D. M. & Swets, J. A.(1966). Signal detection theory and psychophysics. New York: Wiley.

Grice, H.P. (1975). Logic and conversation. In P. Coleand J.L. Morgan (Eds.), Syntax and semantics (Vol. 3): Speech acts (pp. 41-58). New York: Seminar Press.

Grosz, B. and J. Hirschberg (1992) Some intonational characteristics of discourse structure. In J. J. Ohala, T. M. Nearey, B. L. Derwing, M. M. Hodge and G. E. Wiebe (Eds.), Proceedings of the International Conference on Spoken Language Processing (pp. 429-432). Banff, Canada.

Grosz, B.J., & Sidner, C.L. (1986). Attention, intentions, and the structure of discourse. Computational Linguistics, 12 (3), 175-204.

Haberlandt, K. F., & Graesser, A. C. (1985). Component processes in text comprehension and some of their interactions. Journal of Experimental Psychology: General, 114, 357-374.

Halliday, M. (1967). Intonation and Grammar in British English. Mouton, The Hague.

Halliday, M.A.K. & Hasan, R. (1976). Cohesion in English. London: Longman.

Haviland, S.E., & Clark, H.H. (1974). What’s new? Acquiring new information as a process in comprehension. Journal of Verbal Learning and Verbal Behavior, 13, 515-521.

Hempelmann, C.F., Dufty, D., McCarthy, P., Graesser, A.C., Cai, Z., & McNamara, D.S. (2005). Using LSA to automatically identify givenness and newness of noun-phrases in written discourse. In B. Bara (Ed.), Proceedings of the 27th Annual Meetings of the Cognitive Science Society. (pp. 941-946). Mahwah, NJ: Erlbaum.

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.

Hobbs, J.R. (1985). On the coherence and structure of discourse. Technical Report CSLI-85-37, Center for the Study of Language and Information, Stanford University.

Koslin, B.I., Zeno, S., & Koslin, S. (1987). The DRP: An effective measure in reading. New York: College Entrance Examination Board.

Jurafsky, D., & Martin, J.H. (2008). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, NJ: Prentice-Hall.

Just, M. A., & Carpenter, P. A. (1987). The psychology of reading and language comprehension. Boston: Allyn & Bacon.

Just M.A., & Carpenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. Psychological Review, 99, 122-149.

Kintsch, E., Coccamise, D., Franzke, M., Johnson, N., & Dooley, S. (2007). Summary Street: Computer-guded summary writing. In T. Landauer, D.S. McNamara, S. Dennis, and W. Kintsch (Eds.), Handbook of Latent Semantic Analysis (pp. 263-278). Mahwah, NJ: Erlbaum.

Kintsch, W. (1974). The representation of meaning in memory. Hillsdale, NJ: Erlbaum.

Kintsch, W. (1998). Comprehension: A paradigm for cognition. Cambridge, UK: Cambridge University Press.

Kintsch, W., & Van Dijk, T.A. (1978). Toward a model of text comprehension and production. Psychological Review, 85, 363-394.

Klare, G.R. (1974-1975). Assessing readability. Reading Research Quarterly, 10, 62-102

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, & representation of knowledge. Psychological Review, 104, 211-240.

Landauer, T.K., Laham, D., & Foltz, P.W. (2003). Automated essay assessment. Assessment in education: Principles, policy, and practice, 10, 295-308.

Landauer, T, McNamara, D.S., Dennis, S., Kintsch, W. (2007)(Eds.), Handbook of Latent Semantic Analysis. Mahwah, NJ: Erlbaum.

Lee, D. Y. W. (2004). Modeling variation in spoken and written English. London/New York: Routledge.

Lehnert, W. (1997). Information extraction: What have we learned? Discourse Processes, 23, 441-470.

Lehnert, W. G. & Ringle, M. H. (Eds.) (1982). Strategies for natural language processing. Hillsdale, NJ: Lawrence Erlbaum.

Lenat, D. B. (1995). CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM, 38, 33-38.

Lin, D. (1998). Dependency-based Evaluation of MINIPAR. In Workshop on the Evaluation of Parsing Systems, Granada, Spain, May, 1998.

Louwerse, M. M. (2001). An analytic and cognitive parameterization of coherence relations. Cognitive Linguistics, 12, 291–315.

Lowerse, M.M., & Van Peer, W. (2003)(Eds.). Thematics: Interdisciplinary studies. Amsterdam: John Benjamins.

Louwerse, M.M., & Mitchell, H.H. (2003). Toward a taxonomy of a set of discourse markers in dialog: A theoretical and computational linguistic account. Discourse Processes, 35, 199-239.

Mandler, J., & Johnson, N. (1977). Remembrance of things parsed: Story structure and recall. Cognitive Psychology, 9, 111-151.

Mann, W. C, & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. Text, 8 , 243-281.

Marcu, D. (2000). The theory and practice of discourse parsing and summarization. Cambridge: MIT Press.

Marcus, M., Santorini, B., & Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics, 19, 313-330.

National Reading Panel (2000). Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction (NIH Pub. No. 00-4769). Jessup, MD: National Institute for Literacy.

McCarthy, P. M., Myers, J. C., Briner, S. W., Graesser, A. C., & McNamara, D. S. (in press). Are three words all we need? A psychological and computational study of genre recognition. Journal for Computational Linguistics and Language Technology.

McNamara, D.S. (2007)(Ed.), Theories of text comprehension: The importance of reading strategies to theoretical foundations of reading comprehension. Mahwah, NJ: Erlbaum.

McNamara, D.S., Kintsch, E., Songer, N.B., & Kintsch, W. (1996). Are good texts always better? Text coherence, background knowledge, and levels of understanding in learning from text. Cognition and Instruction, 14, 1-43.

McNamara, D.S., Levinstein, I.B. & Boonthum, C. (2004). iSTART: Interactive strategy trainer for active reading and thinking. Behavioral Research Methods, Instruments, and Computers, 36, 222-233.

McNamara, D.S., Louwerse, M.M., and Graesser, A.C. (2007). Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. Unpublished final report on Institute of Education Science grant, University of Memphis.

McNamara, D.S., Ozuru, Y., Graesser, A.C., & Louwerse, M. (2006). Validating Coh-Metrix. In R. Sun & N. Miyake (Eds.), Proceedings of the 28th Annual Conference of the Cognitive Science Society (pp. 573). Mahwah, NJ: Erlbaum.

McEnery, T. (2003). Corpus linguistics. In: R. Mitkov (Ed.), The Oxford encyclopedia of computational linguistics. Oxford: Oxford University Press.

Meyer, B.J.F. (1975). The organization of prose and its effect on memory. New York: Elsevier.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, & K. J. Miller (1990). Introduction to wordnet: An on-line lexical database. Journal of Lexiography, 3, 235-244.

Millis, K.K., Kim, H.J., Todaro, S. Magliano, J., Wiemer-Hastings, K., & McNamara, D.S. (2004). Identifying reading strategies using latent semantic analysis: Comparing semantic benchmarks. Behavior Research Methods, Instruments, & Computers, 36, 213-221.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. Journal of Experimental Psychology Monograph Supplement, 76, 1–25.

Palmer, M., Kingsbury, P., Gildea, D. (2005). The Proposition Bank: An annotated corpus of semantic roles. Computational Linguistics, 31, 71-106.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic Inquiry and Word Count. Austin, TX: LIWC.net (www.liwc.net).

Pennebaker, J. W., & Chung, C. K. (2007). Expressive writing, emotional upheavals, and health. In H. S. Friedman and R. C. Silver (Eds.), Foundations of Health Psychology, pp.263-284. New York, NY: Oxford University Press.

Pennebaker, J.W., & Francis, M.E. (1999). Linguistic inquiry and word count (LIWC). Mahwah, NJ: Erlbaum.

Pennebaker, J. W. & King, L. A. (1999). Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology, 77, 1296-1312.

Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: Language use over the life span. Journal of Personality and Social Psychology, 85, 291-301.

Pickering, M.J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Brain and Behavioral Sciences, 27, 169-190.

Polanyi, L., 1988. A formal model of the structure of discourse. Journal of Pragmatics, 12, 601–638.

Prince, E. F. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), Radical pragmatics (pp. 223-255). New York: Academic Press.

Rieger, C. (1978). GRIND-1: First report on the Magic Grinder story comprehension project. Discourse Processes, 1, 267-303.

Rayner, K. (1998) Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124, 372-422.

Rommetveit, R. (1974). On message structure. New York: Wiley.

Rosenblatt, L.M. (1978). The reader, the text and the poem: The transactional theory of the literary work. Carbondale, IL: Southern Illinois University Press.

Rumelhart, D.E. (1975). Notes on a schema for stories. In D.G. Bobrow and A. Collins (Eds.), Representation and understanding (pp. 211-236). New York: Academic Press.

Rumelhart, D.E., & Ortony, A. (1977). The representation of knowledge in memory. In R.C. Anderson, R.J. Spiro, and W.E. Montague (Eds.), Schooling and the acquisition of knowledge (pp. 99-135). Hillsdale, NJ: Erlbaum.

Rus, V. (2004). A first exercise for evaluating logic form identification systems, Proceedings Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3), at the Association of Computational Linguistics Annual Meeting, July 2004. Barcelona, Spain: ACL.

Rus, V., McCarthy, P.M., & Graesser, A.C. (2006). Analysis of a text entailer. 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.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simple systematic for the organization of turn taking in conversation. Language, 50, 669-735.

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.

Schank, R.C. (1972). Conceptual dependence: A theory of natural language understanding. Cognitive Psychology, 3, 552-631.

Schank, R.C., & Abelson, R.P. (1977). Scripts, plans, goals, and understanding: An inquiry into human knowledge structures. Hillsdale, NJ: Erlbaum.

Schank, R. and Riesbeck, C. (1981). Inside computer understanding. Hillsdale, NJ: Lawrence Erlbaum.

Searle, J.R. (1969). Speech acts. London: Cambridge University Press.

Sekine, S., & Grishman, R. (1995). A corpus-based probabilistic grammar with only two nonterminals. Fourth International Workshop on Parsing Technologies (pp. 260-270). Prague/Karlovy Vary, Czech Republic.

Snow, C. (2002). Reading for understanding: Toward an R&D program in reading comprehension. Santa Monica, CA: RAND Corporation.

Spilich, G.J., Vesonder, G.T., Chiesi, H.L., & Voss, J.F. (1979). Text processing of domain related information for individuals with high and low domain knowledge. Journal of Verbal Learning and Verbal Behavior, 18, 275-290.

Stenner, A.J.(2006). Measuring reading comprehension with the Lexile framework. Durham, NC: Metametrics, Inc. presented at the California Comparability Symposium, October 1996. Retrieved January 30, 2006 from http://www.lexile.com/DesktopDefault.aspx?view=re.

Toglia, M. P., & Battig, W. F. (1978). Handbook of semantic word norms. Hillsdale, NJ: Erlbaum.

Van den Broek, P., Virtue, S., Everson, M.G., & Tzeng, Y., & Sung, Y. (2002). Comprehension and memory of science texts: Inferential processes and the construction of a mental representation. In J. Otero, J. Leon, & A.C. Graesser (Eds.), The psychology of science text comprehension (pp. 131-154). Mahwah, NJ: Erlbaum.

Van Dijk, T.A. (1972). Some aspects of text grammars. The Hague: Mouton.

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.

Williams, P. J. (2007). Literacy in the curriculum: Integrating text structure and content area instruction. In D. S. McNamara (Ed.), Reading comprehension strategies: theories, interventions, and technologies (pp. 199-219). Mahwah, NJ: Erlbaum.

Zipf, G. K. (1949). Human behaviour and the principle of least effort. Cambridge, Mass.: Addison-Wesley Press.

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