Running head: AUTOMATED ANALYSES OF VERBAL RESPONSES

# Automated Analysis of Essays and Open-ended Verbal Responses

Arthur C. Graesser and Danielle S. McNamara

University of Memphis

**Graesser, A. C., & McNamara, D.S. (in preparation).** ***Automated Analysis of Essays and Open-ended Verbal Responses.*** **In H. Cooper, A.T. Panter et al. (Eds.),** ***APA Handbook of Research Methods in Psychology.*** **Washington, DC: American Psychological Association.**

Send correspondence to:

Art Graesser

Department of Psychology & Institute for Intelligent Systems

202 Psychology Building

University of Memphis

Memphis, TN 38152-3230

901-678-4857

901-678-2579 (fax)

a-graesser@memphis.edu

# Automated Analysis of Essays and Open-ended Verbal Responses

One approach to analyzing psychological mechanisms is to perform qualitative and quantitative assessments of verbal content. Interviews, essays, and answers to open-ended questions are sometimes the best window to understanding psychological processes. These verbal protocols are routinely collected in many psychological fields, but notably education, discourse processes, cognitive science, social and personality psychology, survey methods, forensics, and clinical psychology. In the past, human experts have often been recruited to annotate and score these verbal protocols. However, during the last decade there has been ample progress in automated computer analyses of the verbal content. This chapter reviews methods for computer analyses of open-ended verbal responses.

This is a unique point in history because there is widespread access to hundreds of computer tools that quickly analyze texts and large text corpora. This increase in automated text analyses can be attributed to landmark advances in computational linguistics (Jurafsky & Martin, 2008; Shermis & Burstein, 2003), discourse processes (Graesser, Gernsbacher, & Goldman, 2003; McNamara & Magliano, 2009), statistical representations of world knowledge (Landauer, McNamara, Dennis, & Kintsch, 2007; McNamara, in press), corpus analyses (Biber, Conrad, & Reppen, 1998), word dictionaries with psychological attributes (Miller et al., 1990; Pennebaker, Booth, & Francis, 2007), and automated analyses of discourse cohesion (Graesser, McNamara, Louwerse, & Cai, 2004; Graesser & McNamara, 2010). Thousands of texts can be quickly accessed and analyzed on thousands of measures in a short amount of time. Some levels of language and discourse cannot be automated reliability, such as complex novel metaphors, humor, and conversations with cryptic provincial slang. In such cases, human experts need to annotate the texts systematically. However, essays on academic topics can be analyzed by

computers as reliably as expert human graders (Shermis, Burstein, Higgins, & Zechner, 2010), as we will discuss in this chapter.

Some people will remain skeptical of the notion that computers can analyze language and discourse because it is believed that only humans have sufficient intelligence and depth. Errors in computer assessment open the door to litigation when there are high stakes tests or psychological diagnoses of individuals or groups. People complain about computer errors even when trained human experts are no better, if not worse, than computer assessments. It is worthwhile to take stock of the advantages of computer assessments of verbal responses when compared with humans. Computers can provide instantaneous feedback, do not get fatigued, are consistent, are unbiased in assigning scores to particular individuals, provide greater detail on many dimensions, and can apply sophisticated algorithms that humans could never understand and apply (Shermis & Burstein, 2003; Shermis et al., 2010; Streeter, Psotka, Lahan, & MacCuish, 2002). After the up-front costs of developing the computer systems, the application of computer assessments to thousands or millions of people is quite economical compared with human assessments. Human grading and annotation is an expensive, time-consuming enterprise that few humans enjoy. Moreover, the argument can be made that an objective analysis of language and discourse should not rely entirely on human intuitions for scoring and annotation.

The remainder of this chapter is divided into three parts. The first section discusses automated grading of essays of 150 words or longer. This work has enormous implications for high stakes tests and writing portfolios in K12, college, and the workforce. The second section concentrates on shorter verbal responses that range from a few words to 2-3 sentences. Short verbal responses occur in short-answer test questions, conversational interactions, and computerized learning environments with natural language interaction. The third section briefly

describes some recently developed systems that induce psychological attributes from verbal responses, such as emotions, status, and personality characteristics.

## Automated Essay Scoring

Automated essay scoring (AES) has now reached a level of accuracy that the scoring of many classes of written essays is as accurate as expert human raters (Attali & Burstein, 2006; Burstein, 2003; Elliot, 2003; Landauer, Laham, & Foltz, 2003; Rudner, Garcia, & Welch, 2006; Shermis et al., 2010; Streeter et al., 2002; Valenti, Neri, & Cucchiarelli, 2003). This is indeed a remarkable achievement. How do these developers of an AES defend such a claim? The methodological approach to establishing this claim is straightforward. Two or three expert human raters grade a large sample of essays after they receive training on a scoring rubric. The grading scale typically has 5 to 7 levels on an ordinal scale. The essays are divided into a training set and a validation set. The computer program has a set of computational algorithms that are tuned to optimally fit the essays in the training set. The quantitative solution to the training set is typically a linear multiple regression formula or a set of Bayesian conditional probabilities between text characteristics and grade level. The quantitative solution is then applied to the essays in the validation set and these scores are compared to the scores of the human raters. An AES is considered successful if the scores between the computer and humans are about the same as the scores between humans.

### Performance of Automated Essay Scoring

A Pearson correlation coefficient (*r*) is a simple way to assess the performance of an AES. Success of the AES is confirmed if two human raters show correlations of some value R, whereas the computer scores correlate with the raters with the value R, almost as high as R, or even higher than R. Another scoring index is the percentage of exact agreements in the scores.

An adjacent agreement index is the percentage of scores that are within one value of each other on the ordinal grading scale. Shermis et al. (2010) reviewed the performance of the three most successful AES systems: the *e-rater* system developed at Educational Testing Service (Attali & Burstein, 2006; Burstein, 2003), the *Intelligent Essay Assessor* developed at Pearson Knowledge Technologies (Landauer et al., 2003; Streeter et al., 2002), and the *IntelliMetric Essay Scoring System* developed by Vantage Learning (Elliot, 2003; Rudner et al., 2006). These systems have had exact agreements with humans as high as the mid-80's, adjacent agreements in the high mid-90's, and correlations as high as the mid-80's. Just as impressive, these performance measures are slightly higher than agreement between trained human raters.

The performance of these AES systems has been sufficiently impressive to scale them for use in educational applications. They are used in a scoring process for high stakes tests, such as the Analytic Writing Assessment of the Graduate Management Admission Test (GMAT). In this test there are two 30-minute writing tasks to measure the test taker's ability to think critically and communicate ideas. One task involves an analysis of an issue; test takers receive an issue or opinion and are instructed to explain their point of view by citing relevant reasons or evidence. The second task is an analysis of an argument; the test taker reads a brief argument, analyzes the reasoning behind it, and writes a critique of the argument. The AES's are also used in electronic portfolio systems to help students improve writing by giving them feedback on many features of their essays, as in the case of *Criterion* (Attali & Burstein, 2006) and *MY Access* (Elliot, 2003).

Although the practical use of AESs is undeniable, critics do raise questions that challenge the ubiquitous use of these systems without some human expertise. The critics voice concerns about the aspects of writing that the AES systems fail to capture, the ethics of using computers rather than teachers to teach writing, and differences in the criteria that humans versus the

computers use to grade the essays (Calfee, 2000; Ericsson & Haswell, 2006). There is also a persistent third variable that robustly predicts essay scores, namely the number of words in the essay. Although these AES systems do have predictive validity beyond word count, the incremental gain from the advanced computational algorithms is either not reported or is unspectacular in some evaluations that have controlled for number of words.

**How do the AES Systems Grade Essays?**

It is beyond the scope of this chapter to give a precise specification of the computational algorithms in these AESs, particularly because some are proprietary or the published reports do not reflect the current systems. An edited volume by Shermis and Burstein (2003) provides detailed descriptions of these systems to the extent that the corporations are comfortable in sharing the information. The e-rater AES (Attali & Burstein, 2006) scores essays on 6 areas of analysis (12 features) which are aligned with human scoring criteria: errors in grammar, errors in word usage, errors in mechanics, style, inclusion of organizational segments (e.g., inclusion of a thesis statement or some evidence), and vocabulary content. The IntelliMetric AES (Elliot & Mikulis, 2004) matches the words to a vocabulary of over 500,000 unique words, identifies more than 500 linguistic and grammatical features that may occur in the text, and analyzes this content through a word concept net. These text characteristics are then associated with essays in each level of scoring rubric of the training corpus in order discover what essay characteristics are diagnostic of each level.

The Intelligent Essay Assessor AES (Landauer et al., 2003) analyzes the words in the essay with latent semantic analysis (LSA, Landauer et al., 2007) and also sequences of words with an n-gram analysis (e.g., word pairs, word triplets). The algorithm computes the similarity of the words and word sequences between the incoming essay and the essays associated with

each level of the scoring rubric. LSA is an important method of computing the conceptual similarity between words, sentences, paragraphs, or essays because it considers implicit knowledge. LSA is a mathematical, statistical technique for representing knowledge about words and the world on the basis of a large corpus of texts that attempts to capture the knowledge of a typical test taker. LSA captures knowledge in an encyclopedia rather than a dictionary. The central intuition of LSA is that the meaning of a word W is reflected in the company of other words that surround word W in naturalistic documents (imagine 40,000 texts or 11 million words). Two words are similar in meaning to the extent that they share similar surrounding words. For example, the word *glass* will be highly associated with words of the same functional context, such as *cup, liquid, pour, shatter,* and *transparent*. These are not synonyms or antonyms that would occur in a dictionary or thesaurus. 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, entire essay) is computed as the geometric cosine between the values and weighted dimensions of the two text excerpts. The value of the cosine typically varies from approximately 0 to 1.

An observant reader might object that these analyses of language are merely word crunchers and do not construct deep, structured, meaning interpretations. This observation is correct. However, there are two counter arguments to this objection. First, a large proportion of the essays are written under extreme time pressure so there is a high density of content that is ungrammatical, semantically ill-formed, and lacking in cohesion. More sophisticated computational analyses of language and discourse is appropriate for text that has been edited and has passed publication standards. Second, an important distinction is made between a *trin* and

*prox* (Page & Petersen, 1995). A trin is an intrinsic characteristic of writing, such as content, creativity, style, mechanics, and organization. A prox (short for proxy) is a superficial observable countable feature of text that is diagnostic of a trin. One or more prox may be adequate to estimate a trin. Therefore, it is entirely an empirical question whether the prox landscape in an AES is adequate for recovering the essential intrinsic characteristics of writing.

**Characteristics of Writing**

A holistic grade for an essay has some value to the writer as an overall index of writing quality. However, more specific feedback on different characteristics of writing provides more useful information to the student and instructor. Is there a problem with spelling, vocabulary, syntax, cohesion of the message, missing content, elements of style, and so on? The e-rater AES gives this feedback on 12 features in support of *Criterion*, an electronic portfolio of the students' writing. The portfolio of writing samples can be collected over time for students or instructors to track progress. Similarly, the LSA modules in the Intelligent Essay Assessor have been used in a system called *Summary Street* (Franzke, Kintsch, Caccamise, Johnson, & Dooley, 2005), a system that gives feedback to the student on the quality of their summaries of a text. Summary Street identifies sentences that have low LSA relevance scores with other sentences in the text and low scores with expected information in different content categories of an underlying content rubric. An ideal summary would cover the expected content and also have sentences that relate to one another conceptually.

Burstein, Marcu, and Knight (2003) developed an automated scoring technology for the Criterion system at ETS that identifies the extent to which an essay contains particular components of an essay. The targeted categories of the essay include the title, the introductory material, a thesis statement, main ideas with respect to the thesis, supporting ideas, conclusions,

and irrelevant segments. Trained human judges can identify these sections with kappa agreement scores of approximately 0.80 (between 0.86 and 0.95 on three different essay prompts). Kappa scores correct for guessing, adjust for the distribution of decisions, and vary between 0 (chance) and 1.0 (perfect agreement). Kappa scores have an advantage over correlations but in practice the performance metrics lead to identical conclusions in this line of research. The kappa scores between the computer algorithms and human raters are respectable, typically above .70.

In addition to kappa and correlations, researchers in computational linguistics routinely collect recall, precision, and F-measure scores between the computer decision and the decision of a human judge (and also between one judge and another judge). A recall score for a computer system is the proportion of human decisions that receive the same decision as the computer. The precision score is the proportion of computer decisions that agree with a human. The F-measure is 2*recall*precision/(recall + precision), essentially an average between recall and precision scores. Burstein et al. (2003) reported that the scores between computer and human were approximately the same for these three metrics and averaged .76, depending on various parameters and criteria. Agreement between pairs of human judges averaged .91. Although not perfect, these automated systems are clearly making significant progress in identifying components of essays. These categories are important to identify in order to give informative guidance on how students can improve writing.

Coh-Metrix, developed in the Institute for Intelligent Systems at the University of Memphis (Graesser & McNamara, 2010; Graesser et al., 2004; McNamara, Louwerse, McCarthy, & Graesser, in press), is another promising tool for analyzing writing. Coh-Metrix was originally developed to provide rapid automated analyses of printed text on a wide array of linguistic and discourse features, including word information (e.g., frequency, concreteness, multiple senses),

syntactic complexity, semantic relations, cohesion, lexical diversity, and genre. Coh-Metrix is available in both a public version for free on the web (http://cohmetrix.memphis.edu, version 2.1) and there is an internal version as well. The public version provides 63 measures of language and discourse, whereas the internal research version has nearly a thousand measures that are at various stages of testing. According to a principal components analysis conducted on a large corpus of 37,351 texts (Graesser & McNamara, in press), the multiple measures provided by Coh-Metrix funnel into the following five major dimensions:

1. **Narrativity.** Narrative text tells a story, with characters, events, places, and things that are familiar to the reader. Narrative is closely affiliated with everyday oral conversation.
2. **Situation model cohesion.** Causal, intentional, and temporal connectives help the reader to form a more coherent and deeper understanding of the text.
3. **Referential cohesion**. High cohesion texts contain words and ideas that overlap across sentences and the entire text, forming threads that connect the explicit text together for the reader.
4. **Syntactic simplicity**. Sentences with few words and simple, familiar syntactic structures are easier to process and understand. Complex sentences have structurally embedded syntax.
5. **Word concreteness**. Concrete words evoke mental images and are more meaningful to the reader than abstract words.

One of the central purposes of Coh-Metrix is to examine the role of cohesion in distinguishing text types and in predicting text difficulty (Graesser & McNamara, 2010; McNamara et al., in press). Indeed, one underlying assumption of Coh-Metrix is that cohesion is an important component in facilitating comprehension. Cohesion arises from a variety of sources, including explicit referential overlap and causal relationships (Givón, 1995; Graesser,

McNamara, & Louwerse, 2003; Halliday & Hasan, 1975; McNamara, 2001). For example, referential cohesion refers to the degree to which there is overlap or repetition of words or concepts across sentences, paragraphs, or the entire text. Causal cohesion refers to the degree to which causal relationships are expressed explicitly, often using connectives (e.g., *because, so,* and *therefore)* as linguistic cues.

The importance of cohesion to text comprehension begs the question of the relationship between the presence of cohesion cues and essay quality. Current studies are underway that use Coh-Metrix and other computer analyses of text complexity to analyze essays and other writing samples. A recent project by McNamara, Crossley, and McCarthy (2010) used Coh-Metrix to examine the role of cohesion in essays written by undergraduate college students. They found that linguistic features related to language sophistication characterized the essays that were rated as higher quality. The better essays featured less familiar words, more complex syntax, and greater diversity of ideas. By contrast, cohesion cues such as word overlap, conceptual overlap through LSA, causal cohesion, and the use of various types of connectives were not predictive of essay quality. The finding that the wording and syntax increased in complexity with higher quality essays is quite intuitive. Indeed, these results are compatible with the culture of English teachers who encourage more erudite language. However, it is surprising that cohesion played such a small role in explaining essay quality. The role of cohesion and coherence in writing merit greater attention in future studies.

**Challenges in Assessing Writing Instruction**

There are a number of methodological challenges that require attention for those who develop instructional systems designed to track and improve writing over time. One problem is that there are a limited number of standardized tests of writing achievement with norms that

allow gauging of progress over time. A second problem is that the available norm-referenced standardized tests, such as the *Woodcock-Johnson* or the *Wechsler Individual Achievement Test,* cover few writing skills and genres. A third problem is that the writing process is influenced by a number of factors associated with the pragmatic writing context, intended audience, writing prompts, time allotted for writing, mode of writing (handwriting versus keyboard), choice of topics to write about, and characteristics of the writer (Graham & Perin, 2007).

The time-intensive nature of scoring written essays has traditionally limited teachers from giving a large number of writing assignments. This limitation can of course be circumvented by AES systems. There are also other methods other than the use of computers. Teachers can have students assess their own writing performance and progress as writers, a process that improves writing (Andrade & du Boulay, 2003; Graham & Perin, 2007; Ross, Rolheiser, & Hogaboam-Gray, 1999). Teachers can have students assess each others' writing. When learners are taught how to assess and provide feedback to their peers, their writing and the writing of their peers improves (Cho, Schunn, & Wilson, 2006; Graham & Perin, 2007). Future research needs to compare the computerized AES systems with self-assessment and peer-assessment of writing over time.

## Scoring of Short Verbal Responses

Short verbal responses by students can vary from one word to 2-3 sentences. There currently are technologies, such as *C-Rater* developed at ETS (Lealock & Chadorow, 2003) or the AutoTutor system described below that can score answers to short-answer questions that extend beyond single words. The verbal responses may be answers to open-ended questions or they may be contributions in dialogues or multiparty conversations. The computer is expected to score the students' answers or conversational turns on a variety of dimensions: accuracy

compared to an expected answer, relevance, completeness, verbosity, style, and so on. Compared with the scoring of essays, the scoring of short verbal responses is easier in some ways but more difficult in others. It is easier because there is less information for the computer to process and a deeper analysis of the language can be accomplished. It is more difficult because less information implies some degradation in the reliability of statistical approximations of relevant parameters. In essence, there are tradeoffs between the depth of processing and the reliability of parameter estimates.

The scoring of single word answers is not a significant challenge when there are one or a few expected answers to a question. The computer can score exact matches, synonyms, semantic associates, and words that have a close match via LSA and other statistical algorithms (Landauer et al., 2007). This is not a difficult technology for typed input but presents more of a challenge with spoken input (Mostow, 2008). This chapter will not address the scoring of single word answers because this is a mature technology and our goal is to address the more challenging questions about the processing of human verbal responses beyond the word.

This chapter will also not cover spoken verbal responses even though the utility of such systems would be widespread. Unfortunately, the quality of these speech-to-text systems is still evolving and has not reached the level of accuracy for use other than an experimental basis. There are limitations in the accuracy of speech-to-text technologies that have word error rates for conversational speech ranging from 14% to 50%, depending on the system, domain, and testing environment (D'Mello, King, Chipman, & Graesser, in press; Hagen, Pellom, & Cole, 2007; Litman et al., 2006; Zechner, Higgins, Xi, & Williamson,in press; Zolnay, Kocharov, Schluter, & Ney, 2007). The *SpeechRater* system developed at ETS is promising (Zechner et al., in press),

but still has a high error rate and has operational use only for students who volunteer to use the system for on-line practice.

**Scoring of Student Responses in Intelligent Tutoring Systems and Trainers**

Intelligent Tutoring Systems (ITS) are computerized learning environments with computational models that track the subject matter knowledge, strategies, and other psychological states of learners, a process called student modelling (Sleeman & Brown, 1982; Woolf, 2009). An ITS adaptively responds with activities that are both sensitive to these states and that advance the instructional agenda. The interaction between student and computer follows a large, if not an infinite number of alternative trajectories that attempt to fit constraints of both the student and the instructional agenda. This is quite different from learning from a book or a lecture, which unfold in a rigid linear order and is not tailored to individual students.

Assessment of student contributions in this turn-by-turn tutorial interaction is absolutely essential in any ITS. Such assessments are straightforward when the students' responses are selections among a fixed set of alternatives, as in the case of multiple-choice questions, true-false questions, ratings, or toggled decisions on a long list of possibilities. However, challenges arise when the student input is verbal responses in natural language, which is the focus of this subsection.

A number of ITS's and trainers have been developed that hold conversations in natural language. Two of these systems are described in some detail in this section because they have been systematically tested on the extent to which the computer accurately scores the students' verbal responses. These two systems are *AutoTutor* (Graesser, Chipman, Haynes, & Olney, 2005; Graesser, Jeon, & Dufty, 2008; Graesser, Lu et al., 2004) and *iSTART* (Levinstein, Boonthum, Pillarisetti, Bell, & McNamara, 2007; McNamara, Levinstein, & Boonthum, 2004). However, a number of other systems have been developed with similar goals, such as *ITSPOKE*

(Litman et al., 2006), *Spoken Conversational Computer* (Pon-Barry, Clark, Schultz, Bratt, & Peters, 2004), *Tactical Language and Culture Training System* (Johnson & Valente, 2008), and *Why-Atlas* (VanLehn et al., 2007). This section also includes a system called R-SAT that collects think aloud protocols while students comprehend text and scores the extent to which these verbal protocols reflect particular comprehension processes (Millis et al., 2004; Millis & Magliano, in press).

***AutoTutor***. AutoTutor is an intelligent tutoring system that helps students learn about computer literacy, physics, critical thinking skills, and other technical topics by holding conversations in natural language (Graesser, Chipman, Haynes, & Olney, 2005; Graesser, Jeon, & Dufty, 2008; Graesser, Lu et al., 2004). AutoTutor shows learning gains of approximately 0.80 sigma (standard deviation units) compared with pretests or with a condition that has students read a textbook for an equivalent amount of time (Graesser, Lu et al., 2004; VanLehn et al., 2007). The tutorial dialogues are organized around difficult questions and problems that require reasoning and explanations in the answers. The following are examples of challenging questions on the topics of Newtonian physics and computer literacy.

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?

COMPUTER LITERACY QUESTION: When you turn on the computer, how is the operating system first activated and loaded into RAM?

These questions require the learner to construct approximately 3-7 sentences in an ideal answer and to exhibit reasoning in natural language.

When asked one of these questions, the students' answers are short during the first conversational turn, typically ranging from a couple of words to a couple of sentences. It takes a conversation to draw out more of what the student knows even when the student has reasonable subject matter knowledge. The dialogue for one of these challenging questions typically lasts 50-100 conversational turns between AutoTutor and the student. AutoTutor provides *feedback* to the student on what the student types in (positive, neutral, versus negative feedback), *pumps* the student for more information ("What else?"), *prompts* the student to fill in missing words, gives the student *hints*, fills in missing information with *assertions*, *corrects* erroneous ideas and misconceptions, *answers* the student's questions, and *summarizes* answers. These acts of feedback, pumps, prompts, hints, assertions, corrections, answers, and summaries are important *dialogue moves* of AutoTutor. A full answer to the question is eventually constructed during this dialogue.

There are many different ways to score the performance of AutoTutor (Graesser, Penumatsa, Ventura, Cai, & Hu, 2007; Jackson & Graesser, 2006; VanLehn et al., 2007). One method is to score the extent to which the students' verbal contributions match good answers to the question (called *expectations*) versus bad answers (called *misconceptions*). For example, listed below are some of the expectations and misconceptions in the example physics question.

(Expectation E1) The magnitudes of the forces exerted by the two objects on each other are equal.

(Expectation E2) If one object exerts a force on a second object, then the second object exerts a force on the first object in the opposite direction.

(Expectation E3) The same force will produce a larger acceleration in a less massive object than a more massive object.

(Misconception M1) A lighter/smaller object exerts no force on a heavier/larger object.

(Misconception M2) A lighter/smaller object exerts less force on other objects than a heavier/larger object.

Students will receive higher scores to the extent that they express more of the expectations and fewer of the misconceptions in the tutorial dialogue. Such expectation coverage scores and misconception scores can be computed during the first student turn, or alternatively after they have finished the conversational dialogue. It is important to point out that AutoTutor cannot interpret student contributions that have no matches to the anticipated expectations and misconceptions; it can only make comparisons between the student input and these anticipated ideas through pattern matching algorithms.

Students rarely articulate the expectations perfectly because natural language is much too imprecise, fragmentary, vague, ungrammatical, and elliptical. AutoTutor has used a number of semantic match algorithms to evaluate the extent to the students' verbal responses match any given expectation E (Graesser, Penumatsa, et al., 2007). These semantic match algorithms have included keyword overlap scores, word overlap scores that place higher weight on words that have lower frequency in the English language, word overlap scores that consider the order in which words are expressed, latent semantic analysis cosine values, and symbolic procedures that compute logical entailment (Rus & Graesser, 2006). These computer match scores have shown correlations with human expert ratings of $r = 0.29$ to .42 (Graesser, Penumatsa et al., 2007), depending on the automated algorithm tested. Expectation E is considered covered by the student if the verbal responses meet or exceed a threshold value T of a semantic match. Such assessments can be performed on individual student turns, combinations of turns, or the cumulative sequence of turns that led up to any point of measurement in the dialogue. One

metric of performance is a coverage score that compares (a) the proportion of expectations that are covered by the student according to the semantic match scores at an optimal threshold T and (b) the proportion of expectations covered according to expert human judges (Graesser, Penumatsa et al., 2007; P. Wiemer-Hastings, K. Wiemer-Hastings, & Graesser, 1999). These correlations have varied between $r$ = .35 and .50, with most estimates leaning towards .50. Other metrics of agreement, such as kappa scores, recall, precision, and F-measures, also reflect intermediate levels of agreement between computer scores and human experts.

Another method of assessing student performance in AutoTutor is to analyze the number and type of dialogue moves by AutoTutor that were selected to extract information from the student during the evolution of the answer. The system periodically identifies a missing expectation during the course of the dialogue and posts the goal of covering the expectation. When expectation E is posted, AutoTutor attempts to induce the student to articulate it by generating hints and prompts that encourage the student to fill in missing words and propositions. Learners often leave out a content word, phrase, or entire clause within E. Specific prompts and hints are generated that maximize the learner's filling in this content and boosting the match score above threshold. For example, suppose that expectation E1 needs to be articulated in the answer. The following family of candidate prompts is available for selection by AutoTutor to encourage the student to articulate particular content words in expectation E1 (*The magnitudes of the forces exerted by two objects on each other are equal*).

(a) The magnitudes of the forces exerted by two objects on each other are ____.

(b) The magnitudes of forces are equal for the two _____.

(c) The two vehicles exert on each other an equal magnitude of _____.

(d) The force of the two vehicles on each other are equal in _____.

If the student has failed to articulate one of the four content words (*equal, objects, force, magnitude*), then AutoTutor selects the corresponding prompt (a, b, c, and d, respectively).

Performance of a student in AutoTutor can be measured by computing the number of hints and prompts it takes for the student to generate an answer to a question. This was assessed in an analysis of four dialogue move categories that attempt to cover the content of particular expectations: Pumps, hints, prompts, and assertions (Graesser, Penumatsa, et al., 2007; Jackson & Graesser, 2006). The proportion of dialogue moves in these categories should be sensitive to student knowledge of physics (as measured by a pre-test of physics with multiple-choice questions similar to the Force Concept Inventory, Hestenes, Wells, & Swackhamer, 1992). There is a continuum from the student supplying information to the tutor supplying information as we move from pumps, to hints, to prompts, to assertions. The correlations with student knowledge reflected this continuum perfectly, with correlations of .49, .24, -.19, and -.40. For students with more knowledge of physics, AutoTutor can get by with pumps and hints, thereby encouraging the student to articulate the expectations. For students with less knowledge of physics, AutoTutor needs to generate prompts that elicit specific words or to assert the correct information, thereby extracting knowledge piecemeal or merely telling the student the correct information.

These analyses of student verbal responses through AutoTutor support a number of claims. First, there are several automated algorithms that can score whether particular sentences are covered in verbal responses that evolve in conversational turns over the course of a conversation. Second, the computer scores for sentential content matches have a moderate, but unspectacular level of accuracy, at least compared with the scoring of lengthy essays. There is less content in a sentence than a lengthy essay, so this second conclusion is quite expected. On the other hand, it is interesting to note that the scoring of verbal responses is extremely high

when the expectation unit is a single word, intermediate when it is a sentence, and high when it is an essay. Third, the scoring of verbal responses with AutoTutor requires an analysis of expected content and an assessment of the extent to which verbal responses match the expected content. It is beyond the scope of AutoTutor to analyze content that is not on the radar of these expectations.

***iSTART (Interactive Strategy Trainer for Automated Reading and Thinking).*** iSTART (Levinstein et al., 2007; McNamara, Levinstein, & Boonthum,2004) is an interactive tutoring system that helps high school and college students learn and practice strategies to improve comprehension of challenging expository text. Studies evaluating iSTART's impact indicate that both strategy use and science comprehension are enhanced (Magliano et al., 2005; McNamara et al., 2006; McNamara, O'Reilly, Rowe, Boonthum, & Levinstein, 2007). iSTART is particularly effective in helping low knowledge and less skilled comprehenders better understand challenging text.

iSTART contrasts with AutoTutor because it focuses on the detection and training of strategy use rather than the accuracy of content understanding. There are several modules in iSTART. The students are first provided with information about strategies and examples of the use of reading strategies (i.e., bridging inferences and elaborations) in the context of generating self-explanations of text. The student moves on to tutoring modules in which they are asked to type self-explanations of science or history texts. A crucial aspect of iSTART's effectiveness is the feedback provided to students by a pedagogical agent as they type in self-explanations of text using the comprehension strategies. The automated algorithm detects the quality of the self-explanation so that directive feedback can be provided to the student.

Several versions of the iSTART evaluation algorithm have been developed and assessed (McNamara, Boonthum, Levinstein, & Millis, 2007). The ultimate goal was to develop an algorithm that was completely automated and did not rely on any human or hand-coded computations. The resulting algorithm uses a combination of both word-based approaches and LSA (Landauer et al., 2007). Word-based approaches include a length criterion in which the student's explanation must exceed a specified number of content words that are in the text. The LSA-based approach relies on a set of benchmarks from the target text including the title of the passage, the words in the target sentence, and the words in the previous two sentences. The word-based algorithms provide feedback on shallow explanations (i.e., ones that are irrelevant or that simply repeat the target sentence). LSA augments the word-based algorithms by providing a deeper, qualitative assessment. More positive feedback is given for longer, more relevant explanations, whereas increased interactions and support are provided for shorter, less relevant explanations. For example, if the self-explanation appears irrelevant, an animated agent asks the student to add more information that is related to the sentence. Satisfactory explanations might receive feedback such as, "That's really great!" or "That's pretty good. "

The accuracy of the iSTART evaluation algorithms has been assessed by computing linear equations based on a discriminate analysis of one data set and calculating its ability to predict human ratings for a variety of data sets (Boonthum, Levinstein, & McNamara, 2007; McNamara, Boonthum, Levinstein, & Millis, 2007; Millis et al., 2004; Jackson, Guess, & McNamara, 2010). Across a number of evaluations, the iSTART algorithms have corresponded well to human ratings. McNamara, Boonthum et al. (2007) reported that algorithms corresponded highly with human evaluations of the self-explanations on two texts in the initial iSTART practice module; there was a 62-64% agreement between the algorithm and the human

judgments ($r$ = .64-.71; d'= 1.54-1.79). The algorithms also successfully transferred to texts that were on a variety of science topics used in a classroom study that included 549 high school students who engaged in extended practice using iSTART across an academic year (Jackson, Guess, & McNamara, 2010). This study showed a $r$ =.66 correlation between the human evaluations and iSTART's algorithms. This is remarkable given the variety of texts self-explained by the students in this study. Although, this performance appears to be higher than that of AutoTutor, it is important to consider that the two systems target quite different information. iSTART assesses the quality of the student's self-explanation strategies whereas AutoTutor assesses the quality, depth, and accuracy of expected substantive content.

One computational challenge for the iSTART system is to provide the students with rapid and accurate feedback on the quality of their self-explanations. This evaluation challenge is achieved in several steps. First, the response is screened for metacognitive and frozen expressions (such as "I don't understand what they are saying here"; I'm bored"). If the explanation is dominated by the frozen expressions and contains little other content, then the pedagogical agent responds directly to those statements using a pool of responses that are randomly chosen, "Please try to make a guess about what this means" or "Can you try to use one of the reading strategies? Maybe that will help your understanding."

After the frozen statements are removed from the explanation, then the remainder of the explanation is analyzed using both word-based and LSA-based methods (McNamara et al., 2007). If the length of the explanation does not reach a particular threshold T relative to the length of the target text, then the student is asked to add more to the explanation. The agent might then say "Could you add to your explanation? Try to explain how it relates to something you already know." If the explanation does not have sufficient overlap in words or semantically meaning to

the target and surrounding text, then it is assessed as irrelevant. The two examples below show a target sentence (TS) and a self-explanation (SE). The SEs were categorized as irrelevant. The first is completely off topic, while the second is on topic but would not help the student to understand the target text.

TS: Survival depends on the cell's ability to maintain the proper conditions within itself.

SE: no i will not you crazy magic man haha.

TS: What kinds of environmental changes do you notice?

SE: trash on the ground

In both cases, the explanation would receive feedback such as "Please try to add information that is related to the sentence. Explain what the sentence means and how it relates to what you already know." The agent does not usually give feedback that may frustrate the student.

The explanation is further assessed in terms of its similarity to the target text. If it is too close to the target text in terms of the total number of words and the number of overlapping content words, as in the example below, then it categorized as a repetition.

TS: Inherited behavior of animals is called innate behavior.

SE: the inherited behavior of animals is called innate behavior

A repetition might receive feedback such as "Try adding some more information that explains what the sentence means." The goal is to induce the student to go beyond the sentence.

The explanation might also be assessed as being beyond a repetition and into the rhealm of paraphrase, as in the example below.

TS: The goldfish may depend on other living things for food, or it may be food for other life.

SE: The goldfish is either predator or prey.

Paraphrasing is an excellent and optimal way to start an explanation, but the goal is usually to induce the student to go beyond paraphrasing by bringing in prior text or outside knowledge to the explanation. In that case, the student would receive feedback such as "It looks like you've reworded the sentence. Now can you explain it by thinking about what else you know?"

Once the explanation passes the thresholds for length, relevance, and similarity, then feedback is provided on the quality of the explanation. Lower quality explanations are just at the threshold and have very little content that goes beyond the target text.

TS: Energy-storing molecules are produced on the inner folds.

SE: Energy for the plant is produced within the inner folds of the mitochondrion

TS: Inherited behavior of animals is called innate behavior.

SE: if the behavior is inherited the animal has innate behavior

Both of the examples are closer to paraphrases and only slightly go beyond the meaning of the text, if at all. In these cases, either cursory feedback such as OK is provided, or the student is provided with advice on how to do better on the next sentence, "For the next sentence, explain more about how it is related to other sentences or ideas."

For medium and high quality explanations, the student is provided with qualitative feedback only, such as “That’s pretty good.” for a medium quality explanation, and "You’re doing a great job!" for a high quality explanation. Below are examples of medium and high quality explanations. It is evident from these explanations that the student is processing the text more deeply by either bringing information in from prior text or from prior knowledge, which is the objective of iSTART.

*Medium Quality Self-Explanations*

TS: Energy-storing molecules are produced on the inner folds.

SE: The kind of molecules that keep energy are made by the cilia on the inside.

*High Quality Self-Explanations*

TS: They obtain nutrients by eating other organisms.

SE: the consumer called the heterotroph eats the producers such as the grass or when an owl eats a rat

TS: Survival depends on the cell's ability to maintain the proper conditions within itself.

SE: Every cell that's alive keeps a steady balance, no matter what's going on inside or outside the cell. Doing this is what keeps the cell alive.

The current version of iSTART provides only verbal feedback. A new version (iSTART-Motivationally Enhanced or iSTART-ME) is more game-based with points that are contingent on student performance (McNamara, Jackson, & Graesser, in press). This version attempts to enhance student motivation and to provide extended practice of strategies over longer periods of time.

***RSAT (Reading Strategy Assessment Tool).*** RSAT was developed to identify the comprehension strategies that are manifested in think-aloud protocols that students type in (or say aloud) while reading texts (Gilliam, Magliano, Millis, Levinstein, & Boonthum, 2007; Magliano et al., 2009; Millis & Magliano, in press; Millis, Magliano, & Todaro, 2006). The RSAT team has worked closely with the iSTART team in automated analyses of different types of self-explanations

One important comprehension strategy measured by RSAT is to identify content that reflects causal connections or *bridges* between clauses in the text. Whereas iSTART only assesses the quality of self-explanations, RSAT distinguishes between local and distal bridges. Local bridges occur between the target sentence and the immediately prior sentence. Distal

bridges occur between the target sentences and sentences located two or more sentences back. Skilled readers are more likely to make distal bridges whereas less-skilled readers tend to focus more on the immediate context surrounding each sentence (Coté, & Goldman, & Saul, 1998). A second type of strategy is *elaboration*. Elaborative inferences are constructed in a fashion that caters to the constraints of the text but also recruits relevant world knowledge (Graesser, Millis, & Zwaan, 1997; Long, Golding, & Graesser, 1992; McNamara & Magliano, 2009). Unlike bridges, elaborations do not connect sentences. A third strategy is *paraphrasing*. The student articulates explicit text information but in slightly different words. There is some evidence that the amount of paraphrasing in verbal protocols is negatively correlated with comprehension whereas bridging and elaborating is positively correlated (Magliano & Millis, 2003).

RSAT uses a semantic benchmark rubric in identifying the strategies. There are expected responses when a target sentence S in the text is probed with the think aloud protocol. The expectations refer to explicit sentences in the text, whereas other content constitutes inferences. The system counts the number of content words in the think aloud response R for sentence S that matches a benchmark. A local bridging score computes a match to sentence S-1, the sentence immediately prior to sentence S, whereas a distal bridging score is the match to sentences which are more than two sentences back from the sentence S. A paraphrasing score is the match to the target sentence S. An elaboration score is the number of content words in the answer which do not appear in the text. The scores from several target sentences in the text are averaged, thereby computing an overall score for comprehension as well as mean scores for the strategies of local and distal bridging, elaboration, and paraphrasing.

Assessments of RSAT (Magliano et al., 2009; Millis & Magliano, in press) report that RSAT does a reasonable job predicting overall comprehension and also discriminating

comprehension strategies. In Magliano et al. (2009), college students took one of three forms of RSAT and the Gates-MacGinitie test of comprehension, as well as other open-ended experimenter-generated tests of comprehension that served as a gold standard. The correlation between the open-ended test and RSAT was $r = .45$, slightly lower than the $r = .52$ correlation with Gates-MacGinitie. They also reported that the strategy scores predicted a significant 21% of the variance on the open-ended test with positive significant slopes for bridges and elaborations, but with a significant negative slope for paraphrases. Correlations between the automated strategy scores and expert human raters for those strategies varied between $r = .46$ and .70.

***Summary***. The analyses in this section support the claim that automated computer analyses are moderately successful in evaluating the quality of short verbal responses. The algorithms generally compare the short responses to a rubric of expected content. A variety of algorithms have been used to compute semantic matches between student verbal responses and sentence expectations. Most of these algorithms are based on the overlap of content words and inferential content through LSA, but a few consider the order in which words are expressed and even deep symbolic analyses of the natural language. The performance of these computational analyses is moderately successful, but not as impressive as the automatic scoring of lengthier essays. We anticipate that future efforts will perform deeper analyses of the content with more sophisticated natural language processing (Olney, Graesser, & Person, in press; Rus & Graesser, 2006; Rus, McCarthy, McNamara, & Graesser, 2008).

## Inducing Psychological Attributes from Verbal Responses

The previous sections discussed research that assesses how accurately computers can score the quality of information expressed by humans when their verbal responses are compared to a scoring rubric. The focus was on the accuracy or quality of the content in the verbal

protocols. This section takes a very different angle. To what extent can computers infer psychological characteristics of people from the verbal responses? For example, can emotions, leadership, personality, status, familiarity, deception and other psychological characteristics be accurately induced from verbal responses? This angle is aligned with a new research framework called Social Language Processing (SLP), which marries social and psychological theory with computational techniques in order to understand relationships between discourse and social dynamics (Hancock et al., 2010). For example, what are the key words, language, and discourse patterns that identify the leader in a group? What verbal cues are diagnostic of deception, emotions, or familiarity between group members? Are there characteristics of language and discourse that predict whether groups are meeting their goals?

It is beyond the scope of this section to review the large and emerging literature that is relevant to social language processing. Instead, we focus on two topics. First, we discuss the Linguistic Inquiry Word Count (LIWC) system that was developed by Pennebaker and his colleagues (Pennebaker, Booth, & Francis, 2007). The LIWC system has been used to analyze a wide range of phenomena in social language processing, far more than any other effort with automated systems. Second, we will discuss research that infers emotions during learning on the basis of language and discourse.

**Linguistic Inquiry Word Count (LIWC)**

LIWC is an automated word analysis tool that has received considerable attention in the social sciences (Pennebaker et al., 2007). LIWC reports the percentage of words in a given text devoted to grammatical (e.g. *articles, pronouns, prepositions*), psychological (e.g. *emotions, cognitive mechanisms, social*), or content categories (e.g. *home, occupation, religion*). For example, “crying”, and “grief” are words in the *sad* category, whereas “love” and “nice” are

words that are assigned the *positive emotion* category. The mapping between words and word categories is not mutually exclusive because a word can map onto several categories. The 2007 version of LIWC provides roughly 80 word categories, but also groups these word categories into broader dimensions. The broader dimensions are linguistic words (e.g. pronouns, past tense), psychological constructs (e.g. causations, sadness), personal constructs (e.g. work, religion), paralinguistic dimensions (e.g. speech disfluencies), and punctuations (e.g. comma, period). There is a general descriptor category that measures word count, number of words per sentence, and so on. LIWC operates by analyzing a transcript of naturalistic discourse and counting the number of words that belong to each word category. A proportion score for each word category is then computed by dividing the number of words in the verbal response that belong to that category by the total number of words in the text.

LIWC categories have been shown to be valid and reliable markers of a variety of psychologically meaningful constructs (Chung & Pennebaker, 2007; Pennebaker, Mehl, & Niederhoffer, 2003). The relative frequency of psychological words would obviously map onto relevant psychological constructs and these references review such trends. However, the more counterintuitive finding that Pennebaker and his colleagues have documented is the role of the linguistic features of words. LIWC provides 28 linguistic features that comprise function words, various types of pronouns, common and auxiliary verbs, different tenses, adverbs, conjunctions, negations, quantifiers, numbers, and swear words. It is the function words rather than the content words that surprisingly are diagnostic of many social psychological states. Function words are difficult for people to deliberately control so examining their use in natural language samples provides a non-reactive way to explore social and personality processes.

Some of the basic findings of the work on function words have revealed demographic and individual differences in function word production. There are sex, age, and social class differences in function word use (Newman, Groom, Handleman, & Pennebaker, 2008; Pennebaker & Stone, 2003). For example, first-person singular pronouns (e.g., *I, me, my*) have higher usage among women, young people, and people of lower social classes. Pronouns have been linked to psychological states such as depression and suicide across written text, natural conversations, and in published literature (Stirman & Pennebaker, 2001; Rude, Gortner, & Pennebaker, 2004). In an analysis of natural language from personal blogs, language exhibited a social coping model following the 9/11 attacks (Cohn, Mehl, & Pennebaker, 2001). That is, talk about the terrorist attacks increased immediately, with a sharp increase in negative emotion words, and a sharp decrease in positive emotion words. Ironically, positive emotion words remained at an elevated rate of use for weeks after the talks, while negative emotion words returned to baseline levels (i.e. pre-9/11 rates) within a few days after the attacks. In the days after the attacks, "I" use decreased and "we" use increased. These results suggest that after a terrorist attack, people felt more positive and psychologically connected with others.

**Inferring Emotions**

There are a number of different approaches to analyzing the affective content of text samples. One straightforward approach is to identify a small number of dimensions that underlie expressions of affect (Samsonovich & Ascoli, 2006). This research was pioneered decades ago by Osgood and colleagues who analyzed how people in different cultures rated the similarity of various emotional words (Osgood, May, & Miron, 1975). His analyses converged on *evaluation* (i.e., good or bad), *potency* (i.e., strong or weak), and *activity* (i.e., active or passive) as the critical dimensions. These dimensions are very similar to valence and arousal, which today are

considered to be the fundamental dimensions of affective experience (Barrett, Mesquita, Ochsner, & Gross, 2007; Russell, 2003). One could imagine using LIWC categories to map onto these fundamental dimensions as a plausible first-cut computational model.

The second approach is to conduct a more detailed lexical analysis of the text in order to identify words that are predictive of specific affective states of writers or speakers (Cohn, Mehl, & Pennebaker, 2004; Kahn, Tobin, Massey, & Anderson, 2007; Pennebaker, Mehl, & Niederhoffer, 2003). Once again, the LIWC program provides a straightforward approach to conducting such analyses. Other researchers have developed lexical databases that provide affective information for common words. For example, *WordNet-Affect* (Strapparava & Valitutti, 2004) is an extension of WordNet for affective content.

The third approach to affect detection systems go beyond the words and into a semantic analysis of the text. For example, Gill, French, Gergle, and Oberlander (2008) analyzed 200 blogs and reported that texts judged by humans as expressing fear and joy were semantically similar to emotional concept words (e.g. phobia and terror for fear, but delight and bliss for joy). They used LSA (Landauer et al., 2007) and the Hyperspace Analogue to Language (Burgess, Livesay, & Lund, 1998) to automatically compute the semantic similarity between the texts and emotion keywords (e.g., fear, joy, anger). Although this method of semantically aligning text to emotional concept words showed some promise for fear and joy texts, it failed for texts conveying six other emotions, such as anger, disgust, and sadness. D'Mello, Craig, Witherspoon, McDaniel, and Graesser (2008) analyzed whether student emotions could be induced from the language and discourse in tutorial dialogues with AutoTutor. Feedback, speech act categories (such as indirect hints), cohesion, negations, and other linguistic features could predict student

affect states that are frequent during tutoring, such as boredom, frustration, confusion, and engagement.

The fourth and most sophisticated approach to text-based affect sensing involves systems that construct affective models from large corpora of world knowledge and apply these models to identify the affective tone in texts (Wiebe, Wilson, & Cardie, 2005; Breck, Choi, & Cardie, 2007; Pang & Lee, 2008). For example, the word "accident" is typically associated with an undesirable event so the presence of "accident" will increase the assigned negative valence of the sentence "I was held up from an accident on the freeway". This approach is sometimes called sentiment analysis, opinion extraction, or subjectivity analysis because it focuses on valence of a textual sample (i.e., positive or negative; bad or good), rather than assigning the text to a particular emotion category (e.g., angry, sad). Sentiment and opinion analysis is gaining traction in the computational linguistics community (Pang & Lee, 2008).

## Closing Comments

Over the past few decades, enormous progress has been made in our ability to provide automated assessments of natural language and discourse. This progress has been fueled by advances in computational power, statistical techniques, linguistic databases, and theoretical understanding of discourse processes. These developments have undergirded techniques for scoring essays, analyzing characteristics of different types of writing and texts, assessing text difficulty, assessing the accuracy, quality, and type of student contributions in tutoring systems, inferring psychological characteristics of speakers and writers, and detecting affective dimensions in discourse.

Automated analyses of text and discourse are expected to flourish the next decade and beyond. Indeed, this chapter has not covered all of the advances that are at the intersection of

computational modeling and psychology. Some colleagues will continue to have a healthy skepticism of the automated analyses of language and discourse. At the other extreme are those who will continue to discover new algorithms that capture aspects of psychological mechanisms that can be automatically computed from text. Both of these mindsets are needed to converge on automated assessments that are both reliable and valid.

## References

Andrade, H.G., & Boulay, B.A. (2003). Role of rubric-referenced self-assessment in learning to write. *Journal of Educational Research, 97,* 21-34.

Attali, Y. & Burstein, J. (2006). Automated essay scoring with e-rater R V.2. *Journal of Technology, Learning and Assessment, 4,* 1-30..

Barrett, L., Mesquita, B., Ochsner, K., & Gross, J. (2007). The experience of emotion. *Annual Review of Psychology, 58,* 373-403.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Boonthum, C., Levinstein, I., & McNamara, D.S. (2007). Evaluating self-explanations in iSTART: Word matching, latent semantic analysis, and topic models. In A. Kao & S. Poteet (Eds.), *Natural Language Processing and Text Mining* (pp. 91-106). London: Springer-Verlag UK.

Breck, E., Choi, Y., & Cardie, C. (2007). *Identifying expressions of opinion in context*. Paper presented at the Proceedings of the 20th international joint conference on Artificial intelligence.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, and discourse. *Discourse Processes, 25*, 211-257.

Burstein, J., Marcu, D., and Knight, K. (2003). Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems, 18*, 32–39.

Burstein, J. (2003). The E-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective,* 133-122. Mahwah, NJ: Erlbaum.

Calfee, R. (2000). To grade or not to grade. *IEEE Intelligent Systems, 15*, 35-37.

Chung, C., & Pennebaker, J. (2007). The Psychological Functions of Function Words. In K. Fielder (Ed.), *Social Communication* (pp. 343-359). New York: Psychology Press.

Cho, K., Schunn, C.D., & Wilson, R.W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98,* 891-901.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science, 15,* 687-693.

Coté, N., Goldman, S.R., & Saul, E.U. (1998). Students making sense of informational text: Relations between processing and representation. *Discourse Processes, 25,* 1-53.

D'Mello, S. K., Craig, S.D., Witherspoon, A. W., McDaniel, B. T., & Graesser, A. C. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction, 18*(1-2), 45-80.

D'Mello, S., King, B., Chipman, P., & Graesser, A.C. (in press). Towards spoken human-computer tutorial dialogues. *Human Computer Interaction*.

Elliott, S. (2003). IntelliMetric: From here to validity. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective.* Hillsdale, NJ: Erlbaum.

Ericsson, P.F., & Haswell, R. (Ed.)(2006). *Machine scoring of student essays: Truth and consequences*. Logan, UT: Utah State University Press.

Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., & Dooley, S. (2005). Summary street: Computer support for comprehension and writing. *Journal of Educational Computing Research, 33,* 53-80.

Gill, A., French, R., Gergle, D., & Oberlander, J. (2008). Identifying emotional characteristics from short blog texts. In B. C. Love, K. McRae & V. M. Sloutsky (Eds.), *30th Annual Conference of the Cognitive Science Society* (pp. 2237-2242). Washington, DC: Cognitive Science Society.

Gilliam, S., Magliano, J. P., Millis, K. K., Levinstein, I., & Boonthum, C. (2007). Assessing the format of the presentation of text in developing a Reading Strategy Assessment Tool (RSAT). *Behavior Research Methods, Instruments, & Computers*, *39*, 199-204.

Graesser, A. C., Chipman, P., Haynes, B., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education, 48*(4), 612-618.

Graesser, A. C., Gernsbacher, M. A., & Goldman, S. (Eds.). (2003). *Handbook of discourse processes*.
Mahwah, NJ: Erlbaum.

Graesser, A. C., Jeon, M., & Dufty, D. (2008). Agent technologies designed to facilitate interactive knowledge construction. *Discourse Processes, 45*, 298–322.

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. (2010). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*.

Graesser, A.C., & McNamara, D.S. (in press). Technologies that support reading comprehension. In C. Dede and J. Richards (Eds.), *Digital teaching platforms*.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text? In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 82–98). New York: Guilford.

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., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual review of psychology, 48,* 163-189.

Graesser, A. C., Penumatsa, P., Ventura, M., Cai, Z., & Hu, X. (2007). Using LSA in AutoTutor: Learning through mixed initiative dialogue in natural language. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 243–262). Mahwah, NJ: Erlbaum.

Graham, S., & Perrin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99,* 445-476.

Hagen, A., Pellom, B., & Cole, R. (2007). Highly accurate children's speech recognition for interactive reading tutors using subword units. *Speech Communication, 49,* 861-873.

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

Hancock, J.T., Beaver, D.I., Chung, C.K., Frazee, J., Pennebaker, J.W., Graesser, A., & Cai, Z. (in press). Social language processing: A framework for analyzing the communication of terrorists and authoritarian regimes. *International Journal of Language, Culture, and Society*.

Hestenes, D., Wells, M. & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30,* 141-158.

Jackson, G. T., & Graesser, A. C. (2006). Applications of human tutorial dialog in AutoTutor: An intelligent tutoring system. *Revista Signos, 39,* 31–48.

Jackson, G.T., Guess, R.H., & McNamara, D.S. (2010). Assessing cognitively complex strategy use in an untrained domain. *Topics in Cognitive Science, 2,* 127-137.

Johnson, L. W. & Valente, A. (2008). Tactical language and culture training systems: Using artificial intelligence to teach foreign languages and cultures. In Proceedings of the Twentieth Conference on Innovative Applications of Artificial Intelligence. AAAI Press.

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.

Kahn, J., Tobin, R., Massey, A., & Anderson, J. (2007). Measuring emotional expression with the linguistic inquiry and word count. *American Journal of Psychology, 120*(2), 263-286.

Landauer, T.K., Laham, D., & Foltz, P.W. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice, 10(3),* 295-308.

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (2007) (Eds.), *Handbook of latent semantic analysis.* Mahwah, NJ: Erlbaum

Leacock, C. & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities, 37,* 389–405.

Levinstein, I. B., Boonthum, C., Pillarisetti, S. P., Bell, C., & McNamara, D. S. (2007). iSTART 2: Improvements for efficiency and effectiveness. *Behavior Research Methods, 39,* 224–232.

Litman, D.J, Rose, C.P., Forbes-Riley, K., VanLehn, K., Bhembe, D., & Silliman, S. (2006). Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence in Education, 16,* 145-170.

Long, D. L., Golding, J. M., & Graesser, A. C. (1992). A test of the on-line status of goal-related inferences. *Journal of Memory and Language*, *31*, 634–647.

Magliano, J.P., & Millis, K.K. (2003). Assessing reading skill with a think-aloud procedure. *Cognition and Instruction*. *21*, 251-283.

Magliano, J., Millis, K., the RSAT development team, Levinstein, I., & Boonthum, C. (2009). Assessing comprehension during reading with the Reading Strategy Assessment Tool (RSAT). Manuscript submitted for publication.

Magliano, J.P., Todaro, S., Millis, K.K., Wiemer-Hastings, K., Kim, H.J., & McNamara, D.S. (2005). Changes in reading strategies as a function of reading training: A comparison of live and computerized training. *Journal of Educational Computing Research, 32*, 185–208.

McNamara, D.S. (in press). Computational methods to extract meaning from text and advance theories of human cognition. *Topics in Cognitive Science*.

McNamara, D. S. (2001). Reading both high and low coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology*, *55*, 51–62.

McNamara, D.S., Boonthum, C., Levinstein, I.B., & Millis, K. (2007). Evaluating self-explanations in iSTART: comparing word-based and LSA algorithms. In T. Landauer, D.S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 227-241). Mahwah, NJ: Erlbaum.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). The linguistic features of quality writing. *Written Communication. 27*, 57-86

McNamara, D.S., Jackson, G.T., & Graesser, A.C. (in press). Intelligent tutoring and games (ITaG). In Y.K. Baek (Ed.), *Gaming for classroom-based learning: Digital role-playing as a motivator of study*. IGI Global.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (in press). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*.

McNamara, D. S., & Magliano, J. P. (2009). Towards a comprehensive model of comprehension. In B. Ross (Ed), *The psychology of learning and motivation*, vol. 51 (pp 297-384). New York, NY, US: Elsevier Science.

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

McNamara, D.S., Louwerse, M.M., McCarthy, P.M., & Graesser, A.C. (in press). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes.*

McNamara, D.S., O'Reilly, T., Rowe, M., Boonthum, C., & Levinstein, I.B. (2007). iSTART: A web-based tutor that teaches self-explanation and metacognitive reading strategies. In D.S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies (pp. 397–421).* Mahwah, NJ: Erlbaum.

Miller, G.A., Beckwith, R.. Fellbaum, C.D., Gross, D., *Miller, K. (1990). WordNet*: An online lexical database. *International Journal of Lexicography, 3-4,* 235-244.

Millis, K.K., & Magliano, J. (in press). Assessing comprehension processes during reading. In L. Albro and J. Sabatini (Eds.), *Assessing Reading in the 21stCentury: Aligning and Applying Advances in the Reading and Measurement Sciences*.

Millis, K.K., Magliano, J., Todaro, S (2006). Measuring Discourse-Level Processes with Verbal Protocols and Latent Semantic Analysis. *Scientific Studies of Reading, 10,* 225-240.

Millis, K., Kim, H. J., Todaro, S. Magliano, J. P., 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.

Mostow, J. (2008). Experience from a Reading Tutor that listens: Evaluation purposes, excuses, and methods. In C. K. Kinzer, L. Verhoeven (Eds.), *Interactive Literacy Education: Facilitating Literacy Environments Through Technology (pp. 117-148).* Mahwah, NJ: Erlbaum

Newman, M. L., Groom, C. J., Handleman, L. D., & Pennebaker, J.W. (in press). Sex differences in language use: An analysis of text samples from 70 studies. *Discourse Processes*

Olney, A. M., Graesser, A.C., & Person, N.K. (in press). Tutorial dialogue in natural language. In R. Mizoguchi, J. Bourdeau, and R. Nkambou (Eds.) *Advances in Intelligent Tutoring systems*. Amsterdam: Springer.

Osgood, C. E., May, W. H., & Miron, M. (1975). *Cross-cultural universals of affective meaning*. Urbana: University of Illinois Press.

Page, E.B., & Petersen, N.S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76*, 561-565.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135.

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

Pennebaker, J., Mehl, M., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54*, 547-577.

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.

Pon-Barry, H., Clark, B., Schultz, K., Bratt, E. O., & Peters, S. (2004). Advantages of spoken language interaction in tutorial dialogue systems. In J. C. Lester, R. M. Vicari, & F. Paraguacu (Eds.), *Proceedings of the 7th International Conference on Intelligent Tutoring Systems* (pp. 390-400). Berlin: Springer-Verlag.

Ross, J.A., Rolheiser, C., & Hogaboam-Gray, A. (1999). Effects of self-evaluation training on narrative writing. *Assessing Writing, 6,* 107-132.

Rude, S. S., Gortner, E. M., & Pennebaker, J. W. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion, 18,* 1121-1133.

Rudner, L.M., Garcia, V., Welch, C. (2006). An evaluation of the IntelliMetric essay scoring system. *Journal of Technology, Learning and Assessment, 4*, 1-22.

Rus, V., & Graesser, A.C. (2006). Deeper natural language processing for evaluating student answers in intelligent tutoring systems. In the Proceedings of the American Association of Artificial Intelligence. Menlo Park, CA: AAAI.

Rus, V., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2008). A study of textual entailment. *International Journal on Artificial Intelligence Tools*, *17*, 659–685.

Russell, J. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110,* 145-172.

Samsonovich, A., & Ascoli, G. (2006). Cognitive map dimensions of the human value system extracted from natural language. In B. Goertzel & P. Wang (Eds.), *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms* (pp. 111–124). Amsterdam: IOS Press.

Shermis, M.D., & Burstein, J. (2003)(Eds.), *Automated essay scoring: A cross-disciplinary perspective.* Hillsdale, NJ: Erlbaum.

Shermis, M.D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In E. Baker, B. McGaw and N. S. Petersen (Eds.), International Encyclopedia of Education (Third edition). Oxford, UK: Elsevier.

Sleeman D. & J. S. Brown. (1982)(Eds.). *Intelligent Tutoring Systems*. Orlando, Florida: Academic Press, Inc.

Strapparava, C., & Valitutti, A. (2004). *WordNet-Affect: an affective extension of WordNet.* Paper presented at the Proceedings of the International Conference on Language Resources and Evaluation, Lisbon, Portugal.

Stirman, S.W., & Pennebaker, J.W. (2001). Word use in the poetry of suicidal and non-suicidal poets. *Psychosomatic Medicine 63,* 517-522.

Streeter, L., Psotka, J., Laham, D., & MacCuish, D. (2002). The credible grading machine: essay scoring in the DOD. The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC).

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education, 2*, 319-330.

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.

VanLehn, K., Jordan, P., Rosé, C. P., et al. (2002). The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In S. A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Intelligent Tutoring Systems: 6th International Conference* (pp. 158-167). Berlin: Springer.

Wiebe, J., Wilson , T., Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation, 39,* 165-210.

Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. (1999). Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In S. P. Lajoie & M. Vivet (Eds.), *Proceedings of the 9th International Conference on Artificial Intelligence: Artificial intelligence in education* (pp. 535–542). Amsterdam: IOS Press.

Woolf, B.P. (2009). *Building intelligent interactive tutors*. Burlington, MA: Morgan Kaufmann Publishers.

Zechner, K., Higgins,D., Xi, X., &Williamson, D.M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*.

Zolnay, A., Kocharov, D., Schluter, R., & Ney, H. (2007). Using multiple acoustic feature sets for speech recognition. *Speech Communication, 49*, 514-525.

Author Notes

The research on was supported by the National Science Foundation (ITR 0325428, BCS 0904909, ALT-0834847, ALT-0834847, DRK-12-0918409) and the Institute of Education Sciences (R305H050169, R305B070349, R305A080589, R305A080594). 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 or IES. The Tutoring Research Group (TRG) is an interdisciplinary research team comprised of researchers from psychology, computer science, physics, and education (visit http://www.autotutor.org, http://emotion.autotutor.org, http://fedex.memphis.edu/iis/-). 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.