

Assessing *Entailer* with a Corpus of Natural Language From an Intelligent Tutoring System

Philip M. McCarthy, Vasile Rus, Scott A. Crossley, Sarah C. Bigham, Arthur C. Graesser, &
Danielle S. McNamara

Institute for Intelligent Systems
University of Memphis
Memphis, TN 38152
{pmmccrth, vrus, scrossley, sbigham, a-graesser, ds McNamara} @ memphis.edu

Abstract

In this study, we compared *Entailer*, a computational tool that evaluates the degree to which one text is entailed by another, to a variety of other text relatedness metrics (LSA, lemma overlap, and *MED*). Our corpus was a subset of 100 self-explanations of sentences from a recent experiment on interactions between students and iSTART, an Intelligent Tutoring System that helps students to apply metacognitive strategies to enhance deep comprehension. The sentence pairs were hand coded by experts in discourse processing across four categories of text relatedness: *entailment*, *implicature*, *elaboration*, and *paraphrase*. A series of regression analyses revealed that *Entailer* was the best measure for approximating these hand coded values. The *Entailer* could explain approximately 50% of the variance for entailment, 38% of the variance for elaboration, and 23% of the variance for paraphrase. LSA contributed marginally to the entailment model. Neither lemma-overlap nor *MED* contributed to any of the models, although a modified version of *MED* did correlate significantly with both the entailment and paraphrase hand coded evaluations. This study is an important step towards developing a set of indices designed to better assess natural language input by students in Intelligent Tutoring Systems.

Introduction

Over the last three decades, researchers have made important progress in developing Intelligent Tutoring Systems (ITS) that implement systematic techniques for promoting learning (e.g., Alevin & Koedinger, 2002; Gertner & VanLehn, 2000; McNamara, Levinstein, & Boonthum, 2004). Such techniques include fine-grained *student modeling* that track particular knowledge states and conceptual misconceptions of the learners and that adaptively respond to the knowledge being tracked. The accuracy of such responses is critical and depends on the interpretation of the natural language user input. This interpretation is generally calculated through textual measures of relatedness such as latent semantic analysis (LSA; Landauer & Dumais, 1997; Landauer et al., 2006) or content word overlap metrics (Graesser, McNamara, et al., 2004). Metrics such as these have been incorporated into the user modeling component of ITSs based on results

from a wide variety of successful previous applications (e.g., essay grading, matching text to reader, text register disambiguation). However, the user input of ITS needs to be evaluated at a deeper level and to account for word order, including syntax, negation, and semantically well-formed expression. This study compared one deeper approach to evaluating user input of ITS, the *Entailer* (Rus, McCarthy, & Graesser, 2006), with an array of other textual relatedness measures using texts taken from ITS interactions.

Five Major Problems with Assessing Natural Language User Input in ITS

Text length. Text length is a widely acknowledged confound that needs to be accommodated by all text measuring indices. The performance of syntactic parsers critically depends on text length, for example (Jurafsky & Martin, 2000). As another example, lexical diversity indices (such as type-token ratio) are sensitive to text length because as the length of text increases the likelihood of new words being incorporated into the text decreases (McCarthy & Jarvis, in press; Tweedie & Baayen, 1998). This length problem is similar for text relatedness measures such as LSA and overlap-indices: Given longer texts to compare, there is a greater chance that similarities will be found (Dennis, 2006; McNamara, Ozuru, et al., 2006; Penumatsa et al., 2004; Rehder et al. 1998). As a consequence, the analysis of short texts, such as those created in ITS environments, appears to be particularly problematic (Wiemer-Hastings, 1999). The upshot of this problem is that longer responses tend to be judged by the ITS as closer to an ideal set of answers (or *expectations*) retained within the system. Consequently, a long (but wrong) response can receive more favorable feedback than one that is short (but correct).

Typing errors. It is unreasonable to assume that students using ITS should have perfect writing ability. Indeed, student input has a high incidence of misspellings, typos, grammatical errors, and questionable syntactical choices. Current relatedness indices do not cater to such eventualities and assess a misspelled word as a very rare

word that is substantially different from its correct form. When this occurs, relatedness scores are adversely affected, leading to negative feedback based on spelling rather than understanding of key concepts.

Negation. For measures such as LSA and content word overlap, the sentence *the man is a doctor* is considered very similar to the sentence *the man is not a doctor*, although semantically the sentences are quite different. Antonyms and other forms of negations are similarly affected. In ITSs, such distinctions are critical because inaccurate feedback to students can seriously affect motivation (Graesser, Person, & Magliano, 1995).

Syntax. For both LSA and overlap indices, *the dog chased the man* and *the man chased the dog* are viewed as identical. ITSs are often employed to teach the relationships between ideas (such as causes and effects), so accurately assessing syntax is a high priority for computing effective feedback.

Asymmetrical issues. Asymmetrical relatedness refers to situations where sparsely-featured objects are judged as less similar to general- or multi-featured objects than *vice versa*. For instance, *poodle* may indicate *dog* or *Korea* may signal *China* while the reverse is less likely to occur (Tversky, 1977). The issue is important to text relatedness measures, which tend to evaluate lexico-semantic relatedness as being equal in terms of reflexivity. Intelligent Tutoring Systems need to understand such differences and distinguish the direction of relationships. Thus, accurate feedback can be given to students depending on whether they are generalizing a rule from specific points (summarizing) or making a specific point from a general rule (elaborating).

Computational Approaches to Assessing Text Relatedness

Established text relatedness metrics such as LSA and overlap-indices have proven to be extremely effective measures for a great variety of the systems we have developed that analyze natural language and discourse, such as Coh-Matrix (Graesser, McNamara et al., 2004), iSTART (McNamara, Levinstein, & Boonthum, 2004), and AutoTutor (Graesser, Chipman et al, 2005; VanLehn, Graesser et al., in press). Despite such successes, there remains the potential for new measures of textual assessment to augment existing measures and thereby better assess textual comparisons. In this study, we assess a variety of textual relatedness assessment metrics. Each of these measures provides unique approaches to assessing the relatedness between text fragments.

Latent Semantic Analysis. LSA is a statistical technique for representing world knowledge based on large corpora of texts. LSA uses a general form of factor analysis (singular value decomposition) to condense a very large corpus of texts to 300-500 dimensions. These dimensions represent how often a word (or group of words) co-occurs across a range of documents within a large corpus (or space). Unlike content overlap indices, LSA affords

tracking words that are semantically similar, even when they may not be morphologically similar.

Content Overlap Indices. Content overlap indices assesses how often a common noun exists between two sentences. While such measures may appear shallow and lack the semantic relatedness qualities of LSA, they are used widely and have been shown to aid in text comprehension and reading speed (Kintsch & Van Dijk, 1978). As a measure of co-referentiality, content overlap indices also measure redundancy between sentences, which is important in constructing linguistic connections between sections of text (Haber & Haber, 1981). In this study, we focus on lemma-overlap, which allows plural and singular noun forms to be treated as one lexical item.

Minimal Edit Distances (MED). *MED* is a computational tool designed to evaluate text relatedness by assessing the similarity of strings across texts. *MED* is a combination of measuring Levenshtein distances (1966) and string theory matching (Dennis, 2006). Essentially, *MED* functions like a spellchecker; that is, it looks for the shortest route through which to match two strings. The evaluations work through a set of *costs*: shifting the string (right or left) has a cost of one; deleting a character costs one; and inserting a character costs one. *MED* scores are continuous, with a score of zero representing an identical match. For example, Table 1 shows a variety of string matching evaluations.

| | Mean words | Mean string | <i>MED</i> |
|--|------------|-------------|------------|
| The dog chased the cat. | 5.0 | 44.0 | 0.0 |
| The cat chased the dog. | 5.0 | 44.0 | 6.0 |
| The cats chased the dogs. | 5.0 | 52.0 | 13.0 |
| The cat didn't chase the dog. | 5.5 | 68.0 | 23.0 |
| Elephants tend to be larger than mice. | 6.0 | 104.0 | 43.0 |

Table 1. *MED* Evaluations of Five Input Sentences to a Target Sentence of "The dog chased the cat."

MED has a number of advantages and disadvantages. Chief among the disadvantages is that *MED* recognizes highly similar graphic representations of words to be highly similar in semantic terms. Thus, *Med* judges *elephant* and *elegant* as more similar than *woman* and *lady*. Incorporating an online dictionary may address this issue in future developments (as with *Entailer*, see below). A second problem is text length: the longer the text, the greater the potential for differences. Consequently, *MED* values are highly correlated with text length. Addressing this problem, we hypothesize that once a *MED* value has passed a certain point (just beyond the mean of a typical corpus), that no meaningful relatedness exists between the two texts, regardless of the *MED* value given. Thus, only low *MED* values are predicted to be meaningful.

Despite such problems, *Med* has two major advantages. As Dennis (2006) points out, the primary benefit of comparing strings is that syntactical variation can be assessed. Thus, for *MED*, *the cat chased the dog* is

different from *the dog chased the cat* (see Table 1). A second advantage for *MED* directly addresses our task at hand: assessing authentic natural language input. *MED*'s weakness for recognizing *elephant* and *elegant* as similar is its strength for recognizing misspellings as being highly similar to target terms. Thus *elegant/elephant* rate a minimal difference for *MED*, whereas overlap indices and LSA would judge the two tokens a maximally different. This point is of particular importance when dealing with ITS where important terms and ideas are often difficult to spell; yet whether such ideas have been learned by the student may often end up being judged primarily by the spelling.

Entailer. The purpose of *Entailer* is to evaluate the degree to which one text is entailed by another text. *Entailer* is based on the industry approved testing ground of the *recognizing textual entailment* corpus (RTE; Dagan, Glickman, & Magnini, 2004-2005). *Entailer* uses minimal knowledge resources and delivers high performance compared to similar systems. The approach encompasses lexico-syntactic information, negation handling, and synonymy and antonymy embedded in a thesaurus (WordNet; Miller, 1995). *Entailer* addresses two forms of negation: *explicit* and *implicit*. Explicit negation is indicated in the text through surface clues such as *n't*, *not*, *neither*, and *nor*. Implicit negation, however, has no direct representation at surface level so we incorporate antonymy relations between words as encoded in WordNet. *Entailer* functions by having each pair of text fragments (assigned as text [T] and hypothesis [H]) mapped into two graphs, one for T and one for H, with nodes representing main concepts and links indicating syntactic dependencies among concepts as encoded in T and H, respectively. An entailment score, $\text{entail}(T,H)$, is then computed quantifying the degree to which the T-graph subsumes the H-graph. The score is the weighted sum of one lexical and one syntactic component. The lexical component expresses the degree of subsumption between H and T at word level, (i.e. vertex-level) while the syntactic component work does the same thing at syntactic-relationship level (i.e. edge-level)¹.

The derived *Entailer* score is so defined as to be non-reflexive, such that $\text{entail}(T,H)$ does not entail $\text{entail}(H,T)$. Our results in earlier studies have been promising and better than state-of-the-art solutions that use the same array of resources (e.g. Rus, Graesser et al, 2005; Rus, McCarthy et al, 2006). Our formula to obtain an overall score aims to deliver both a numerical value for the degree of entailment between T and H and a degree of confidence in our decision. The scores range from 1 (meaning TRUE entailment with maximum confidence) to 0 (meaning FALSE entailment with maximum confidence). There are three important components of the score: lexical or node matching, syntactic or relational matching, and negation. The score also plays the role of a confidence score necessary to compute a proposed confidence weighted

score metric (CWS). The CWS varies from 0 (no correct judgments at all) to 1 (perfect score), and rewards the system's ability to assign a higher confidence score to the correct judgments. Accuracy in terms of the fraction of correct responses is also reported.

For the purposes of natural language assessment in ITS, *Entailer* offers a number of advantages over current text relatedness measures such as LSA and overlap indices. First, because lexical/word information acts only as a component of the overall formula, *Entailer* is less susceptible to the problem of text length. In addition, as *Entailer* addresses both syntactical relations and negation, the tendency for higher relatedness results over lengthier texts is reduced. Second, *Entailer* addresses asymmetrical issues by evaluating text non-reflexively, so $\text{entscore}(H, T) \neq \text{entscore}(T,H)$. As such, the evaluation of a response (self explanation) to a stimulus (source text) will be different from the evaluation of the stimulus to the response. Third, *Entailer* handles negations so it offers the opportunity of providing more accurate feedback. Currently, *Entailer* is not equipped to handle problems such as misspellings and typos any more than other text relatedness measures. However, the current study provides evidence to suggest that results from *Entailer* may be sufficiently robust to render such concerns negligible.

ELIMENT: Elaboration, Implicature and Entailment

In order to test the four textual relatedness approaches outlined above, we created a natural language corpus of ITS user input statements (hereafter, the *ELIMENT* corpus). The corpus comprises a subset of data taken from the *Interactive Strategy Trainer for Active Reading and Thinking* (iSTART, McNamara et al, 2004). The primary goal of iSTART is to help high school and college students learn to use a variety of reading comprehension strategies. iSTART training culminates with students reading two short science passages during which they are asked to apply their newly learned strategies by typing *self-explanations* of key sentences. The iSTART *stimulus sentences* and the corresponding student *self-explanations* forms *the pairs* we refer to in this study.

The data *pairs* used to make the *ELIMENT* corpus were generated from a typical iSTART experiment. The experiment in question was run on 90 Shelby County Tennessee high-school students drawn from four 9th grade Biology classes (all taught by the same teacher). Overall, the experiment generated 826 sentence pairs, from which the *ELIMENT* corpus consisted of 100 randomly selected pairs. The average length of the combined sentence pairs was 16.65 words ($SD = 5.63$).

The terms we used to categorize the *ELIMENT* sentence pairs were based on general, linguistic definitions for *elaboration*, *implicature*, and *entailment*, hence *ELIMENT*. To these three primary aspects of textual relatedness assessment we also add an evaluation for *paraphrase* and *error* (see Table 2 for examples). Our criteria and definitions were based on the operational requirements of

¹ For a complete discussion see Rus, McCarthy, et al, 2006

the iSTART system (McNamara et al., 2004). It is important to make clear that the terms used in these criteria (such as *entailment* and *implicature*) remain the subject of discussion, and it is not the purpose of this study to settle such disputes. Examples of our terms used in *ELIMENT* are provided in Table 2.

| Category | Student Statement | Relationship to Source Sentence |
|-------------|--|---------------------------------|
| Entailment | John went to the store. | Explicit, logical implication |
| Implicature | John bought some supplies. | Implicit, reasonable assumption |
| Elaboration | He could have borrowed stuff. | Non-contradictory reaction |
| Paraphrase | He took his car to the store to get things that he wanted. | Reasonable re-statement |
| Error | John walked to the store. | Contradiction |

Table 2. Showing how Various Responses Would Be Categorized According to *ELIMENT* for the Sentence of "John drove to the store to buy supplies."

Entailment and Implicature. The distinction we employ between *entailment* and *implicature* is critical to providing accurate and appropriate feedback from the tutoring systems in which textual relatedness indices are incorporated. Essentially, we use entailment to refer to *explicit* textual reference whereas we use the term implicature to refer to references that are only *implied*. Our definition of implicature is similar to the *controlled knowledge elaborative inference* definition given in Kintsch (1993). Kintsch argued that the sentence pair "Danny wanted a new bike / He worked as a waiter," (pp 194-195) does not supply the specific (explicit) information in the text to know (to entail) that *Danny is working as a waiter to buy a bike*. However, Kintsch also argues that it would be quite typical for a reader to draw such a conclusion (inference) from the sentence pair.

An authentic example of the importance of the distinction was recently supplied during the 2006 Israeli/Lebanon conflict. At a news conference, the U.N. Secretary General, Kofi Annan, released the following statement: "While Hezbollah's actions are deplorable, and as I've said, Israel has a right to defend itself, the excessive use of force is to be condemned." Within the hour, the BBC reported Annan's statement as "[Annan] condemned Hezbollah for sparking the latest violence in the country, but also attacked Israel for what he called its 'excessive use

of force'". Asked to comment on the reports, White House spokesman, Tony Snow, pointed out that Annan's statement did not *entail* the BBC's remarks. That is, according to Snow, Annan had remarked only that "the excessive use of force is to be condemned", but he had not said that Israel, explicitly, was itself guilty of committing such excess. Such a distinction is reflected in *ELIMENT*, where the BBC's commentary would be considered *implicature* rather than *entailment*.

Elaboration, Paraphrase, and Error. The remaining categories of *ELIMENT* are less controversial. We use *elaboration* to refer to any recalled information that is generated as a *response* to the stimulus text without being a case of entailment or implicature. An elaboration may differ markedly from its textual pair provided it does not *contradict* either the text or world knowledge. In such an event, the text is considered under the category of *error*.

A *paraphrase* is a reasonable restatement of the text. Thus, a paraphrase tends to be an entailment, although an entailment does not have to be a paraphrase. For example, a sentence of *the dog has teeth* is entailed by (but not a paraphrase of) the sentence *the dog bit the man*.

An *error* is a response statement that *contradicts* the text or *contradicts* world knowledge. Thus, even if a statement differs substantially in theme or form from its corresponding sentence pair, it is evaluated as elaboration rather error. In this study we concentrate on the four categories of *relatedness*. We plan to address the *error* category in future research.

Methods

To assess the 100 pairs from the *ELIMENT* corpus, five experts working in discourse processing at the University of Memphis evaluated each sentence pair on the five dimensions of *ELIMENT*. Each pair (for each category) was given a rating of 1 (min) to 6 (max). A Pearson correlation for each inference type was conducted between all possible pairs of raters' responses. If the correlations of any two raters did not exceed .70 (which was significant at $p < .001$) the ratings were reexamined until scores were agreed upon by all the raters. Thus, the 100 pairs' corpus comprising *ELIMENT* were rated across the four categories of textual relatedness and a single mean score of the evaluations was generated for each of the four categories.

Results

Our evaluations of the *four* text relatedness indices consisted of a series of multiple regressions. The *ELIMENT* corpus of hand coded evaluations of *entailment*, *implicature*, *elaboration*, and *paraphrase* were dependent variables and the four relatedness indices were the independent variables. The results for the dependent variable of *Hand coded Entailment* from the *ELIMENT* corpus showed *Entailer* to be the most significant predictor. Using the *forced entry method* of linear regression, selected as a conservative form of multivariate analysis, a significant model emerged, $F(4, 95) = 26.15, p$

< .001. The model explained 50.4% of the variance (Adjusted $R^2 = .504$). *Entailer* was a significant predictor ($t = 9.61, p < .001$) and LSA was a marginal predictor ($t = -1.90, p = .061$). Neither Lemma nor *MED* were significant predictors. The results for the dependent variable of *Hand coded Implicature* were not significant, and no significant model emerged, $F(4, 95) = 0.40, p = .824$. The results for the dependent variable of *Hand coded Elaboration* again showed *Entailer* to be the most significant predictor. The model significantly fit the data, $F(4, 95) = 16.14, p < .001$, explaining 38.0% of the variance (Adjusted $R^2 = .380$). The *Entailer* was a significant predictor ($t = -7.98, p < .001$) whereas LSA, Lemma, and *MED* were not significant predictors. The model for the dependent variable of *Hand coded Paraphrase* also significantly fit the data, $F(4, 95) = 8.58, p < .001$, explaining 23.4% of the variance (Adjusted $R^2 = .234$). *Entailer* index was a significant predictor ($t = 5.62, p < .001$), whereas LSA, Lemma, and *MED* were not significant predictors. The results suggest *Entailer* is the most significant predictor of three of the four categories of *ELIMENT* textual relatedness. The remaining category of *implicature* was not well identified by the computational indices. The reason for this can be attributed to each of the computational indices inclusion of surface level text relatedness rather than the solely implicit relatedness assessed by the category of *implicature*.

Post Hoc Analyses

Addressing LSA Results. The relatively poor performance of the LSA measure might be explained by its previously mentioned sensitivity to text length. The correlation between Raw LSA and text length was $r = .33$ ($p = .001$). To address this problem, we factored out the sentence length effect using log equations similar to Maas (1972). As shown in McCarthy and Jarvis (in press), the log formula can be quite effective in redressing metric problems caused by short text lengths. Using $LSA_{log} = \text{raw LSA}/\log(\text{words})$ to correct for text length, we found that LSA_{log} correlated with raw LSA ($r = .96, p < .001$) but did not correlate with text length ($r = .11, p = .292$). However, substituting LSA log for raw LSA did not improve the model. Thus, more research is needed to assess whether LSA can significantly contribute to the textual assessment conducted in this study.

Addressing MED Results. The relatively poor performance of *MED* is probably also caused by its sensitivity to text length ($r = .55, p < .001$). In practice (and as shown above in Table 1), we know that *MED* scores can be quite informative if texts are relatively short and genuine lexical similarities exist. As such, we can have more confidence in lower *MED* scores being meaningfully representative of similarities than we can be in higher *MED* scores being representative of differences. That is, high *MED* scores are quite uninformative, whereas lower ones may provide useful information about the relatedness of the texts. As a *post hoc* analysis, we converted *MED* values to z-scores and adjusted the parameters for *MED*

such that increasingly lower values of *MED* were removed from our analyses. The results revealed that *MED* may indeed be quite informative. Specifically, when all values above 1SD were removed (14% of the data), *MED* significantly correlated with the hand coded entailment value ($r = -.32, p < .05$), the hand-coded paraphrase value ($r = -.27, p < .05$), and the *Entailer* output ($r = -.32, p < .05$). However, regression analyses focusing on this subset of *MED* values did not produce a significant difference in any of the models. Once again then, more research is needed to assess the degree to which *MED* can significantly contribute to the kind of textual assessment conducted in this study.

Discussion

In this study, we compared *Entailer*, a computational tool that evaluates the degree to which one text is entailed by another, to a variety of other text relatedness metrics (LSA, lemma-overlap, and *MED*). Our corpus (*ELIMENT*) was formed from a subset of 100-sentence self-explanations from a recent iSTART experiment. The *ELIMENT* sentence pairs were hand coded by experts in discourse processing across four categories of text relatedness: *entailment*, *implicature*, *elaboration*, and *paraphrase*. A series of regression analyses suggested that the *Entailer* was the best measure for approximating these hand coded values. The *Entailer* explained approximately 50% of the variance for entailment, 38% of the variance for elaboration, and 23% of the variance for paraphrase. LSA marginally predicted entailment. Neither lemma-overlap nor *MED* predicted any of the four categories of text relatedness although a modified version of *MED* did correlate significantly with both entailment and paraphrase hand coded evaluations.

Previous research has shown that *Entailer* delivers high performance analyses when compared to similar systems in the industry approved testing ground of *recognizing textual entailment* tasks (Rus, McCarthy, et al., 2006; Rus, Graesser, et al., 2005). However, the natural language input from the *ELIMENT* corpus (with its spelling, grammar, asymmetrical, and syntax issues) provided a far sterner testing ground. The results of this study suggest that in this environment too, the performance of *Entailer* has been significantly better than comparable approaches.

In future research, we will seek to better assess the parameters of the measures discussed in this study. That is, certain measures are geared more to evaluate certain categories of similarities over others. As such, we want to assign confidence values to measures so as to better assess the accuracy of our models. In addition, establishing parameters will more easily accommodate categories of prediction for our indices that will allow the reporting of *recall* and *precision* evaluations.

This study builds on the recent major developments in assessing text relatedness indices, particularly the focus of incorporating strings of indices designed to better assess natural language input in Intelligent Tutoring Systems (Dennis, 2006; Landauer et al 2006; Rus et al., 2006).

More accurate assessment metrics are necessary so as to better assess input, and from this input supply the most optimal feedback to students. This study offers promising developments in this endeavor.

Acknowledgements

This research was supported by the Institute for Education Sciences (IES R305G020018-02) and partially by the National Science Foundation (REC 0106965 and ITR 0325428). The authors also acknowledge the contributions made to this project by Stephen W. Briner, Erin J. Lightman and Adam M. Renner.

References

- Aleven, V., and Koedinger, K. R. 2000. An effective meta-cognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science* 26: 147-179.
- Dagan, I., Glickman, O., and Magnini, B. 2004 - 2005. Recognizing textual entailment. *Pattern Analysis, Statistical Modelling and Computational Learning*. Retrieved February, 14, 2006 from <http://www.pascalnetwork.org/Challenges/RTE>.
- Dennis, S. 2006. Introducing word order in an LSA framework. In *Handbook of Latent Semantic Analysis*. T. Landauer, D. McNamara, S. Dennis and W. Kintsch eds.: Erlbaum.
- Gertner, A. S., and VanLehn, K. 2000. Andes: A coached problem solving environment for physics. In *Intelligent Tutoring Systems: 5th International Conference, ITS 2000*. G. Gauthier, C. Frasson, and K. VanLehn. eds. 133-142. New York: Springer.
- Graesser, A.C., Chipman, P., Haynes, B.C., and Olney, A. 2005. AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions in Education* 48, 612-618.
- Graesser, A., McNamara, D., Louwerse, M., and Cai, Z. 2004. *Coh-Matrix*: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers* 36, 193-202.
- Graesser, A. C., Person, N. K., and Magliano, J. P. 1995. Collaborative dialogue patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology* 9: 495-522.
- Haber, R. N., and Haber, L. R. 1981. Visual components of the reading process. *Visible Language* 15: 147-182.
- Jurafsky, D. S., and Martin, J. H. 2000. *Speech and language processing*. Englewood, NJ: Prentice Hall.
- Kintsch, W. 1993. Information Accretion and Reduction in Text Processing: Inferences. *Discourse Processes* 16: 193-202.
- Kintsch, W., and Van Dijk, T.A. 1978. Toward a model of text comprehension and production. *Psychological Review* 85: 363-394.
- Landauer, T. K., and Dumais, S. T. 1997. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review* 104: 211-240.
- Landauer, T., McNamara, D. S., Dennis, S., and Kintsch, W. eds. 2006. *LSA: A road to meaning*. Mahwah, NJ: Erlbaum.
- Maas, H.D. 1972. Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. *Zeitschrift für Literaturwissenschaft und Linguistik* 8: 73-79.
- McCarthy, P.M., and Jarvis, S. (in press). A theoretical and empirical evaluation of *vocd*. *Language Testing*.
- McNamara, D. S., Levinstein, I. B., and Boonthum, C. 2004. iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instrument, & Computers* 36: 222-233.
- McNamara, D. S., O'Reilly, T., Best, R. and Ozuru, Y. 2006. Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research* 34:147-171.
- Miller, G.A. 1995. WordNet: A lexical database for English. In *Communications of the ACM* 38: 39-41.
- Penumatsa, P., Ventura, M., Graesser, A.C., Franceschetti, D.R., Louwerse, M., Hu, X., Cai, Z., and the Tutoring Research Group 2004. The right threshold value: What is the right threshold of cosine measure when using latent semantic analysis for evaluating student answers? *International Journal of Artificial Intelligence Tools* 12: 257-279.
- Rehder, B., Schreiner, M., Laham, D., Wolfe, M., Landauer, T., and Kintsch, W., 1998. Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes* 25: 337-354.
- Rus, V., Graesser, A.C., McCarthy, P.M., and Lin, K. 2005. A Study on Textual Entailment, *IEEE's International Conference on Tools with Artificial Intelligence*. Hong Kong.
- Rus, V., McCarthy, P.M., and Graesser, A.C., 2006. Analysis of a textual entailment, *International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico.
- Tversky, A. 1977. Features of similarity. *Psychological Review* 84: 327-352.
- Tweedie F. and Baayen R. H. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities* 32: 323-352.
- VanLehn, K., Graesser, A.C., Jackson, G.T., Jordan, P., Olney, A., and Rose, C.P. (in press). When are tutorial dialogues more effective than reading? *Cognitive Science*.
- Wiemer-Hastings, P., Wiemer-Hastings, K., and Graesser, A. 1999. Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In S.P. Lajoie Lajoie and M. Vivet, *Artificial Intelligence in Education*. 535-542. Amsterdam: IOS Press.