# Computational Replication of Human Paraphrase Assessment

Philip M McCarthy, Zhigiang Cai, & Danielle S. McNamara

University of Memphis

{pmmccrth, zcai, dsmcnamr}@memphis.edu

**Abstract**

Two sentences are paraphrases if their meanings are equivalent but their words and syntax are different. Paraphrasing can be used to aid comprehension, stimulate prior knowledge, and assist in writing skills development. While *automated paraphrase assessment* is both common-place and useful, research has centered solely on artificial, edited paraphrases and has used only binary dimensions (i.e., *is* or *is-not* a paraphrase). In this study, we use 1998 natural paraphrases generated by high school students that have been assessed along 10 dimensions of paraphrase (e.g., semantic completeness). This study investigates the components of paraphrase quality emerging from these dimensions, and examines whether computational approaches (e.g. LSA, MED) can simulate those human evaluations. The results suggest that semantic and syntactic evaluations are the primary components of paraphrase quality, and that computationally light systems such as LSA (semantics) and MED (syntax) present promising approaches to simulating human evaluations of paraphrases.

## Introduction

Paraphrasing is the restating of a sentence such that the meaning of both sentences would generally be recognized as lexically and syntactically different while remaining semantically equal. Paraphrasing is an important issue in fields that center on reading and writing. For example, paraphrasing text can facilitate reading comprehension by transforming the text into a more familiar construct or by activating relevant prior knowledge (McNamara 2004; McNamara et al. 2007). And, in the field of composition, paraphrasing allows writers to restate ideas from other works or their own drafts so that the reformatted language may better suit a voice, flow, or line of argument (Hawes 2003).

Paraphrasing is undoubtedly useful to developing reading and writing skills. Not surprisingly then, intelligent tutoring systems (ITS) that aim to teach reading and writing strategies have seen the need to develop some level of computational paraphrase assessment. Thus, we need computational algorithms that can judge the quality and other characteristics of a user's attempts to paraphrase sentences. But these algorithms need to be both fast and accurate. A system that operates too slowly in providing assessment and subsequent feedback can frustrate users, leading to lower engagement with the system (Rus et al. 2008). More specifically, users typically expect systems to respond within the boundaries of a normal conversational turn, about one-second (Lockelt et al. 2007). Such a constraint severely limits programming options that might lead to greater accuracy. Nonetheless, the accuracy of the judgment is equally important. Accuracy is important because misleading or misdirected system feedback based on the evaluation risks compromising user motivation and metacognitive awareness of the system's learning goals (Millis et al. 2007). Thus, a paraphrase assessment must operate within a trade off between speed and accuracy.

While speed and accuracy remain key elements of paraphrase assessment, a potentially greater problem facing system developers is the lack of appropriate paraphrase data upon which to train systems. Most research on computational assessment of paraphrasing (e.g., Rus, Lintean, et al. 2008) has centered on *edited* paraphrases stemming from professional writers in data collections such as the Microsoft Paraphrase Corpus (Dolan, Quirk, and Brocket 2005). Data such as this has a rich history of utility for developing approaches to paraphrase assessment in applications such as *natural language generation* (Iordanskaja, Kittredge, and Polgere 1991), *question answering* (Ibrahim, Katz, and Lin 2003), and *summarization* (Inderjeet 2001). While such research is undoubtedly valuable, we cannot escape the fact that these paraphrase systems are trained *on* edited text for application *to* edited text. ITS input is far from edited. Indeed, the primary characteristic of ITS input is its propensity for unusual typographical and grammatical choices (McCarthy and McNamara 2008). Indeed, as McCarthy, Guess, and McNamara (in press) point out, less than 12% of student input can be assumed to be free of any form of written error.

Our final and possibly most important concern with existing paraphrase data sets is that their expert (human) evaluations tend to be coarse-grained. Specifically, existing paraphrase data tends to be binary coded as either *is* a paraphrase or *is-not* a paraphrase. Such categorization is perhaps understandable if the purpose of paraphrase identification is question answering, data retrieval, or text

Copyright © 2009, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved

summarization, in which a system may retrieve many possible candidate texts for further action and allow a (presumably expert) user to choose among a list of options. In the case of an ITS, however, paraphrasing is often the subject being taught, and so the system may have to choose what is the best candidate among a list of possible candidates and/or supply feedback to the (presumably *not* expert) user as to why such a selection was made.

In sum, we can assert the following: Paraphrasing is a useful strategy for both reading and writing development. There are ITSs that seek to teach students how to paraphrase. Facilitative feedback to students based on ITS training depends on accurate and timely computational assessment. And, the development and training of computational techniques for this assessment of paraphrasing has been based on text data that is far from characteristic of the input typical to ITSs.

Such assertions led us to two major research questions: (1) What are the components of paraphrase? That is, which paraphrase dimensions constitute paraphrase quality?; and (2) Can computationally light systems (i.e., systems that can process and evaluate input within one-second) assess paraphrase quality to a similar degree as that of humans?

## Experiment One: Human

To begin to address the issues and questions outlined above, we use in this study the User-Language Paraphrase Challenge corpus (ULPC; McCarthy and McNamara 2008). The ULPC is a corpus of 1998 paraphrases written by American high-school students using an intelligent tutoring system. The paraphrases in the corpus are rated on a 6-point scale by human experts across 10 dimensions of paraphrase. In this study, we focus on four of these dimensions, leaving out less directly related dimensions such as *garbage content*, *irrelevant response*, *degree of entailment*, *elaborative response*, and *presence of frozen expressions* (e.g. *The sentence is saying that …*).

**Semantic completeness.** Semantic completeness refers to the student's paraphrase having the same meaning as the sentence targeted for paraphrasing. Semantic completeness is evaluated without regard to word- or structural-overlap between sentences. For example, *During vigorous exercise, the heat generated by working muscles can increase total heat production in the body markedly***,** was evaluated highly for the user response of *exercising vigorously icrease mucles total heat production markely in the body*.

**Lexical similarity.** Lexical similarity is the degree to which the same words are in both sentences, regardless of syntax or semantics. Thus, for this index, *the dog chased the cat* is identical to *the cat chased the dog*.

**Syntactic similarity.** Syntactic similarity is the degree to which similar parts of speech and phrase structures are employed in the user response, regardless of the words used. For example, the sentence *the bad dog chased the quiet cat* would be syntactically the same as *the large elephant thumped the little mouse*.

**Paraphrase Quality.** Paraphrase quality refers to an overarching evaluation of the user response. Evaluators could take into account any of the dimensions of paraphrase (or yet other qualities) to any degree they thought appropriate.

## Results

Following ULPC guidelines, the data was divided into training sets (67%) and test sets (33%). Correlations (see Table 1) were computed for the human evaluated training set to examine the relationships between the variables and to determine which variables showed the strongest relationships with paraphrase quality. The correlations indicated that semantic completeness and lexical similarity show the strongest relationships to paraphrase quality.

Table 1: Correlation matrix for the variables of paraphrase quality, Semantic completeness, lexical similarity, and syntactic similarity between each sentence in the paraphrase

| | Semantic | Lexical | Syntactic |
|---|---|---|---|
| Quality | 0.774* | 0.451* | 0.035 |
| Semantic | | 0.673* | 0.421* |
| Lexical | | | 0.579* |

Note: * = Significant at $p < .001$

The results also indicated that paraphrase quality was not significantly correlated with syntactic similarity. This lack of correlation was due to a curvilinear relationship between syntactic similarity and paraphrase quality, for which an S-curve best fit the data ($R^2 = .12$, $d.f. = 1010$, $F = 134.57$, $p < .001$). This curvilinear relationship contrasts with the linear relationships of semantic completeness and lexical similarity to paragraph quality. The curvilinear relationship between paraphrase quality and syntactic similarity suggests that both low and high evaluations of paraphrase quality are associated with low values of syntactic similarity. Thus, a paraphrase that is not at all related to a target sentence is syntactically different from the attempted paraphrase; and a paraphrase that is of high quality is also syntactically different from the target sentence.

We conducted a hierarchical multiple regression analysis to determine the amount of variance associated with paraphrase quality that was explained by the three predictor variables. Semantic completeness was entered as the first predictor variable, which accounted for 60% of the variance associated with paraphrase quality. When lexical similarity was included as a second variable, it predicted only 0.1% additional variance. This is likely due to the high correlation

between lexical similarity and semantic completeness. The best model emerged when syntactic similarity was entered as the second predictor variable. A significant model emerged: $F$ (2, 1009) = 1190.325, $p$ < 001, $r$ = .838; adjusted $R^2$ = .702, explaining 70% of the variance. Thus, syntactic similarity predicted 10% additional variance after semantic completeness was entered.

Semantics and syntax are prominent and explicit textual components of paraphrase quality evaluation that are likely to feature in any or most definitions. However, other components of paraphrase evaluation are also possible and may have to be considered. For example, the perception of the *quality of the writing* or the *length of paraphrase relative to the target sentence* may affect ratings. Such factors are not likely to be ignored in an evaluation of quality, even while they may not explicitly feature in a definition of paraphrase. Writing quality could feature because poor spelling or grammar may imply the meaning of the paraphrased sentence is more distant from the target sentence. Indeed, examining the corpus of paraphrases, 1761 of the 1998 (or 88%) reported some kind of grammatical or spelling error. And, length of response relative to the target sentence could also affect ratings because obviously longer or shorter responses are unlikely to yield the same meaning.

Correlation results supported our hypotheses of these two additional features, with significant results for both paraphrase quality and writing quality ($r$ = .509, $p$ < .001) and paraphrase quality and length difference between sentences ($r$ = -.374, $p$ < .001). The positive correlation for writing quality suggests that people who are better writers may paraphrase better. The negative correlation for the latter indicates that the greater the difference in length between sentences in the paraphrase, the lower the rating.

Given these significant correlations, both variables were added to the model. The contribution of writing quality to the model was significant; however, the $R^2$ change was small (.016). The length difference variable was not significant. With the addition of the writing quality component, the model explained 72.2% of the variance (adjusted $R^2$ = .722).

### Testing the Validity of the Model

To test the validity of the model, we generated a new composite variable based on the B-weights of the model generated from the training set data. Lexical similarity was not included because its role appeared to be subsumed by semantic completeness. The length difference variable was not significant but was retained in the model. We retained length because our goal is to replicate the human model with computational variables, and the length variable is highly objective computationally. Ideally, we would use a computational variable for writing quality; however, no simple solution for that variable was available. We discuss this issue further in the computational section.

The new composite variable (i.e., the training model applied to the test set data) significantly correlated with paraphrase quality ($r$ = .866, $p$ < .001, $n$= 649), explaining 75% of the variance. Removing the writing quality and length variables from the model did not result in a significant change ($r$ = .857, $p$ < .001, $n$= 649). The high correlations from the test set data results suggest that paraphrase quality may largely comprise the components of semantics and syntactical change, with the components of writing quality and length being minor factors. The result is important because computationally measuring a construct such as paraphrase quality presents challenges, foremost simply in definition. However, if the components are more easily defined (e.g., semantics and syntax) then computational assessment becomes more easily directed.

## Experiment Two: Computational

Our second research question asked *Can computationally light systems such as LSA assess paraphrase quality to a similar degree to that of humans?* The UPLC corpus provides evaluations of 10 computational indices, although several of them are quite shallow (e.g., sentence length). For this study, we were primarily interested in the computational indices with a rich history of textual similarity assessment (i.e., LSA: Landauer et al. 2007) and the Entailer (Rus et al. 2008); and two approaches to syntactic similarity assessment Minimal Edit Distances (MED: McCarthy et al. 2008) and the Coh-Metrix Structural Similarity Index (STRUT, Graesser et al. 2004). The ULPC invites researchers to consider all measures given in the challenge and/or new measures. In this study we extend MED to two new indices (see below) and include STRUT as a syntax index to compare to MED. Given that our human analysis suggested that paraphrase quality was primarily a construct of semantics and syntax, these indices seem to be an appropriate place to assess computational replication. Brief descriptions of each of these indices and their reason for inclusion in this study are given below.

**Latent Semantic Analysis.** LSA is a statistical technique for representing word similarity. It is based on occurrences of lexical items within a large corpus of text. LSA is able to judge semantic completeness even while morphological similarity may differ markedly. LSA is an ideal candidate for paraphrase quality evaluation because it can assess the semantic similarity of any two texts.

**The Entailer.** The Entailer is a lexico-syntactic approach to entailment evaluation. The Entailer is based on word and structure similarities that are evaluated through graph subsumption. The approach has been highly successful in standardized entailment testing, in edited paraphrase testing assessment (Rus et al. 2008), and even in user-language paraphrase assessment testing (McCarthy et al. 2008). As *an entailment measure*, we can assume that the second sentence in any pair is shorter or the same length as the target

sentence because that which is entailed is likely to contain less information than that which is entailing. In paraphrasing, the reverse is more likely to be the case. That is, the first sentence is a *more or less* ideal form of the sentence. To rephrase the sentence (especially considering that *the rephraser* is a non-expert) often requires more lexicon (and maybe more information) than is present in the target sentence. As such, in this study we use the Reverse Entailment index. This index assesses the degree to which the second sentence (i.e., the paraphrase attempt) entails the target sentence.

**Minimal Edit Distances (MED).** MED (McCarthy et al. 2008) assesses differences between any two sentences in terms of the words in the sentences and the position of the words in the sentences. As such, sentences with the same words may not be considered identical if the position of those words is different (see Table 2 for examples). Because MED assesses word similarity in terms of sentential positions, MED is hypothesized to be an ideal approach for syntactic similarity evaluation.

In this study, we extend MED from word position similarity assessment to syntax position similarity assessment. When MED considers only the lexical items in the sentence we refer to this index as MED (L). When the syntax in the sentence is considered, we refer to it as MED (S). When a combination of lexicon and syntax is used, we refer to this index as MED (LS). The syntax for the MED assessment was gathered for the 1998 paraphrases using the Charniak parser (Charniak, 2000). Having parsed the paraphrases, each item was analyzed using the MED tool. As such, the two sentences *the dog chased the cat* and *the cat chased the dog* receive low scores because there are fewer differences than between, for example, the two sentences *the dog chased the cat* and *why don't we go to the zoo?*, which receive high scores because they are both lexically and syntactically different.

Table 2: Examples of values for MED (L), MED (S), MED (LS), and Structural overlap (STRUT) for the target sentence, *The dog chased the cat*.

| Sentence | L | S | LS | STRUT |
|---|---|---|---|---|
| The dog chased the cat. | 0 | 0 | 0 | 0 |
| The cat chased the dog. | 0.2 | 0 | 0.1 | 0 |
| The cats chased the dogs. | 0.4 | 0.2 | 0.3 | 0.3 |
| The cat didn't chase the dog. | 0.7 | 0.4 | 0.5 | 0.6 |
| Elephants tend to be larger than mice. | 1 | 0.8 | 0.9 | 0.8 |

**Structural Overlap.** One previously unconsidered index of syntax is Coh-Metrix's *structural overlap* (STRUT). This index compares the syntactic tree structures of two adjacent sentences in a text; here, the sentences are the paraphrases. The algorithm builds a maximum intersection tree between two syntactic trees. The value of the index is the proportion of nodes in the two tree structures that are intersecting nodes (see Table 2). To compare with other indices, we used the formula $y = 1 - x$ to reverse the STRUT score to a difference index.

## Results

Using the training set data (n = 1012), results suggested moderate to high correlations for all the computational candidate indices (see Table 3).

Table 3: Correlations between leading indices (LSA, Entailer, MED, and STRUT) and the paraphrase dimensions of semantic completeness, lexical similarity, syntactic similarity, and paraphrase quality in rank order of highest to lowest correlation.

| | **Order of highest correlation** | | |
|---|---|---|---|
| Dimension | First | Second | Third |
| Paraphrase Quality | LSA 0.427 | Entailer 0.319 | MED (S) -0.162 |
| Semantic Completeness | Entailer 0.581 | LSA 0.575 | MED (S) -0.416 |
| Lexical Similarity | LSA 0.818 | Entailer 0.800 | MED (L) -0.580 |
| Syntactic Similarity | MED (L) -0.742 | STRUT 0.602 | Entailer 0.584 |

The best performing variable for the dimension of semantic completeness was the Entailer, $r = .58$; however, the correlation was significantly lower than that produced by human expert-to-expert correlations (compare human: $r = .74$; z-diff = 5.72, $p < .001$). The best performing variable for the dimension of lexical similarity was LSA, $r = .82$; a result that was significantly higher than that produced by human expert-to-expert correlations (compare human: $r = .67$; z-diff = 7.02, $p < .001$). The best performing variable for the dimension of syntactic similarity was MED (L), $r = -.74$ (note that MED calculates *differences*, hence the negative correlation); another result that was significantly higher than that produced by human expert-to-expert correlations (compare human: $r = .51$; z-diff = 7.83, $p < .001$). STRUT also correlated with syntactic similarity ($r = .602$); however, its result was significantly lower than MED (*z-diff* = 5.420, $p < .001$), although the two computational indices themselves correlated highly ($r = .620$, $p < .001$).

Thus, our first computational finding of importance is that MED (Lexical) outperforms rival syntax indices and even human evaluations of syntactic similarity. Finally, the best performing variable for the overall dimension of paraphrase Quality was LSA, $r$ = .43, a result that was significantly lower than that produced by human expert-to-expert correlations (compare human: $r$ = .59; z-diff = 4.35, $p$ < .001). Therefore, initial correlations suggest that computational tools are *comparable* to human expert evaluations; and specifically, that in terms of lexical and syntactical similarity, they are significantly better than human expert agreement, but that human agreement is significantly higher for the qualities of semantic completeness and overall paraphrase quality.

To attempt to replicate the human ratings, we conducted a hierarchical multiple regression analysis with paraphrase quality as the dependent variable and the computational indices of LSA (for semantics) and MED (for syntax) as predictor variables. Using LSA as the first predictor variable and MED (L) as the second predictor variable, a significant model emerged: $F(2, 1009) = 146.359$, $p < 001$. The model explained 22% of the variance (Adjusted $R^2$ = .223). The predictor variable of LSA contributed 18.0% of the variance and the variable MED (L) contributed a further 4.3% of the variance. The computational model was encouraging when compared to that of human experts evaluations (compare average human raters: $r$ = .590; model: $r$ = .474). Although the human inter-correlations (averaged across all raters) are significantly higher (z-diff = 3.24, $p$ = .001) than this model, it should be noted that the UPLC gives one pair of raters (G1) a correlation of $r$ = .520, and there was no significant difference between our model and the agreement of the G1 raters. As such, our initial results can be described as promising.

To verify our model, we also examined whether replacing MED (L) with MED (S) improved performance; however, it did not improve the model. The contribution of MED (S) was 0.5% as compared to MED (L) at 4.3%. Similarly, when Entailer was tested as a second variable, it did not significantly contribute to the model.

In the analysis of the human raters' data, writing quality and length differences were added to the model. At this stage, we have no computational variable to replicate writing quality; however, length of sentence differences (in terms of words) correlates moderately with writing quality ($r$ = .464) and so it was used to further develop and test the model. With the addition of differences of length, a significant and improved model emerged: $F$ (2, 1009) = 124.744, $p < 001$. This revised model increased the amount of the variance predicted from 22% to 27% (Adjusted $R^2$ = .269).

### Testing the Computational Model's Validity

To test the validity of the computational model, we generated a new composite variable based on the model generated from the training set data. Applied to the test set, the new variable significantly correlated with paraphrase quality ($r$ = .462, $p$ < .001, $n$= 649). This correlation increased to $r$ = .505 ($p$ < .001) when used on the entire data set ($N$ = 1998). The result is encouraging, and the correlation does not significantly differ from the agreement reached by the G1 pair of raters in the ULPC, although it is significantly lower than the average of the raters z-diff = 2.71, $p$ = .007). The syntactic similarity index of MED is impressive, and outperforms human agreement; however, the semantic completeness indices (LSA and Entailer) perform significantly below human agreement.

## Conclusion

Our findings suggest that the components of a paraphrase include an assessment of semantic completeness, syntactic similarity, and may be also evaluations of writing quality and/or differences in sentence length. Raters' judgments of semantic completeness appear to play the largest role in judging overall paraphrase quality. While lexical similarity would seem to be an important component of paraphrase evaluation, and does indeed correlate with paraphrase quality, the results of this study suggest that its role appears to be subsumed by that of semantic completeness. That is, the semantic similarity of two sentences and the lexical similarity of two sentences are highly related.

Because highly trained human raters demonstrated significantly higher agreement for semantic completeness and syntactic similarity than overall paraphrase quality, it seems reasonable to assume that individually assessing semantic completeness and syntactic similarity could lead to more reliable evaluations of paraphrase quality for both human raters and computational approaches. That is, not all tasks are equal in terms of assessment, and raters may find evaluating paraphrase quality overly complex, leading to lower reliability. Similarly, computational indices may be more easily developed if their role is better defined (i.e., syntax assessment or semantic assessment). The computational indices of LSA and MED (L) correlated highly with expert evaluations of semantic completeness and syntactic similarity respectively. Thus, we can posit that these indices offer substantial potential for computational evaluation of the quality of paraphrases, although improvement for the semantic component seems desirable.

Writing quality appears to be a small but significant component of paraphrase evaluation. One possible approach for improving the model produced in this study would be to correct the writing quality. That is, the typographical and grammatical errors produced in the paraphrases may affect the raters' assessment of the paraphrase and thus affect ratings. One potential avenue of research is to examine whether assessing or correcting typographical and grammatical errors affects raters' (and even automated algorithms') evaluations of paraphrase quality.

Establishing a fast and accurate evaluation of user-language paraphrases may facilitate appropriate feedback such that the assessment would be comparable to one or more trained human raters. This study offers an important step towards that goal in that it offers compelling evidence for the primary components and relative contributions of those components to paraphrase quality: namely, semantic completeness and syntactic differences. This study also demonstrates that computational indices such as LSA and MED go a long way towards producing a model that replicates human performance of these assessments.

## Acknowledgments

This research was supported by the Institute for Education Sciences (IES; R305G020018-02, IIS-0735682 and R305A080589). The authors also acknowledge Vasile Rus, Rebekah Guess, Angela Freeman, and John Meyers.

## References

Charniak, E. 2000. A maximum-entropy-inspired parser. Proceedings of the First Conference on North American Chapter of the Association For Computational Linguistics (pp. 132-139). San Francisco, CA: Morgan Kaufmann Publishers.

Dolan, B., Quirk, C., and Brockett, C. 2005. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. *Proceedings of the 20th International Conference on Computational Linguistics* (pp. 350-356). Geneva, Switzerland: Coling 2004.

Hawes, K. 2003. *Mastering academic writing: Write a paraphrase sentence*. Memphis, TN: University of Memphis.

Ibrahim, A., Katz, B., & Lin, J. 2003. Extracting structural paraphrases from aligned monolingual corpora. *Proceedings of the Second International Workshop on paraphrasing* (pp.57-64). Sapporo, Japan: ACL 2003.

Inderjeet, M. 2001. *Automatic summarization*. *Natural Language Processing*, 3: John Benjamins.

Iordanskaja, L., Kittredge, R., and Polgere, A. 1991. lexical selection and paraphrase in a meaning-text generation model. In C.L. Paris, W.R. Swartout, W.C. Mann, (Eds.), *Natural Language Generation in Artificial Intelligence and Computational Linguistics* (pp.293-312). Norwell, MA: Kluwer Academic.

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

Lockelt, M., Pfleger, N., and Reithinger, N. 2007. Multi-party conversation for mixed reality. *The International Journal of Virtual Reading, 6*, 31-42.

McCarthy, P.M. and McNamara, D.S. 2008. The user-language paraphrase challenge. umdrive.memphis.edu/pmmccrth/public/Paraphrase%. Retrieved 1/10/2008.

McCarthy, P.M., Guess, R., and McNamara, D.S. in press. The components of paraphrase. *Behavior Research Methods*.

McCarthy, P.M., Rus, V., Crossley, S.A., Graesser, A.C., and McNamara, D.S. 2008. Assessing forward-, reverse-, and average-entailment indices on natural language input from the intelligent tutoring system, iSTART. In D. Wilson and G. Sutcliffe (Eds.), *Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference* (pp. 165-170). Menlo Park, CA: The AAAI Press.

McNamara, D.S. 2004. SERT: Self-explanation reading training. *Discourse Processes, 38*, 1-30.

McNamara, D. S., Ozuru, Y., Best, R., and O'Reilly, T. 2007. The 4-Pronged Comprehension Strategy Framework. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 465–496). Mahwah, NJ: Erlbaum.

Millis, K., Magliano, J., Wiemer-Hastings, K., Todaro, S., and McNamara, D.S. 2007. Assessing and improving comprehension with Latent Semantic Analysis. In T. Landauer, D.S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 207-225). Mahwah, NJ: Erlbaum.

Rus, V., Lintean, M., McCarthy, P.M., McNamara, D.S., and Graesser, A.C. 2008. Paraphrase identification with lexico-syntactic graph subsumption. In D. Wilson & G. Sutcliffe (Eds.), *Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference* (pp. 201-206). Menlo Park, CA: The AAAI Press.

Rus, V., McCarthy, P.M., McNamara, D.S., and Graesser, A.C. (2008). Natural language understanding and assessment. In J.R. Rabuñal, J. Dorado, A. Pazos (Eds.). *Encyclopedia of Artificial Intelligence*. Hershey, PA: Idea Group, Inc.