Computational Considerations in Correcting User-Language

Adam M. Renner, Philip M. McCarthy, & Danielle S. McNamara

University of Memphis
Departments of Psychology and English
Institute for Intelligent Systems
Memphis, TN, USA
adam.der.renner@gmail.com

Abstract

This study evaluates the robustness of established computational indices used to assess text relatedness in user-language. The original User-Language Paraphrase Corpus (ULPC) was compared to a corrected version, in which each paraphrase was corrected for typographical and grammatical errors. Error correction significantly affected values for each of five computational indices, indicating greater similarity of the target sentence to the corrected paraphrase than to the original paraphrase. Moreover, misspelled target words accounted for a large proportion of the differences. This study also evaluated potential effects on correlations between computational indices and human ratings of paraphrases. The corrections did not yield assessments that were any more or less comparable to trained human raters than were the original paraphrases containing typographical or grammatical errors. The results suggest that although correcting for errors may optimize certain computational indices, the corrections are not necessary for comparing the indices to expert ratings.

Introduction

Intelligent Tutoring Systems (ITSs) are computerized tools that apply systematic procedures for enhancing learning (e.g., Aleven and Koedinger 2002; Gertner and VanLehn 2000). These systems engage with a user in one-on-one tutoring, an effective means of promoting active knowledge building supplemental to textbooks and conventional classroom environments (Bloom 1984; Corbett 2001). A subgroup of ITSs also employ elements of conversational dialogue that use computational linguistic algorithms to translate and respond to natural language input from the user by connecting the appraised text to particular feedback actions. Thus, the efficacy of an ITS relies on the supporting algorithmic architecture that allows the system to assess learners' input and adaptively respond, so that the ITS proficiently scaffolds instruction for the learner (Rus et al. 2008 [a]). The accuracy of the ITS feedback response to the student essentially depends on the precision of its underlying Natural Language Processing (NLP) system (McCarthy et al. 2007).

The last two decades have seen major advances in the NLP technologies that provide the backbone for ITSs (Jurafsky and Martin 2008). Among recent developments are text relatedness metrics such as Latent Semantic Analysis (LSA: Landauer et al. 2007), overlap indices (McNamara, Boonthum, et al. 2007), and entailment algorithms (McCarthy et al. 2007; Rus et al. 2008 [b]). Many of these indices are designed to evaluate natural language input (e.g., iSTART: McNamara, Levinstein, and Boonthum 2004).

The main focus of many of these tools and indices has been to evaluate edited, polished texts, for which they have met with considerable success. By contrast, research on the computational assessment of textual relatedness in ITS user-language is relatively sparse. Thus, the focus of this study is on natural user-language in the context of an ITS.

User-language is defined here as the natural raw input of a user interacting with an ITS. The challenge of evaluating these short, ill-formed sentences can be daunting (McCarthy and McNamara 2008). One source of problems in their evaluation, and the main focus of the current research, is user typographical errors. It is unrealistic to presuppose that students using ITSs will write or type flawlessly. In practice, student input has a high rate of misspellings, typographical mistakes, and dubious syntactic choices. Furthermore, students sometimes enter complaints (e.g., *I don't want to do this*), gibberish (e.g., *mustard is playing golf*), or random keying (e.g., *awerijasdfhy*). Suffice it to say, such users are more likely to produce statements that are neither grammatically nor typographically correct. Conventional text relatedness tools may have limited accommodation for such issues. For instance, the general approach for assessing a misspelled word is to label it as a rare word that is substantially different from its correct form. When this occurs, similarity scores are negatively affected, leading to unhelpful feedback based on spelling rather than understanding of key concepts (McCarthy et al. 2007). These consequences are of primary concern, because many interventions have been shown to be of greatest benefit for low domain knowledge learners, who make more of these types of errors (McNamara 2004; VanLehn et al. 2007).

The present research focuses on characterizing and evaluating the User-Language Paraphrase Corpus (ULPC;

McCarthy and McNamara 2008), according to the types of errors commonly found in typed natural user-language. Our goal is to appraise the robustness of computational indices that guide the feedback in ITSs, and to assess whether these indices' association to trained human tutors is affected by typographical and grammatical input errors.

Corpus

The ULPC corpus comprises 1998 target-sentence/student response text pairs, collected in the context of a paraphrase-training component of iSTART (McNamara et al., 2004). iSTART is an ITS designed to improve students' ability to use reading strategies, one of which is paraphrasing. As such, this corpus comprises high school students' attempts to paraphrase target sentences. The ULPC corpus provides human ratings of the paraphrases as well as computational measures of relatedness that compare target sentence to the student paraphrase (see McCarthy and McNamara 2008).

Error Assessment

Two trained expert coders identified and categorized the errors in a subset of the ULPC corpus, and one of the raters completed this procedure for the full corpus. Inter-rater agreement was assessed for the subset of the data (i.e., 10%; *n* = 200). Cohen's Kappa for overall error identification was κ = .70. Judgments were based on validated models of grammaticality (e.g., Foster and Vogel 2004; Schneider and McCoy 1998). Each error was labeled according to its type and the error was corrected. The following revisions were made such that the corrected version preserved the original intent of the paraphrase to the best extent possible.

1. *Spelling – Internal* (κ = .770, *n* = 665): A misspelled word corresponding to a word in the target sentence.
2. *Spelling – External* (κ = .679, *n* = 386): A misspelled word that is not similar to a word in the target sentence was replaced by a word similar in spelling and contextually appropriate.
3. *Capitalization* (κ = .835, *n* = 1157): A word that ought to begin with a capital letter, e.g., the first word in the sentence, or a word that ought not to be capitalized. In some cases user paraphrases were entered in all caps. The frequency for such cases was coded to indicate the number of words in the sentence, and not every single character.
4. *Agreement* (κ = .683, *n* = 367): If the word is a noun, pronoun, or verb, it can be exchanged by the same word with a different value for an agreement feature, e.g. first person plural *are* with singular person *am*, singular *dog* with *dogs*. This error type also includes corrections of verb forms and auxiliaries.
5. *Spacing* (κ = .804, *n* = 174): Inappropriate spacing is corrected by removing, adding, or moving a space. As noted in McCarthy and McNamara (2008), the corpus was originally evaluated by replacing double spaces between words with single spaces. This was required by some computational measures to assess the paraphrases without error. Thus, the true frequency of this type of mistake is not represented.
6. *Punctuation* (κ = .483, *n* = 344): Inappropriate use of commas, periods, or other punctuation is corrected by removing, adding, or moving the punctuation mark. As noted in McCarthy and McNamara (2008), a period was added to the end of the original input if one did not exist previously. This was the second of two necessary changes to the original data.
7. *Article agreement* (κ = .585, *n* = 75): An article not in agreement with its noun is replaced, e.g. *a* elephant with *an* elephant.
8. *Preposition agreement* (ns, *n* = 53): A preposition may be replaced by any other more appropriate preposition, e.g. *for* with *of.*
9. *Determiner agreement* (κ = .290, *n* = 59): A determiner may be replaced by any other more appropriate determiner. For the purposes of this study, determiner is meant as a noun modifier that expresses the reference of a noun or noun phrase and is not an article, e.g., *this*, *that*, *these*, *those*, *which*, etc.
10. *Conjunction agreement* (ns, *n* = 43): If the word is a conjunction it may be replaced by another conjunction, e.g., *then* can be replaced by *and* in a sentence not beginning with *once*.
11. *Possessive agreement* (κ = .886, *n* = 71): A separate case error involving the incorrect use of the genitive case on pronouns and nouns, e.g., the possessive *heart's* instead of the plural *hearts*.
12. *Extraneous word* (ns, *n* = 72): Error corrections involving the deletion of a word, either because it was repeated or unnecessary; includes deletion of garbage input.
13. *Omission* (κ = .433, *n* = 98): Error corrections involving the addition of a word if it was needed and could be implied from the context of the sentence, restricted to content words only. Adding missing articles or auxiliaries are included within their own agreement categories.
14. *Substitution* (ns, *n* = 60): An erroneous content word is replaced by a word more suitable to the context of the sentence. Replacement of two similarly spelled words, e.g., *raise* and *rise*, are categorized as a spelling error if the word was in the target sentence.

It is worth noting that some error types did not reach significant agreement levels (*preposition agreement*, *conjunction agreement*, *extraneous word*, *substitution*). This lack of agreement could be attributable to the notion that there is often more than one correction possible for a given error (Foster and Vogel 2004). That is, two raters may correct the same error with different solutions that likely express the same meaning. Let us take *conjunction agreement* as an example, which was not significantly agreed upon. We observed that this disagreement owed in part to one coder correcting such an error by substituting a different conjunction, while the other coder corrected the same error by replacing the conjunction with a semicolon. We anticipated some of these disparities and attempted to correct them by extended training. Overall however, we contend that since overall agreement is established, we can be relatively confident that a single rater's evaluation of the data is consistent and valid, even for those error types that did not reach significant agreement.

Computational Measures

Each of the computational indices we examined in the present study has been validated as a method of representing similarity between two bodies of text. Many of them employ overlap techniques, and therefore rely heavily on individual words and stems to be well-formed. Thus, we expected that the corresponding indices of the user-language paraphrases would be optimized when reanalyzed in a corrected form. For a detailed explanation of these indices, see McCarthy and McNamara (2008).

Latent Semantic Analysis. LSA as a local measure calculates a vector cosine value between adjacent pairs of sentences to represent their degree of semantic overlap (values range from 0 to 1). Because this technique is based on a large corpus of well-formed text, the values of the revised paraphrases should increase in comparison to their ill-formed counterparts. Also, spelling errors, which account for a large proportion of the total observed errors (n = 665), should contribute largely to this improvement.

Overlap indices. Stem-overlap judges two sentences as overlapping if a common stem of a content word occurs in both sentences (McNamara et al. 2006). The exact measure in this corpus is binary, so it is difficult to predict the degree to which these values would change.

Minimal Edit Distances (MED). MED indices assess differences between any two sentences in terms of the words and the position of the words in their respective sentences. The final MED value gives a range of 0 to 1, 0 indicating greater similarity and 1 indicating greater difference. Although we can expect some change in the values, this change would not be large, because syntax errors were less frequent and we aimed to preserve sentences as close to their original structure as possible.

Type Token Ratio. TTR is a shallow NLP approach, where each unique word in a text is a word type, and how frequently it occurs is a token (Graesser et al. 2004). In the ULPC version, TTR counts only for content words. TTR is derived by dividing the number of unique words (types) by the number of total words (tokens). The 0 to 1 ratio in comparing paraphrases to target sentences indicates that a lower value represents greater similarity. We might expect this measure to see the largest improvement, because identifying the words that co-occur is likely to yield a higher number of matched tokens, and thus a lower value.

The Entailer. Entailer indices are based on a lexico-syntactic approach to sentence similarity and are used to evaluate the degree to which one text is entailed by another text (Rus et al. 2008). The scores are the weighted sum of one lexical and one syntactic component. These components utilize the Charniak probabilistic parser (2000), which is robust because it is trained on a large body of data where no precise distinction is made between the grammatical and the ungrammatical. Therefore, one might expect this measure to be more robust to ill-formed input. However, it is quite possible that the range of errors in the present corpus is not represented in the data in which the parser was trained. So we expect that this measure will be improved, but to what degree is uncertain.

One primary goal in evaluating the User-language Paraphrase Corpus is so that ITSs may provide users with assessment and feedback comparable to human raters. For the second round of hypotheses in this study, the computational measures are to be compared to the human gold standard paraphrase dimensions. The degree to which the computational measures of the original paraphrase corpus correlate with the human dimensions varies from weak to strong. Specific predictions on how these correlations might improve are difficult to formulate, but in general we predicted that computational measures that currently exhibit strong correlations with the human coded dimensions would improve. However, we cannot predict that the improvements would be large because the computational measures may only represent a proportion of overall quality as humans would rate them.

Table 1: Computational index means of original and edited user-language in the complete ULPC

	Original $\chi(\sigma)$	Corrected $\chi(\sigma)$	r	t	part. η^2
LSA	.642 (.271)	.694 (.266)	.915	20.828	.178
Stem	.915 (.274)	.921 (.268)	.942	2.774	.004
TTR	.747 (.125)	.720 (.129)	.929	-25.094	.240
MED	.746 (.237)	.721 (.251)	.959	-15.813	.111
Entailer	.442 (.222)	.494 (.232)	.931	27.017	.268

N = 1998; $p < .01$ for all

Results

First, we examined the effect of typographical and grammatical errors on computational measures of textual relatedness. Paired-samples t-tests were conducted to evaluate the impact of error correction on five computational measures provided in the ULPC (LSA, Stem overlap, MED, TTR, and Entailer).

As predicted, all of the measures improved significantly as a function of correcting typographical and grammatical errors. Note that some indices increased in value whereas others decreased. String-matching approaches (i.e., TTR, MED) emphasize differences rather than similarity, and thus the lower MED and TTR values indicate greater similarity to the target sentence. By contrast, greater similarity is indicated by higher values for LSA, Stem overlap, and Entailment. However, somewhat contrary to our predictions, Entailer values showed the most significant change, followed by TTR, LSA, MED, and Stem overlap.

The second goal of our study was to evaluate which amongst our error correction rules effected these changes in values. We conducted linear regressions of each index for two-thirds of the data, in which the original and corrected version values were regressed onto error type. Because the indices rely on matching principles and a large proportion of our corrections were typographical in nature, we expected spelling to contribute most. Table 2 indicates that several of the error categories predicted a significant and large variance of these difference scores.

Of the 14 error types, and as predicted, *Spelling-Internal* was the most influential, accounting for nearly 70% of the predictive power for the MED regression model and over 90% for each of the three other measures. Assessments of Stem-overlap could not be appropriately evaluated, because the value changed for only 13 of the 1332 cases. This finding (or lack thereof) is consistent with the measure itself, given that stem-overlap assigns a value of 0 or 1 by judging two sentences as overlapping if a common stem of a content word appears in both sentences (McNamara et al. 2006). When comparing a paraphrase to its target sentence, chances are high that at least one stem will overlap. Therefore, we would not expect these values to change because we can be confident that the revised scores would be repeatedly equal to the original scores.

Cross-validation. We cross-validated the regression models by generating predicted difference scores for the remaining 33% of the cases. Difference scores were

Table 2: Effects of Error Type on Computational Indices of Text Relatedness

Index	Predictor (Error Type)	Adjusted R^2	ΔR^2	*F* Change	ß	*b* (*SE*$_b$)	*P*
LSA	Spelling (Internal)	0.35	0.35	716.141	-0.576	-.081 (.003)	<.001
F(4, 1362) = 216.42*	Spelling (External)	0.384	0.035	75.688	-0.174	-.034 (.004)	<.001
	Punctuation	0.389	0.005	11.643	-0.077	-.017 (.005)	<.001
	Spacing	0.393	0.004	9.722	-0.022	-.022 (.007)	0.002
MED	Spelling (Internal)	0.166	0.166	265.462	0.392	.037 (.002)	<.001
F(9, 1321) = 47.41*	Spacing	0.193	0.028	45.928	0.173	.038 (.005)	<.001
	Omission	0.19	0.02	34.421	0.128	.030 (.006)	<.001
	Subject-Verb Agreement	0.221	0.009	14.816	0.085	.012 (.003)	<.001
	Spelling (External)	0.226	0.006	10.02	-0.08	-.011 (.003)	0.001
	Capitalization	0.23	0.004	7.748	0.067	.003 (.001)	0.006
	Article Agreement	0.234	0.004	7.185	0.06	-.022 (.009)	0.013
	Preposition Agreement	0.237	0.004	6.289	0.06	.029 (.012)	0.012
	Extraneous Word	0.239	0.003	4.697	0.052	.011 (.005)	0.03
TTR	Spelling (Internal)	0.459	0.459	1129.876	0.664	.043 (.001)	<.001
F(6, 1324) = 222.33*	Spacing	0.475	0.017	42.282	0.131	.020 (.003)	<.001
	Subject-Verb Agreement	0.491	0.016	41.003	0.122	.012 (.002)	<.001
	Possessive Agreement	0.496	0.005	14.458	0.075	.019 (.005)	<.001
	Article Agreement	0.498	0.002	6.244	0.049	.012 (.005)	0.012
	Spelling (External)	0.5	0.002	5.707	0.047	.004 (.002)	0.017
Entailer	Spelling (Internal)	0.451	0.451	1092.418	0.655	.074 (.002)	<.001
F (8, 1322) =158.25*	Spacing	0.469	0.019	47.67	0.142	.038 (.005)	<.001
	Subject-Verb Agreement	0.475	0.007	16.494	0.072	.012 (.003)	<.001
	Article Agreement	0.479	0.004	9.53	0.061	.027 (.009)	<.002
	Capitalization	0.481	0.003	6.873	0.05	.003 (.001)	0.012
	Omission	0.483	0.003	6.473	0.05	.014 (.006)	0.012
	Punctuation	0.485	0.002	5.132	0.04	.007 (.004)	0.041
	Determiner Agreement	0.486	0.002	4.099	0.04	.022 (.011)	0.043

Note: * p < .001

Table 3: Original vs. Corrected Paraphrases on Correlations between Computational Indices and Human-Rated Dimensions

Dependent Variable	LSA mean	Stem	TTR	MED	Entailer
Irrelevant	-0.473 (-0.438)*	-0.502 (-0.497)*	0.350 (0.326)*	0.205 (0.198)*	-0.390 (-0.361)*
Elaboration	-0.170 (-0.175)*	0.027 (0.022)	0.176 (0.178)*	0.150 (0.145)*	-0.219 (-0.213)*
Semantic Completeness	0.570 (0.555)*	0.457 (0.461)*	-0.538 (-0.529)*	0.396 (0.397)*	0.434 (0.430)*
Entailment	0.548 (0.535)*	0.483 (0.49)*	-0.507 (-0.497)*	-0.359 (-0.360)*	0.429 (0.427)*
Syntactic Similarity	0.475 (0.465)*	0.316 (0.325)*	-0.540 (-0.513)*	-0.750 (-0.739)*	0.487 (0.465)*
Lexical Similarity	**0.829 (0.804)****	0.608 (0.614)*	**-0.773 (-0.740)****	-0.592 (-0.578)*	0.731 (0.710)*
Paraphrase Quality	0.422 (0.41)*	0.421 (0.427)*	-0.312 (-0.313)*	-0.053 (-0.059)	0.245 (0.245)*
Writing Quality	0.495 (0.498)*	0.534 (0.536)*	-0.398 (-0.410)*	-0.262 (-0.281)*	0.370 (0.392)*

Note: All computational measures of original and edited paraphrases correlate at r >.9, p<.001
Pearson r for original paraphrases appear in parentheses; N=1998 for all correlations; * p <.001
** for Fisher's z to r transformation of significant differences, p<.05

computed by using the coefficients in the regression equations by the true values for each case. These predicted scores were then correlated with the actual calculated differences between the original and revised paraphrases. There was a significant correlation between predicted and verified values of the validation set for each of the four computational measures: LSA, r (666) = .625, p < .001; MED, Pearson r (666) = .393, p < .001; TTR, Pearson r (666) = .695, p < .001; Entailer, Pearson r (666) = .675, p < .001. Our results indicate that the regression models reliably predict differences in the computational measures.

Human ratings. The second purpose of this study was to assess whether changed computational values improved the ability to predict human gold standards of 10 paraphrase dimensions. Provided in the ULPC (McCarthy and McNamara 2008), correlations assessed the degree to which the computational approaches correlated with human raters. Fisher's r to z transformations were performed to identify the significant differences between the correlations for the corrected and original paraphrases. Two dimensions, *garbage* and *frozen expressions*, were excluded from analysis because neither is independently applicable as assessments of text relatedness.

As shown in Table 3, several of the correlations between the computational measures and the human evaluations saw non-significant improvement. The only significant difference between any two correlations were LSA to Lexical Similarity (z = 2.37, p = .018, 2-tailed) and TTR to Lexical Similarity (z = -2.44, p = .014, 2-tailed). Taken together, these results indicate that although correcting ill-formed user input yields stronger computational scores, the changes do not significantly improve the paraphrase scores' comparisons to human ratings.

Discussion

In this study, 1998 user-language paraphrases were corrected for errors, to examine potential effects of errors on computational measures of text relatedness. Our results suggest that computational measures generally improve when errors are corrected, therefore also improving their potential application to ITSs. Our results demonstrate that misspelled words that are from the target sentence (i.e., internal spellings) had the largest influence in the improvements, presumably because the algorithms' string matching principals are not tuned for minor spelling errors of the type examined here. Spacing and agreement errors also consistently affected the indices, although their contribution was smaller. The largest changes were observed for the Entailer, presumably because the lexico-syntactic index was improved by both spelling and grammatical corrections. Moreover, the results seem to reflect the sensitivity of each measure to error correction.

A second finding of this research is that correcting for errors does not significantly increase correlations between computational measures and human ratings of paraphrases. Only the dimension Lexical Similarity saw a significant difference, the effect of internal spellings appearing to account for this change. This result supports the construct underlying this dimension, because correcting a misspelled word would make it easier for a human rater to identify it, which would facilitate the rater's ability to determine word commonality of the text pair.

Plagiarism detection and automated essay rating are two related areas of study that could potentially draw implications from these results. For instance, FindFraud (Xi et al. 2003) is a compression-based software that detects similarity between documents, but does not use pre-processing. We speculate that a similar study on plagiarism detection may yield similar results. However, most essays and academic documents that would be relevant for such a system would probably have been doctored in a word processor prior to assessment.

Overall, the majority of differences between the indices and human ratings were not significant, indicating that the computational measures for the original paraphrases are sufficient to assess their relationship to human rated dimensions of paraphrase. For that reason, the laborious process of identifying and correcting erroneous input may be unnecessary in future studies of natural user-language paraphrase evaluation. In sum, preprocessing of natural user-language may not be necessary because the indices are just as comparable to human ratings despite the errors. Nonetheless, future research will examine the computational expense of correcting internal errors to assess whether feedback to users may change.

In summary, the responsibility of interpreting a student's input to respond appropriately falls upon NLP within ITSs. Inappropriate responses can affect users' learning and motivation. Indeed, assessing user input to determine appropriate responses is essential to establish intelligent human-computer dialogue. Thus, established ITSs are expected to be sufficiently robust. However, contemporary NLP systems are far from perfect, and each ITS system may employ different types of preprocessing techniques along with its own unique algorithmic architecture. While much work in this area still lies ahead, this study is a major step forward in evaluating the potential sturdiness of ITSs that evaluate user-language and developing the field of natural language assessment and understanding.

Acknowledgements

This research was supported in part by the National Science Foundation (IIS-0735682) and the Institute for Education Sciences (IES R305A080589). The authors would also like to acknowledge the contributions made to this paper by Vasile Rus, Ben Duncan, Christina Sinquefield, and Sidney D'Mello.

References

Aleven, V., and Koedinger, K. R. 2002. An effective meta-cognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science* 26: 147-179.

Bloom, B. S. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher* 13: 4-16.

Charniak, E. 2000. A maximum-entropy-inspired parser. In Proceedings of the North American Chapter of the Association for Computational Linguistics, 132-139.

Corbett, A. T. 2001. Cognitive computer tutors: Solving the two-sigma problem. In Proceedings of the Eighth International Conference of User Modelling, 137-147.

Foster, J. and Vogel, C. 2004. Parsing ill-formed text using an error grammar. *Artificial Intelligence Review: Special AICS 2003 Issue* 21: 269-291.

Gertner, A. S. and VanLehn, K. 2000. Andes: A coached problem solving environment for physics. In G. Gauthier, C. Frasson, K. VanLehn eds., Proceedings of the 5th International Conference on Intelligent Tutoring Systems, ITS 2000, 133-142. Montreal, Canada.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers* 36: 193-202.

Jurafsky, D., and Martin, H. 2008. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics, 2nd Ed. Prentice-Hall.

Landis, J. R., and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1): 159-174.

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

McCarthy, P. M., Rus, V., Crossley, S. A., Bigham, S. C., Graesser, A. C., and McNamara, D. S. 2007. Assessing entailer with a corpus of natural language. In D. Wilson and G. Sutcliffe eds. Proceedings of the twentieth International Florida Artificial Intelligence Research Society Conference, 247-252. Menlo Park, CA: The AAAI Press.

McCarthy, P. M. and McNamara, D. S. 2008. The user-language paraphrase challenge. Special ANLP topic of the 22nd International Florida Artificial Intelligence Research Society Conference. Retrieved February 10, 2009 from *https://umdrive.memphis.edu/pmmccrth/public/Paraphrase Corpus/Paraphrase_site.htm.*

McNamara, D. S., Ozuru, Y., Graesser, A. C., and Louwerse, M. 2006. Validating Coh-Metrix. In R. Sun and N. Miyake eds., Proceedings of the 28th Annual Conference of the Cognitive Science Society, 573-578. Austin, TX: Cognitive Science Society.

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

McNamara, D. S., and Kintsch, W. 1996. Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes* 22: 247-287.

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 and G. Sutcliffe eds., Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference, 201-206). Menlo Park, Calif: 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.

Rus, V., McCarthy, P. M., McNamara, D. S., and Graesser, A. C. 2007. A study of textual entailment. *International Journal on Artificial Intelligence Tools.*

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