A Linguistic Analysis of Student-Generated Paraphrases

Vasile Rus[1], Shi Feng[2], Russell Brandon[2], Scott Crossely[2], and Danielle S. McNamara[2]

[1]Department of Computer Science
[2]Department of Psychology
The University of Memphis
Memphis, TN, 38152
vrus@memphis.edu

Abstract

Paraphrase identification is a core Natural Language Processing task that involves assessing the semantic similarity of two texts. To foster systematic studies of this task, standardized datasets were created on which various approaches could be compared more fairly. However, a better understanding and more precise operational definition of a paraphrase are needed before any further datasets or systematic evaluations of the task of paraphrase identification are proposed. This study develops the concept of paraphrasing as a writing strategy. Six types of paraphrases are defined through the creation of a relatively large corpus of student-generated paraphrases. These paraphrases are analyzed along several dozen linguistic dimensions ranging from cohesion to lexical diversity. The most significant indices from these dimensions were then used to build a prediction model that could identify true and false paraphrases and each of the six paraphrase types.

Introduction

Paraphrase is a text-to-text relation between two non-identical text fragments that express the same idea in different ways. As an example of a paraphrase, we show below a pair of sentences from the Microsoft Research (MSR) Paraphrase Corpus (Dolan, Quirk, & Brockett, 2004) in which Text A is a paraphrase of Text B and vice versa.

Text A: *York had no problem with MTA's insisting the decision to shift funds had been within its legal rights.*
Text B: *York had no problem with MTA's saying the decision to shift funds was within its powers.*

Paraphrase identification is the task of deciding whether two given text fragments have the same meaning. We focus in this article on providing a more precise definition of a paraphrase and also on identifying paraphrase relations between text fragments in a natural language setting.

Paraphrase identification is an important task in a number of applications including Question Answering (Ibrahim, Katz, & Lin, 1991), Natural Language Generation (Iordanskaja, Kittredge, & Polgere, 1991), and Intelligent Tutoring Systems (Graesser et al., 2005). For Intelligent Tutoring Systems with natural language input, paraphrases are useful to assess whether student's articulated answers to deep questions (e.g., conceptual physics questions) are similar to paraphrases of ideal answers (Graesser et al., 2005).

In this paper, we examine the concept of paraphrase in the context of a writing strategy training tutoring system, Writing-Pal (W-Pal; McNamara, Raine, Roscoe, et al., in press). One of the strategies in this tutoring system is paraphrasing. As the system is supposed to prompt students to paraphrase and then provide feedback on their paraphrases, a better understanding of student-generated paraphrases is needed. To address this need, we created a large corpus of student-generated paraphrases and analyzed them along several dozen linguistic dimensions ranging from cohesion to lexical diversity. We use the most significant indices to build a prediction model that can identify true and false paraphrases and also several categories of paraphrase types.

What Is a Paraphrase

Most writer style guides define a sentence as a paraphrase of another sentence if it conveys the same meaning using different words (at minimum). However, definitions of paraphrases do not specify how many new words there should be - i.e., should all of the original word be replaced with new ones or only a proportion? If a paraphrase includes a proportion of different words, is there an acceptable threshold above which the new text is considered different enough in form but not meaning from the original? This paper is a first step towards providing answers to some of these questions and make progress towards a more operational definition of a paraphrase.

An intriguing aspect of recent research on automated paraphrase identification focusing primarily on sentence-level paraphrases (texts the size of a sentence) is that these paraphrases do not seem to follow the traditional definitions discussed above as evidenced by a quick analysis of the existing paraphrase data sets.

For instance, in the Microsoft Research (MSR) Paraphrase corpus the paraphrases tend to have many words in common as opposed to using different words to express the same meaning. On average, there is 68% word overlap between the two sentences in each paraphrase instance in the corpus (69.5% after lemmatization).

In a similar fashion, student generated paraphrases also demonstrate strong lexical overlap. For instance, the iSTART corpus (McCarthy and McNamara, in press) exhibits 57.65% word overlap between two sentences (the original text and student-generated paraphrase of it). The iSTART corpus was collected from high-school students prompted to paraphrase science textbook sentences while interacting with the intelligent tutoring system iSTART (McNamara et al., 2007). These paraphrases show that students, who are not in general very knowledgeable about the science domain, tend to reuse much of the words from the original science sentence.

Thus, for sentential paraphrases, the expectation of different words seems to be too restrictive, although not impossible. Indeed, given an isolated sentence it would be quite challenging to omit/replace some core concepts when trying to paraphrase. An example of a sentence (instance 735 in MSR corpus) which is hard to paraphrase using many other/different words is *Counties with population declines will be Vermillion, Posey and Madison*. The difficulty is due to the large number of named entities in the sentence. Actually, its paraphrase in the corpus is *Vermillion, Posey and Madison County populations will decline.*, which retains all the named entities from the original sentence.

One effect of the high-world overlap in these sentential paraphrase datasets is that a baseline approach that implements a simple word overlap method leads to competitive results with sophisticated approaches. For instance, on the MSR corpus, a simple overlap method yields an accuracy of 72.30% which is comparative to accuracy reported by more complex approaches, e.g., 71.50% (Corley & Mihalcea, 2005) or 72.00% (Qiu et al., 2008).

Another interesting aspect of sentential paraphrasing, at least as learned from the existing paraphrase datasets, is the fact that there seem to be two different ways to judge them. On one hand, two sentences are considered paraphrases of each other if and only if they are semantically equivalent, by conveying the same message with no additional information present in one sentence but not the other. Thus, in order to detect whether two sentences are not paraphrases of each other, we only need to find one concept that is present in one sentence but not in the other. On the other hand, two sentences can be judged as forming a paraphrase if they convey roughly the same message (differences in minor details are acceptable). In this latter case, the paraphrase relation can be looked at as a bidirectional entailment relation (Rus et al., 2008b). To exemplify such loose paraphrases, we show below a pair of sentences that has been tagged as paraphrase in the MSR Paraphrase Corpus:

Text A: *Ricky Clemons' brief, troubled Missouri basketball career is over.*
Text B: *Missouri kicked Ricky Clemons off its team, ending his troubled career there.*

In this example, the first sentence specifies that the career of Mr. Clemons was brief, while the second sentence specifies the reason why Mr. Clemons' career is over. This characteristic of the MSR corpus impacts the performance of general approaches, such as those proposed by (Lintean et al., 2010), to paraphrase identification that are not biased towards a particular judging style (for instance, the MSR judging style instructs judges to ignore exact numbers and consider them generic NUMBER semantic categories).

One obvious conclusion from the discussion above is that evidence from existing datasets suggests a high word re-use between sentences in a paraphrase relation which is at odds with the standard definition of a paraphrase. There is need for further research to understand what a paraphrase is. In this paper, we make progress towards this goal by analyzing student-generated paraphrases along several linguistic dimensions in order to provide a more precise definition of a paraphrase. The outcome of this work will inform future researchers focusing on the task of paraphrase identification.

Methods

Corpus

To create our paraphrase corpus we selected a set of texts, called original texts hereafter. The texts were excerpted from 184 essays taken from undergraduate students at Mississippi State University (MSU; see McNamara, Crossley, and McCarthy, 2010, for more details). The original texts consisted of 100 passages from these essays, half from high proficiency essays (50) and half from low proficiency essays (50). An equal number of passages were picked from the major elements of an essay: introductions, theses, topic sentences, evidential sentences, conclusions. The passages range in size from 1 to 4 sentences.

We then asked human subjects, i.e. college students, to paraphrase the samples into six different types of paraphrases. The first type, *free paraphrase*, was unguided. The last five types, *changed words*, *changed structure*, *changed words and structure*, *condensed*, and *improved*, were guided. These types of paraphrases are described below.

Free Paraphrase. Also called unguided paraphrases, these paraphrases are generic paraphrases of the original passage. The role of this condition was to understand students' natural tendencies when asked to paraphrase. This condition should help us identify what are the characteristics of a students' natural paraphrase. Students were simply given the traditional definition of a paraphrase: paraphrasing means restating an idea that you have read about using your own words. They were then instructed that when you paraphrasing more than just one word, they would need to change the structure as well. However, it was pointed out that they do not always have to change the words and structure. If there were a better way or different way to say something, then it would be fine to use that way.

Changed Words. In this condition, students were explicitly asked to *change the words* in the original passage. That is to say, they were supposed to change as many of the words as possible while still preserving the meaning of the original passage. The role of this condition is to see how many words students do change when explicitly told to do so. For space reasons, we do not provide examples of this type of paraphrase or the following ones.

Changed Structure. Students were asked to change the structure of the original passage while generating the paraphrase. That is, they were supposed to change the major phrases and clauses in the original passage such that the meaning was preserved. We were interested to see how much structural change students make when explicitly asked to do so.

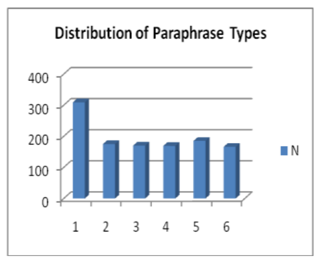
Changed Words and Structure. In this case, students were asked to change both the words and structure of the original passage while still preserving its meaning. In this case, we wanted to understand what students do when explicitly asked to change the words and structure.

Condensed. In this condition, students were asked to restate the original passage in a more condensed form, i.e. summarize, while still preserving the meaning of the original passage. Summarization is an important writing strategy which explains our interest in this type of paraphrase.

Improved. In this condition, students were asked to generate an improved paraphrase of the original passage which meant editing, revising, or rewarding the original passage to make it better. This is another specific condition to our context, the writing strategy training, in which students need to paraphrase their own initial sketches of (essay) ideas in an improved form in order to improve their final essay.

Each subject in our pool of 40 subjects was prompted to generate unguided paraphrases for 10 different texts randomly selected from the original texts set of 100 passages. Then, each subject was asked to generate 5 guided paraphrases using one strategy (condensed, reword, restructure, reword and restructure, and improve) for 5 different passages randomly selected from the original text set. The paraphrase collection process aimed at creating a balanced data set across the different types of paraphrases and the set of original texts. To this end, each passage to be paraphrased was selected randomly from the list of 100 original texts. If a passage had already been paraphrased using an unguided strategy, the program would select another passage that has not been paraphrased. After all passages have been used once, the random selection began again and added a second example to each passage (and then a third, forth, etc.). Furthermore, paraphrase types were counterbalanced with unguided always at the beginning. Word and Structure are counterbalanced with each other. Word, Structure, and Both are counterbalanced as a group with Condensing and Improved.

The above data collection effort resulted in a corpus of 1174 pairs of (original passage, student paraphrase) data points. This number was obtained after discarding the compromised pairs (e.g., students typing random characters) from the collected data set of 1400 pairs (400 unguided, 200 each for the 5 guided types of paraphrases). The distribution of the data is shown in Figure 1.



Paraphrase Type	N
Unguided	309
Words	175
Structure	170
Words & Struct.	169
Condensed	185
Improved	166

Figure 1. Distribution of Paraphrase Types.

Expert Analysis

We had experts analyze manually the collected student paraphrases with respect to whether or not students preserved the meaning of the original passage. This was needed because we had to know which of the student paraphrases were indeed preserving the meaning of the original passage (i.e. which were true paraphrases, and which were not). Only linguistic profiling of true paraphrases can help us better understand what a paraphrase is. The analysis demonstrated that the corpus contained 154 false-paraphrases (meaning not preserved) and 1029 true-paraphrases (meaning preserved).

Linguistic Analysis

Our goal was to understand the different linguistic characteristics of each type of paraphrase and then build predictors that would allow us to detect what type of paraphrase students generate when prompted to do so. This linguistic analysis serves two purposes. First, it allows us to better understand what a paraphrase is. Second, the analysis allows us to develop a paraphrase algorithm for W-Pal to be used when students are asked to generate a particular type of paraphrase, e.g. changing the words. The system must assess whether the student's paraphrase is of the expected type and based on that assessment, the system will provide appropriate feedback such as "You did not change enough words." when the student paraphrase is almost identical with the original passage. Positive feedback will be provided when the students changed enough words.

We conducted our linguistic analysis using Coh-Metrix, which is a computation tool that introduces over 600 indices on cohesion and language (Graesser, McNamara, Louwerse, and Cai, 2004). Coh-Metrix is able to analyze deep levels of linguistic features including language and discourse characteristics. It accomplishes this task using syntactic parsing techniques, latent semantic analysis (LSA), and other common computational linguistic features. There are several categories of these indices that Coh-Metrix provides, the key categories include:

Causal Cohesion. Coh-Metrix calculates causal cohesion using the ratio of causal verbs to causal particles. Examples in our corpus for causal verbs include: *change*, *cause*, and *make*. Causal particles are items such as *because*, *for*.

Coreferential Cohesion. Coh-metrix calculates coreferential cohesion using noun overlap, argument overlap, stem overlap, and LSA-based semantic overlap.

Connectives. Connectives form cohesive links between separated sentential ideas. They include positive-additive connectives, such as *also*, *moreover*, negative-additive connectives, such as *however*, *but*, positive-temporal connectives, such as *after, before*, and negative-temporal connectives, such as *until*.

Density of Major Parts of Speech. Coh-Metrix reports the incidence of parts of speech within a text. These include pronouns, nouns, verbs, adjectives, adverbs, cardinal numbers, determiners, and possessives.

Polysemy and Hypernymy. Coh-Metrix incorporates WordNet (Fellbaum, 1998) to calculate values for the number of senses and the number of levels in a conceptual and taxonomic hierarchy.

Syntactic Complexity. Coh-Metrix measures syntactic complexity in terms of structural density and syntactic ambiguity.

Word Information and Frequency. Coh-Metrix incorporates four metrices: familiarity, concreteness, imagability, and meaningfulness. Coh-Metrix obtains scores for these matrices from the MRC Psycholinguistic database (Coltheart, 1981).

Situation Model Dimension. An important level of text comprehension consists of constructing a situation model, which is the referential content of what a text is about (e.g., Graesser, Millis, & Zwaan, 1997). These indices include causation, intentionality, time, and space.

Basic Indices. Coh-Metrix also provides information for syllable count, word length, sentence length, number of words per sentence/paragraph/text, and various other indices in determining mode, genre, and style.

Because Coh-Metrix analyzes individual texts, we altered the standard Coh-Metrix output in order to observe differences between two texts (student paraphrase versus original passage). In our case, we need to capture what is changed in the student paraphrase with respect to the original passage. Therefore, we first analyzed the original passages and their corresponding student paraphrases using Coh-Metrix. Then, we mapped the reported values into normalized z-scores before we examined the difference between the scores for original passages and student paraphrases. The obtained difference scores are called Z-diff-INDEX, where INDEX is any of the standard Coh-Metrix indices. In total, we used 82 Coh-Metrix indices. It should also be noted that some Coh-Metrix indices are not defined for single-sentence texts (e.g. the LSA average score between adjacent sentences). We replaced such indices with proxies, e.g. the LSA average between adjacent sentences was replaced with the LSA paragraph-to-paragraph index where the first paragraph was considered the original passage and the second the student paraphrase.

Results

Study 1: Paraphrase Type Analysis

The purpose of our first study was to investigate which linguistic indices significantly differ across the six types of paraphrases. We also explored how well these indices can predict the 6 paraphrase types. To examine differences, we ran ANOVAs (Multivariate Analysis of Variance) on all Coh-Metrix indices with the paraphrase type as a fixed factor. To predict paragraph types, we conducted Discriminant Function Analyses (DFA). A total of 785 instances from the original data set of 1174 instances were selected for training, amounting to 66.9% of the original set, and 389 instances were held for testing (33.1%).

Out of the 82 indices 36 were significant ($p<0.05$). Before our DFA, we further filtered out the significant indices to avoid problems with multicollinearity. Therefore, bivariate correlations among all 36 significant indices were computed. Based on correlation results, we filtered out indices that correlated greater than .70 with another index whose effect size (revealed by the previous ANOVA step) was greater. This step filtered out 16 indices, leaving us with 20 predictors with which to use in DFA to predict group membership. The six paraphrase types differed along several linguistic dimensions including basic counts, situation model, lexical diversity,

Index	Unguided	Words	Structure	Wds-Strc.	Improved	Condensed	F(5,779)	η^2
READNW	.440	.531	.547	.898	.912	.971	6.550	.040
READASL	.256	.388	.220	.453	.839	.355	6.538	.040
TEMPtrs	.143	.157	.308	.457	.450	.679	5.412	.033
LexDensity	.194	-.070	.003	.202	.495	.113	5.095	.032
LSAGN	.300	.237	.345	.610	.545	.697	4.027	.025

Table 1. Significant predictors with largest effect sizes (Study 1).

syntactic complexity, and coreferential cohesion. The five predictors with the greatest effect size are shown in Table 1.

A discriminant analysis (forced entry) was performed with paraphrase type as the dependent variable and the selected 20 indices as predictors. Univariate ANOVAs confirmed that the six paraphrase types differed significantly on each of the 20 predictors in the training set. Five discriminant functions were calculated of which the first two were significant ($p<0.001$). These two functions accounted for a cumulative variance of 71.2%.

The first significant variate (χ^2=229.643, df=100, $p<0.001$) had a canonical correlation of .359 and explained 47.5% of the variance, while for the second significant variate (χ^2=123.145, df=76, $p<0.001$) the correlation was lower at .262 and explained 23.7% of the variance. The first discriminant function had the strongest positive correlation with average sentence length (READASL) and the number of words (READNW) and strongest negative correlation with type-token ratio (TYPTOKc) and verb incidence (VERBi). The second variate correlated positively with paragraph-to-paragraph semantic overlap (LSAppa) and negatively with additive connective incidence (CONADDi). Accuracy on the held-out test set was 24.9% (chance for this analysis was 17%) with a fair kappa statistics of 0.10 ($p<.001$). A post-hoc analysis showed that the condensed condition was the only condition that could be significantly distinguished from the others. The modest performance results together with the outcome of the post-hoc analysis suggest that a new approach or set of indices may be needed to improve the overall performance of paraphrase type predictors. We will explore such avenues in the future as described in the Discussion and Future Work section.

Study 2: Appropriateness Analysis

We also wanted to understand the differences between true and false paraphrases along various linguistic dimensions. As the True-False division of the collected data set is greatly skewed towards True paraphrase (1029-154), we derived a balanced data set of 154-154 True-False paraphrases. A balanced data set is a precondition of ANOVAs. The data set was split into 66.9-33.1% train-test subsets. Similar to the previous experiment, we ran ANOVAs on all 82 Coh-Metrix indices resulting in 16 significant predictors. To avoid multicollinearity when running the DFA, we computed correlations among the 16 predictors. Among indices that correlated greater than .70 we kept the index with largest effect size. Furthermore, following a rule of having at least 20 data points available per predictor (Field, 2005), we decided to select only 5 predictors, with the largest effect sizes, for the next phase of building predictor models based on DFA. These five predictors are shown in Table 2. The DFA revealed one significant function (χ^2=27.743, df=5, $p<0.001$) which explained 100% of the variance. The function had a canonical correlation of .524. Correlations between predictor variables and the discriminant function suggested that semantic overlap between original passage and student paraphrase (LSAppa), temporal cohesion (TEMPta), and editing distance (MEWawm) were the best predictors of paraphrase polarity (true vs. false). The obtained accuracy obtained on the test set was 72.1% with a kappa of .442 ($p<0.001$). Precision was 75.00% and Recall was 71.79%.

Index	True Paraph.	False Paraph.	F(1,217)	η^2
LSAppa	.632	.469	35.936	.142
MEDawm	.116	.545	12.056	.052
TEMPta	.096	.564	11.315	.049
LEXDIVTD	.091	.518	8.043	.035
INTEi	.077	-.454	6.905	.030

Table 2. Significant predictors with largest effect sizes (Study 2).

Discussion and Future Work

A closer analysis of the linguistic indices along which the 6 types of paraphrases differed revealed that true paraphrases tend to have sentences similar in length with sentences in the original text, while false paraphrases differ more in length. Also, true paraphrases are more similar to original texts in terms of content word overlap and semantic overlap. The lexical diversity of true paraphrases is also more similar to that of original texts as compared to bad paraphrases. The true paraphrases are structurally more different from original texts while bad paraphrases are more similar. False paraphrases tend to use more pronouns compared to original texts. Based on the above observation we can conclude that when students are asked to paraphrase they tend to have sentences that are similar in length, have a considerable degree of content word and semantic overlap, and are structurally different from the original texts. These findings confirm some of the characteristics observed on existing sentential paraphrase datasets. In conclusion, when students are asked to paraphrase and they preserve the meaning of the original texts they do reuse much of the words but not structure of the original texts. The obvious question arises about the linguistic profile of expert-generated paraphrases. We will investigate in the future this issue and compare the results with the ones on student paraphrases described here.

One other avenue for future research is exploring new indices and classification models to improve the performance of the paraphrase type identification task described in Study 1. We will explore indices that quantify directly the overlap at various levels between student generated paraphrases and original texts instead of using indices which were designed for individual texts. Furthermore, we will explore binary classifiers instead of 6-way classifiers. For each type of paraphrase we will design one classifier that detects whether a pair of student paraphrase and original text is of that type or not. This will address case in which a student paraphrase is of several paraphrase types at the same time, which happens.

Acknowledgments

This research was supported in part by the Institute for Education Sciences (R305A100875, R305A080589) and National Science Foundation (0938239). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsoring agencies.

References

Barzilay, R., and Lee, L. (2003). Learning to Paraphrase: An Unsupervised Approach Using Multiple Sequence Alignment. In *Proceedings of NAACL* 2003.

Brockett, C., and Dolan, W. B. (2005). Support Vector Machines for Paraphrase Identification and Corpus Construction. In *Proceedings of the 3rd International Workshop on Paraphrasing*.

Corley, C., and Mihalcea, R. (2005). Measuring the Semantic Similarity of Texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment.* Ann Arbor,MI.

Dagan, I., Glickman, O., and Magnini B. (2006). *The PASCAL Recognising Textual Entailment Challenge*. In Quionero-Candela, J. Dagan, I. Magnini, B. d'Alch-Buc, F. (Eds.), Machine Learning Challenges. Lecture Notes in Computer Science, Vol. 3944, 177–190, Springer, 2006.

Dolan, B. Quirk, C. and Brockett, C. (2004). Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In Proceedings of COLING, Geneva, Switzerland.

Graesser, A.C. Olney, A. Haynes, B. and Chipman, P. (2005). Cognitive Systems: Human Cognitive Models in Systems Design. Erlbaum, Mahwah, NJ. chapter AutoTutor: A cognitive system that simulates a tutor that facilitates learning through mixed-initiative dialogue.

Ibrahim, A. Katz B. and Lin, J. (2003). Extracting Structural Paraphrases from Aligned Monolingual Corpora. In *Proceeding of the Second International Workshop on Paraphrasing*, (ACL 2003).

Iordanskaja, L. Kittredge, R. and Polgere, A. (1991). Natural Language Generation in Artificial Intelligence and Computational Linguistics. Lexical selection and paraphrase in a meaning-text generation model, Kluwer Academic.

Lintean, M., Moldovan, C., Rus, V., & McNamara, D. (2010). The Role of Local and Global Weighting in Assessing The Semantic Similarity Of Texts using Latent Semantic Analysis. Proceedings of the *23st International Florida Artificial Intelligence Research Society Conference*. Daytona Beach, FL.

McCarthy, P.M., & McNamara, D.S. (in press). The User Language Paraphrase Corpus. In P. M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution*. Hershey, PA: IGI Global.

McNamara, D.S. Boonthum, C. Levinstein, I. B. and Millis, K. (2007). Handbook of Latent Semantic Analysis. Erlbaum, Mahwah, NJ. chapter Evaluating self-explanations in iSTART: comparing word-based and LSA algorithms, 227–241.

McNamara, D. S., Raine, R., Roscoe, R., Crossley, S., Jackson, G. T., Dai, J., Cai, Z., Renner, A., Brandon, R., Weston, J., Dempsey, K., Lam, D., Sullivan, S., Kim, L., Rus, V., Floyd, R., McCarthy, P. M., & Graesser, A. C. (in press). The Writing-Pal: Natural language algorithms to support intelligent tutoring on writing strategies. In P. M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution*. Hershey, PA: IGI Global.

Qiu, L. Kan M. Y. and Chua T. S. (2006). Paraphrase Recognition via Dissimilarity Significance Classification. In *Proceeding of EMNLP, Sydney 2006.*

Rus, V. McCarthy, P. M. Lintean, M. McNamara, D. S. and Graesser, A. C. (2008a). Paraphrase Identification with Lexico-Syntactic Graph Subsumption. In *Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS-2008).*

Rus, V. McCarthy, P. M. McNamara, D. S. and Graesser, A. C. (2008b). A Study of Textual Entailment. International Journal of Artificial Intelligence Tools, August 2008.

Zhang, Y., and Patrick, J. (2005). Paraphrase Identification by Text Canonicalization, In *Proceedings of the Australasian Language Technology Workshop 2005, 160–166.*