The Components of Paraphrase Evaluations

Philip M. McCarthy
Rebekah H. Guess
Danielle S. McNamara

University of Memphis

Contact Information:
FedEx Institute of Technology
4th Floor, RM 410
Institute for Intelligent Systems (IIS)
University of Memphis
Tennessee
38152
pmmccrth@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. As such, paraphrasing is a feature of fields as diverse as discourse psychology, composition, and computer science. Although 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 an extensive database (N = 1998) of natural paraphrases generated by high school students that have been assessed along 10 dimensions (e.g., semantic completeness, lexical similarity, syntactical similarity). This study investigates the components of paraphrase quality emerging from these dimensions, and examines whether computational approaches 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 latent semantic analysis (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 (e.g., McNamara, 2004; McNamara, Ozuru, Best, & O’Reilly, 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 (Golightly & Sanders, 1997; 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, McCarthy, McNamara, & Graesser, 2008[a]). More specifically, users typically expect systems to respond within the boundaries of a normal conversational turn, about one-second (Cavazza, Perotto, & Cashman, 1999; Lockelt et al. 2007). This time 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 may risk
compromising user motivation and metacognitive awareness of the system’s learning goals (Graesser, Person, & Magliano, 1995; Millis et al., 2007). Thus, a paraphrase assessment must operate within a trade off between speed and accuracy.

Although 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, McCarthy, McNamara, & Graesser, 2008) has centered on edited paraphrases stemming from professional writers in data collections such as the Microsoft Paraphrase Corpus (Dolan, Quirk, & 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, & Polgereg 1991), question answering (Ibrahim, Katz, & 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 et al., 2007, McCarthy & McNamara, 2008; Renner, McCarthy, & McNamara, in press). Indeed, as Renner and colleagues demonstrate, 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 course-grained. Specifically, existing paraphrase data sets tend 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 summarization. That is, a system may retrieve any number of 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.

The need for appropriate feedback based on the analysis of the paraphrase means that binary coded paraphrase data is insufficient. Instead, paraphrase data are needed that are coded for all (or at least many) of the possible components that comprise paraphrases. Recalling our opening definition of a paraphrase, 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, we can see that a quality paraphrase may contain at least three distinct components of paraphrase: semantic completeness, lexical difference, and syntactic difference. Establishing a better understanding of which and how much of these dimensions contribute to overall paraphrase quality would facilitate paraphrase teaching. And importantly, training computational systems to recognize these features would facilitate automatic paragraph evaluation and subsequent feedback.

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 lead us to two major research questions that are the foci of this study.

1. What are the components of paraphrase? That is, which dimensions of paraphrase (and to what degree) constitute paraphrase quality?

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?

User-Language Paraphrase Challenge (The ULPC Corpus)

To begin to address the issues and questions outlined above, McCarthy and McNamara (2008) devised the User-Language Paraphrase Challenge (ULPC). The ULPC comprises a corpus of 1998 paraphrases written by American high-school students using the Intelligent Tutoring System, iSTART (McNamara, Levinstein, & Boonthum, 2004; see Table 1 for example responses when asked to paraphrase a sentence). Each paraphrase is evaluated by trained experts along 10 paraphrase dimensions (e.g., semantic completeness, lexical similarity, syntactical similarity) and is further evaluated computationally by 10 paraphrase assessment indices (e.g., latent semantic analysis, the Entailer, MED). The challenge invites researchers to use some, any, or all of the human paraphrase dimensions and some, any, or all of the computational assessments either alone or in combination with other approaches to better inform the research community regarding ITS user input paraphrase assessment. The ULPC corpus is unique in as much as it is the only large-scale, publically available collection of paraphrase data collected from an ITS, comprising text that is typical of ITS input and is evaluated for the
improvement of ITS assessment. This study represents the first analysis of the ULPC corpus.

*** Table 1 about here ***

Paraphrase Dimensions

As mentioned earlier, the UPLC corpus consists of 10 paraphrase dimensions. However, to address our research question in this study, we initially require just four of these dimensions: semantic completeness, lexical similarity, syntactic similarity, and paraphrase quality. Each of these dimensions is briefly described below.

Semantic completeness. Semantic completeness refers to the student’s paraphrase (user response) having the same meaning as the sentence targeted for paraphrasing (target sentence). For example, the following target sentence *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 increase muscles total heat production markedly in the body*. Semantic completeness is evaluated without regard to word- or structural-overlap between sentences. As such, if the user response is exactly the same as the target sentence then it is also semantically the same. The completeness element of the dimension accounts for the possibility that only part of the target sentence may have been targeted. Thus, a user paraphrase that only addresses one clause of a two clause target sentence will receive a partial evaluation, even if what the student has written is accurate.
Lexical similarity. Lexical similarity refers to the degree to which the same words were employed in the user response, regardless of syntax. For example, given the target sentence *Scanty rain fall, a common characteristic of deserts everywhere, results from a variety of circumstances*, one user response was *a common characteristic of deserts everywhere, results from a variety of circumstances, Scanty rain fall*. Note that the definition of lexical similarity means that a target sentence of *the dog chased the cat* would be identical to a student response of *the cat chased the dog* because each lexical item is present in both sentences. That is, lexical similarity may not necessarily account for semantics, just as semantic completeness may not necessarily account for lexical choice.

Syntactic similarity. Syntactic similarity refers to the degree to which similar syntax (i.e., parts of speech and phrase structures) is employed in the user response, regardless of the words used. For example, given the target sentence *An increase in temperature of a substance is an indication that it has gained heat energy*, one user response was *a raise in the temperature of an element is a sign that it has gained heat energy*. Similar to the lexical similarity dimension, the syntactic similarity dimension accounts for neither lexical choice nor semantics. As such, 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. For example, given the target sentence *Scanty rain fall, a common characteristic of deserts everywhere, results from a variety of circumstances*, one user response was *small amounts of rain fall, a normal trait of deserts everywhere, is caused*
from many things. Evaluators could take into account any of the other dimensions of paraphrase (and feasibly yet other dimensions) to any degree they thought appropriate.

**Rating the Paraphrases**

Each of the 1998 paraphrases in the corpus is rated on a 1-6 interval scale. Raters were informed that a value of 1 (minimum) to 3 (lower-median) could be interpreted as meaning *no, absent, or wrong*, with a rating of 1 having higher confidence than a rating of 2 and 3, respectively. Conversely, a rating of 4 (upper-median) to 6 (maximum) could be interpreted as *yes, present, or correct*, with a rating of 6 having higher confidence than a rating of 5 and 4, respectively. This rating scheme provides a scale that can be used as a continuous variable (1-6), a binary variable (1-3 vs 4-6), or a three part scale 1-2, 3-4, 5-6). Further, by evaluating each of the 10 dimensions, the raters could judge a response as, for instance, a 1 for paraphrase quality and a 6 for irrelevant, as is the case for the final example given in Table 1.

The raters were three advanced undergraduate students working in a cognitive science laboratory. All three students were majoring in the fields of either cognitive science or linguistics. Each rater completed 50 hours of training on a data set of 198 paraphrase sentence pairs from a similar experiment. The raters were given extensive instruction on the meaning of the paraphrase dimensions and given multiple opportunities to discuss interpretations. Numerous examples of each paraphrase type were highlighted to act as anchor-evaluations for each paraphrase type. Each rater was assessed on their evaluations and provided with extensive feedback. Following training, the 1998 paraphrases were randomly divided into three groups. Raters 1 and 2 evaluated Group 1 \((n = 655)\); Raters 1 and 3 evaluated Group 2 \((n = 680)\); and Raters 2 and 3 evaluated
Group 3 \((n = 653)\). The raters were given 4 weeks to evaluate the 1998 protocols across the 10 dimensions, for a total of 39,960 individual assessments.

**Inter-rater agreement**

Establishing an “acceptable” level of inter-rater agreement is no simple task. Although many studies report various inter-rater agreements as being *good, moderate,* or *weak,* such reporting can be highly misleading because it does not take into account the task at hand (Douglas Thompson, & Walter, 1988). For instance, assessing whether and the degree to which a paraphrase contains *garbage* is a far easier task than assessing whether and the degree to which a paraphrase is syntactically similar. As such, the inter-rater agreements for the ULPC should be interpreted for what they are: *the degree of agreement that has been reached by raters who have received 50 hours of extensive training.*

At this point, it is also important to note the over-arching goal of the ULPC. The purpose of establishing evaluations of user-language paraphrase is so that *ITSs may provide users with accurate, rapid assessment and subsequently facilitative feedback, such that the assessments are comparable to human raters.* However, as any student knows, even experienced and established teachers differ as to how they grade. Consequently, our goal in evaluating the paraphrases is to establish a reasonable gold standard for paraphrases and to have researchers attempt to replicate those standards computationally or statistically such that the assessments of user-language are comparable to raters who may not be perfect, but who are, at least, extensively trained and demonstrate reasonable and consistent levels of agreement.
The most practical approach to assessing the reliability of an approach is to report correlations of that approach with the human gold standards. If an approach correlates with human raters to a similar degree as human raters correlate with each other then the approach can be regarded as being as reliable as an extensively trained human. For this reason, we emphasize the correlations between raters in reporting here the inter-rater agreement and establishing the gold standard (see Table 2).

### Table 2 about here

#### Computational Indices

As mentioned earlier, the UPLC corpus was evaluated by 10 computational indices. Several of these indices are shallow measures (though feasibly diagnostic) and do not play a roll in this study\(^{\text{v}}\). For this study, we were primarily interested in computational indices with a richer history of textual similarity assessment. These included Latent Semantic Analysis (LSA; Landauer et al., 2007), the Entailer (Rus, McCarthy, McNamara, & Graesser, 2008[b]), and Minimal Edit Distances (MED: McCarthy et al., 2007, McCarthy et al., 2008). LSA has provided an effective assessment evaluation within many of the systems that analyze user-language (e.g., iSTART: McNamara et al., 2004; AutoTutor: Graesser et al., 2005); however, in more recent studies, *entailment* approaches (McCarthy et al., 2007, 2008; Rus et al., 2008[a], 2008[b]) have reported significant success, often out-performing LSA. In addition, the string-matching approach of MED, an index that emphasizes differences rather than similarities between paraphrases, has also been successful in recent studies (McCarthy et al., 2008).
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. For an example of entailment, consider the following two sentences:

1) The man drove to the store to buy some bread
2) The man went to the store.

The first sentence *entails* the second sentence: if *the man drove to the store*, regardless of why he drove there, then it must be true that *he went to the store*. However, note that the reverse is not true: if he *went to the store* then we do not know that he *drove* there, nor do we know why he went there.

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 such as the RTE challenge (Rus et al., 2008[b]), and successful also in paraphrase testing assessment (McCarthy et al., 2007; Rus 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., 2007, 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 3 for examples). Because MED assesses word similarity in terms of sentential positions, MED is hypothesized to be an ideal approach for syntactic similarity evaluation.

Recall that the ULPC invites existing and new approaches to assess paraphrases. 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 this index 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 *the elephant drank the water*, which receive high scores because they are both lexically and syntactically different (see examples in Table 3).
Results

Rater Analysis

The 1998 ULPC paraphrases are divided into training (1012 items, 50.7%), validation (337 items, 16.9%), and test sets (649 items, 32.5%). The purpose of this division is so that researchers may experiment with any number of possible hypotheses using the training set data, tweak those hypotheses using the validation data, and then test models using the test set data (see Witten & Frank, 2005). In this study, we take advantage of these divisions of data and report procedures using the training set data and final validation using the test set data.

Correlations (see Table 4) were computed for the 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.

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 (see Figure 1), 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.

*** Figure 1 about here ***

We conducted a hierarchical multiple regression analysis to determine the amount of variance associated with paraphrase quality that was accounted for 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 $R^2 = .702$, thus accounting for 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 also have to be considered. For example, the perception of the quality of the writing or the length of
Components of Paraphrase Evaluations BSC906

paraphrase relative to the target sentence may affect ratings. Writing quality is one of the 10 dimensions of the ULPC, and as with the other dimensions, it received a rating between 1-6 based on such features as grammatical correctness and spelling. The length of the paraphrase was not a dimension, but is easily calculated as the number of words in the response (calculated as the number of white spaces between words). These factors (writing quality and length of response) are not likely to be ignored in the raters’ evaluation of paraphrase quality, even while they may not explicitly feature in a typical 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 (see Renner, McCarthy, & McNamara, in press). 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 (see Table 5). 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.

*** Table 5 about here ***

The new composite variable (i.e., the 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 components such as writing quality 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 (semantics and syntax, and possibly length), then computational assessment becomes more easily directed.

**Computational Analysis**

Our second research question was: *Can a computationally light system* (e.g. LSA, the Entailer, MED) *assess paraphrase quality to a similar degree to that of humans*? The human analyses suggested that semantics and syntax were the primary components of paraphrase quality. As such, to address our second question, we assessed the indices of LSA and the Entailer (for semantics) and MED (for syntax). As with the human assessment, we used the ULPC division of the training and test set data.

Using the training set data (*n* = 1012), initial results suggested moderate to high correlations for all the computational candidate indices (see Table 6). 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).

***Table 6 about here***

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: \( r = .51; z\text{-diff} = 7.83, p < .001 \)).

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: \( r = .59; z\text{-diff} = 4.35, p < .001 \)). 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 and MED as predictor variables. Although previous research has consistently shown the Entailer to outperform LSA (Rus et al., 2008 McCarthy et al., 2008), the correlation values in this study suggest that LSA is a better predictor of paraphrase quality than the Entailer (LSA: \( r = .427 \); Entailer: \( r = .319 \); \( z\text{-diff} = 2.880, p = .004 \)). With LSA also producing the highest correlation for lexical similarity, we included LSA (rather than the Entailer) in the first stage (or block) of the analysis. A number of computational variables were potential candidates for the syntactic similarity component of the model. Nonetheless, the index with the highest correlation with syntactic similarity was MED (L) (\( r = -.742 \)) and thus it was included as the second variable.
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% 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 for one pair of raters (G1 human raters: $r = .52$), there was no significant difference between our model and their agreement. As such, our initial results can be described as promising.

To verify our reasoning, we 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 differences correlate 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 Validity of the Computational Model

To test the validity of the computational model, we generated a new composite variable based on the model generated from the training set data (see Table 7).

*** Table 7 about here ***

Applied to the test set data, 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 between one of the pairs of raters, 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, an assessment of syntactic similarity, and may also include evaluations of writing quality and/or differences in sentence length. Raters’ judgment of semantic completeness appears 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 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 computational 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 to producing a model that replicates human performance of these assessments.
References


Acknowledgments

This research was supported in part by the Institute for Education Sciences (IES: R305GA080589; R305G020018-02; and R305G040046), and in part by the National Science Foundation (NSF: IIS-0735682). The views expressed in this paper do not necessarily reflect the views of the IES or the NSF. The authors acknowledge the contributions made to this project by Angela Freeman, John Myers, Zhiqiang Cai, and Arthur Graesser.
Table 1: Examples of target sentences and student responses in the context of being asked to paraphrase the sentence.

<table>
<thead>
<tr>
<th>Target Sentence</th>
<th>Student Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sometimes blood does not transport enough oxygen, resulting in a condition called anemia.</td>
<td>Anemia is a condition that is happens when the blood doesn't have enough oxygen to be transported</td>
</tr>
<tr>
<td>During vigorous exercise, the heat generated by working muscles can increase total heat production in the body markedly.</td>
<td>If you don't get enought exercsie you will get tired</td>
</tr>
<tr>
<td>Plants are supplied with carbon dioxide when this gas moves into leaves through openings called stomata.</td>
<td>so u telling me day the carbon dioxide make the plant grows</td>
</tr>
<tr>
<td>Flowers that depend upon specific animals to pollinate them could only have evolved after those animals evolved.</td>
<td>the flowers in my yard grow faster than the flowers in my friend yard,i guess because we water ours more than them</td>
</tr>
<tr>
<td>Plants are supplied with carbon dioxide when this gas moves into leaves through openings called stomata.</td>
<td>asoyaskljgt&amp;Xgdjkjndcndvshhjaale johnson how would you llike some ice creacm</td>
</tr>
</tbody>
</table>
Note: Student paraphrases are recorded as the student entered them. That is, no corrections are made to the text.
Table 2: Correlations between ratings for semantic completeness, lexical similarity, syntactic similarity, and paraphrase quality for each of the Rater Groups (G1, G2, G3).

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Semantic</th>
<th>Lexical</th>
<th>Syntactic</th>
<th>Paraphrase Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>662.67</td>
<td>0.74</td>
<td>0.67</td>
<td>0.51</td>
<td>0.59</td>
</tr>
<tr>
<td>G1</td>
<td>655.00</td>
<td>0.69</td>
<td>0.76</td>
<td>0.57</td>
<td>0.52</td>
</tr>
<tr>
<td>G2</td>
<td>680.00</td>
<td>0.77</td>
<td>0.58</td>
<td>0.61</td>
<td>0.62</td>
</tr>
<tr>
<td>G3</td>
<td>653.00</td>
<td>0.76</td>
<td>0.66</td>
<td>0.35</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Notes: All $p < .001$
Table 3: Examples of values for MED (L), MED (S), and MED (LS), for the
target sentence, *The dog chased the cat.*

<table>
<thead>
<tr>
<th>Sentence</th>
<th>MED (L)</th>
<th>MED (S)</th>
<th>MED (LS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The dog chased the cat.</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>The cat chased the dog.</td>
<td>0.200</td>
<td>0</td>
<td>0.067</td>
</tr>
<tr>
<td>The cats chased the dogs.</td>
<td>0.400</td>
<td>0.200</td>
<td>0.267</td>
</tr>
<tr>
<td>The cat didn't chase the dog.</td>
<td>0.727</td>
<td>0.391</td>
<td>0.543</td>
</tr>
<tr>
<td>Elephants tend to be larger than mice.</td>
<td>1</td>
<td>0.774</td>
<td>0.867</td>
</tr>
</tbody>
</table>

Note: L refers to lexical, S to syntax, LS to both lexical and syntax
Table 4: Correlation matrix for the variables of paraphrase quality, Semantic completeness, lexical similarity, syntactic similarity, and Length difference between each sentence in the paraphrase

<table>
<thead>
<tr>
<th></th>
<th>Semantic</th>
<th>Lexical</th>
<th>Syntactic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality</td>
<td>0.774</td>
<td>0.451</td>
<td>0.035</td>
</tr>
<tr>
<td>Semantic</td>
<td></td>
<td>0.673</td>
<td>0.421</td>
</tr>
<tr>
<td>Lexical</td>
<td></td>
<td></td>
<td>0.579</td>
</tr>
</tbody>
</table>

Note: All correlations were significant at p < .001 except between syntactic similarity and paraphrase quality
Table 5: The unstandardized and standardized regression coefficients for the variables included in the human model

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>SE</th>
<th>Beta</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.649</td>
<td>0.119</td>
<td>5.456</td>
<td>&lt; .001</td>
<td></td>
</tr>
<tr>
<td>Semantic completeness</td>
<td>0.665</td>
<td>0.017</td>
<td>0.832</td>
<td>39.164</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Syntactic similarity</td>
<td>-0.390</td>
<td>0.020</td>
<td>-0.368</td>
<td>-19.710</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Writing quality</td>
<td>0.217</td>
<td>0.026</td>
<td>0.163</td>
<td>8.322</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Length difference</td>
<td>-0.008</td>
<td>0.005</td>
<td>-0.031</td>
<td>-1.625</td>
<td>0.104</td>
</tr>
</tbody>
</table>
Table 6: Correlations between computational indices (LSA, Entailer, and MED) and the paraphrase scores for semantic completeness, lexical similarity, syntactic similarity, and paraphrase quality in rank order of highest to lowest correlation.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paraphrase Quality</td>
<td>LSA</td>
<td>Entailer</td>
<td>MED (S)</td>
</tr>
<tr>
<td></td>
<td>0.427</td>
<td>0.319</td>
<td>-0.162</td>
</tr>
<tr>
<td>Semantic Completeness</td>
<td>Entailer</td>
<td>LSA</td>
<td>MED (S)</td>
</tr>
<tr>
<td></td>
<td>0.581</td>
<td>0.575</td>
<td>-0.416</td>
</tr>
<tr>
<td>Lexical Similarity</td>
<td>LSA</td>
<td>Entailer</td>
<td>MED (L)</td>
</tr>
<tr>
<td></td>
<td>0.818</td>
<td>0.800</td>
<td>-0.580</td>
</tr>
<tr>
<td>Syntactic Similarity</td>
<td>MED (L)</td>
<td>Entailer</td>
<td>LSA</td>
</tr>
<tr>
<td></td>
<td>-0.742</td>
<td>0.584</td>
<td>0.485</td>
</tr>
</tbody>
</table>
Table 7: The unstandardized and standardized regression coefficients for the variables included in the computational model

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>SE</th>
<th>Beta</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.806</td>
<td>0.240</td>
<td>3.366</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>LSA</td>
<td>2.366</td>
<td>0.190</td>
<td>0.438</td>
<td>12.458</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>MED (L)</td>
<td>1.883</td>
<td>0.202</td>
<td>0.305</td>
<td>9.325</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Length difference</td>
<td>-0.085</td>
<td>0.011</td>
<td>-0.267</td>
<td>-7.963</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
Figure 1: Semantic completeness, lexical similarity, and syntactic similarity in relation to paraphrase quality
For an example of reading see iSTART: http://csep.psyc.memphis.edu/istart/ and for an example of writing see W-Pal: http://w-pal.memphis.edu/

Full details of the ULPC and its associated corpus can be viewed at http://tinyurl.com/5bwo64. Note that the website does not contain analysis of the data but does offer an extensive description of the corpus. Because the description is extensive, and because we believe the corpus may be of value to many other researchers, we have chosen to publicize the data online and present the information as a challenge.

The remaining six dimensions may appear to be less directly connected to paraphrase quality. For instance, they include garbage: the degree to which a user has entered random text such as awgbieg3g73; frozen expressions: a binary evaluation as to whether a paraphrase begins with a statement such as “This sentence is saying …”; and elaboration: and an evaluation as to how thematically rather than semantically the sentences relate.

The frozen expression dimension is binary: includes/does not include a frozen expression.

The remaining, less sophisticated indices, are reported as providing lower correlations with the paraphrase dimensions. These indices include type-token ratio values, simple overlap values, and sentence length values (based on number of words).