Behavior Research Methods
2009, 41 (3), 682-690
doi:10.3758/BRM.41.3.682

The components of paraphrase evaluations

Philip M. McCarthy, Rebekah H. Guess, and Danielle S. McNamara
University of Memphis, Memphis, Tennessee

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 commonplace 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 = 1,998) 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 minimal edit distances (syntax) present promising approaches to simulating human evaluations of paraphrases.

Paraphrasing is the restating of a sentence such that 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 in fostering reading and writing skills. Not surprisingly, then, implementers of Intelligent Tutoring Systems (ITSs) that aim to teach reading and writing strategies[1] 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, 2008b). More specifically, users typically expect systems to respond within the boundaries of a normal conversational turn—about 1 sec (Cavazza, Perotto, & Cashman, 1999; Lockelt, Pfleger, & Reithinger, 2007). This time constraint severely limits its programming options that might lead to greater accuracy. Nonetheless, the accuracy of the judgment is equally important, because misleading or misdirected system feedback based on the evaluation may compromise user motivation and metacognitive awareness of the system's learning goals (Graesser, Person, & Magliano, 1995; Millis, Magliano, Wiemer-Hastings, Todaro, & McNamara, 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, & Brockett, 2005). Data such as these have a rich history of utility for developing approaches to paraphrase assessment in applications such as *natural language generation* (Iordanskaja, Kittredge, & Polgere, 1991), *question answering* (Ibrahim, Katz, & Lin, 2003), and *summarization* (Mani, 2001). Although 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 & McNamara, 2008; McCarthy et al., 2007; Renner, McCarthy, & McNamara, 2009). Indeed, as Renner and colleagues have demonstrated, 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)

P. M. McCarthy, pmmccrth@memphis.edu

evaluations tend to be coarse 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 from a list of options. In the case of an ITS, however, paraphrasing is often the subject being taught, so the system may have to choose the best candidate from 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 are insufficient. Instead, paraphrase data are needed that are coded for all (or at least many) of the possible components that constitute paraphrases. Recalling our opening definition of a paraphrase—*Paraphrasing is the restating of a sentence such that 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 have been based on text data that are far from characteristic of the input typical to ITSs. Such assertions lead us to the two major research questions that are the foci of this study.

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

2. Can computationally light systems (i.e., systems that can process and evaluate input within 1 sec) assess paraphrase quality to a similar degree as humans?

USER-LANGUAGE PARAPHRASE CHALLENGE (THE ULPC CORPUS)

To begin to address the issues and questions just outlined, McCarthy and McNamara (2008) devised the User-Language Paraphrase Challenge[2] (ULPC). The ULPC comprises a corpus of 1,998 paraphrases written by American high school students using the ITS iSTART (McNamara, Levinstein, & Boonthum, 2004; see Table 1 for example responses when subjects were 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, or LSA, the Entailer, and minimal edit distances, or 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 inasmuch as it is the only large-scale, publicly 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.

Paraphrase Dimensions

As mentioned earlier, the ULPC corpus consists of 10 paraphrase dimensions. However, to address our research question in this study, we initially require just four[3] 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 degree to which a student's paraphrase (user response) has the same meaning as the sentence targeted for paraphrasing (target sentence). For example, the target sentence *During vigorous exercise, the heat generated by*

Table 1
Examples of Target Sentences and Student Responses in the Context of Being Asked to Paraphrase the Sentence

Target Sentence	Student Response
Sometimes blood does not transport enough oxygen, resulting in a condition called anemia.	Anemia is a condition that is happens when the blood doesn't have enough oxygen to be transported
During vigorous exercise, the heat generated by working muscles can increase total heat production in the body markedly.	If you don't get enught exercsie you will get tired
Plants are supplied with carbon dioxide when this gas moves into leaves through openings called stomata.	so u telling me day the carbon dioxide make the plant grows
Flowers that depend upon specific animals to pollinate them could only have evolved after those animals evolved.	the flowers in my yard grow faster than the flowers in my friend yard,i guess because we water ours more than them
Plants are supplied with carbon dioxide when this gas moves into leaves through openings called stomata.	asoyaskljgt&Xgdjkjndcndvshhjaale johnson how would you llike some ice creacm

Note—Student paraphrases are recorded as the student entered them. That is, no corrections have been made to the text.

working muscles can increase total heat production in the body markedly was evaluated highly for the user response of *exercising vigorously icrease mucles total heat production markely in the body*. Semantic completeness is evaluated without regard to word or structural overlap between sentences. Thus, 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 was targeted. Thus, a user paraphrase that addresses only one clause of a two-clause target sentence will receive a partial evaluation, even if what the student wrote was accurate.

Lexical similarity. *Lexical similarity* refers to the degree to which the same words were employed in the user response, regardless of syntax or semantics. 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.* Such a response would be rated very highly for lexical similarity, regardless of the fact that the word order has been changed. 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*. Thus, the user response is highly similar in terms of syntax (and therefore, not a good example in terms of paraphrase quality). As with 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 that was judged highly was *Small amounts of rain fall,a normal trait of deserts everywhere, is caused from many things*. For this dimension, evaluators could take into account any of the other dimensions of paraphrase (and feasibly even other dimensions) to any degree that they thought appropriate.

Rating the Paraphrases

Each of the 1,998 paraphrases in the corpus is rated on a 1–6 interval scale.[4] 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 or 3. 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 or 4. 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 and 2, 3 and 4, 5 and 6). Furthermore, 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 h 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 his or her evaluations and was provided with extensive feedback. Following training, the 1,998 paraphrases were randomly divided into three groups. Raters 1 and 2 evaluated Group 1 ($n = 665$); 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 1,998 protocols across the 10 dimensions, for a total of 39,960 individual assessments.

Interrater Agreement

Establishing an acceptable level of interrater agreement is no simple task. Although many studies report various interrater agreements as being *good*, *moderate*, or *weak*, such reporting can be highly misleading, because it does not take into account the task at hand (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. Thus, the interrater 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 h of extensive training.

At this point, it is also important to note the overarching goal of the ULPC. The purpose of establishing evaluations of user-language paraphrase is so that ITSs can provide users with accurate, rapid assessment and subsequently facilitative feedback, such that the assessments are comparable to those of 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 with those of raters who may not be perfect but who, at least, are 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

Table 2
Correlations Between Ratings for Semantic Completeness, Lexical Similarity, Syntactic Similarity, and Paraphrase Quality for Each of the Rater Groups (G1, G2, and G3)

Group	*N*	Semantic	Lexical	Syntactic	Paraphrase Quality
G1	665.00	.69	.76	.57	.52
G2	680.00	.77	.58	.61	.62
G3	653.00	.76	.66	.35	.63
Average	662.67	.74	.67	.51	.59

Note—All *ps* $< .001$.

with the human gold standard. 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 the interrater agreement here and establishing the gold standard (see Table 2).

Computational Indices

As mentioned earlier, the ULPC corpus was evaluated by 10 computational indices. Several of these indices are shallow measures (though feasibly diagnostic) and do not play a role in this study.[5] For this study, we were primarily interested in computational indices with a richer history of textual similarity assessment. These included LSA (Landauer, McNamara, Dennis, & Kintsch, 2007), the Entailer (Rus, McCarthy, McNamara, & Graesser, 2008a, 2008b), and MED (McCarthy et al., 2007; McCarthy, Rus, Crossley, Graesser, & McNamara, 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, Chipman, Haynes, & Olney, 2005); however, in more recent studies, entailment approaches (McCarthy et al., 2007; McCarthy et al., 2008; Rus, McCarthy, et al., 2008a, 2008b) have reported significant success, often outperforming 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.

LSA. 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 *recognizing textual entailment* challenge (Dagan, Glickman, & Magnini, 2004–2005; Rus, McCarthy, et al., 2008b), and successful also in paraphrase assessment (McCarthy et al., 2007; Rus, Lintean, et al., 2008). As an entailment measure, we can assume that the second sentence in any pair is shorter than, 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 nonexpert) often requires more lexicon (and maybe more information) than is present in the target sentence. Therefore, 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.

MED. MED (McCarthy et al., 2007; McCarthy et al., 2008) assesses differences between any two sentences in terms of the words in the sentences and the position of the words in the sentences. As such, sentences with the same words may not be considered identical if the position of those words is different (see Table 3 for examples). Because MED assesses word similarity in terms of sentential positions, it is considered 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

Table 3
Examples of Values for MED (L), MED (S), and MED (LS) for the Target Sentence *The dog chased the cat*

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

Note—L, lexical; S, syntax; LS, both lexical and syntax.

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 1,998 paraphrases using the Charniak parser (Charniak, 2000). After the paraphrases had been parsed, each item was analyzed using the MED tool. The two sentences *The dog chased the cat* and *The cat chased the dog* receive low scores, because there are fewer differences between them than between, for example, the 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 1,998 ULPC paraphrases are divided into training (1,012 items, 50.7%), validation (337 items, 16.9%), and test (649 items, 32.5%) sets. The purpose of this division is so that researchers could 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.

Table 4
Correlation Matrix for the Variables of Paraphrase Quality, Semantic Completeness, Lexical Similarity, Syntactic Similarity, and Length Difference Between All Sentences in the Paraphrase

	Semantic	Lexical	Syntactic
Quality	.774	.451	.035
Semantic		.673	.421
Lexical			.579

Note—All correlations were significant at $p < .001$ except between syntactic similarity and paraphrase quality.

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 showed 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 relation-

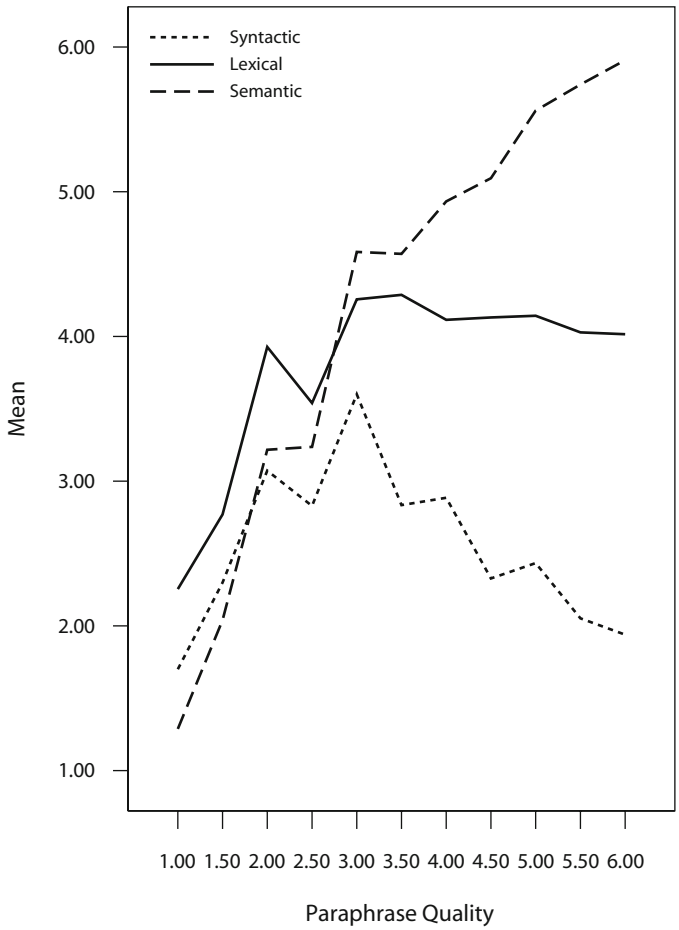

Figure 1. Semantic completeness, lexical similarity, and syntactic similarity in relation to paraphrase quality.

ship between syntactic similarity and paraphrase quality (see Figure 1), for which an S-curve best fit the data [$F(1010) = 134.57, p < .001$; $R^2 = .12$]. This curvilinear relationship contrasts with the linear relationships of semantic completeness and lexical similarity to paragraph quality. The curvilinear relationship between paraphrase quality and syntactic similarity suggests that both low and high evaluations of paraphrase quality are associated with low values of syntactic similarity. Thus, a paraphrase that is not at all related to a target sentence is syntactically different from the attempted paraphrase; and a paraphrase that is of high quality is also syntactically different from the target sentence.

We conducted a hierarchical multiple regression analysis to determine the amount of variance associated with paraphrase quality that was 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 [$F(2,1009) = 1{,}190.325$, $p < .001$, $r = .838$; adjusted $R^2 = .702$], 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 possible and may also have to be considered. For example, the perception of the quality of the writing or the length of the 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 and 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 that the meaning of the paraphrased sentence is more distant from the target sentence. Indeed, examining the corpus of paraphrases, 1,761 of the 1,998 (or 88%) contained some kind of grammatical or spelling error (see Renner et al., 2009). And, length of the 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.

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.

Table 5
Unstandardized and Standardized Regression Coefficients for the Variables Included in the Human Model

	B	*SE*	Beta	*t*	*p*
Constant	0.649	0.119		5.456	<.001
Semantic completeness	0.665	0.017	0.832	39.164	<.001
Syntactic similarity	−0.390	0.020	−0.368	−19.710	<.001
Writing quality	0.217	0.026	0.163	8.322	<.001
Length difference	−0.008	0.005	−0.031	−1.625	=.104

Table 6
Correlations Between Computational Indices (LSA, Entailer, and MED) and the Paraphrase Scores for Paraphrase Quality, Semantic Completeness, Lexical Similarity, and Syntactic Similarity, in Rank Order From Highest to Lowest Correlation

	Order of Highest Correlation		
Dimension	First	Second	Third
	LSA	Entailer	MED (S)
Paraphrase quality	0.427	0.319	−0.162
	Entailer	LSA	MED (S)
Semantic completeness	0.581	0.575	−0.416
	LSA	Entailer	MED (L)
Lexical similarity	0.818	0.800	−0.580
	MED (L)	Entailer	LSA
Syntactic similarity	−0.742	0.584	0.485

Computational Analysis

Our second research question was as follows: *Can a computationally light system* (e.g., LSA, the Entailer, MED) *assess paraphrase quality to a similar degree as humans can?* The human analyses suggested that semantics and syntax were the primary components of paraphrase quality. Thus, 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 = 1,012), 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, $p < .001$).

The best-performing variable for the dimension of lexical similarity was LSA (r = .82), a result that was significantly higher than that produced by human expert-to-expert correlations (compare human: r = .67; z_{diff} = 7.02, $p < .001$).

The best-performing variable for the dimension of syntactic similarity was MED (L) (r = −.74; note that MED calculates *differences*, hence the negative correlation); another result that was significantly higher than that produced by human expert-to-expert correlations (compare human: r = .51; z_{diff} = 7.83, $p < .001$).

Finally, the best-performing variable for the overall dimension of paraphrase quality was LSA (r = .43), a result that was significantly lower than that produced by human expert-to-expert correlations (compare human: r = .59; z_{diff} = 4.35, $p < .001$). Therefore, initial correlations suggest that computational tools are comparable with 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 (McCarthy et al., 2007; Rus, McCarthy, Lintean, Graesser, & McDaniel, 2007), 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_{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 with that of human experts' evaluations (compare average human raters, r = .590; model, r = .474). Although the human intercorrelations (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. Thus, 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 with 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
Unstandardized and Standardized Regression Coefficients for the Variables Included in the Computational Model

	B	*SE*	Beta	*t*	*p*
Constant	0.806	0.240		3.366	<.001
LSA	2.366	0.190	0.438	12.458	<.001
MED (L)	1.883	0.202	0.305	9.325	<.001
Length difference	−0.085	0.011	−0.267	−7.963	<.001

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 ($N = 1{,}998$). 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.

CONCLUSIONS

Our findings suggest that the components of a paraphrase include an assessment of semantic completeness, an assessment of syntactic similarity, and possibly 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. Although 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 for 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 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, so that the assessment would be comparable to that of one or more trained human raters. This study offers an important step toward 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 toward producing a model that replicates human performance of these assessments.

AUTHOR NOTE

This research was supported in part by the Institute for Education Sciences (Grants R305GA080589, R305G020018-02, and R305G040046), and in part by the National Science Foundation (Grant IIS-0735682). The views expressed in this article 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. Correspondence concerning this article should be addressed to P. M. McCarthy, FedEx Institute of Technology, 4th Floor, Room 410, Institute for Intelligent Systems, University of Memphis, Memphis, TN 38152 (e-mail: pmmccrth@memphis.edu).

REFERENCES

Cavazza, M., Perotto, W., & Cashman, N. (1999). The "virtual interactive presenter": A conversational interface for interactive television. In M. Diaz, P. Owezarsji, & P. Senac (Eds.), *Proceedings of the 6th International Workshop on Interactive Distributed Multimedia Systems and Telecommunications Services, IDSM '99* (pp. 235-243). Toulouse: Springer.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the First Conference of the North American chapter of the Association for Computational Linguistics* (pp. 132-139). San Francisco: Morgan Kaufmann.

Dagan, I., Glickman, O., & Magnini, B. (2004–2005). Recognizing textual entailment. www.pascal-network.org/Challenges/RTE.

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

Golightly, K. B., & Sanders, G. (Eds.) (1997). *Writing and reading in the disciplines*. Boston: Pearson.

Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, **48**, 612-618.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, **9**, 495-522.

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

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

Iordanskaja, L., Kittredge, R., & Polgere, A. (1991). Lexical selection and paraphrase in a meaning-text generation model. In C. L. Paris, W. R. Swartout, & W. C. Mann (Eds.), *Natural language generation in artificial intelligence and computational linguistics* (pp. 293-312). Norwell, MA: Kluwer.

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.) (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

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

Mani, I. (2001). *Automatic summarization* (Natural Language Processing, 3). Philadelphia: John Benjamins.

McCarthy, P. M., & McNamara, D. S. (2008). *The user-language paraphrase challenge*. Retrieved January 10, 2008, from https://umdrive.memphis.edu/pmmccrth/public/Paraphrase%20Corpus/Paraphrase_site.htm.

McCarthy, P. M., Rus, V., Crossley, S. A., Bigham, S. C., Graesser, A. C., & McNamara, D. S. (2007). Assessing Entailer with a corpus of natural language. In D. Wilson & G. Sutcliffe (Eds.), *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference* (pp. 247-252). Menlo Park, CA: AAAI Press.

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

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

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

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

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

Renner, A. M., McCarthy, P. M., & McNamara, D. S. (2009). Computational considerations in correcting user-language in an ITS environment. In C. H. Lane & H. W. Guesgen (Eds.), *Proceedings of the 22nd International Florida Artificial Intelligence Research Society Conference* (pp. 278-283). Menlo Park, CA: AAAI Press.

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

Rus, V., McCarthy, P. M., Lintean, M. C., Graesser, A. C., & McDaniel, D. (2007). Deep natural language processing for evaluating student self-explanations in iSTART. In *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference* (pp. 422-427). Menlo Park, CA: AAAI Press.

Rus, V., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2008a). Natural language understanding and assessment. In J. R. Rabuñal, J. Dorado, & A. Pazos (Eds.), *Encyclopedia of artificial intelligence* (pp. 1179-1184). Hershey, NY: IGI Global.

Rus, V., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2008b). A study of textual entailment. *International Journal on Artificial Intelligence Tools*, **17**, 659-685.

Thompson, W. D., & Walter, S. D. (1988). A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology*, **10**, 949-958.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann.

NOTES

1. 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/.

2. Full details of the ULPC and its associated corpus can be viewed at http://tinyurl.com/5bwo64. Note that the Web site does not contain an analysis of the data but does offer an extensive description of the corpus. Because the description is extensive, and because we believe that 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.

3. 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*, an evaluation as to how thematically rather than semantically the sentences relate.

4. The frozen expression dimension is binary; it includes/does not include a frozen expression.

5. 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 (on the basis of number of words).

(Manuscript received November 11, 2008;
revision accepted for publication February 19, 2009.)