A Linguistic Analysis of Expert-Generated Paraphrases

Russell D. Brandon[1], **Scott A. Crossley**[2], and **Danielle S. McNamara**[1]

[1]Department of Psychology, Learning Science Institute, Arizona State University, Tempe, AZ 85287
[2]Department of Applied Linguistics/ESL, Georgia State University, Atlanta, GA 30302
russell.brandon@asu.edu, scrossley@gsu.edu, dsmcnamara1@gmail.com

Abstract

The authors used the computational tool Coh-Metrix to examine expert writers' paraphrases and in particular, how experts paraphrase text passages using condensing strategies. The overarching goal of this study was to develop machine learning algorithms to aid in the automatic detection of paraphrases and paraphrase types. To this end, three experts were instructed to paraphrase by condensing a set of target passages. The linguistic differences between the original passages and the condensed paraphrases were then analyzed using Coh-Metrix. The condensed paraphrases were accurately distinguished from the original target passages based on the number of words, word frequency, and syntactic complexity.

Introduction

Paraphrasing is the restatement of a passage such that both the original passage and the restated passage are recognized as lexically and syntactically different, but both passages contain the same propositional meaning (McCarthy, Guess, & McNamara 2009). Although paraphrasing is defined as the simple restatement of a passage, paraphrase identification and classification is computationally challenging. It is difficult to identify paraphrases automatically using only word-to-word correspondence because paraphrases often involve the alteration of entire clauses rather than just words (Barzilay & Lee, 2003). This challenge has inspired research on the use of natural language processing (NLP) to automate paraphrase recognition (Rus, McCarthy, & Lintean, 2008). In this study, we examine the linguistic features that characterize good paraphrases, specifically, paraphrases produced by skilled writers.

The automatic identification of paraphrases is important for many NLP applications. In reading, paraphrasing is an ideal means for students to begin a self-explanation for deeper text comprehension (McNamara, 2004). When students practice self-explaining in an intelligent tutoring system (such as iSTART; McNamara, Levinstein, & Boonthum, 2004), it is necessary to identify when a student has paraphrased to give appropriate feedback. Within the Writing Pal tutoring system (W-Pal; Dai, Raine, Roscoe, Cai, & McNamara, 2011), students are provided with writing strategy training followed by opportunities to practice the strategies. NLP algorithms must be able to accurately identify and assess students' paraphrasing strategies to provide appropriate feedback and guidance.

In this study, we analyze and model paraphrases written by expert writers. Differences between condensed paraphrases and original passages were assessed via the computational tool, Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004). Coh-Metrix analyses text on several dimensions of lexical cohesion and reports on a variety of lexical indices such as psycholinguistic information about words (e.g., familiarity and imagability from the MRC database; Coltheart, 1981), semantic word features and relations (e.g., polysemy and hypernymy values from WordNet; Fellbaum, 1998), word frequency (from the CELEX database; Baayen, Piepenbrock, & Gulikers, 1995), and lexical diversity indices (McCarthy & Jarvis, 2010). Coh-Metrix also provides indices related to syntax using a parser based on Charniak (2000).

Method

We collected text passages and instructed expert writers to paraphrase these passages using condensing strategies. The linguistic differences between the condensed expert paraphrases and the original passages were analyzed.

Corpora

Expert raters were given a set of 99 passages excerpted from 184 essays written by undergraduate students. The passages were categorized by type (introduction, conclusion, evidence, and topic) and by

essay quality (low or high). The experts were given the following instructions for paraphrasing:

> Please provide a condensed paraphrase for the original passage above. Condensed paraphrasing means restating an idea that you have read using fewer words or sentences. When you paraphrase and condense you must make sure that it all still makes sense to your reader. However, you don't always have to change the words and structure. If you know a better way or different way to say something, then it's fine to use that way.

The final corpus contained 99 original passages and 297 condensed paraphrases (i.e., 100 paraphrases from each expert). We used Coh-Metrix to compute linguistic features for the passages and paraphrases. We then examined which linguistic features showed differences between the passages and the paraphrases using an ANOVA. A discriminant function analysis (DFA) was used to classify the passages and paraphrases.

Coh-Metrix Analysis

For this analysis, we selected Coh-Metrix indices related to cohesion, lexical sophistication, and syntactic complexity. These indices were selected because they have theoretical overlap with common paraphrasing strategies (i.e., changing the structure and vocabulary found in a sample).

Number of words. Coh-Metrix calculates the number of words in a text and other basic properties.
Word frequency. Word frequency indices measure how often words occur in the English language (CELEX).
Word polysemy. Polysemy measures the number of senses a word has (WordNet).
Word hypernymy. Hypernymy assesses the specificity of words in a text (WordNet).
Word concreteness. Words that refer to an object, material, or person generally receive a higher concreteness score than abstract words (MRC Psycholinguistic Database).
Word familiarity. Familiarity measures how likely words are to appear in spoken discourse (MRC Psycholinguistic Database).
Word imageability. Indicates whether words easily evoke imagery (MRC Psycholinguistic Database).
Word meaningfulness. Measures the strength of association a word has to other words (MRC Psycholinguistic Database).
Connectives and Logical Operators. Coh-Metrix reports connectives for two dimensions: positive versus negative connectives and additive, temporal, and causal connectives. The logical operators include variants of *or, and, not,* and *if-then* combinations.
Causality. Coh-Metrix calculates causality through the ratio of causal verbs to causal particles. The causal verb count is based on the number of main causal verbs (WordNet).
Syntactic complexity. Coh-Metrix measures the mean number of modifiers per noun phrase or the mean number of high-level constituents per word and per noun phrase.

Statistical Analysis

An ANOVA was conducted to select variables that best distinguished original passages and condensed paraphrases. A DFA was then used to predict group membership (the original passages or the condensed paraphrases) using a series of independent variables (the linguistic indices selected from the ANOVA). We used the DFA first on the entire set. A leave-one-out cross-validation model was then used in which one instance in turn is left out with 395 remaining instances as the training set. We report the findings of the DFA using an estimation of the accuracy of the

Table 1

ANOVA results for selected linguistic indices: Means, standard deviations, f value, p value, and hp^2

	Original	Condensed	*f* value	*p* value	hp^2
Number of words	41.980 (28.317)	16.686 (11.654)	449.865	< .001	0.605
CELEX word frequency in sentence	2.796 (0.238)	2.472 (0.373)	230.891	< .001	0.440
Word hypernymy	1.738 (0.537)	2.086 (0.661)	83.685	< .001	0.222
Number of causal verbs and particles	52.245 (44.244)	78.814 (73.317)	57.147	< .001	0.163
Word familiarity every word	592.703 (8.799)	587.686 (13.396)	46.349	< .001	0.136
Word imagability every word	325.334 (23.764)	338.464 (36.407)	46.403	< .001	0.136
Number of words before main verb	5.387 (4.669)	3.524 (3.081)	39.222	< .001	0.118
Word polysemy	4.239 (1.259)	3.919 (1.529)	9.267	< .010	0.031
Incidence of all connectives	94.381 (46.623)	80.964 (68.770)	9.492	< .010	0.031

analysis by plotting the correspondence between the actual texts and the predictions made by the DFA model. Results are in terms of recall, precision, and F1 score. Precision scores are computed by tallying the number of hits over the number of hits + misses. Recall is the number of correct predictions divided by the sum of the number of correct predictions and false positives. The F1 score is a weighted average of the precision and recall results.

Analysis and Results

ANOVA

Repeated measures ANOVAs were conducted using the selected Coh-Metrix indices as the independent variables and the original passages and condensed paraphrases as the dependent variables. Descriptive statistics for the selected Coh-Metrix indices are presented in Table 1. The indices from each measure with the greatest effect size were selected for the DFA.

Collinearity

Two of the 11 variables demonstrated multicollinearity ($r > .70$). These variables were *concreteness scores every word* and *meaningfulness scores every word*, both correlated highly with *word imagability every word.* Because *word imagability every word* demonstrated a greater effect size with our dependent variables, it was retained while the other indices were removed from the analysis. The tolerance checks of VIF values and tolerance levels for remaining indices were approximately 1, indicating that the model data did not suffer from multicollinearity. Descriptive statistics for the final variables are located in Table 1.

DFA

The Wilks's Lambda for the discriminant function was significant, $\Lambda = .66$, $\chi^2 = 161.825$, $p < .001$. The results demonstrate that the DFA correctly classified 320 of the 396 texts in the total set ($df = 1$, $n = 39$, $\chi^2 = 98.502$, $p < .001$) for an accuracy of 88.8%. For the cross-validated set, the DFA correctly allocated 316 of the 396 samples for an accuracy of 79.8% (see Table 2). The measure of agreement between the actual text type and the model produced a Cohen's Kappa of 0.498. The precision and recall scores for predicting sample type are presented in Table 2. The accuracy for the total set was .749 and the cross-validated set was .738.

The discriminant function coefficients (DFC) from the discriminant analysis correspond to the partial contributions in the discriminant function. The indices that most strongly contributed to classifying the texts as either original passages or condensed paraphrases were the number of words, word frequency, and the mean number of words before the main verb. The standardized discriminant function coefficients are reported in Table 3.

DFA without Text Length Index

We conducted a post-hoc analysis to examine the predictive strength of the Coh-Metrix indices in the absence of the text length index to assess how well purely linguistic indices distinguished original passages from condensed passages. We used the same corpus of passages and paraphrases as our dependent variables and used all indices reported by Coh-Metrix except the text length index as independent variables.

Table 2

Classification results for among grade level analyses

		Original	Condensed
Total set	Original	64	35
	Condensed	41	256
Percentage	Original	64.60	35.40
	Condensed	13.80	86.20
Cross-Validated	Original	63	36
	Condensed	44	253
Percentage	Original	63.60	36.40
	Condensed	14.80	85.20

Table 3

Standardized Discriminant Function Coefficients

Index	Coefficient
Number of words	0.802
CELEX word frequency	0.315
Syntactic complexity	0.202
Word polysemy	0.181
Word imagability	0.052
Causal verbs and particles incidence	0.017
Word familiarity	0.015
Connectives incidence	-0.028
Word hypernymy	-0.130

The Wilks's Lambda for the discriminant function was significant, $\Lambda = .815$, $\chi^2 = 79.822$, $p < .001$. The results demonstrate that the DFA correctly classified 274 of the 396 texts in the total set ($df = 1$, $n = 395$, $\chi^2 = 50.535$, $p < .001$) for an accuracy of 69.2%. For the cross-validated set, the DFA correctly allocated 271 of the 396 samples for an accuracy of 68.4%. The strongest classifiers were word frequency and syntactic complexity (with discriminant function coefficients of .736 and .449 respectively).

Discussion and Conclusion

The goal of this study was to develop machine learning algorithms to aid in the automatic detection of paraphrases and paraphrase types. The results indicated that original passages and condensed paraphrases can be discriminated using linguistic features related to text length, word sophistication and syntactic complexity.

Number of words made the largest contribution in the DFA. Word frequency and syntactic complexity in the absence of a number of words indices remained strong predictors of whether a passage was original or condensed. These two indices classified 69% of the paraphrases and passages correctly.

Content word frequency demonstrated that the process of paraphrasing involves the production of more infrequent words. The use of more infrequent words is likely related to the removal of frequent words when condensing and the condensing of longer phrases into more concise phrases using rarer words.

Syntactic complexity was also an important contributor in the DFA with original passages containing greater syntactic complexity than condensed paraphrases. This is likely because number of words before the main verb was reduced, shortening the length of the sentence.

The ANOVA and DFA results demonstrated that many lexical indices and indices of cohesion were indicators of condensing paraphrases. The lexical indices demonstrated that experts used more specific words that were more imagable and less ambiguous but also less familiar. Expert paraphrasers also used fewer connectives and greater causality. Experts condense by using more distinct words that are less familiar and by reducing connected phrases. However, cohesion is increased through the use of more causal verbs and particles.

These analyses show that differences exist at the lexical and syntactic level as well as with text length. These results have multiple applications. These features may be used to identify paraphrases with automated algorithms. They may also inform processes in which experts engage when they paraphrase (Petrić & Czárl, 2003). The algorithm developed can also be extended to intelligent tutoring systems to assess when a passage has been paraphrased using condensing strategies.

To further this research, future studies might focus on other types of paraphrases (paraphrasing when words and structure are changed, but the passage is not condensed) and on other indicators of paraphrases (e.g., topic continuity and rhetorical features). The application of the current findings should also be tested in pedagogical settings. Finally, while the results of this study indicate that assessing the type of paraphrase produced using machine learning algorithms is successful, the success of such algorithms in assessing the quality of a paraphrase is as yet untested.

Acknowledgments

This research was supported in part by the Institute for Education Sciences (R305A080589, R3056020018-02). We thank Vasile Rus, Art Graesser, and Zhiqiang Cai for their assistance in this research. We also thank Brad Campbell, Meg Henderson, and Tyler Trimm for providing the paraphrases used in this analysis.

References

Baayen, R. H., Piepenbrock, R., & Gulikers, L. 1995. *CELEX*. Philadelphia, PA: Linguistic Data Consortium.

Barzilay, R., & Lee, L. 2003. Learning to paraphrase : an unsupervised Approach Using Multiple-Sequence Alignment. Proceedings of HLT-NAACL. 16-23. Edmonton, Canada.

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

Coltheart, M. 1981. The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology* 33: 497-505.

Dai, J., Raine, R. B., Roscoe, R., Cai, Z., & McNamara, D. S. 2011. The Writing-Pal tutoring system : Development and design. *Computer*, *2* 1-11.

Fellbaum, C. 1998. *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

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

McCarthy, P. M., Guess, R. H., & McNamara, D. S. 2009. The components of paraphrase evaluations. Behavior Research Methods. 41:682-690.

McCarthy, P. M., Jarvis, S. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*. 42.

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 training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers* 36: 222-233.

Petrić, B., & Czárl, B. 2003. Validating a writing strategy questionnaire. System 31:187-215.

Rus, V., McCarthy, P. M., Lintean, M. C., McNamara, D. S., & Graesser, A. C. 2008. Paraphrase identification with Lexico-Syntactic graph subsumption. *Artificial Intelligence*, 201-206.

Wilson, M. D. 1988. The MRC psycholinguistic database: Machine readable dictionary, version 2. *Behavioural Research Methods, Instruments, and Computers*, 20: 6–11.