Listening. Learning. Leading.®

Research Report
ETS RR–12-18

Examining Linguistic Characteristics of Paraphrase in Test-Taker Summaries

Jill Burstein

Michael Flor

Joel Tetreault

Nitin Madnani

Steven Holtzman

October 2012

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Examining Linguistic Characteristics of Paraphrase in Test-Taker Summaries

Jill Burstein, Michael Flor, Joel Tetreault, Nitin Madnani, and Steven Holtzman
ETS, Princeton, New Jersey

October 2012

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Associate Editor: James Carlson

Reviewers: Paul Deane and Beata Beigman Klebanov

Abstract

This annotation study is designed to help us gain an increased understanding of paraphrase strategies used by native and nonnative English speakers and how these strategies might affect test takers' essay scores. Toward that end, this study aims to examine and analyze the paraphrase and the types of linguistic modifications used in paraphrase in test-taker responses and differences that may exist between native and nonnative English speakers. We are also interested in how these factors might influence final essay score. Outcomes discussed in this report can be used to inform the development of new *e-rater*® scoring engine features that capture information related to paraphrase, specifically in nonnative speaker responses to the *TOEFL*® exam integrated writing task.

Key words: summary writing, annotation study, paraphrase strategies, nonnative English speaker writing

Acknowledgments

We would first like to thank the TOEFL Research Program for funding this research. We would like to thank Adjua McNeil and Jennifer Lentini for completing the annotation work in this project. We are very grateful to Bill Dolan and Chris Brockett at the Microsoft Corporation for early discussions that informed the annotation protocol, and to Nitin Madnani for discussions throughout this project—in the beginning toward the development of the annotation scheme and at completion for reviews and very helpful discussion.

Table of Contents

List of Tables

Page

List of Figures

Page

A significant body of paraphrase research in computational linguistics investigates paraphrase detection, paraphrase generation, and corpus construction, especially with regard to machine translation (Bannard & Callison-Burch, 2005; Barzilay & Lee, 2003; Barzilay & McKeown, 2001; Bond, Nichols, & Appling, 2008; Callison-Burch, 2008; Callison-Burch, Koehn, & Osbourne, 2006; Cohn, Callison-Burch, & Lapata, 2008; Dolan, Quirk, & Brockett, 2004; Madnani, Ayan, Resnick, & Dorr, 2007; Madnani, Resnick, Dorr, & Schwartz, 2008; Marton, Callison-Burch, & Resnik, 2009; Pang, Knight, & Marcu, 2003). This research has traditionally investigated paraphrase based on well-formed text (e.g., Associated Press newswire). At ETS, paraphrase detection research has been ongoing for noisy, test-taker response text in primarily two assessment contexts: Essay Similarity Detection (ESD) research that addresses possible plagiarism in test-taker essay responses and *c-rater*™ that addresses short-answer content evaluation (Leacock & Chodorow, 2003; Sukkarieh & Blackmore, 2009; Sukkarieh & Bolge, 2008). There is a body of work around latent semantic analysis (LSA) techniques (Foltz, Gilliam, & Kendall, 2000; Landauer, Laham, Rehder, & Schreiner, 1997) that is used in Pearson writing evaluation products and automated essay scoring contexts to evaluate the topical coverage and, to some extent, correct content. In contrast to content vector analysis methods for vocabulary modeling used at ETS in the *e-rater*® scoring engine, LSA can detect different but related words that are likely to appear in a text on a particular topic. Therefore, one could argue that LSA is capturing paraphrase to some extent. LSA techniques, of course, handle content only at the word level, so this method does not take into account larger text units, such as multiword phrases or sentences.

None of the work cited above examines paraphrase in noisy data, nor does it investigate the linguistic modification strategies that writers use to produce paraphrase. However, a small body of existing classification and annotation research—to support the development of paraphrase detection systems designed for noisy data collected in instructional and assessment settings—is relevant to our specific interest in paraphrase. Culicover (1968), for instance, developed a linguistically based framework to generate and detect paraphrase to support research in information retrieval. His research contributed to methods in which one might automatically generate semantically similar information in variant linguistic forms (e.g., “John has a car and his wife, Mary, has a car” versus “John and his wife, Mary, have two cars between them”). In the context of his information retrieval research, Culicover developed a linguistic classification

framework that can be easily applied to the annotation of paraphrase in test-taker writing. We will discuss later in this report how we adapted Culicover's framework for this study.

More recently, Keck (2006) investigated paraphrase in native speaker and nonnative speaker summary writing. This work grew from research geared to examine teaching methods that would support student writers' understanding about the difference between plagiarism (i.e., verbatim copying from text) and paraphrase (Purdue University Writing Lab, 2006, Yamada, 2003). Keck drew on previous work on plagiarism (Campbell, 2000; Shi, 2004) to develop an annotation scheme for her research. Similar to our research interest, Keck was interested in the extent of paraphrase versus verbatim copying from a source text between native English speakers (L1) and nonnative English speakers (L2), and also the difference in types of paraphrase between the two groups, using four categories that capture a continuum of verbatim copying to significant paraphrase. Keck's analysis framework captured the continuum of *copying-to-paraphrase* in student summaries with these categories: near copy, minimal revision, moderate revision, and substantial revision. These four *paraphrase categories* essentially capture the extent of word overlap between the source and summary response texts. Categories that imply less copying, such as moderate or substantial revision, are intended to indicate that some lexical or syntactic change has occurred in the paraphrase. Keck's study is the closest to the work described in this paper. Her study used the four categories mentioned above to classify *attempted paraphrase* in L1 and L2 undergraduate summary writing, where students had 45 minutes to summarize a 1,000-word text. Two main coding categories were then used to label the reuse of words or strings of words from the source text that appeared in a student's summary. *Unique links* are words or strings of words that appear only once in the source text, and *general links* are words or strings of words that appear multiple times in the source text. In Keck's study, unique and general links were coded in sentences if the sentence was also identified as an attempted paraphrase, meaning that some kind of change was observed in a sentence. After annotation was completed, the summary data were categorized into the four paraphrase categories based on the proportion of overlap determined by the number of unique and general links between the source and summary text (refer to Keck, 2006, for details). Outcomes from Keck's study indicated that nonnative speakers used more of the minimal revision types (i.e., more instances of copying), while native speakers made more substantial revisions (i.e., more paraphrase was observed).

As discussed earlier, there is very little research along the lines of Keck (2006) that would be related to our research interest, yet this strand of research is important in the context of automated essay scoring to help us understand the extent to which nonnative speakers are able to paraphrase, what linguistic structures they may rely on for producing paraphrase, and how these factors might influence essay scores. In the ETS context, studies that support paraphrase detection are critical to improving automated essay scoring, and especially so for the *TOEFL*® exam integrated writing task. This task requires test takers to summarize (paraphrase) a lecture and a reading passage that contain contrasting points of view. The writing construct for scoring these responses assesses a test taker's ability to paraphrase the contrasting viewpoints in the reading and lecture stimuli, respectively. While e-rater is currently used to score responses from this task, the system has no explicit feature that captures paraphrase in test-taker responses.

This annotation study is designed to help us gain an increased understanding of paraphrase strategies used by native and nonnative English speakers and how these strategies might affect test takers' essay scores. Toward that end, this study aims to examine and analyze the extent (amount) of paraphrase and the types of linguistic modifications used in paraphrase in test-taker responses, and differences that may exist between native and nonnative English speakers. We are also interested in how these factors might influence final essay score. What we learn from the outcomes of this exploratory research can then be used to inform the development of new e-rater features that capture information related to paraphrase, specifically in nonnative speaker responses to the TOEFL integrated writing task.

Details of the annotation framework and procedure, the analyses of the paraphrase annotations, and the analyses of relationships between paraphrase and response score are discussed in the report.

Annotation

Annotators

Two annotators were identified in the ETS Philadelphia office. Both annotators work in literacy research and have formal linguistic knowledge; one of the annotators has a master's degree in linguistics. A 2-day training session took place.

Data Sets

The existing TOEFL integrated data set that was used contained native and nonnative English speaker responses (see Gurevich & Deane, 2007). All data are summaries in response to one prompt (see Appendix A).[1]

For purposes of training, a set of 47 summary responses was randomly selected from across the full data set. The set included 25 native and 22 nonnative test-taker responses. The remaining data were used for additional annotation after the training was completed. When training was completed, annotators labeled 100 new responses (50 native and 50 nonnative). These 100 responses were annotated without discussion between the annotators. The analyses are based on this set of 100. All responses were double-annotated. Data set sizes are illustrated in Table 1.

Table 1

Data Set Sizes

Data set description	*N*
Nonnative	
Training	25
Posttraining	50
Subtotal	75
Native	
Training	22
Posttraining	50
Subtotal	72
Total	77

Scheme Framework and Tool Development

Scheme framework. As discussed earlier, there appears to be very little work addressing the annotation of paraphrase and its correspondence to specific linguistic structures. Keck's (2006) paraphrase labeling categories looks primarily at overlap between words and reference texts summarized by native and nonnative speaker writers. However, Keck does not examine specific linguistic structures used to generate paraphrase. We are not aware of any previous work that addresses a detailed linguistic analysis of paraphrase in test-taker writing and that also examines differences between native and nonnative English speaker data with regard to the

extent of paraphrase production and kinds of linguistic structures used by writers to produce paraphrase.

After reviewing the limited amount of research in this area, we adapted the Culicover (1968) framework for use in the annotation task. This framework is discussed in Madnani and Dorr (2010). This framework was built into ParaMarker, a graphical user interface developed specifically for paraphrase annotation in this study.[2] (See Appendix C.)

Annotator tasks and training. In this study, annotators performed two tasks: (a) to identify a text segment from the prompt stimuli, along with its corresponding paraphrase in the test-taker response, and (b) to classify the paraphrased response text in a linguistic modification category: lexical, syntactic, or conceptual. A training protocol was developed. (See Appendix D.) One of the authors conducted a formal 2-day training session at ETS's Philadelphia office with both annotators. Annotators did subsequent training on their own, and the authors continued to communicate regularly but remotely with the annotators.

During the formal training sessions, annotators worked with one of the authors to learn how to use the tool. Annotators also worked through the protocol to understand the meaning of a paraphrase unit (P-unit) and to practice not only labeling the units of paraphrased text between the prompt and the summary, but also assigning appropriate linguistic modification categories to each P-unit. During this time, a training set of five summary responses was annotated using paper and pencil and an additional set of 10 to 15 summary responses was annotated collaboratively using the tool. In both sets, the author and annotators discussed all labeling decisions. Once the formal training was completed, the annotators began using the tool to complete the training set data. During this time, they annotated sets of five to 10 responses and agreement was computed after each set. Agreement for labeling of paraphrases was relatively low, so annotators were asked to discuss the differences on summary data that they had already annotated before moving onto the next set. Agreement was computed until the training data were completed. Agreement did remain relatively low for P-unit identification, but this did not seem to adversely affect the patterns of results reported later in the paper, especially with regard to how different aspects of paraphrase appear to affect final essay score. Interannotator agreement for the posttraining annotation sets is reported later in this paper.

Here we present a high-level description of how the annotators were asked to approach the task. (See Appendix D for details.) As the training proceeded, the consensus was that some

changes to the protocol needed to be implemented. For instance, while the annotation scheme requested that annotators assign all P-units and then subsequently assign the linguistic classifications, the annotators found it easier to assign the classification immediately following the identification and assignment of an individual P-unit.

A P-unit is a pair of text segments that contains (a) a text segment(s) from the prompt stimuli and (b) a text segment(s) that paraphrases the text segment in the test-taker response from the prompt. A P-unit can be of any length. For instance, the annotator could identify a complete sentence in the prompt text that was paraphrased in the essay as multiple sentences or even as a smaller fragment of a single sentence. As well, part of a sentence in the prompt could be paraphrased in the essay in one or more sentences in the test-taker response. In some cases, annotators believed that a sentence might summarize (paraphrase) the entire higher-level idea in the reading, the lecture, or both. In these cases, they could not find a specific segment of text in the prompt that matched but wanted to be able to categorize this summary. Therefore, for annotation purposes, they selected only the text segment in the test-taker response (essay) and classified it based on the section of the prompt stimuli that it summarized: global paraphrase—reading, global paraphrase—lecture, or global paraphrase—reading and lecture.

Identifying the P-units was a difficult and time-intensive task. All P-units identified were then classified as lexical, syntactic, or conceptual. Note that these categories are not mutually exclusive. For instance, a P-unit might be classified in lexical and syntactic categories. High-level and fine-grained linguistic modification categories are listed and described in Table 2.

Table 2

Linguistic Classifications

Classification	Description
Syntactic paraphrase	
Active-passive	An active sentence has been paraphrased as a passive sentence or vice versa
Declarative-question	A declarative sentence in the prompt has been paraphrased as a question or vice versa
Verb aspect shift	Paraphrase from the prompt text involves verb aspect shift (e.g., *can work* to *work*)
Verb tense shift	Paraphrase from the prompt text involves verb tense shift (e.g., *work* to *will work*)

Classification	Description
Finite-nonfinite VP	Paraphrase from the prompt text involves finite to nonfinite verb phrase or vice versa (e.g., *managed to become* to *became*)
Pronoun-NP	Paraphrase from the prompt text involves pronominalization of a noun phrase or vice versa (e.g., *the project* to *it*)
Relative clause-NP	Paraphrase from the prompt text involves a transformation from a relative clause to a noun phrase or vice versa (e.g., *directions that might not work* to *the wrong directions*)
Relative clause-VP	Paraphrase from the prompt text involves a transformation from a relative clause to a verb phrase or vice versa (e.g., *managed to become influential over what their group did* to *who sort of take over everything*)
Reordering of complements	Paraphrase from the prompt text involves exchanging placement of the sentence elements (e.g., *John arrived yesterday* to *Yesterday, John arrived*)
Unspecified syntactic reordering	Cases of paraphrase from the prompt text in which phrases or clauses have similar meaning and are reordered, but the reordering cannot be described by a formal syntactic transformation (e.g., *creative solutions come about because a group* to *more people involved does promote more creative ideas*)
Lexical paraphrase	
Synonyms	Paraphrase from the prompt text involves the use of synonyms (e.g., *moving in the wrong direction* to *heading in the wrong direction*)
Morphology	Cases in which paraphrase is attempted in morphologically variant forms (e.g., *make the team responsible* to *the group's responsibility*)
Multiple word units	Cases where one word is paraphrased by expansion to a multiple word unit or a multiple word unit is reduced to a smaller unit or even one word. This also covers cases where one multiple word unit is paraphrased with another multiple word unit of the same size. The original or the paraphrased text may be an idiom or collocation (e.g., *come up with* to *create*)
Unspecified lexical substitution (may overlap with conceptual paraphrase)	Paraphrase involves some other lexical substitution (e.g., *that will never work* to *their opinions*)

Classification	Description
Conceptual paraphrase (e.g., *the recognition for a job well done went to the group as a whole, no names were named to they got the same amount of recognition as the members who actually worked*)	Paraphrase that cannot be easily characterized by any syntactic or word-based classification
Global paraphrase	
Reading	Paraphrase of the gist of the reading that could not be isolated to specific text segments in the passage.
Lecture	Paraphrase of the gist of the lecture that could not be isolated to specific language segments in the stimuli.
Reading and lecture	Paraphrase of the gist of the reading and the lecture that could not be isolated to specific text segments or language segments in the stimuli.

Note. NP = noun phrase; VP = verb phrase.

Analyses

Interannotator Agreement

Interannotator agreement for P-units. Table 3 shows interannotator agreement for P-units for all data and for native and nonnative speaker data, independently. Agreement was measured using different *matching thresholds*. We established these matching thresholds for the following reason: When annotators selected the paraphrased text segment in the test-taker response, we noticed that sometimes they would match exactly, except for a single word at the beginning or the end (such as *a*, *an*, or *the*), or they may agree on the core part of the paraphrase but include a few words for which it may be fuzzy as to whether or not they are part of the paraphrased text. For instance, Figure 1 is an example in which annotators agree on the core paraphrase, and so we do want to consider this to be agreement.

Table 3

Interannotator Agreement for Paraphrase Units (P-Units)

Threshold	Number of agreed matches	Dice coefficient
All data (n = 100)		
Total P-units = 6,278		
1	1,014	0.323
0.7	1,325	0.422
0.6	1,552	0.494
0.5	1,759	0.560
Nonnative (n = 50)		
Total P-units = 3,202		
1	538	0.336
0.7	691	0.431
0.6	803	0.501
0.5	889	0.555
Native (n = 50)		
Total P-units = 3,076		
1	476	0.309
0.7	634	0.412
0.6	749	0.486
0.5	870	0.565

Prompt Text, source text segment: "group members who worked especially well and who provided a lot of insight on problems and issues"

Annotator A, paraphrased text segment: *The real excellent and creative member's work*

Annotator B, paraphrased text segment: *real excellent and creative member's work the*

Figure 1. **Annotators agree on the core paraphrase.**

In Figure 1, Annotator B excluded the sentence initial determiner *The* and probably inadvertently selected (highlighted in the interface) the final *the*. This occurrence of *the* is most likely the beginning of the next noun phrase in the sentence and does really belong in this paraphrase. However, because annotators do agree on the core paraphrase, we want to consider this example as agreement.

The matching thresholds work as follows: If the annotators agreed on every word in the P-unit (i.e., both the text segment from prompt and the paraphrased text segment in the test-taker essay), then this *matching threshold* would be 1; if annotators agreed on at least 70% of words in the text segment from the prompt and the paraphrased, 0.7; if annotators agreed on *at least* 60% of words in the text segment from the prompt and the paraphrased, 0.6; if annotators agreed on at least 50% of words in the text segment from the prompt and the paraphrased, 0.5.

For each of the matching thresholds described above, the *Dice coefficient* is used to calculate interannotator agreement in tasks where annotators can select from a set of labels for a particular observation. This method is discussed in Bentivogli et al. (2010). In Bentivogli et al., the Dice coefficient was used to compute agreement in a task where annotators had to identify Wikipedia links that corresponded to nonpronominal links in a data set. As in Bentivogli et al.'s task, our annotators could identify a paraphrase as being any segment of text from the test-taker response, or even multiple segments, so we found that this was an appropriate way to measure agreement of P-units.

The formula for the Dice co-efficient is as follows, where X is Annotator A and Y is Annotator B:

$$Dice = 2 \frac{|X \cap Y|}{|X| + |Y|}.$$

Agreement seems to be similar for native and nonnative English speaker data. Note that for the two annotators there were a total of 6,278 P-units identified. Annotator A found 3,211 P-units, and Annotator B found 3,067 P-units. Note that this is an average of more than 30 P-units per essay, indicating a large amount of paraphrase in test-taker summaries. The remainder of the paper discusses how paraphrase is created in test-taker responses.

Interannotator agreement for linguistic classifications. Annotators were asked to select from among *multiple linguistic classifications* to describe the linguistic modification used

to create a paraphrase. For example, the test-taker response paraphrase in Figure 2 was classified as verb tense shift and multiple word units. (See Table 2).

Prompt Text, source text segment: "didn't contribute much at all"
Summary, paraphrased text segment: *does nothing*

Figure 2. **Test-taker paraphrase.**

Figure 3 illustrates the distribution of high-level linguistic classifications across matching thresholds assigned to paraphrased text segments in the essay data. Lexical modifications appear to be the most frequent, but there also appears to be usage of all types.

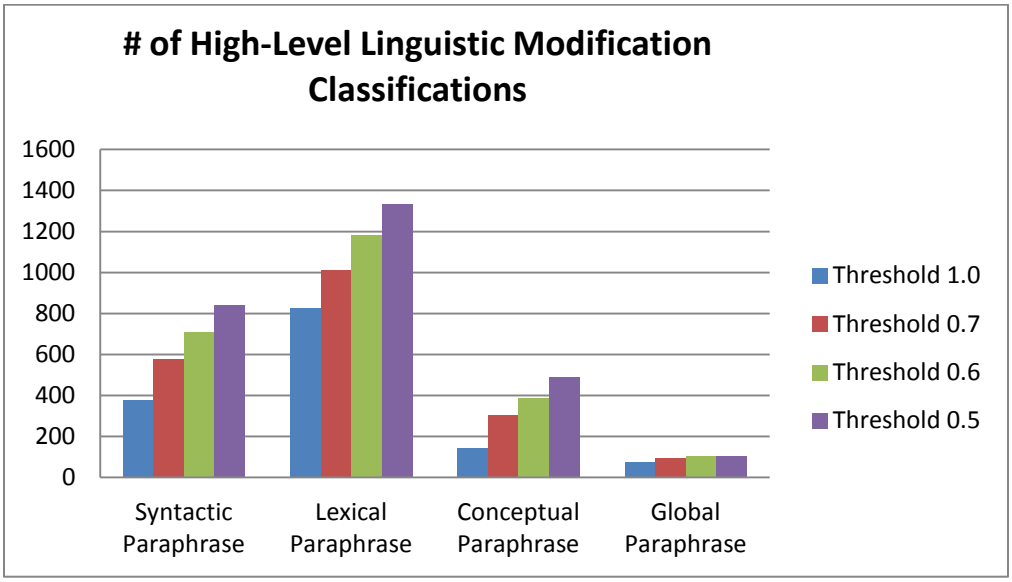

Figure 3. **Distribution of high-level linguistic classification labels for P-units.**

Figure 4 illustrates the distribution of fine-grained linguistic modification classifications. This would suggest that we made reasonable choices of classification categories. Declarative-question is the only category with a very small number of occurrences. The multiple word units category seems to be the most frequently used linguistic modification. Examples are shown in Table 4.

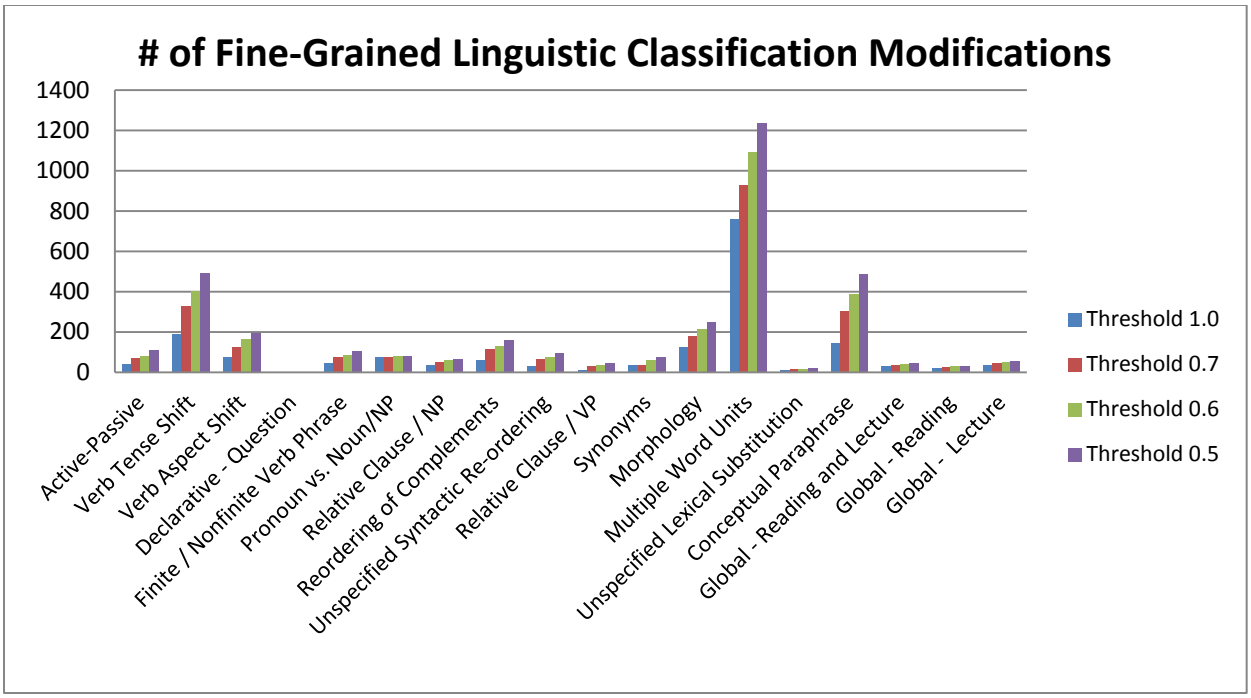

Figure 4. **Distribution of fine-grained linguistic classification labels for P-units.**

Table 4

Examples of Linguistic Modification Using Multiple Word Units

Original word/phrase	Paraphrase
they	a team member
on the other hand	secondly
the group	team
one or two people	one or two influential or persuasive people
a group of people	team work
a job well done	work
didn't contribute much	does nothing
a group of people	working in a group
a wider range of	more
knowledge, expertise, and skills	minds to think
if the decision turns out to be wrong	a difficult situation

Figures 5 and 6 illustrate interannotator agreement for high-level linguistic classifications at the P-unit thresholds. Agreement was relatively high for at the higher-level categories but widely varied at the fine-grained categories, as can be seen in the full analyses in Appendix E.

Agreement was computed only for P-units for which the annotators agreed (at a specific matching threshold). The formula for classification agreement is below (the denominator includes all classifications assigned a label by X and Y):

$$\text{Classification Agreement} = \frac{|X \cap Y|}{|X| + |Y|}.$$

Detailed tables in Appendix E illustrate that annotator agreement was lower in some of the fine-grained linguistic classification categories. All of the analyses that follow were computed for both annotators, and outcomes are typically consistent between both annotators throughout the analyses.

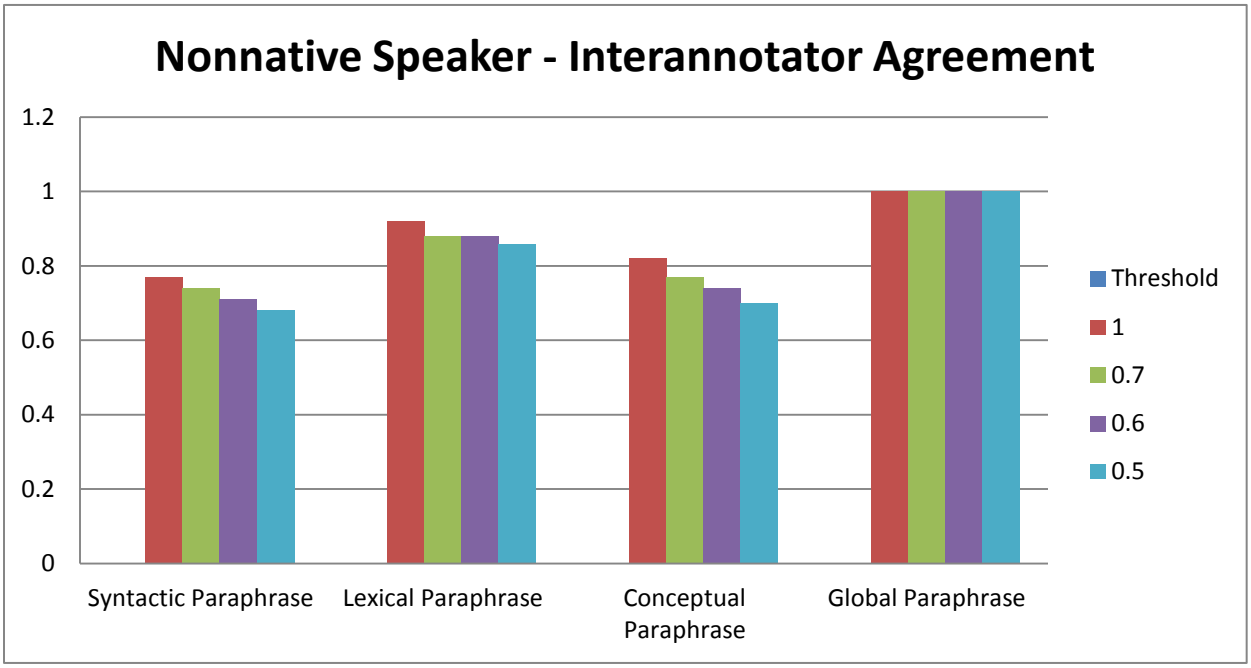

Figure 5. **Interannotator agreement on nonnative speaker data for high-level categories.**

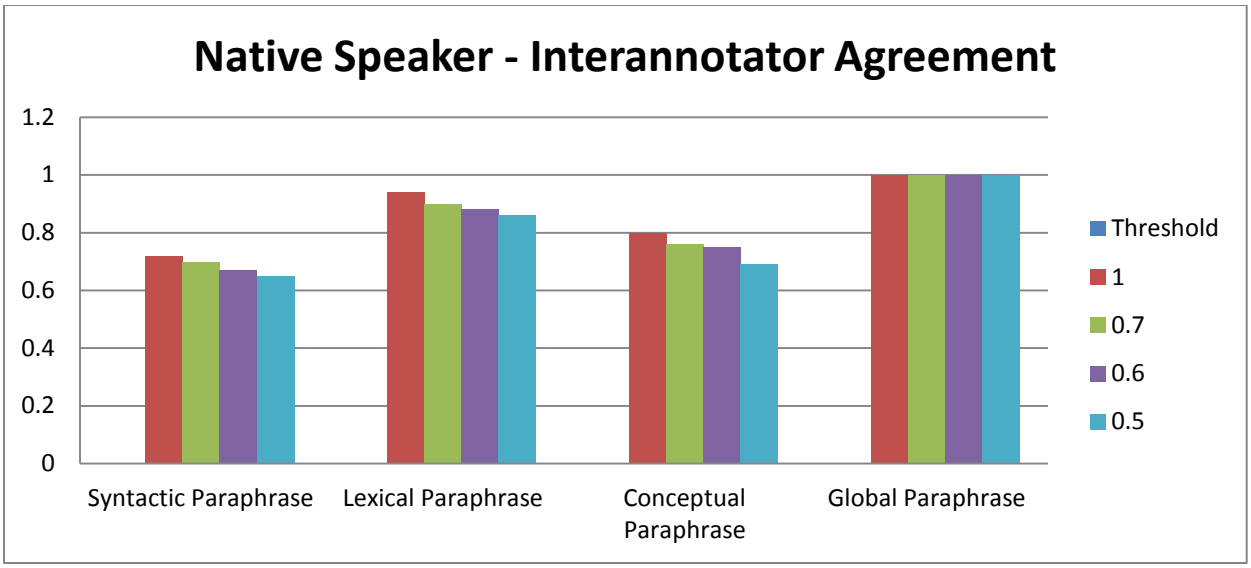

Figure 6. **Interannotator agreement on native speaker data for high-level categories.**

Frequency of Paraphrase

Table 5 illustrates that a statistically significant difference is found across test-taker populations and between the mean number of paraphrases from the lecture and from the reading, where lecture paraphrase is significantly higher. Results are relatively consistent across the two annotators (A and B). The simple explanation for the larger proportion of lecture paraphrase is that of the 24 sentences in the stimuli, a smaller proportion is from the reading than from the lecture: 9/24 (37.5%) and 15/24 (62.5%), respectively. Therefore, the test taker would have to paraphrase much more information from the lecture stimuli to offer a thorough summary. However, Table 5 also illustrates that compared to native speakers, nonnative speakers actually had a slightly lower mean number of lecture paraphrases and a noticeably higher mean number of reading-based paraphrases. This is not unexpected because the text of reading is available for the duration of the test and the lecture is not. Greater recall is required to paraphrase from the lecture. It makes sense then that nonnative English speakers would struggle more with content recall from the lecture stimuli and, as a result, paraphrase less often from that part of the stimuli.

Table 5

Paired T-Tests Comparing Number of Lecture and Reading Paraphrases (P-Units) for Both Raters (A and B)

Annotator	Test taker	Lecture paraphrases			Reading paraphrases			T-test	P-units
		N	Mean	*SD*	*N*	Mean	*SD*		
A	All	100	20.41	11.59	100	10.10	8.87	6.54	0.00
A	Native	50	21.64	10.05	50	8.22	7.60	7.24	0.00
A	Nonnative	50	19.18	12.94	50	11.98	9.70	2.89	0.00
B	All	100	20.34	12.15	100	8.94	8.55	7.29	0.00
B	Native	50	20.84	11.11	50	7.48	7.25	7.09	0.00
B	Nonnative	50	19.84	13.19	50	10.40	9.53	3.80	0.00

Paraphrase and response score. In Tables 6 and 7, a trend appears that is consistent with the findings in Table 5. Specifically, a higher correlation is noted between score and number of P-units when test takers produce a paraphrase that is drawn from the lecture stimuli. This finding is consistent for both annotators. In addition, correlations for nonnative speaker paraphrase of the reading are negatively correlated with score and strongly positively correlated with the lecture. Consistent with the discussion above, this finding also suggests that compared to the native speaker responses, nonnative speaker responses appear to have been more harshly penalized for paraphrasing from the reading but more strongly rewarded for paraphrasing from the lecture. These findings are consistent with outcomes in Gurevich and Deane (2007). Using the same data, Gurevich and Deane performed a comparison of language in test-taker responses and their similarity to language in the reading and lecture portions of the stimuli. Similar to our findings, their results indicated that nonnative speakers who relied mostly on the reading received lower scores, and those nonnative speakers who included more of the lecture received higher scores. In addition, in our study, the occurrence of global paraphrase is more strongly correlated with score for nonnative speakers. Again, this is true for both annotators. Recall that global paraphrase requires that the test taker summarize in a single sentence the full perspective of the reading, lecture, or both. The ability to produce a global paraphrase would suggest that a test taker would need to have a stronger ability to comprehend and make meaningful connections between multiple ideas in the stimuli.

Table 6

Correlations Between Score and Number of Paraphrases by Type for Native English Speakers

Annotator	Category	*N*	Correlation
A	Reading paraphrases	50	0.196
A	Lecture paraphrases	50	0.351
A	Global paraphrases	50	0.419
B	Reading paraphrases	50	0.295
B	Lecture paraphrases	50	0.326
B	Global paraphrases	50	0.199

Table 7

Correlations Between Score and Number of Paraphrases by Type for Nonnative English Speakers

Annotator	Category	*N*	Correlation
A	Reading paraphrases	50	-0.113
A	Lecture paraphrases	50	0.766
A	Global paraphrases	50	0.499
B	Reading paraphrases	50	-0.021
B	Lecture paraphrases	50	0.793
B	Global paraphrases	50	0.382

Word overlap and response score. Table 8 indicates that a negative correlation is present between word overlap and nonnative English speaker response scores. On the other hand, for native speakers, correlations are positive. This direction of the correlation is consistent across annotators.

Table 8

Correlations Between Score and Average Word Overlap for Lecture and Reading Paraphrases

Annotator	Test takers	*N*	Correlation
A	All	100	-0.034
A	Native	50	0.123
A	Nonnative	50	-0.102
B	All	100	-0.078
B	Native	50	0.166
B	Nonnative	50	-0.232

Note. Word overlap is not relevant for global paraphrase because it is a broad summary of multiple concepts from the prompt stimuli.

One interpretation of the results in Table 8 is extended in Table 9, which shows that native speakers have a greater range (higher maximum number) of words that overlap with the lecture stimuli and nonnative speakers have a greater range (higher maximum number) of words that overlap with the reading stimuli. This suggests that test takers will receive a higher score for recalling lecture stimuli, even if verbatim. Perhaps this is because the lecture stimuli are not available for the duration of the test (as noted earlier). By contrast, verbatim copying of the available reading text is not rewarded.

Table 9

Maximum Word Overlap in Paraphrase From Lecture and Reading

Annotator	Test takers	Lecture: max. # words overlap	Reading: max. # words overlap
A	Native	15	12
A	Nonnative	12	19
B	Native	21	16
B	Nonnative	19	22

Note. Minimum word overlap is always 0. Range is always from 0 to maximum number reported.

Sentences Frequently Paraphrased From the Stimuli

Since the lecture content seems to be driving the summary score, we were curious to see the distribution of paraphrased sentences based on the lecture and reading stimuli. Specifically, we wanted to know the following: Are test takers paraphrasing from a variety of sentences from the stimuli, or do they repeatedly paraphrase from a small subset of sentences?

Figure 7 illustrates that the distribution of paraphrased sentences in test-taker responses is quite uneven, suggesting that some sentences were paraphrased more frequently than others. Patterns are reasonably consistent between annotators. Here is what we found: From the reading (Sentences 1 through 9), Sentences 3 and 4 appear to be paraphrased most frequently, and Sentences 6 and 9 are loosely tied for third place. From the lecture[3] (Sentences 10 through 24), Sentences 12 and 18 seem to be paraphrased at noticeably higher frequencies, and Sentences 19 and 24 are closely tied for third place. Overall, about 8/24 sentences (33%) are paraphrased at a noticeably higher frequency. (See the bold sentences in the prompt text in Appendix A.)

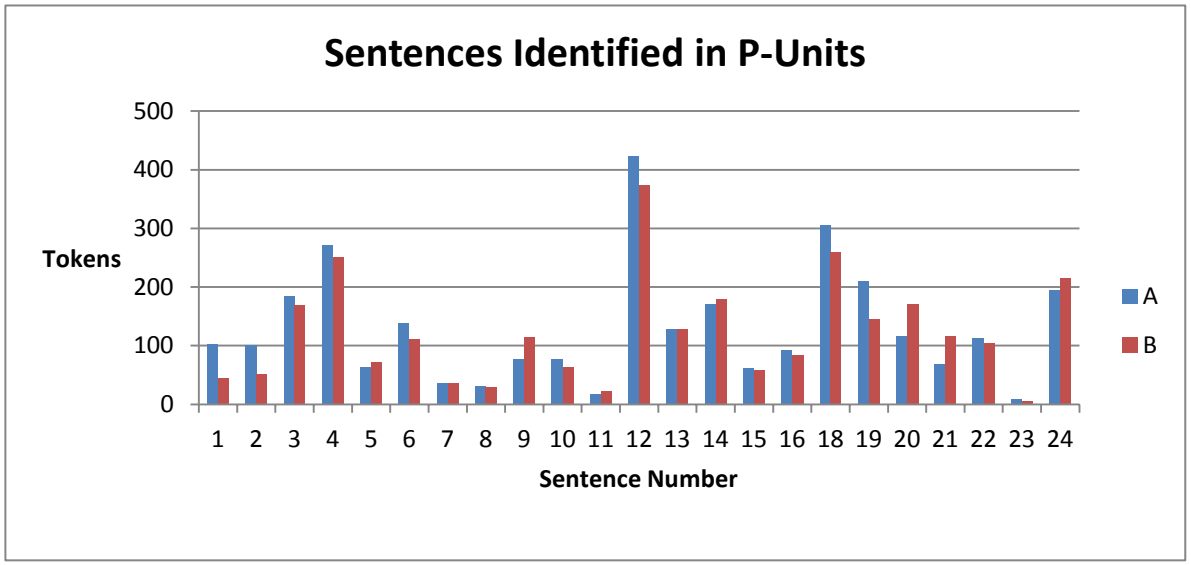

Figure 7. **Prompt sentences identified by annotators (A and B) in P-units.**

Characteristics of frequently paraphrased sentences. The most frequently paraphrased sentences appear to be those that introduced opposing viewpoints or undesirable outcomes between the reading and the lecture perspectives. To illustrate this point, try the following. If you insert the conjunction *however* in between the reading sentence and the lecture sentence in the

rows in Table 10, the two opposing perspectives are emphasized. The italicized text in each sentence represents the claim from the reading and the opposition from the lecture.

Table 10

Set of Frequently Paraphrased Sentences

Sentence #	Reading (proteamwork)	Sentence #	Lecture (antiteamwork)
3	First of all, *a group of people has a wider range of knowledge, expertise, and skills than any single individual* is likely to possess.	12	On virtually every team, *some members got almost a "free ride" . . . they didn't contribute much at all, but if their team did a good job, they nevertheless benefited from the recognition the team got.*
4	Also, because of the numbers of people involved and the greater resources they possess, *a group can work more quickly* in response to the task assigned to it and can come up with highly creative solutions to problems and issues.	18	Because it *took so long to reach consensus* . . . it took many, many meetings to build the agreement among group members about how they would move the project along.
6	This is because the group spreads responsibility for a decision to all the members and *thus no single individual can be held accountable* if the decision turns out to be wrong.	24	When the project failed, *the blame was placed on all* the members of the group.
9	Also, the individual team member has a much better chance to “shine,” to get his or her contributions and ideas not only recognized but recognized as highly significant, because a *team’s overall results can be more far-reaching and have greater impact than what might have otherwise been possible for the person to accomplish or contribute working alone.*	19	On the other hand, there were other instances where *one or two people managed to become very influential over what their group did.*

Note. The italicized text in each sentence represents the claim from the reading and the opposition from the lecture.

Paraphrased sentences and response score. As discussed above, test takers appear to paraphrase some sentences more than others. The next logical question addresses if a relationship is present between response score and paraphrase of particular sentences from the stimuli. Results between the two populations showed the following: For native speakers, the only significant correlation across both annotators was Sentence 12, where $p <= 0.05$. So, what this suggests is that both annotators identified paraphrases in Sentence 12, and these were significantly, positively correlated with response score, where $p <= 0.05$. Interestingly, Sentence 12 is the first sentence in the lecture stimuli that introduces an opposing point of view. Specifically, Sentence 12 is where the opposing view to the reading is first introduced. It follows that this sentence sets the stage for the expression of opposition in the test-taker response.

For nonnative speakers, moderate to strong positive correlations, where $p <= 0.05$, were found for Sentence 1 and Sentences 12 through 19. This finding was consistent for both annotators. Sentence 1 is the introductory sentence in the reading and expresses the initial perspective that sets the tone for the perspective presented in the reading. Sentence 12 (as discussed) is the first expression of opposition presented in the lecture.

Sentences 13 through 19 are all lecture sentences. A subset—Sentences 12, 18, and 19—were among the most frequent (see Figure 7).

What this might suggest is that while nonnative speakers seem to paraphrase less frequently from the lecture than native speakers (see Table 7), when they did paraphrase lecture points, they tended to receive higher scores.

Differences in Linguistic Classification

Native versus nonnative English speakers. As shown in Figures 3 and 4, a wide array of linguistic structures was used by both populations to produce paraphrases in summary responses. However, was there a statistically significant difference between the two populations? Table 11 shows classifications types for which z-scores[4] indicate statistical significance related to the proportions of linguistic classifications assigned to paraphrases identified in the native and nonnative speaker response data. Results indicated a few statistically significant outcomes across all P-unit thresholds and for P-units on which annotators agree at the different matching thresholds. Nonnative responses show greater use of verb aspect shift, lexical paraphrase (over all fine-grained categories in the lexical paraphrase class), and specifically, morphology at

statistically significant levels, where $p <= .05$. Native responses appear to use the relative clause/verb phrase (VP) transformation.

Native responses tend to use global paraphrase, and specifically for lecture content, at statistically significant levels, where $p <= .05$. Recall that instances of global paraphrase refer to cases where the writer summarizes the gist of the reading, the lecture, or both. In these cases, the P-units in the response cannot be assigned a correspondence with a particular unit of text in the prompt stimuli because they summarize multiple, global concepts from the prompt that may extend over several sentences. Native speakers appear to do this more frequently for the lecture.

Table 11

Z-Test for Proportion of Paraphrase Classifications in Native and Nonnative Responses

Classifications	Threshold of 1		Threshold of 0.7		Threshold of 0.6		Threshold of 0.5	
	Z-statistic	P-value	Z-statistic	P-value	Z-statistic	P-value	Z-statistic	P-value
Verb aspect shift	-1.83	0.07	-2.25	0.02	-1.83	0.07	-2.49	0.01
Relative clause/VP	1.19	0.23	2.91	0.00	2.52	0.01	2.76	0.01
Lexical paraphrase	-1.09	0.28	-1.45	0.15	-1.96	0.05	-1.82	0.07
Morphology	-1.74	0.08	-2.38	0.02	-1.82	0.07	-0.84	0.40
Global paraphrase	2.48	0.01	1.61	0.11	1.70	0.09	1.52	0.13
Lecture	1.95	0.05	1.21	0.23	1.40	0.16	1.06	0.29

Note. Shaded cells indicate statistical significance. VP = verb phrase.

Linguistic classifications and response score. We now know that there are a few significant differences between linguistic modification used by native and nonnative speakers when they paraphrase. However, do the differences in linguistic modification have any relationship to the final response score? Our findings indicate that for native speakers, only a few linguistic classifications are significantly, positively correlated ($p <= .05$) with response score across all thresholds, specifically, overall lexical paraphrase, morphology, and multiple word units. Conceptual paraphrase is also significantly, positively correlated at 0.7, 0.6, and 0.5

thresholds. By contrast, for nonnative speakers, the use of almost all linguistic classifications for paraphrase is significantly, positively correlated to response score.

The only classifications not correlated to response score are the following: pronoun versus noun/noun phrase (NP), unspecified syntactic reordering, relative clause/VP, unspecified lexical substitution, and declarative-question (which no one used). All other categories were significantly, positively correlated with response score across thresholds, and most correlations were moderately or highly correlated, where $p <= 0.02$.

The results in Table 11 indicated that there was little difference between the proportions of linguistic modification types used by the two populations. Curiously, use of almost any linguistic modification by nonnative test takers does appear to significantly and positively affect response score. This would suggest that all types of in nonnative texts are highly rewarded by annotators.

Discussion and Future Directions

The purpose of this annotation study was to gather and compare information about the characteristics of paraphrase production in native and nonnative speaker summary responses to a TOEFL integrated writing task. To our knowledge, this is the first attempt to *systematically* investigate different paraphrase production strategies in the native and nonnative English speaker populations through the identification of paraphrase and the subsequent classification of paraphrased text into linguistic categories. To this end, we developed an annotation scheme that instructed annotators how to identify text segments in test-taker summary responses that had been paraphrased from the prompt stimuli and how to label those paraphrased text segments with a linguistic classification. Our hypothesis was that findings from the analysis of the annotated data would inform the direction of paraphrase research as it pertains to e-rater feature development for scoring TOEFL integrated summary responses. Findings from our analysis are discussed below, and each of our original research interests is addressed.

Across test-taker populations, our analyses indicate a statistically significant difference between the mean number of paraphrases from the lecture and from the reading, where lecture paraphrase in test-taker responses is significantly higher. While this may be attributed to the fact that there are almost twice as many lecture sentences, Table 5 illustrates that nonnative speakers have a lower mean number of paraphrases that are drawn from the lecture stimuli than do native speakers.

Consistent with this finding, Tables 6 and 7 illustrate that the number of paraphrases drawn from the lecture stimuli were positively correlated with essay score, and these correlations were even stronger with the nonnative speaker data. This finding is consistent with Gurevich and Deane (2007), who also found for the same data set that summaries with more evidence of paraphrase drawn from the lecture stimuli received higher scores. These consistent findings may be possibly due to the fact that the lecture is unavailable for the full duration of the exam, and so paraphrase from the lecture stimuli relies on accurate note-taking during the lecture portion and recall from memory.

Findings in Table 8 also indicate that word overlap (verbatim copying) is negatively correlated with essay score in nonnative speaker essays but positively correlated with essay score for native speaker essays. Results in Table 9 suggest that the reason for this might be that mean word overlap for nonnative speakers is higher in the reading, while for native speaker essays, it is higher in the lecture. In other words, nonnative speakers are copying directly from the reading passage that is available for the duration of the exam while native speakers are using verbatim language from the lecture stimuli. In the latter case, test takers need to take notes or recall exactly what the lecturer has said, since the lecture stimuli are not available for the duration of the exam. It makes sense that raters would be less likely to reward the presence of exact wording from the reading than the lecture because the reading is available for the duration of the test. Copying from the reading would not be an indicator of the test takers' facility to summarize information from the prompt stimuli, and so it probably would not help a test taker get a higher score.

A related finding is that paraphrasing from specific sentences in the lecture is also positively correlated with essay score. For both native and nonnative speaker responses, the sentence (Sentence 12) that appears be the discourse move (change in polarity) in the prompt stimuli that triggers the opposition perspective presented in the lecture is positively correlated with essay score. This is an intriguing finding that is potentially related to the argument structure of the test-taker essay. If further research indicated that this phenomenon were found in responses across other prompts, then the ability to automatically identify the presence of a paraphrase of this trigger sentence in essay responses might become important because it could be a new predictor of essay score. In other words, essays that contained the sentence associated with this particular discourse move may be a factor in essay scoring.

In addition, if upon further investigation using additional prompts we were able to show that paraphrase of these trigger sentences is consistently correlated with essay score, then this would imply that perhaps opinion detection could be useful in identifying when discourse move and subsequent evaluation of paraphrase could be used to ensure that this change in polarity was, in fact, the trigger sentence—that is, the sentence that introduces the opposing perspective. The ability to identify this trigger sentence in an essay could serve as part of score prediction in an automated scoring engine. This problem is potentially interesting as we think about how to develop our paraphrase and as opinion detection approaches to support new e-rater features, such as these trigger sentences, which might require the use of both approaches.

Results indicate a few statistically significant outcomes with regard to native and nonnative speaker use of linguistic modification for paraphrase. Specifically, native responses show greater use of verb aspect shift, lexical paraphrase, and, especially, morphology at statistically significant levels. Native responses appear to use the relative clause/VP transformation. Further, native responses tend to incorporate global paraphrase, specifically for lecture content. Related to this, our findings also indicate that for native speakers, only a few linguistic classifications are significantly, positively correlated with response score across all thresholds, specifically, overall lexical paraphrase, morphology and multiple word units, and conceptual paraphrase.

By contrast, for nonnative speakers, use of almost all linguistic classifications for paraphrase is significantly, positively correlated to response score. The only classifications not correlated to response score are the following: pronoun versus noun/NP, unspecified syntactic reordering, relative clause/VP, unspecified lexical substitution, and declarative-question (which was used infrequently). All other categories were significantly, positively correlated with response score across thresholds, and most correlations were moderately or highly correlated. These findings may be relevant when developing methods to identify paraphrase. The majority of paraphrase classifications for nonnative speakers does appear to be lexical paraphrase (Figure 5) and, specifically, multiple word units (Table E1). An additional challenge would be to use natural language processing (NLP) to automate the capture of global paraphrase.

Building on earlier work by Culicover (1968) and Keck (2006), this research is informative, not only to examine differences between the two populations, but also to identify useful methods that help us to differentiate between proficiency levels in nonnative speaker

writing using automated methods, including automated essay scoring. The outcomes of this work reveal linguistic characteristics of paraphrase production in native and nonnative speaker writing, as well as aspects of the argument structure in the prompt stimuli that appear to be highly valued. To this end, our findings reveal that paraphrase of the specific sentence from the lecture stimuli where opposition is introduced is correlated with higher essay scores.

Moving forward, we suggest the following lines of future research:

- Our findings suggest that both NLP research in opinion detection and paraphrase recognition could support the development of new e-rater features that expand the writing construct represented in e-rater and potentially improve e-rater scoring performance. The annotated data created during this study are a useful resource for opinion and paraphrase detection research. The data that label lexical categories of paraphrase, such as synonyms, and multiple word units could potentially be used in paraphrase research that addresses synonym use. The annotated data are available upon request for ETS research.
- Our analyses indicate that paraphrased text segments associated with specific sentences in the prompt stimuli are correlated with essay score (for a single prompt). This finding is potentially important, and replication of this phenomenon across more prompts could inform the development of additional e-rater features related to the presence of important content in responses. In light of this finding, additional research that investigates this phenomenon on different data sets should be pursued. As manual annotation was time-consuming and difficult, we should consider automated ways to identify paraphrases in responses that are associated with trigger sentences. Perhaps an annotation task asking annotators to confirm that an automatically selected text segment is a relevant paraphrase, as opposed to asking annotators to find the sentences, might be a useful and easier task, and significantly more data could be annotated. In a simpler annotation task, it is possible that crowd sourcing can be used to collect more data in a shorter period of time. Using additional annotations for more prompts, we can then look at relationships between essay score for essays that contain paraphrases of trigger sentences and essays that do not. We will also have additional data on which to develop and evaluate our paraphrase and opinion detection systems.

References

Bannard, C., & Callison-Burch. C. (2005). Paraphrasing with bilingual parallel corpora. *In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (pp. 597–604). New Brunswick, NJ: Association for Computational Linguistics.

Barzilay, R., & Lee, L. (2003). Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 16–23). Stroudsburg, PA: Association for Computational Linguistics.

Barzilay, R., & McKeown, K. (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (pp. 50–57). Retrieved from http://www.aclweb.org/anthology-new/P/P01/P01-1008.pdf

Bentivogli, L., Forner, P., Giuliano, C., Marchetti, A., Pianta, E., & Tymoshenko, K. (2010). Extending English ACE 2005 corpus annotation with ground-truth links to Wikipedia. In *23rd International Conference on Computational Linguistics* (pp. 19–26). Beijing, China: Coling 2010 Organizing Committee.

Bond, F., Nichols, E., & Appling, D. S. (2008). Improving statistical machine translation by paraphrasing the training data. In *Proceedings of IWSLT* (pp. 150–157). Waikiki, HI: Consortium for Speech Translation Advanced Research.

Callison-Burch, C. (2008). Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 196–205). Stroudsburg, PA: Association for Computational Linguistics.

Callison-Burch, C., Koehn, P., & Osborne, M. (2006). Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference* (pp, 17–24). New York, NY: Association for Computational Linguistics.

Campbell, C. (1990). Writing with others' words: Using background reading text in academic compositions. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 211–230). Cambridge, UK: Cambridge University Press.

Cohn, T., Callison-Burch, C., & Lapata, M. (2008). Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics, 34,* 597–614.

Culicover, P. W. (1968). Paraphrase generation and information retrieval from stored text. *Mechanical Translation and Computational Linguistics, 11*(1–2), 78–88.

Dolan, W., Quirk, C., & Brockett, C. (2004, August). *Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources.* Paper presented at the meeting of the International Conference on Computational Linguistics, Geneva, Switzerland.

Foltz, P., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments, 8*(2), 111–127.

Gurevich, O., & Deane, P. (2007). Document similarity measures to distinguish native vs. non-native essay writers. In *Proceedings of NAACL* (pp. 49–52). Rochester, NY: Association for Computational Linguistics.

Keck, C. (2006). The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing, 15,* 261–278.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th Annual Meeting of the Cognitive Science Society* (pp. 412–417). Mahwah, NJ: Erlbaum.

Leacock, C., & Chodorow, M. (2003). C-rater: Scoring of short-answer questions. *Computers and the Humanities, 37,* 389–405.

Madnani, N., Ayan, N. F., Resnik, P., & Dorr, B. J. (2007). Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation* (pp. 120–127). Prague, Czech Republic: Association for Computational Linguistics.

Madnani, N., & Dorr, B. (2010). Generating phrasal & sentential paraphrases: A survey of data-driven methods. *Computational Linguistics, 36,* 341–387.

Madnani, N., Resnik, P., Dorr, B. J., & Schwartz, R. (2008). Are multiple reference translations necessary? Investigating the value of paraphrased reference translations in parameter optimization. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA)* (pp. 143–152). Waikiki, HI: AMTA.

Marton, Y., Callison-Burch, C., & Resnik, P. (2009, August). *Improved statistical machine translation using monolingually-derived paraphrases*. Paper presented at the Conference on Empirical Methods in Natural Language Processing (EMNLP), Suntec, Singapore.

Pang, B., Knight, K., & Marcu, D. (2003). Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of the NAACL-HLT* (pp. 102–109). Stroudsburg, PA: Association for Computational Linguistics.

Purdue University Online Writing Lab. (2006). *Paraphrase: Write it in your own words.* Retrieved from http://owl.english.purdue.edu/handouts/research/r_paraphr.html

Shi, L. (2004). Textual borrowing in second-language writing. *Written Communication, 21*, 171–200.

Sukkarieh, J. Z., & Blackmore, J. (2009). C-rater: Automatic content scoring for short constructed responses**.** In *Proceedings of the 22nd- International FLAIRS Conference.* Sanibel Island, FL: Association for the Advancement of Artificial Intelligence.

Sukkarieh, J. Z., & Bolge, E. (2008). Leveraging c-rater's automated scoring capability for providing instructional feedback for short constructed responses. In B. P. Woolf, E. Aimeur, R. Nkambou, & S. Lajoie (Eds.), *Lecture notes in computer science: Vol. 5091. Proceedings of the 9th International Conference on Intelligent Tutoring Systems, ITS 2008* (pp. 779–783). New York, NY: Springer-Verlag.

Yamada, K. (2003). What prevents ESL/EFL writers from avoiding plagiarism? Analyses of 10 North-American college websites. *System, 31,* 247–258.

Notes

[1] The responses used to develop the annotation guidelines were in response to the prompt in Appendix B.

[2] Michael Flor implemented the annotation tool.

[3] Note that Sentence 17 is excluded from the analysis because it was only one word: *Why?*

[4] The z-score is computed as native proportion minus the nonnative proportions, so a positive z-statistic indicates that native speaker data has a higher mean for a particular feature and a negative z-statistic indicates that nonnative speaker data has a higher mean for a particular feature.

List of Appendices

Appendix A
Prompt Text (VB531093)

Reading

In many organizations, perhaps the best way to approach certain new projects is to assemble a group of people into a team. Having a team of people attack a project offers several advantages. **First of all, a group of people has a wider range of knowledge, expertise, and skills than any single individual is likely to possess. Also, because of the numbers of people involved and the greater resources they possess, a group can work more quickly in response to the task assigned to it and can come up with highly creative solutions to problems and issues.** Sometimes these creative solutions come about because a group is more likely to make risky decisions that an individual might not undertake. **This is because the group spreads responsibility for a decision to all the members and thus no single individual can be held accountable if the decision turns out to be wrong.**

Taking part in a group process can be very rewarding for members of the team. Team members who have a voice in making a decision will no doubt feel better about carrying out the work that is entailed by that decision than they might doing work that is imposed on them by others. **Also, the individual team member has a much better chance to "shine," to get his or her contributions and ideas not only recognized but recognized as highly significant, because a team's overall results can be more far-reaching and have greater impact than what might have otherwise been possible for the person to accomplish or contribute working alone.**

Lecture

Now I want to tell you about what one company found when it decided that it would turn over some of its new projects to teams of people, and make the team responsible for planning the projects and getting the work done. After about six months, the company took a look at how well the teams performed.

On virtually every team, some members got almost a "free ride" . . . they didn't contribute much at all, but if their team did a good job, they nevertheless benefited from the recognition the team got. And what about group members who worked especially well and who provided a lot of insight on problems and issues? Well . . . the recognition for a job well

done went to the group as a whole, no names were named. So it won't surprise you to learn that when the real contributors were asked how they felt about the group process, their attitude was just the opposite of what the reading predicts.

Another finding was that some projects just didn't move very quickly. Why? **Because it took so long to reach consensus . . . it took many, many meetings to build the agreement among group members about how they would move the project along. On the other hand, there were other instances where one or two people managed to become very influential over what their group did.** Sometimes when those influencers said "that will never work" about an idea the group was developing, the idea was quickly dropped instead of being further discussed. And then there was another occasion when a couple influencers convinced the group that a plan of theirs was "highly creative." And even though some members tried to warn the rest of the group that the project was moving in directions that might not work, they were basically ignored by other group members. Can you guess the ending to this story? **When the project failed, the blame was placed on all the members of the group.**

Appendix B

Used for Examples in Guidelines (VB157528)

Reading

An interesting new development has taken place in the United States: The number of self-employed entrepreneurs has suddenly started to grow after several years of steady decline. Why this sudden interest in running one's own business? Some analysts say that the real cause of this increase lies in the fact that jobs at large corporations, once highly prized, have lost their appeal. Working for a large corporation has simply become less attractive than owning a business.

First, a number of people who have become entrepreneurs after leaving large corporations say that what frustrated them most was the amount of bureaucracy they had to face in their work. New ideas had little chance of being implemented because they had to make their way through countless committees; one had to spend so much time obtaining permission to do anything out of the ordinary that there was little incentive to even try.

Second, one of the major attractions of corporate jobs was job security. Traditionally in the United States, corporate jobs have been very secure, but now corporations are increasingly laying off employees, even employees who have worked for the companies for many years. Under these circumstances, it is not surprising that more and more people prefer to run their own business.

Third, corporate jobs have become less attractive for Americans because of changes in benefits. Corporations have traditionally provided their employees extra benefits in addition to salaries, such as retirement pensions and inexpensive medical insurance. In recent years, however, American corporations have cut back on employee pensions and are asking their employees to pay more money for medical insurance.

Lecture

The statistics are correct: In recent years the number of Americans running their own small businesses has gone up. But the explanation can't be that running your own business is all of a sudden much more attractive than working for a big corporation. I'll tell you why.

First, OK, it can be hard at times to get your ideas accepted at a corporation because many things have to be approved by managers at several levels, but I doubt that anyone expects that setting up and running your own business involves less bureaucracy. Do you have any idea

how many regulations there are that a small business has to meet? And to do that, you have to deal with all kinds of authorities at several levels. You have to get your plans approved. You have to get your facilities inspected and reinspected, etc., etc., etc. All that takes time and is very frustrating.

Second, even if it's true that corporate jobs have become less secure than they used to be, does that make them less secure than running one's own business? It's common knowledge that the risk of failure for new small businesses is much higher than the risk of being laid off from a corporate job.

And third, people who run their own business have to put money aside for retirement, and the cost of buying medical insurance on their own is steep. So even in a reduced form, the extra benefits corporations offer their employees are probably still a major attraction for most Americans.

Appendix C

ParaMarker Main Screen

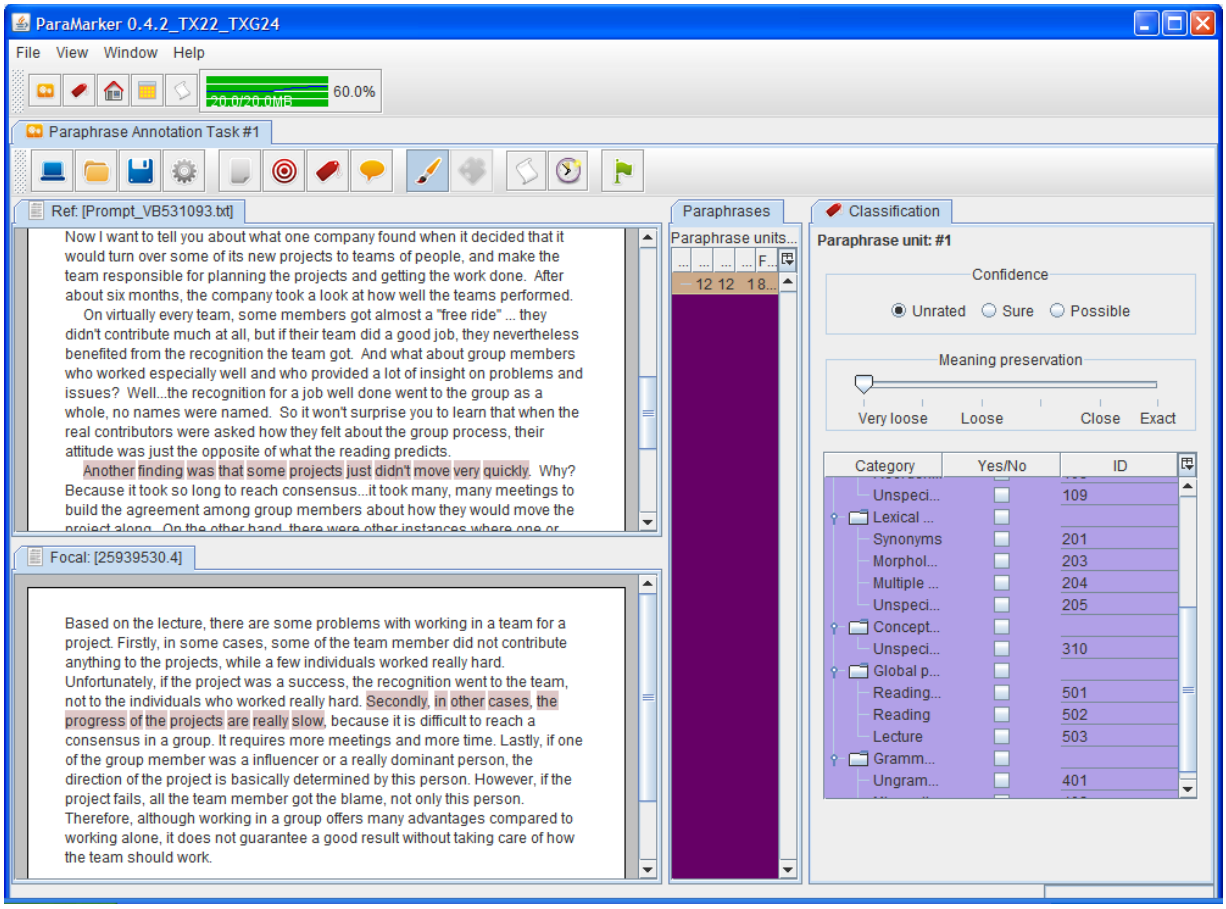

***Figure C1*. ParaMarker main screen. Refer to the Using-ParaMarker.ppt user guide for details.**

Appendix D

Annotation Guidelines for Identification Classification of Paraphrase in Test-Taker Responses

For this annotation task, you will identify, label and classify paraphrased text across a text (document) pair. Specifically, you will be asked to do the following: (a) label text segments in test-taker essays that you have identified as paraphrases of prompt text, and (b) classify the labeled paraphrased text segments relative to your confidence about "if it is a paraphrase"; the extent to which the paraphrase of the prompt text segment preserves meaning of that segment; the linguistic device used to create the paraphrase (syntactic, lexical, or conceptual); and the grammaticality of the paraphrase. Directions about how to implement these tasks are described in this document.

To complete the tasks, you will be using ParaMarker, a paraphrase annotation interface tool (see Figure D1). A companion document, *Using-ParaMarker.ppt*, will walk you through the use of ParaMarker.

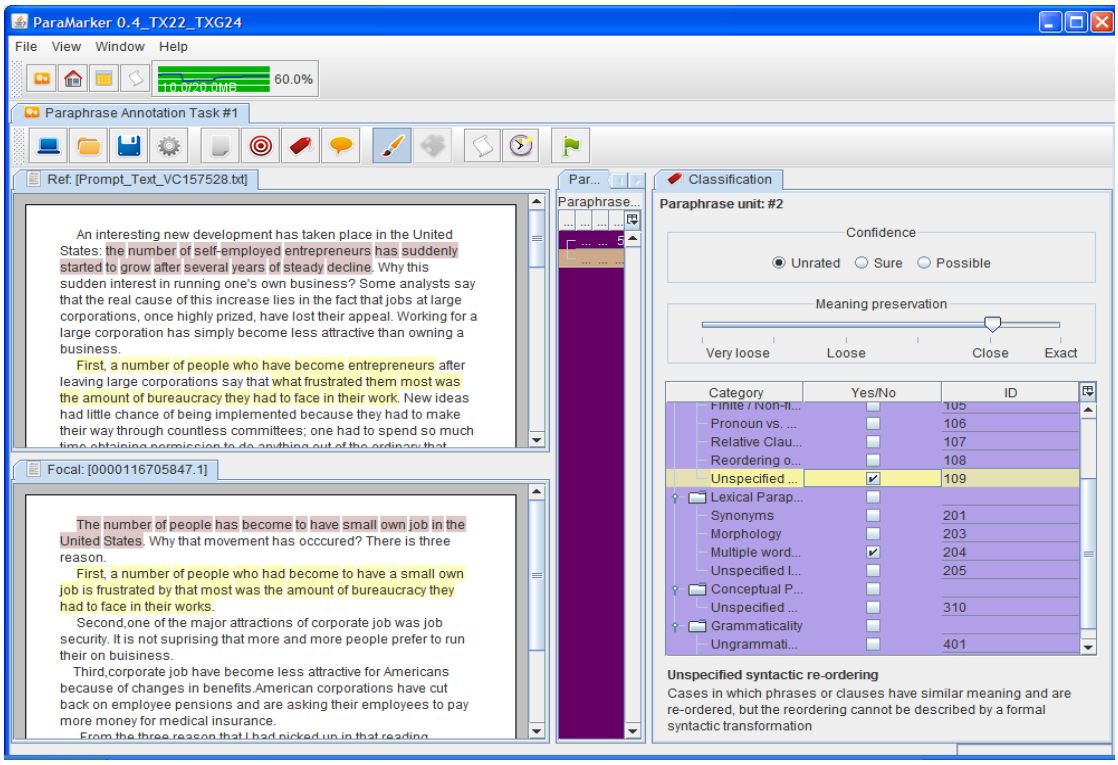

Figure D1. **ParaMarker Interface. Refer to the Using-ParaMarker.ppt user guide for details.**

Paraphrase Annotation Task Steps

Step 1. Get to Know the Reference Text

The first step is to read and familiarize yourself with the reference text. You will be loading it into the interface for annotation purposes, but you might consider printing out a paper copy and keeping it next to you. After 5 or 10 annotations, you will probably know it by heart! The reference text for the actual task can be found in the Appendix A in this document. (The prompt used for examples throughout this guidelines document is in Appendix B; this prompt will not be used for the actual annotation task).

Step 2. Highlight P-Units

Using ParaMarker, highlight text segment pairs in the reference text and focal text, where you have identified text in the focal text that paraphrases text in the reference text. These text pairs are P-units. They can be full sentences, or parts of sentences in which the two text segments (in the focal text and reference text) convey the same or very similar meanings but are worded differently. See Figures D2 and D3.

Reference Text Excerpt

"An interesting new development has taken place in the United States: **the number of self-employed entrepreneurs has suddenly started to grow** after several years of steady decline. Why this sudden interest in running one's own business? Some analysts say that the real cause of this increase lies in the fact that jobs at large corporations, once highly prized, have lost their appeal. Working for a large corporation has simply become less attractive than owning a business . . . "

Focal Text Excerpt

"The number of people has become to have small own job in the United States. Why that movement has occurred? There is three reason.

First, a number of people who had become to have a small own job is frustrated by that most was the amount of bureaucracy they had to face in their works . . . "

Figure D2. **P-unit pairs involving full and partial sentences. Boldface indicates P-unit pairs between the reference text and focal text.**

Reference Text Excerpt

"Second, one of the major attractions of corporate jobs was job security. Traditionally in the United States, corporate jobs have been very secure, but now corporations are increasingly laying off employees, even employees who have worked for the companies for many years. Under these circumstances, **it is not surprising that more and more people prefer to run their own business . . . "**

Focal Text Excerpt

"Second, one of the major attractions of corporate job was job security. **It is not suprising that more and more people prefer to run their on buisiness . . . "**

Figure D3. **P-unit pairs involving full and partial sentences. Boldface indicates P-unit pairs between the reference and focal texts. Underlined text indicates a close copy.**

Note that in Figure D3, the focal text also contains a sentence (underlined) that is almost an exact copy of a sentence from the reference text.

> *Second, one of the major attractions of corporate jobs was job security. / Second, one of the major attractions of corporate job was job security.*

Above, the test-taker copies the sentence exactly from the reading passage with the exception of one grammar error or typo ("jobs" becomes "job"). We do not consider text segments this close to be paraphrases. A rule of thumb might be that if the syntactic form (order) and the lexical stems remain the same, then the focal text wording would not be a paraphrase of the reference text.

Step 3. Label Confidence

For each P-unit, rate your confidence with regard to your certainty about whether the P-unit text represents paraphrased text by selecting Sure or Possible. The default button is unrated, so you need to select either Sure or Possible.

- Select Sure if you are absolutely certain that the focal text is a paraphrase of the reference text.

- Select Possible if you have any uncertainty that focal text is a paraphrase of the reference text.

Step 4. Label Meaning Preservation

For each P-unit, use ParaMarker's Meaning Preservation Slider to rate how close in meaning the focal text is to the reference text.

Step 5. Classify Paraphrase Devices: Syntactic, Lexical, or Conceptual Paraphrase Classifications

Under these high-level categories are more fine-grained categories from which you can select. These are described in the Tables D1 and D2 and Figures D4 and D5. Please note that for this task, we require that you make a selection only if you are absolutely certain of your decision.

Step 6. Label Grammaticality Categories

- *Ungrammaticality*: Paraphrased focal text segment contains grammar errors
- *Misspelling*: Paraphrased focal text segment contains only spelling errors

Table D1

Syntactic Paraphrase

	Reference text	Focal text
Active ↔ passive	When the project failed, **the blame was placed on all the members of the group**.	Moreover, when **the team fails to complete the project, all of the team members got the blame**, even if one member tries to warn everyone that the project is headed the wrong direction.
Verb aspect shift	Also, because of the numbers of people involved and the greater resources they possess, a group **can work** more quickly in response to the task assigned to it and can come up with highly creative solutions to problems and issues.	Second, members **work** faster and create better solutions.

	Reference text	Focal text
	On virtually every team, some members **got** almost a "free ride" . . . they **didn't contribute** much at all, but if their team did a good job, they nevertheless benefited from the recognition the team got.	Firstly, being in a group is not always beneficial to some people because some people **may get** a "free ride" and only some members of the team did all the work.
Verb tense shift	On virtually every team, some members got almost a "free ride" . . . **they** didn't contribute much at all, but if their team did a good job, they nevertheless benefited from the recognition the team got.	Moreover, not every member of the group **had contributed** to the work, some had had a free ride.
Finite ↔ non-finite VP	On the other hand, there were other instances where one or two people **managed to become** very influential over what their group did.	The agreement between every member did not happen with ease, therefore one or two **became** leaders and modified the other's opinion.
Pronominalization ↔ NP	When **the project** failed, the blame was placed on all the members of the group.	When **it** is proven to be a failure, the blame is then put on the whole group, not on the misleading influential member.
Pronoun ↔NP	On virtually every team, some members got almost a "free ride" … **they** didn't contribute much at all, but if their team did a good job, they nevertheless benefited from the recognition the team got.	Moreover, **not every member of the group** had contributed to the work, some had had a free ride.
Relative clause ↔ NP	And even though some members tried to warn the rest of the group that the project was moving in **directions that might not work**, they were basically ignored by other group members.	Moreover, when the team fails to complete the project, all of the team members got the blame, even if one member tries to warn everyone that the project is headed **the wrong direction**.

	Reference text	Focal text
Reordering of complements	**On virtually every team, some members got almost a "free ride"** . . . they didn't contribute much at all, but if their team did a good job, they nevertheless benefited from the recognition the team got.	At least **one person has a free ride in the team** and that makes the rest of the team angry.
Unspecified syntactic reordering: Cases in which phrases or clauses have similar meaning and are re-ordered, but the reordering cannot be described by a formal syntactic transformation.	First of all, **a group of people has** a wider range of knowledge, expertise, and skills than any single individual is likely to possess.	First, **in a team, you have** a wider range of knowledge and expertise.
	Sometimes **these creative solutions come about because a group** is more likely to make risky decisions that an individual might not undertake.	Furthermore, **the solutions of the teamwork are more creative**.
	… because **a team's overall results** can be more far-reaching …	At last, **the result of the project** is belonged to the whole team and more far-reaching.

Note. A syntactic paraphrase is a paraphrase characterized by change in syntax. ↔ means that the transformation can work in both directions. NP = noun phrase; VP = verb phrase.

Table D2

Lexical (Word-Based) Paraphrase

Fine-grained labels	Reference text	Focal text
Synonyms: Similar words	And even though some members tried to warn the rest of the group that the project was **moving** in directions that might not work, they were basically ignored by other group members.	Moreover, when the team fails to complete the project, all of the team members got the blame, even if one member tries to warn everyone that the project is **headed** the wrong direction.

Fine-grained labels	Reference text	Focal text
Morphology: Cases in which paraphrase is attempted using nominalization or other morphologically variant forms.	Now I want to tell you about what one company found when it decided that it would turn over some of its new projects to teams of people, and **make the team responsible** for planning the projects and getting the work done.	In the lecture, the professor talked about **the group responsibility**. (*Nominalization*)
Multiple word units: Cases where one word is paraphrased by expansion to a multiple word unit, or a multiple word unit is reduced to a smaller unit or even one word. This also covers cases where one multiple word unit is paraphrased with another multiple word unit of the same size. The original or the paraphrased text may be an idiom or collocation.	Also, because of the numbers of people involved and the greater resources they possess, a group can work more quickly in response to the task assigned to it and can **come up with** highly creative solutions to problems and issues.	Second, members work faster and **create** better solutions.
	Also, because of the numbers of people involved and the greater resources they possess, a group can work more quickly in response to the task assigned to it and can come up with **highly creative solutions** to problems and issues.	Second, members work faster and create **better solutions.**
	On the other hand, there were other instances where one or two people **managed to become** very influential over what their group did.	In addition to that, some group members **were** much more influential than others and usually got things their own way, causing a bitter attitude on the other group members.

Note. A lexical (word-based) paraphrase is a paraphrase characterized by modification to a single word or multiple word units.

Reference Text

"Well . . . the recognition for a job well done went to the group as a whole, no names were named. "

Focal Text

"However, they got the same amount of recognition as the members who actually worked, and this situation also caused a feeling of frustration."

Figure D4. **Conceptual Paraphrase Example 1.**

Reference Text

"Also, the individual team member has a much better chance to "shine," to get his or her contributions and ideas not only recognized but recognized as highly significant, because a team's overall results can be more far-reaching and have greater impact than what might have otherwise been possible for the person to accomplish or contribute working alone. "

Focal Text

"Every one of the team has the chance to shine, that is, to show a bright idea that may be accepted by the others."

Figure D5. **Conceptual Paraphrase Example 2.**

Appendix E

Interannotator Agreement for Linguistic Classifications for P-Units

Table E1

Nonnative Speaker Data: Interannotator Agreement on Syntactic, Lexical, and Conceptual Classifications of P-Units for 50 Responses

Linguistic modification classification	Thresholds							
	1		0.7		0.6		0.5	
	N	Agr	*N*	Agr	*N*	Agr	*N*	Agr
Syntactic paraphrase	203	0.77	303	0.74	355	0.71	414	0.68
Active-passive	24	0.46	41	0.41	44	0.41	53	0.36
Verb tense shift	100	0.76	165	0.7	193	0.67	231	0.66
Verb aspect shift	48	0.46	76	0.46	97	0.43	116	0.4
Declarative-question	0	0	1	0	2	0	2	0
Finite/ nonfinite VP	27	0.41	43	0.35	47	0.36	57	0.33
Pronoun vs. noun/NP	37	0.78	37	0.78	37	0.78	38	0.76
Relative clause/NP	15	0.4	23	0.35	27	0.3	30	0.27
Relative clause/VP	3	0.33	7	0.14	12	0.08	14	0.07
Reordering of complements	28	0.5	59	0.41	66	0.39	76	0.36
Unspecified syntactic reordering	21	0.1	40	0.08	46	0.07	56	0.05
Lexical paraphrase	446	0.92	540	0.88	629	0.88	691	0.86
Synonyms	22	0.82	22	0.82	38	0.47	43	0.42

Linguistic modification classification	Thresholds							
	1		0.7		0.6		0.5	
	N	Agr	*N*	Agr	*N*	Agr	*N*	Agr
Morphology	76	0.55	107	0.51	122	0.5	132	0.49
Multiple word units	404	0.91	484	0.87	569	0.84	628	0.81
Unspecified lexical substitution	4	0	6	0	7	0	9	0
Conceptual paraphrase	82	0.82	162	0.77	196	0.74	241	0.7
Global paraphrase	29	1	41	1	44	1	46	1
Reading and lecture	13	0.77	17	0.65	18	0.61	18	0.61
Reading	11	0.55	15	0.6	16	0.63	17	0.59
Lecture	12	0.75	19	0.74	21	0.71	23	0.74

Note. N = number of instances that both annotators classified in a category; Agr = interannotator agreement for *N* classifications in a category; NP = noun phrase; VP = verb phrase.

Table E2

Native Speaker Data: Interannotator Agreement on Syntactic, Lexical, and Conceptual Classifications of P-Units for 50 Responses

Linguistic modification classification	Thresholds							
	1		0.7		0.6		0.5	
	N	Agr.	*N*	Agr.	*N*	Agr.	*N*	Agr.
Syntactic paraphrase	176	0.72	277	0.7	354	0.67	427	0.65
Active-passive	17	0.41	27	0.41	38	0.42	56	0.36
Verb tense shift	87	0.69	161	0.66	210	0.6	261	0.57
Verb aspect shift	28	0.54	47	0.47	69	0.43	81	0.42

Linguistic modification classification	Thresholds							
	1		0.7		0.6		0.5	
	N	Agr.	*N*	Agr.	*N*	Agr.	*N*	Agr.
Declarative-question	0	0	0	0	0	0	0	0
Finite/ nonfinite VP	18	0.11	32	0.19	38	0.24	47	0.26
Pronoun vs. noun/NP	38	0.66	38	0.66	42	0.62	44	0.61
Relative clause/NP	18	0.5	25	0.44	31	0.35	36	0.31
Relative clause/VP	6	0.33	21	0.19	26	0.15	32	0.13
Reordering of complements	34	0.26	56	0.29	65	0.28	83	0.28
Unspecified syntactic reordering	10	0	23	0	31	0	41	0
Lexical paraphrase	382	0.94	474	0.9	555	0.88	644	0.86
Synonyms	15	0.8	15	0.8	24	0.54	30	0.43
Morphology	50	0.36	70	0.3	90	0.29	117	0.28
Multiple word units	358	0.93	445	0.89	522	0.85	607	0.82
Unspecified lexical substitution	5	0.2	7	0.14	10	0.1	12	0.08
Conceptual paraphrase	61	0.8	143	0.76	189	0.75	248	0.69
Global paraphrase	45	1	52	1	57	1	60	1
Reading and lecture	17	0.82	21	0.81	24	0.71	26	0.73
Reading	11	0.91	11	0.91	12	0.92	12	0.92
Lecture	21	0.81	25	0.8	29	0.72	30	0.73

Note. N = number of instances that both annotators classified in a category; Agr = interannotator agreement for *N* classifications in a category; NP = noun phrase; VP = verb phrase.