Research Report

Applying Content Similarity Metrics to Corpus Data: Differences Between Native and Non-Native Speaker Responses to a TOEFL® Integrated Writing Prompt

Paul Deane

Olga Gurevich

October 2008

ETS RR-08-51

Applying Content Similarity Metrics to Corpus Data: Differences between Native and Non-Native Speaker Responses to a TOEFL® Integrated Writing Prompt

Paul Deane
ETS, Princeton, NJ

Olga Gurevich[1]
Powerset, Inc., San Francisco, CA

October 2008

Abstract

For many purposes, it is useful to collect a corpus of texts all produced to the same stimulus, whether to measure performance (as on a test) or to test hypotheses about population differences. This paper examines several methods for measuring similarities in phrasing and content and demonstrates that these methods can be used to identify population differences between native and non-native speakers of English in a writing task.

Key words: Corpus, NLP, TOEFL, content scoring, EFL, BLEU scores, CVA (content vector analysis)

Table of Contents

Page

List of Tables

Page

List of Figures

Introduction

People produce text (whether spoken or written) in many situations in response to the same stimulus. In real world applications, these situations include tests, where a student's score depends at least partially upon successfully producing the correct content; in linguistic studies, on the other hand, holding the target content constant in order to study differences among populations is often useful. For instance, discourse studies traditionally use the same pictorial stimuli to obtain samples of narrative across language groups and populations (Chafe, 1980; cf. MacCabe & Peterson, 1991; Passonneau, Goodkind, & Levy, 2007).

However, some of the dimensions of linguistic relevance are precisely the choices language users make with respect to content. For instance, if language users are required to present or discuss content from another source, educators may wish to know how closely they have replicated the original phrasing, whether they have chosen to paraphrase or summarize material from their sources, whether they have chosen to introduce material irrelevant to or unmentioned in their source, and so forth. A variety of methods for studying similarity of text content have been developed within the fields of computational linguistics and information retrieval, including content vector analysis (CVA; Salton, 1989), latent semantic analysis (LSA; Landauer, Folz, & Latham, 1998), and BLEU scores (Papineni, Roukos, Ward, & Zhu, 2002),[2] but these methods have usually been applied instrumentally for such purposes as information retrieval, automated essay scoring, and other natural language processing applications. In this report, we examine ways in which these and a range of more sophisticated techniques can be used to support inferences about differences between populations.

In particular, we analyzed results from the new Test of English as a Foreign Language™ (TOEFL®) integrated writing task, described in the next section.[3] All task participants received the same set of prompts and were asked to summarize them. The resulting essays all tried to express the same underlying, or gist, content, so any measurable differences between them had to be due to differences in individual language ability and style. Thus the task was uniquely suited to measuring differences in linguistic behavior between populations, notably between native and non-native speakers.

One of the challenges facing any attempt to measure similarity of document content is the need to account for the complex mapping from surface form to intended meaning. Two utterances considered by native speakers to express the same content may use different but equivalent words, may use the same words arranged in different syntactic patterns, may omit information inferable from other parts of the text, or may combine any of these devices in a variety of ways. Given the limitations of natural language processing (NLP) techniques for disambiguating word senses, representing synonyms and paraphrases, and defining other deeper semantic properties of texts, methods for representing gist similarity, rather than surface text similarity, are necessarily imperfect. One of the goals of this study was therefore to explore multiple techniques for characterizing the similarity of text content, with some techniques representing standard methods for assessing similarity of vocabulary and phrasing and others representing experimental methods intended to provide an approximate measurement of similarity in gist, or semantic, content.

TOEFL Integrated Writing Task and Scoring

The TOEFL is administered to foreign students wishing to enroll in U.S. or Canadian universities. It aims to measure the extent to which a student has acquired English; thus native speakers should, on average, perform better on the test regardless of their analytical abilities. The TOEFL now includes integrated tasks intended to measure overall communicative competence, and pilot studies were conducted with native as well as non-native speakers.

One of the writing components is an integrated writing task. Students first read an expository passage, which remains on the screen throughout the task. Students then hear a segment of a lecture concerning the same topic, however, the lecture contradicts and complements the information contained in the reading. Students hear the lecture once, then summarize the lecture and the reading and describe any contradictions between them.

Human raters score the resulting essays on a scale of 0 to 5, with 5 being the best possible score. The highest scoring essays express ideas from both the lecture and the reading using correct grammar; the lowest scoring essays rely on only one of the prompts for information and have grammatical problems; scores between the two extremes reflect both partially correct content and the presence of linguistic errors.

The test prompt for our study contained passages about the advantages and disadvantages of working in groups. The reading was 260 words long; the lecture, 326 words. In 2004, ETS tested 540 non-native speakers and 950 native speakers[4] and collected essential demographic data such as native language, educational level, and so on, for each student.[5] For later validation, we excluded one third of each set, selected at random, thus involving 363 non-native speakers and 600 native speakers.

Figure 1 shows that the distribution for non-native speakers was positively skewed, with the most common score being 1. By contrast, native speaker scores centered around 3 and were slightly negatively skewed. The difference in distributions confirms that the task is effective at separating non-native speakers by skill level and is easier for native speakers.[6] Potential sources of difficulty include comprehension of the reading passage, listening ability and memory for the lecture, and the analytical ability to find commonalities and differences between the content of the reading and the lecture.

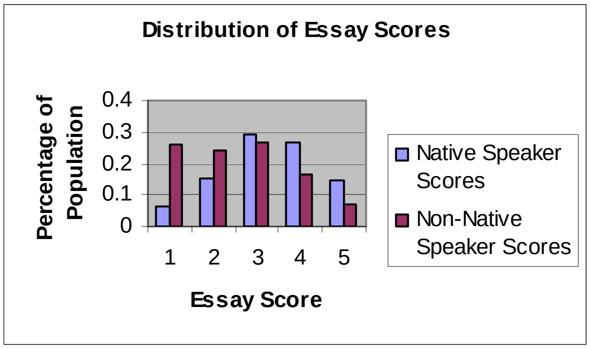

***Figure 1*. Relative score distributions.**

Document Similarity Measures

Due to the design of the TOEFL task, the content of the student essays is highly constrained. A successful student response (e.g., one that receives a score of 5) must successfully summarize the content of both the reading and the lecture, and must correctly characterize the relationship between them. The scoring rubric is also sensitive to grammatical and lexical errors, so that the essay score cannot be attributed solely to content but also cannot be ranked at the top of the scale without content accuracy. Nothing in the design of the task constrains speakers to use exact quotation or close or loose paraphrase, nor is there anything in the prompt or the scoring rubric that requires a particular response style, organizational pattern, or other discourse or structural feature.

Most obviously, the task requires subjects to coordinate reading and listening comprehension skills. Non-native speakers at low skill levels frequently have more trouble with listening than with reading (cf. Ferris & Tagg, 1996, for a discussion focusing on the student population for which TOEFL is designed), which might bias their responses toward the reading material. In addition, previous studies of native versus non-native speakers' text summarization (cf. Campbell, 1987; Keck, 2006) suggest that native speakers are much more likely to paraphrase the prompts while keeping the same gist, whereas non-native speakers are likely to either repeat the prompts close to verbatim or diverge from them in ways that do not preserve the gist.

Writing fluency (as reflected by the response document's length in words) can be expected to be associated with language skill among non-native speakers but not native speakers, and thus we expect to see greater correlations between document length and score level for the non-native speaker population (see Wolfe-Quintero, Inagaki, & Kim, 1998, which notes mixed results in the literature, with word length functioning as a predictor most reliably to compare populations at different proficiency levels). Since document length in words is a strong predictor of essay scores in general (Page & Peterson, 1995), we need to examine the extent to which the various applied metrics account for variance in human scores in addition to document length.

We thus hypothesize various ways in which the responses of native speakers might differ from the responses of non-native speakers. Native speakers can be expected to be more fluent, to use looser paraphrase patterns, and to display little indications of comprehension difficulty with either reading or lecture. Greater differences in fluency of text production can be expected across score levels; non-native speakers can be expected to repeat more often the exact phrasing of the stimulus texts, even at high score levels; and non-native speakers can be expected to rely more on the reading than on the lecture, at least at lower score levels.

To examine how native and non-native speakers perform and to determine whether construct-significant differences exist between the populations in the ways that they summarize content on tasks of this type, multiple measures of document similarity are needed that reflect different degrees of divergence from literal repetition, allowing us to measure how much subjects rely upon either source. To this end, we made use of several methods: CVA, which provides a measure of literal overlap of vocabulary; BLEU scores and the variant adopted by the National Institute of Standards and Technologies (NIST scores, cf. NIST, 2002), which provide a measure of exact phrasal overlap; co-occurrence vector measures of document similarity (in the general family of methods as LSA); and two novel methods, to be described below, which support measurement of document similarity sensitive to variations grammatical structure and to semantic relations among words.

Content Vector Analysis

The student essays and the prompts were compared using CVA, where each document was represented as a vector consisting of the words in it (Salton, 1989).[7] The *tf*idf*-weighted vectors were compared by a cosine measure.

The distribution of responses, regardless of population, closely followed the trend line of equal similarity to lecture and similarity to reading, as illustrated in Figure 2.

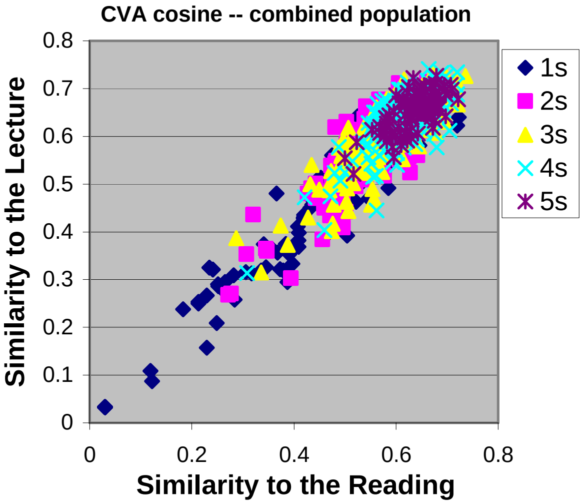

Figure 2. **Content vector analysis (CVA) similarity to reading and lecture by score level for combined native and non-native responses.**

Analysis of the native and non-native subgroups revealed similar trends and very little difference in group behavior on this measure.

For non-native speakers, a noticeable trend was observed. At higher score levels (where the score is determined by a human rater), student essays showed more similarity to both the reading and the lecture prompts. Both the reading and lecture similarity trends were significant (linear trend; $F = \text{MS}_{\text{linear trend}}/\text{MS}_{\text{within-subjects}} = 63$ for the reading; $F = 71$ for the lecture at.05 significance level[8]). Thus, the rate of vocabulary retention from both prompts increases with higher essay scores.

Native speakers showed a similar pattern of increasing cosine similarity between the essay and the reading (F = 35 at.05 significance for the trend), and the lecture (F = 35 at the.05 level).

We calculated Pearson and Spearman correlations between the human score and the CVA similarity of student response to both the reading and the lecture, and obtained the results in Table 1.

Table 1

Correlations Between Content Vector Analysis (CVA) Similarity and Human Scores

	Pearson correlation to human score
CVA similarity to the reading (native speakers)	.44
CVA similarity to the lecture (native speakers)	.41
CVA similarity to the reading (non-native speakers)	.56
CVA similarity to the lecture (non-native speakers)	.59

Note: All values are $p < .01$.

The correlations obtained (between .37 and .61) fell into the range generally observed when CVA is correlated with human essay scores for large scale testing programs such as Graduate Record Examinations® (GRE®) or TOEFL (see, for instance, the range of correlations reported in Attali & Burstein, 2005), despite the fact that only a single reference essay was used in the comparison. The difference between the CVA scores obtained when comparing the target passage to the reading and the lecture did not appear to be significant either for native or non-native speakers, though the somewhat higher CVA similarity for non-native speakers was consistent with the hypothesis that non-native speakers make use of more literal repetition of content than do native speakers.

We then examined the interaction between these variables and document length.[9] We performed multiple linear regression and built models in which word count was used alone and in combination with CVA cosine similarity to the reading and to the lecture. Content vector analysis accounted for some variance above and beyond document length (about .05 difference in R^2),though the patterns for native and non-native speakers were quite different. For native speakers, the best model loaded on document length plus similarity to the reading, though there was very little difference among the models that loaded on similarity to the reading, similarity to the lecture, and similarity to both (R^2 of .26, .25, and .25 respectively). For non-native speakers, the best model loaded on document length plus similarity to both reading and lecture, with an R^2 of .41. Table 2 presents the best models for each.

Table 2

Results of Regression Analysis for Content Vector Analysis (CVA) Similarity

	Native speakers	Non-native speakers
R^2 (document length alone)	.18	.43
R^2 for best model	.23	.48
p-values for best model	< .01	< .01
Partial correlation for document length	.28	.41
Partial correlation for similarity to the reading	.11	-.08
Partial correlation for similarity to the lecture	.03	.20

These results show that CVA provides some additional measurement above and beyond document length for this dataset. It also indicates relatively small population differences between the native and non-native speaker populations:

- Document length is a stronger predictor of overall score for non-native speakers, consistent with the hypothesis that writing fluency will reflect underlying fluency for that population.
- Though the analysis of variance (ANOVA)-style trend analysis did not reveal a significant effect when document length was not considered, the regression model suggests that non-native speakers replicate more material from the lecture and less from the reading at higher score levels.

Document Similarity Based on Co-occurrence (Semantic Space) Vectors

Content vector analysis uses literal equivalence of words to calculate document similarity. The other current major method to measure document similarity involves vector representations based on co-occurrence data from a corpus. Perhaps the best known method of this type is LSA (Landauer et al., 1998), though a number of similar techniques have also been developed, most notably the Word Space method (Schütze, 1993) and the Hyperspace Analog to Language (Lund & Burgess, 1996; Lund, Burgess, & Atchley, 1995). This class of methods continues to be developed (see Rohde, Gonnerman, & Plaut, 2005) and, in its LSA version, has been extensively used to provide scoring of content similarity, most notably in the Summary Street system (Wade-Stein & Kintsch, 2004). In this class of methods, similarity between documents can be represented by summing the vectors of the words contained in the document and then calculating cosine similarities between the resulting document vectors.

We calculated document similarity scores on a freely available vector space model (the correlated occurrence analogue to lexical semantics [COALS] model of Rohde et al., 2005, available from http://dlt4.mit.edu/~dr/COALS/ as of June 25, 2006) and the LSA general vocabulary space available from the University of Colorado, Boulder, Latent Semantic Analysis Web site (http://lsa.colorado.edu/), using the general reading (up to first year college) space with 300 factors and document-document comparison. Neither of these models yielded strong predictions of human score. For the COALS vectors, all correlations were below an absolute value of .1. Almost all responses, whether high or low scoring, had high cosine similarity to the lecture and to the reading. The LSA vectors' predictive power was somewhat stronger, but none was in excess of .29.

It appears that the representation in terms of underlying dimensions employed by LSA may have adversely affected predictive power by eliminating the discrimination afforded by the specific words used in each document. Given the relatively low correlations we found using latent semantic methods for this prompt, we did not conduct regression analyses. The correlation between LSA scores and document length for these texts was .55.

NIST and BLEU Scores

To measure the extent to which whole chunks of text from the prompt were reproduced in the student essays, we used NIST and BLEU scores, known from studies of machine translation (NIST, 2002; Papineni et al., 2002). NIST and BLEU scores are a measure of overlap of *n*-grams (whole phrases or word sequences) with various adjustments; for instance, BLEU scores apply a penalty to correct for response length effects. We used whole essays as sections of text rather than individual sentences, following the applications of *n*-gram methods to summary evaluation by Lin and Hovy (2003). We examined both methods for their efficacy in predicting human score and for their interaction with document length. Table 3 presents the raw correlations with human score for both native and non-native speakers.

Table 3

Correlations Between BLEU- and NIST-Score Similarity and Human Scores

	NIST scores	BLEU scores
Similarity to the reading (native speakers)	.42	.12
Similarity to the lecture (native speakers)	.37	.34
Similarity to the reading (non-native speakers)	.65	-.10
Similarity to the lecture (non-native speakers)	.65	.40

NIST Scores

The distribution of responses, regardless of population, closely followed the trend line of equal similarity to lecture and similarity, as illustrated in Figure 3.

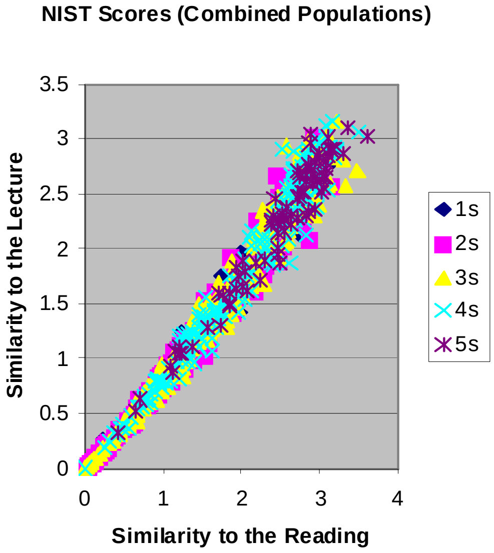

Figure 3. **NIST similarity to lecture and reading by score level for combined native and non-native responses.**

Analysis of the native and non-native subgroups revealed similar trends and very little difference in group behavior on this measure.

For non-native speakers, the trend was similar to that found with CVA: At higher score levels, the overlap between the essays and both prompts increased (*F* = 52.4 at the .05 level for the reading; *F* = 53.6 for the lecture).

Native speakers again showed a similar pattern, with a significant trend toward more similarity to the reading (*F* = 35.6) and the lecture (*F* = 31.3). These results were confirmed by a simple *n*-gram overlap measure. An ANOVA-style trend analysis did not reveal a significant difference between the native and the non-native speaker populations.

However, there was a larger difference in the strength of the correlation for non-native than for native speakers, as Table 3 indicates. Using NIST scores instead of CVA yielded correlations in the same range as CVA.

We then examined the interaction between these variables and document length. We performed multivariate linear regression and built models in which word count was used alone and in combination with NIST score similarity to the reading and to the lecture. For both native and non-native speaker populations, the best model used document length in combination with both similarity to the reading and similarity to the lecture. Table 4 presents the results of these analyses.

Table 4

Regression Analysis for NIST Scores for Native and Non-Native Speakers

	Native speakers	Non-native speakers
R^2 for document length alone	.18	.43
R^2 for combined model	.19	.45
p-values for combined model	< .01	< .01
Partial correlation for document length in combined model	.13	.25
Partial correlation for similarity to the reading in combined model	.13	-.07
Partial correlation for similarity to the lecture in combined model	-.06	.18

The overall performance of the regression model using document length plus NIST scores was actually worse than that for simple CVA, consistent with the argument of Lin and Hovy (2003) that unigram overlap was preferable to longer *n*-gram sequences for determining quality of summaries.

BLEU Scores

Consider Table 5. BLEU scores had significantly smaller correlations with human scores than NIST scores, a fact which appears partly due to the fact that BLEU scores were more frequently zero, recognizing no match at all between student responses and the stimulus materials. However, when document length was factored in, BLEU scores explained significantly more variance than NIST scores and slightly outperformed the use of CVA comparison to the stimulus materials for the non-native speaker population. The reason for this increased performance appears to be the negative partial correlation between human scores and the similarity of the student responses to the reading. This result is consistent with the CVA analysis, but the trend is more pronounced. BLEU scores do not appear to detect the high native speaker unigram similarity to the reading detected using CVA cosines.

Table 5

Regression Analysis for BLEU Scores for Native and Non-Native Speakers

	Native speakers	Non-native speakers
R^2 for document length alone	.18	.43
R^2 for combined model	.21	.49
p-values for combined model	< .01	< .01
Partial correlation for document length in combined model	.12	.61
Partial correlation for similarity to the reading in combined model	.001	-.28
Partial correlation for similarity to the lecture in combined model	.19	.14

Lexico-Grammatical Similarity

Neither BLEU scores nor CVA similarity indicated a strong and clear-cut difference in behavior between the two sources (reading and lecture), though both regression analyses suggested that such an effect may be present. However, neither measure is sensitive to grammatical structure or to word meaning. The succeeding sections of this paper examine a series of efforts to construct measures of content similarity that are more sensitive to gist similarity rather than to exact repetition of content.

We began by measuring lexico-grammatical similarity between each essay and the two prompts. Each essay was represented as a set of features derived from its lexico-grammatical content, as described below. The resulting comparison measure goes beyond simple word or *n*-gram overlap by providing a measure of structural similarity as well. In essence, our method measures to what extent the essay expressed the content of the prompt in the same words, used in the same syntactic positions.

C-rater™ Tuples

To get a measure of syntactic similarity, we relied on *c-rater* (Leacock & Chodorow, 2003), an automatic scoring engine developed at ETS. *C-rater* includes several basic NLP components, including part-of-speech tagging, morphological processing, anaphora resolution, and shallow parsing. The parsing produces *tuples* for each clause, which describe each verb and its syntactic arguments. *C-rater* does not produce full-sentence trees or prepositional phrase attachment. However, the tuples are reasonably accurate on non-native input.

Lexical and Syntactic Features

C-rater produces tuples for each document, often several per sentence. For the current experiment, we used the main verb, its subject, and its object. We then converted each tuple into a set of features, which included the following:

- the verb, subject (pro)noun, and object (pro)noun as individual words
- all of the words together as a single feature
- the verb, subject, and object words with their argument roles

Each document could then be represented as a set of tuple-derived features, or feature vectors.

Document Comparison

Two feature vectors derived from tuples can be compared using a cosine measure. The closer to 1 the cosine, the more similar the two feature sets. To compensate for different frequencies of the features and for varying document lengths, the feature vectors were weighted using standard *tf*idf* techniques.

In order to estimate the similarity between two documents, we used the following procedure. For each tuple vector in Document A, we found the tuple in Document B with the maximum cosine to the tuple in Document A. The maximum cosine values for each tuple were then averaged, resulting in a single scalar value for Document A. We called this measure average maximum cosine (AMC).

We calculated AMCs for each student response versus the reading, the lecture, and the reading and lecture combined. This procedure was performed for both native and non-native essays.

Results and Discussion

Overall similarity to reading and lecture. The AMC similarity measure, which relies on syntactic as well as lexical similarity, produced somewhat different results from simpler bag-of-word or *n*-gram measures. In particular, we found a difference in behavior between native and non-native speakers: Non-native speakers showed increased structural similarity to the lecture with increasing scores, but native speakers did not.

However, the AMC similarity measures did not predict overall score well. The AMC similarity between the reading and the student response accounted for 18% of the variance; the AMC similarity between the lecture and the student response accounted for 27% of the variance, slightly less than the difference between the AMC cosines for reading and lecture. This relatively low level of agreement may be due to the fact that the tuple features made use of only a small portion of the structure of each passage (essentially, only words occupying the subject, object, and verb positions in each clause). Despite this relatively low performance, a significant group trend was apparent.

For non-native speakers, the trend of increased AMC between the essay and the lecture was significant (F = 10.9). On the other hand, no significant increase in AMC was found between non-native essays and the reading (F = 3.4). Overall, for non-native speakers the mean AMC was higher for the reading than for the lecture (.114 versus .08).

Native speakers, by contrast, showed no significant trends for either the reading or the lecture. Overall, the average AMCs for the reading and the lecture were comparable (.08 versus .075).

We know from results of CVA and BLEU analyses for both groups of speakers that higher-scoring essays are more lexically similar to the prompts. Thus, the lack of a trend for native speakers must be due to lack of increase in structural similarity between higher scoring essays and the prompts. Because better essays are presumably better at expressing the content of the prompts, these results are consistent with the hypothesis that native speakers paraphrase the content more than non-native speakers.

To examine how these variables played out in further detail, we performed multivariate linear regression and built models in which document length was used alone and in combination with AMC cosine similarity to the reading and to the lecture. These results confirmed the trend analysis. For native speakers, AMC cosines provided no prediction above and beyond document length. The best model for non-native speakers improved R^2 by .03 and yielded the following partial correlations: .61 partial correlation for document length, -.13 for AMC similarity to the reading, and 15 for AMC similarity to the lecture. This finding was consistent with the regression models for CVA and BLEU scores.

Difference between lecture and reading. Given that the regression models suggested a difference between the native and non-native speaker populations, we examined whether any measure based on AMC cosines could demonstrate such a difference more strongly. The most informative measure of speaker behavior was the difference between the AMC with the reading and the lecture, calculated by subtracting the lecture AMC from the reading AMC. Here, non-native speakers showed a significant downward linear trend with increasing score (F = 6.5; partial eta-squared .08), whereas the native speakers did not show any trend (F = 1.5). The AMC differences are plotted in Figure 4.

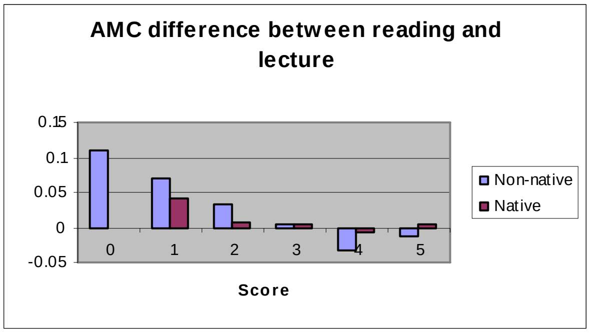

Figure 4. **Average maximum cosine (AMC) difference in similarity to reading and lecture by score point.**

Non-native speakers with lower scores rely more on the reading to produce their response, whereas speakers with higher scores rely somewhat more on the lecture than on the reading. By contrast, native speakers show no correlation between score and reading versus lecture similarity. Thus, there is a significant difference in the overall distribution and behavior between native and non-native speaker populations. This difference also indicates that human raters rely on information other than simple verbatim similarity to the lecture in assigning the overall scores.

Combining Syntactic and Lexical Similarity

Thus far we have examined methods involving strictly lexical comparisons (CVA), lexical comparisons with latent semantic dimensions (LSA-like methods), phrasal matching (BLEU scores), and a combination of strict lexical matching with syntactic information (AMC similarity based on *c-rater* tuples). We know, however, that responses can vary simultaneously in wording (involving use of synonyms and topically related words) and in syntactic form, and thus it would be useful to have a measure of the extent to which two documents are similar by virtue of using similar words in similar syntactic relations.

In particular, we are interested in combining vector-based methods of representing similarity in word meaning with syntactic information about the relationships among words in context. Such a method, while not capable of directly representing paraphrase relations in text, would capture important aspects of literal document similarity.

Tensor Products and Convolution: Connectionist Methods and Their Limitations

The problem of combining relational information with vector representations has been most directly confronted in the connectionist literature on language processing, where a number of mechanisms have been proposed for enabling a connectionist system to represent relational information in combination with vector-based representations of lexical content. Smolensky (1990) proposed the use of tensor products for this purpose, that is, he proposes representing the relation between two vectors as the outer product of the vector pair. This method has the disadvantage of requiring high-dimensional matrix representations to describe complex structure, and thus alternatives have been proposed. In particular, Plate (2003) examined the mathematical properties of *circular convolution*—a method for compressing the outer product of two vectors into a vector of the same dimensionality as the original vectors—and proposes that circular convolution provides a viable mechanism by which neural systems can encode relational information.

In the system that Plate (2003) devised, a predicate is represented by the sum of a set of convolved vectors. For instance, if *dog* is represented by the vector v_{dog} and the semantic role for the subject of *bark* is represented by the vector $v_{bark,subj}$ then the relationship between the two is represented by the convolved vector

$$v_{dog} \circ v_{bark\ subj}$$

and a complex set of propositions can be represented by the sum of these vectors. Plate demonstrates not only that sets of propositions can be encoded in this format but also that the similarity (dot product) between these combined vectors reflects the structure that went into them.

While vector convolution as presented by Plate (2003) has several very attractive properties, technical issues make it difficult to implement a document similarity metric based directly upon it. Perhaps the most important of these is the fact that the vector-based representations in Plate's system are assumed to be independent of (i.e., orthogonal to) one another. This assumption maximizes the information capacity of a convolution-based system but is unrealistic when applied to lexical vector representations derived from co-occurrence data. When typical lexical co-occurrence vectors were convolved and summed, we observed significant crosstalk and information loss. That is, the information stored in the vectors tended to disappear when too many vectors were superimposed on each other, making the method's application to essay-length documents problematic.

We therefore developed a modified method designed to combine lexical vectors with syntactic relations while avoiding the crosstalk and information loss observed when all vectors associated with a document were combined into a single sum. This method is described in detail in the section that follows.

A System for Combining Lexical Vectors With Syntactic Relations

We developed a system that processed parsed text, extracted grammatical relations, translated individual words into vector representations, and then encoded that information in a set of convolved vectors. The details are as follows:

Input format. We parsed responses to the TOEFL integrated writing prompt using the OpenNLP parser (http://opennlp.sourceforge.net/) to obtain parse trees in Penn Treebank format.

Extraction of functional relations. Functional relations such as subject, object, head, and complement were extracted from the parse trees using a parse tree interpreter written in Java. The final output of this part of the system was a set of relations such as *(dog is the subject of barked)* that directly represented the grammatical dependency structures for each sentence.

Stemming and vector lookup. Words were mapped to inflectional stem and root forms using a simple lexical lookup algorithm. Lexical data (both COALS vectors and WordFit vectors, cf. Deane, 2003) were then obtained both for stem and root forms.

A semantic hash for functional relations. In order to avoid crosstalk between unrelated words, each word vector was mapped to a small set of semantic classes defined by a vector centroid for a set of paradigmatic words. These classes were defined by extracting classes from the results of a clustering algorithm applied to the lexical vector data. The classes so defined included about 20 different semantic classes of nouns, including human, animal, concrete, moveable object, and landscape element, and a similar but smaller number of classes for verbs, adjectives, and adverbs.

Functional relations were then mapped to one of a small set of high-level classes based upon assignment of their component words to semantic classes. For instance, the relations *(dog is the subject of barked)* mapped to the relationship between a class of nouns (nouns for types of animals) and a class of verbs (verbs of bodily action). All grammatical relations involving an animal as argument and a bodily action as predicate mapped to the same relation that was then used as a key by which to store and retrieve information associated with instances of that category. Thus a document was represented as a set of vectors—one for each relational class present in the parsed input—rather than a single vector.

Encoding of functional relations. In order for the system to function as desired, it was important that relationships should count as similar to the extent (a) that the head/predicate words were similar, (b) that the argument words were similar, and (c) that the grammatical relations were similar. We used the lexical vectors to account for (a) and (b), and created artificial vector relationships for roles like subject and object, designed to guarantee (for instance) that passive subjects are more like direct objects than active subjects are and that all three of these are more similar to one another than to predicate roles. We recentered the lexical vectors around the class prototype by subtracting the class centroid vector, applied convolution to combine the vectors, and added the resulting vector to the vector stored under the appropriate relational key. The resulting vector thus represented the central tendency of the document with respect to the particular relational class it encoded.

Calculation of document similarity. The representation of document content induced by these methods can be viewed as a summary of the network of predicate-argument relations present in a text. Two documents will be similar to the extent that they contain similar vectors belonging to the same relational class.

We experimented with several methods for calculating document similarity. The method that appeared to work best involved three steps:

1. Each vector v_{Ai} was compared to the corresponding vector in the other document V_{Bi}, and the cosine similarity between vectors was calculated.
2. The Euclidean lengths $|V_{Ai}|$ (for shared vectors) and $|V_{Aj}|$ (for all vectors) in the reference document were calculated.
3. The following formula was applied:

$$\frac{\sum \cos(v_{Ai}, v_{Bi}) * | v_{Ai} |}{\sum | v_{Aj} |}$$

The resulting value is an asymmetric measure that indicates to what extent the word-word relations present in the reference document are matched with similar relations in the target document. It is not a direct representation of paraphrase relations or content equivalencies in a deep semantic sense, but it provides a fairly direct measure of whether one document uses similar words in similar (but not necessarily identical) syntactic relations to those employed in another document.

Results and Discussion

When the combined response set was plotted against syntactic/semantic space similarity to the reading and to the lecture, we found once again a main trend line of equal similarity to the reading and to the lecture, but there was also a subgroup of responses that fell off the main trend line by exhibiting greater similarity to the reading. These almost exclusively lower-scored responses (1s and 2s and a few 3s) are illustrated in Figure 5.

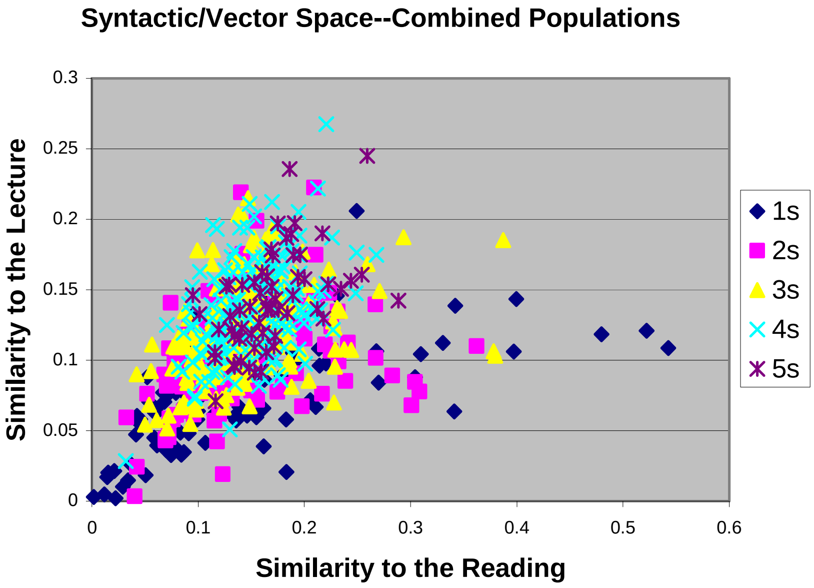

Figure 5. **Combined responses for syntactic/semantic space similarity to the reading and the lecture by score point.**

The most obvious explanation for these off-trend responses is that a subgroup of low-ability respondents adopted a strategy in which they copied material from the reading (which was continuously available while respondents were formulating their responses). Note, however, that the lack of off-trend-line responses with CVA and *n*-gram-based measures indicates that these strategies did not involve preferential copying of exact words and phrases from the reading but rather a repetition of similar words in similar grammatical relationships.[10] A closer examination of the two subpopulations reveals, in addition, that native speakers cannot be employing any such strategy, because responses well off the main trend line are predominantly produced by non-native speakers. In particular, of the 16 essays with scores of 1 or 2 in which similarity to the reading was above .28, 15 were produced by non-native respondents.

However, for the non-native group as a whole, similarity of student response to the reading was not predictive of score level, and a majority of 1- and 2-scored responses were close to the main trend line. Thus it appears likely that though a small group of non-native speakers employed a strategy of repeating material from the reading, this pattern is secondary to a general trend in which respondents attempt to replicate content from both prompts.

Further analysis indicates that there are additional differences between the native and non-native speaker populations. In the native speaker population, similarity of lecture and reading material to student responses predicts human score equally well (but at a low overall level). In the non-native speaker population, on the other hand, similarity of student response to the lecture is reasonably predictive of the human essay score. This confirms the trend observed with AMC cosines in section 3.4. Table 6 below presents correlations in both directions, given that the similarity measure is asymmetric.

Table 6

Correlations Between Syntactic/Semantic Space Similarity Measure and Human Scores

Correlation between human scores and the similarity of	Native speakers	Non-native speakers
Reading passage to essay response	.16	-.12
Essay response to reading passage	-.20	-.37
Lecture to essay response	.21	.55
Essay response to lecture	-.04	.09

Note. Correlations are all $p < .01$.

The best correlation we observed was very close to the corresponding CVA correlation; thus, it appears that including syntactic information in the matching algorithm did not improve prediction and, in most cases, reduced it, relative to use of simple content vectors or BLEU features. On the other hand, the correlations in Table 6 were stronger than the corresponding COALS correlations and roughly comparable to those using LSA. Thus, it appears that, at least for this prompt, including syntactic information improved or at least did not harm performance relative to use of semantic vector spaces.

We then examined the interaction between these variables and document length. We performed multivariate linear regression and built models in which word count was used alone and in combination with syntactic/semantic space similarity to the reading and to the lecture. As with AMC cosines, the syntactic/semantic space measure yielded no significant improvements in performance over document length alone for the native speaker population. With the non-native speaker population, however, as shown in Table 7, there were large increases in R^2, which increased from .43 for document length alone to .53 for the best model, yielding a better performance than that obtained with BLEU scores or with CVA, and with a lower partial correlation for document length than with BLEU scores.[11]

Table 7

Regression Analysis for Syntactic/Semantic Similarity Measure

	Native speakers	Non-native speakers
R^2 for document length alone	.18	.43
R^2 for 3-feature model	.22	.53
p-values for best model	< .01	< .01
Partial correlation for document length	.19	.41
Partial correlation for similarity of student responses to the reading	-.15	-.30
Partial correlation for similarity of the lecture to student responses	.28	.31

Combining BLEU and syntactic/semantic space similarity demonstrated that the two measures captured some independent variance. The combined model increased R^2 to .54 for non-native speakers and to .25 for native speakers. Table 8 indicates the distribution of variance across the five relevant features. These results are consistent with trends we have seen throughout this paper: (a) measures reflecting word order as well as lexical content consistently perform more effectively for the non-native speaker population; and (b) the difference between high and low performance among non-native speakers is strongly connected to the writer's ability to replicate content from the lecture.

Table 8

Multiple Regression Analysis for Combined BLEU and Syntactic/Semantic Measures

	Native speakers	Non-native speakers
R^2 for document length alone	.18	.43
R^2 adding BLEU scores	.21	.49
R^2 adding syntactic/semantic space measure	.22	.53
5-feature model R^2	.25	.54
Partial correlation for document length	.20	.38
Partial correlation for BLEU score similarity to the reading	.05	-.11
Partial correlation for BLEU score similarity to the lecture	.16	.12
Partial correlation for syntactic/semantic space similarity of responses to the reading	-.16	-.17
Partial correlation for syntactic/semantic space similarity of the lecture to responses	.16	.29

Correlations to TOEFL Scale Scores

It is striking that the content-scoring methods examined in this paper worked worse with the native speaker population than with the non-native speakers. Since both the native and non-native speaker essays were scored using the same scoring rubric and under similar operational conditions, the results suggest significant qualitative differences in the performance of native and non-native speakers. To confirm whether our results can reasonably be viewed in this light, we examined how three of the measures correlate with overall TOEFL scaled scores and with TOEFL scaled subscores. In particular, we examined CVA cosines, BLEU scores, and syntactic/semantic space similarity.

For non-native speakers, we got the correlations shown in Table 9. Note the negative correlations across the board in the third and fifth rows of the table, corresponding to the negative correlations we observed for both syntax-sensitive measures when the reading passage was treated as the reference document.

Table 9

Non-Native Speakers: Correlations of Content Vector Analysis (CVA), BLEU, and Syntactic/Semantic Space Measures of Content to TOEFL Scale Scores and Subscores

	Overall	Reading	Writing	Speaking	Listening
CVA (passage)	.46	.39	.45	.41	.41
CVA (lecture)	.47	.39	.44	.40	.42
BLEU (passage)	-.15	-.14	-.11	-.12	-.16
BLEU (lecture)	.36	.29	.38	.30	.30
Syntactic/semantic space (passage)	-.40	-.33	-.38	-.35	-.37
Semantic/semantic space (lecture)	.55	.49	.56	.45	.49

In the non-native speaker data, similarity between the reading passage and the student response predicts a lower TOEFL score, both on the overall scale and on each subscale, if we use either the syntactic/semantic space or BLEU scores, though the correlations for BLEU scores are very small. This contrasts with positive correlations if we use CVA to make the same measurement. Since the same pattern also appeared when we used these features to predict human scores, we may reasonably conclude that the syntactic/semantic space method is detecting a difference between high- and low-scoring students that cannot be detected by CVA, which is not sensitive to grammatical structure, or by BLEU scores, which only detect and measure grammatical structure less directly through exact phrasal overlap.

Table 10 shows the pattern of correlations for native speakers. Note that all correlations are positive, that they are smaller than for the non-native speakers, and that instead of being fairly similar across the board (with a slight trend toward higher correlations in the writing subscale), correlations to the writing subscale are much higher. While the third and fifth rows (for BLEU score and syntactic/semantic space similarity to the reading passage) display lower correlations, they are all positive, rather than being negative as for non-native speakers.

The correlations with TOEFL scale scores and subscores thus display exactly the same patterns we observed with human scores.

Table 10

Native Speakers: Correlations of Content Vector Analysis (CVA), BLEU, and Syntactic/Semantic Space Measures of Content to TOEFL Scale Scores and Subscores

	Overall	Reading	Writing	Speaking	Listening
CVA (passage)	.49	.41	.56	.36	.37
CVA (lecture)	.47	.40	.55	.35	.36
BLEU (passage)	.08	.07	.15	.04	.03
BLEU (lecture)	.39	.34	.42	.28	.27
Syntactic/semantic space (response to passage)	.16	.07	.23	.14	.05
Semantic/semantic space (lecture to response)	.37	.34	.53	.35	.26

E-rater® *Modeling: Capturing Content Similarity and General Writing Skills*

The results we obtained thus far indicate that CVA and two new types of content similarity features—BLEU scores and syntactic/semantic space similarity—may provide useful measurement of student responses, especially for non-native speakers; in particular, they present contrasting patterns that arguably reflect salient differences between the native and non-native speaker populations. However, the TOEFL integrated writing prompt is not scored purely for content. The rubric refers both to accuracy of content and to other features of writing skill. Thus a complete automated scoring model for this kind of prompt would need to combine scoring for content with scoring for writing quality.

ETS's *e-rater* automated scoring engine (Attali & Burstein, 2005; Burstein et al., 1998) is designed precisely to measure writing quality. The current version of this software, *e-rater* V. 2.0, calculates features to measure the following dimensions of writing skill: organization, development, vocabulary, grammar, usage, mechanics, and style. In addition, *e-rater* can be trained using prompt-specific CVA content features. One CVA feature aggregates information over all top-scoring essays in an essay training set and scores responses for their similarity to the resulting centroid vector. A second CVA feature calculates the CVA similarity of a response to the pooled CVA features for all essays at each score point and assigns the response to the score point that yields the highest cosine. *E-rater* models are built on top of this measurement foundation by applying multiple regression to train a model to replicate human holistic scores.

Content features such as BLEU scores and the syntactic/semantic space measure can be viewed within the context of *e-rater* as potential supplements to the baseline content-scoring capability that *e-rater* already provides. In addition, the *e-rater* feature set, by providing measurement both of content and of writing quality features, may allow us to pinpoint differences between the two populations. We therefore conducted regression analyses to determine how these two measures contributed to predicting human holistic scores within an *e-rater* model.

Table 11 presents the baseline performance of the *e-rater* features on the prompt analyzed in this report, with CVA features deleted. The important point to note is that these writing quality features by themselves provide similar R and R^2 values for both populations but perform somewhat worse for non-native than they do for native speakers, which is not particularly surprising given the fact that *e-rater* features were not specifically designed to account for second-language error patterns.

Table 11

Baseline Regression Models (E-rater Features Without Content Vector Analysis [CVA])

	Native speakers	Non-native speakers
Multiple *R*	.60	.65
R^2	.36	.43
Exact agreement	.47	.44
Adjacent agreement	.86	.71
p-values for best model	< .01	< .01
Partial correlations for writing quality features		
Organization	.23	.32
Development	.10	.31
Vocabulary	.16	.13
Grammar	.08	.03
Usage	.10	.07
Mechanics	.31	.21
Style	.09	.13

Table 12 presents the baseline performance of the *e-rater* features on the prompt analyzed in this report. The most striking feature of this analysis is the very large difference in the amount of variance accounted for by the two models. While the model for the native speaker data had an R^2 of .38 and assigned less then half of response essays to the correct score point, the model for non-native data had an R^2 of .78 and assigned two-thirds of response essays to the correct score point. Almost all the difference in performance can be attributed to a single feature, which uses CVA similarity to the *aggregate* vocabularies characteristic of each score point to predict where a student response is likely to fall. This feature jumped from a partial correlation of .10 for the native speaker data to a partial correlation of .68 for the non-native speaker data. In short, this aggregated CVA feature is providing a major increase in measurement for the non-native speaker population for this prompt.

Table 12

Baseline Regression Models (E-rater Features Only)

	Native speakers	Non-native speakers
Multiple *R*	.62	.88
R^2	.38	.78
Exact agreement	.46	.67
Adjacent agreement	.87	.95
p-values for best model	< .01	< .01
Partial correlations for content features		
Similarity to top essays (CVA)	.10	.18
Similarity to essays at all score points (CVA)	.10	.68
Partial correlations for writing quality features		
Organization	.19	.25
Development	.09	.25
Vocabulary	.12	.04
Grammar	.05	-.05
Usage	-.01	.04
Mechanics	.22	.02
Style	.08	.06

Note. CVA = content vector analysis.

It is not clear whether the aggregated CVA feature is actually measuring content in the strict sense, as it is directly measuring differences in vocabulary characteristically used by more or less skilled writers. Insofar as acquisition of flexible use of vocabulary is one concomitant of second language skill, it is possible that the successful prediction obtained by the baseline *e-rater* model is primarily due to characteristic vocabulary patterns. For our purposes, however, the major effect is that the two models present two very different baselines of performance for the content scoring features with which we are concerned.

Next we conducted analyses in which we added BLEU and the syntactic/semantic similarity space measures to the baseline model. Since these measures compare single documents (the reading and lecture) to student responses, they are at a disadvantage relative to the aggregated CVA features used in *e-rater* in that they cannot use patterns of word frequency over multiple essays to extract generalizations about the relative importance of particular words to the prompt. Interestingly, though, the combined model, shown in Table 13, performed about the same as the standard *e-rater* model with native speakers, and for non-native speakers it yielded a .10 increase in R^2 over the baseline shown in Table 11, an increase that was reflected in a .07 increase in exact+adjacent agreement over that baseline. Nonetheless, this model performed much worse than the standard *e-rater* model for this prompt, which benefits from the strong performance of the aggregate CVA feature.

It is worth considering, however, that since the CVA features built into *e-rater* are aggregate features, representing variance captured over hundreds of essays in the training set, such training sets will not be available in some contexts. A few reference documents will be available in other cases, however, particularly in low-stakes assessments where immediate feedback and development of large numbers of items are important considerations. While the additional content features we examined did not equal the performance of the aggregate CVA feature for this prompt for non-native speakers, they did provide additional measurement over the baseline.

We next considered whether BLEU scores and syntactic/semantic space features would add prediction to the standard *e-rater* models that include aggregated CVA features. The resulting regression analysis yielded the results shown in Table 14.

Table 13

Regression Models (E-rater and Content Features Without Aggregated Content Vector Analysis [CVA])

	Native speakers	Non-native speakers
Multiple *R*	.63	.73
R^2	.38	.53
Exact agreement	.47	.44
Adjacent agreement	.86	.77
p-values for best model	< .01	< .01
Partial correlations for content features		
BLEU similarity to the lecture	.12	.16
BLEU similarity to the reading	-.01	-.08
Syntactic/semantic space similarity to the lecture	.09	.24
Syntactic/semantic space similarity to the reading	-.09	-.17
Partial correlations for writing quality features		
Organization	.14	.21
Development	.05	.18
Vocabulary	.15	.10
Grammar	.07	.06
Usage	-.02	.05
Mechanics	.29	.13
Style	.08	.13

This analysis indicates that the additional content features had a positive impact on the resulting model: a .01 increase in *R* for both populations, yielding 2% and 3% increases in exact agreement. These results suggest that for some purposes, such as low-stakes practice examinations where extensive training materials are not available, it may be possible to use BLEU and syntactic/semantic space features to provide content measurement. For this prompt, at least, these content features did not capture large amounts of additional variance over the aggregated CVA features already in use in *e-rater,* even though they outperformed CVA similarity comparison to single reference documents.

Table 14

Regression Models (E-rater and Content Features With Aggregated Content Vector Analysis [CVA])

	Native speakers	Non-native speakers
Multiple *R*	.63	.89
R^2	.38	.78
Exact agreement	.48	.70
Adjacent agreement	.86	.94
p-values for best model	< .01	< .01
Partial correlations for content features		
Similarity to top essays	.10	.19
Similarity to essays at all score points	.05	.62
BLEU similarity to the lecture	.09	.01
BLEU similarity to the reading	-.04	-.05
Syntactic/semantic space similarity to the lecture	.03	.07
Syntactic/semantic Space similarity to the reading	-.07	-.06
Partial correlations for writing quality features		
Organization	.14	.22
Development	.06	.21
Vocabulary	.12	.03
Grammar	.05	-.04
Usage	-.02	.04
Mechanics	.21	.00
Style	.08	.08

These results appear to be consistent with the hypothesis, suggested in earlier sections, that non-native speakers tend to paraphrase their answers much less freely than native speakers. The regression models using *e-rater* features are consistent with that hypothesis, assuming that similar results can be obtained over a wide range of similar prompts. The higher predictive value of the aggregated cosine features for non-native speakers would follow directly from a pattern in which even high-scoring non-native speakers tended to use a smaller vocabulary than native speakers when summarizing text, which would of necessity force them to paraphrase more literally.

Conclusions

This study had two primary goals: to explore the efficacy of various document-similarity methods to predict differences in performance on a complex summarization task and to evaluate the extent to which these same methods can be used to reveal differences between populations.

With regard to predicting human score on this particular (and particularly complex) summarization task, the results indicate that a method that makes use of syntactic structure and lexical (semantic space) information about words can perform significantly better than simple (non-aggregated) CVA comparison to a reference document, with BLEU scores (which at least take phrase-level performance into account) displaying similar patterns at lower overall levels of performance. However, the differences between populations suggest that the methods' efficacy depends critically on population behaviors.

In particular, differences in how the various features perform between native and non-native speakers suggest the following hypotheses:

- Fluency of comprehension and production play a larger role among the non-native speaker population, accounting for much more of the variance than for native speakers.
- Differences in the ability to replicate the content of oral material play a larger role among the non-native speaker population.
- The failure of the syntax-sensitive methods to work well with the native speaker populations suggests that native speakers are less likely to reproduce content by using similar words in similar syntactic relations and are more likely to use looser forms of paraphrase.
- The success (for the non-native speaker population) of the aggregated CVA features used in *c-rater* would be consistent with the above hypotheses but suggests that differences within the non-native speaker population in vocabulary patterns may also play an important role.

Confirming these hypotheses will require replication of the results reported in this paper over many more prompts and additional studies to obtain convergent evidence, but it seems clear, at the very least, that NLP features used to identify similarity of document content are highly sensitive to the differences between native and non-native speakers.

References

Attali, Y., & Burstein, J. (2005). *Automated essay scoring with* e-rater *v.2.0* (ETS Research Rep. No. RR-04-45). Princeton, NJ: ETS.

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., et al. (1998). *Computer analysis of essay content for automated score prediction: A prototype automated scoring system for GMAT analytical writing assessment essays* (ETS Research Rep. No. RR-98-15). Princeton, NJ: ETS.

Campbell, C. (1987). *Writing with others' words: Native and non-native university students' use of information from a background reading text in academic compositions* (Technical Rep. No. 4). Los Angeles: UCLA, Center for Language Education and Research.

Chafe, Wallace L. (Ed.). (1980). *The pear stories: Cognitive, cultural and linguistic aspects of narrative production*. Norwood, NJ: Ablex Publishing.

Cline, F., & Powers, D. (2007). *The TOEFL (iBT): Evaluating its use with native speakers of English*. Manuscript in preparation.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL test* (TOEFL Monograph No. MS-30). Princeton, NJ: ETS.

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2005). *A teacher-verification study of speaking and writing prototype tasks for a new TOEFL test* (TOEFL Monograph No. MS-26). Princeton, NJ: ETS.

Deane, P. (2003). Cooccurrence and constructions. In L. Lagerwerf, W. Spooren, & L. Desgand (Eds.), *Determination of information and tenor in texts: Multidisciplinary approaches to discourse* (pp. 277–304). Ámsterdam: Stichting Neerlandistiek.

Ferris, D., & Tagg, T. (1996). Academic listening/speaking tasks for ESL students: Problems, suggestions, and implications. *TESOL Quarterly*, *30*(2), 297–320.

Higgins, D., & Burstein, J. (2006). Sentence similarity measures for essay coherence. In *Proceedings of the seventh international workshop on computational semantics (IWCS-7). Tilburg, The Netherlands.* Retrieved July 15, 2008, from http://www1.ets.org/Media/Research/pdf/erater_sentence_similarity.pdf

Keck, C. (2006). The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing, 15*(4), 261–278.

Landauer, T., Foltz, P. W., & Latham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284.

Leacock, C., & Chodorow, M. (2003). C-rater*:* Scoring of short-answer questions. *Computers and the Humanities, 37*(4), 389–405.

Lee, Y. W., & Kantor, R. (2005). *Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes* (TOEFL Monograph No. MS-31). Princeton, NJ: ETS.

Lin, C. Y., & Hovy, E. H. (2003). Automatic evaluation of summaries using *n*-gram co-occurrence statistics. In W. Dealemans & M. Osborne (Eds.), *Proceedings of the 2003 conference of the North American chapter of the Association for Computational Linguistics on Human Language Technology: Vol. 1. NAACL '03* (pp. 71–78). Morristown, NJ: Association for Computational Linguistics.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers, 28,* 203–208.

Lund, K., Burgess, C., & Atchley, R. (1995). Semantic and associative priming in high-dimensional semantic space. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the 17th annual conference of the Cognitive Science Society* (pp. 660–665). Hillsdale, NJ: Lawrence Erlbaum Associates.

MacCabe, A., & Peterson, C. (Eds.). (1991). *Developing narrative structure*. Mahwah, NJ: Lawrence Erlbaum Associates.

National Institute of Standards and Technology. (2002). *Automatic evaluation of machine translation quality using n-gram co-occurrence statistics.* Retrieved July 15, 2008, from http://www.nist.gov/speech/tests/mt/2008/doc/ngram-study.pdf

Page, E. B., & Peterson, N. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan*, *76*(7), 561–565.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In I. Pierre (Ed.), *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311–318). Morristown, NJ: Association for Computational Linguistics.

Passonneau, R. J., Goodkind, A., & Levy, E. T. (2007). Annotation of children's oral narrations: Modeling emergent narrative skills for computational applications. In D. Wilson & G. Sutcliffe (Eds.), *Proceedings of 20th annual meeting of the Florida Artificial Intelligence Research Society* (FLAIRS-20) (pp. 253–258). Menlo Park, CA: AAAI Press.

Plate, T. A. (2003). *Holographic reduced representation: Distributed representation for cognitive structures* (CSLI Lecture Notes no. 150). Stanford, CA: CSLI Publications.

Rohde, D. L. T., Gonnerman, L. M., & Plaut, D. C. (2005). *An improved model of semantic similarity based on semantic co-occurrence.* Retrieved June 25, 2006, from http://www.cnbc.cmu.edu/~plaut/papers/pdf/RohdeGonnermanPlautSUB-CogSci.COALS.pdf

Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Reading, MA: Addison-Wesley.

Schütze, H. (1993). Word space. In S. Hanson, J. Cowan, & C. Giles. (Eds.), *Advances in neural information processing systems 5* (pp. 895–902). San Mateo, CA: Morgan Kaufmann Publishers.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence, 46*(1–2), 159–216.

Wade-Stein, D., & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction, 22*(3), 333–362.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity* (Technical Rep. No. 17). Manoa, HI: Second Language Teaching and Research Center.

Notes

[1] This research was funded while the second author was a research postdoctoral fellow at ETS in Princeton, NJ.

[2] BLEU stands for bilingual evaluation understudy, reflecting the method's original use to evaluate the accuracy of machine translations.

[3] For documentation of the integrated writing task and studies examining its linguistic and psychometric properties, see Cumming et al., 2006; Cumming, Grant, Mulcahy-Ernt, & Powers, 2005; and Lee & Kantor, 2005.

[4] See Cline & Powers (2007) for details of the study from which this data was drawn.

[5] We examined the results of our study as reported here for the presence of correlations to population variables such as gender and educational level, but none of these variables had significant correlations with our measures of document similarity, so we do not discuss them further in this paper.

[6] Native speaker essays were initially scored with possible half-grades such as 2.5 reflecting disagreement between raters. Figure 1 displays the distribution that results when half-point scores were randomly distributed to adjacent scores. To calculate regressions and correlations, the originally values were used. For purpose of calculating relative score distributions in the section *Lexico-Grammatical Similarity*, scores were truncated, so that a score of 2.5 was treated equivalently to a 2. The resulting shift in scores toward a somewhat more normal distribution did not materially affect the trends reported in the section *Lexico-Grammatical Similarity*.

[7] If data is aggregated over a large training set as in *e-rater* (Burstein et al., 1998), CVA is capable of matching human ratings quite accurately; in particular, with the prompt we are examining here, a CVA feature based upon comparison with a 500-essay training set accounts for a very large percentage of the variance from human scores. We are concerned in this paper with comparisons with single essays rather than grouped sets. For practical purposes, such comparisons may be useful when a large training set is not available; in addition, we are trying specifically to measure how different populations (native and non-native) summarize a source text, so that multiple reference document methods are not directly applicable. We do,

however, examine how the methods primarily explored in this paper combine with aggregate CVA features in section 3.6.

[8] These statistical calculations were performed as ANOVA-style trend analyses using the standard statistical package, SPSS, originally called the Statistical Package for the Social Sciences.

[9] Strong correlations with document length have been observed for various cooccurrence-vector based methods of measuring similarity among documents (Derrick Higgins, personal communication, June 2007).

[10] We also examined repeated word sequences manually. Very few phrases more than three words in length were repeated in student essays, supporting the conclusion that students were not repeating material from the reading verbatim.

[11] Following the strength of the correlations in Table 6, the best regression model used similarity of lecture to student response and similarity of student response to the reading as the primary features, excluding the inverse similarities.