Syntagmatic, Paradigmatic, and Automatic N-gram Approaches to Assessing Essay Quality

Scott A. Crossley[1], Zhiqiang Cai[2], and Danielle S. McNamara[3]

[1]Department of Applied Linguistics/ESL, Georgia State University
[2]Institute for Intelligent Systems, The University of Memphis
[3]Department of Psychology, Arizona State University

scrossley@gsu.edu, zhiqiang.cai@gmail.com, dsmcnamara1@gmail.com

Abstract

Computational indices related to n-gram production were developed in order to assess the potential for n-gram indices to predict human scores of essay quality. A regression analyses was conducted on a corpus of 313 argumentative essays. The analyses demonstrated that a variety of n-gram indices were highly correlated to essay quality, but were also highly correlated to the number of words in the text (although many of the n-gram indices were stronger predictors of writing quality than the number of words in a text). A second regression analysis was conducted on a corpus of 88 argumentative essays that were controlled for text length differences. This analysis demonstrated that n-gram indices were still strong predictors of essay quality when text length was not a factor.

Writing Practice and Assessment

Writing is a critical skill related to academic and professional success (Kellogg & Raulerson, 2007). However, large-scale assessments often show that writing proficiently is difficult for many students (National Commission on Writing, NCW, 2003). One method from which to evaluate effective writing and better understand writing quality is to examine human judgments of writing quality (i.e., the scores assigned to a writing sample by trained readers). These scores can have important consequences to the writer. Such consequences are especially evident in the values attributed to writing samples used in student evaluations (i.e., class assignments) and high stakes testing such as the Scholastic Aptitude Test and the Graduate Record Examination.

Once a better understanding of writing quality is reached, opportunities for extended practice guided by individual feedback can be given to students in targeted areas. However, teachers are often limited in their opportunities to provide feedback on student writing due to limited time and large class sizes (National Commission on Writing, 2003). One solution has been the use of automated essay scoring (AES) systems. AES systems utilize sophisticated software to evaluate the structure, content, and overall quality of written samples (Shermis & Burstein, 2003). By automating portions of the grading and feedback process, students have more opportunities for writing practice, with fewer burdens placed on instructors.

Automated Essay Scoring

As noted above, AES systems allow students to practice writing and receive feedback, without adding to teachers' burdens (Dikli, 2006). Writing can be assessed via combinations of statistical modeling, natural language processing (NLP) tools, artificial intelligence (AI), machine learning techniques, and other similar methods.

Systems such as e-rater (Burstein, Chodorow, & Leacock, 2004), IntelliMetric (Rudner, Garcia, & Welch, 2006), and Writing-Pal (W-Pal, Dai, Raine, Roscoe, Cai, & McNamara, 2011) rely primarily on NLP and AI. In these systems, expert readers rate a corpus of essays to identify the overall quality of individual essays. Essays are then automatically analyzed along many linguistic dimensions, and statistically analyzed to extract features that discriminate between higher and lower-quality essays. Finally, weighted statistical models combine the features into algorithms that assign grades to essays.

For instance, a study by McNamara, Crossley, and McCarthy (2010) indicated that human judgments of essay quality are best predicted at the linguistic level by linguistic indices related to lexical sophistication (i.e., word frequency and lexical diversity) and syntactic complexity (i.e., the number of words before the main verb). These indices accurately classified 67% of essays as being either of low or high quality. Crossley, McNamara, Weston, and McLain-Sullivan (2011) examined differences

in the production of linguistic features as a function of grade level (ninth grade, 11th grade, and college freshmen). The purpose of this study was to examine linguistic differences in the writing samples of adolescents and young adult students at different grade levels to investigate whether writing styles changed in predictable patterns as writing proficiency develops. Crossley et al. found that nine indices (*number of words*, *number of paragraphs*, *word concreteness*, *CELEX word frequency for all words*, *incidence of positive logical connectives*, *lexical diversity*, *average number of modifiers per noun phrase*, *content word overlap*, and *average word polysemy*) distinguished essays based on grade level. These indices accurately classified 76% of essays into their appropriate grade level.

Such AES systems have successfully increased opportunities for student writing to receive quick and accurate feedback. However, AES systems lack humanist sensitivity and detection is limited by the algorithms available (Hearst, 2002). As a result, many AES systems do not capture all elements of a writers' style, voice, or other individual expressive differences. Thus, despite progress, automated scoring systems are still under development, with opportunities to expand in many areas. One such area that has not been thoroughly investigated is the production of multi-word units (i.e., n-grams). Such units are of interest because they provide both lexical and syntactic information about a text.

Automatically Assessing N-gram Production

N-grams are commonly used in computational linguistics to model language based on co-occurrences. N-grams have been used in a number of language models such as determining the probability of a sequence of words, speech recognition models, spelling correction, machine translation systems, and optical character recognizers.

From a lexical perspective, n-grams fall under the general term *phraseology*. Phraseology is the study of multi-word lexical units in both written and spoken language. Such units are easily accessible and form a crucial part of language fluency (Cowie 1998). The superordinate term phraseology includes other types of multi-word units such as formulaic sequences, prefabricated patterns, and collocations. Unlike these multiword units, n-grams refer to groups of two or more words that repeatedly appear in language as fixed items more frequently than expected by chance and much more frequently than phrasal verbs and idioms (Hyland, 2008). It is the frequency of these multi-word lexical units that provides meaning in word combinations (Sinclair, 1991). N-grams are frequent enough that they occur in over 80% of the words produced in spoken English with bigrams being the most common (Altenberg, 1998).

Knowledge of multi-word lexical units is an important component of communicative competence (Fontenelle, 1994) and the acquisition of n-grams is one of the most important types of lexical units that language learners acquire (Farghal & Obiedat 1995). The importance of n-grams in language acquisition and production is premised on the frequent reoccurrence of these forms and the difficulty of approximating the conceptual meaning of multi-word units as compared to single words (Nesselhauf & Tschichold 2002).

One reason for the centrality of n-gram knowledge as a mark of language acquisition is the notion that n-grams contain both the paradigmatic and syntagmatic features. Thus, the accurate production of n-grams requires both lexical and syntactic knowledge. For instance, the n-grams *there is* and *it is* are common in written language and require not only lexical knowledge of existential constructions, but also the syntactic knowledge to produce dummy subjects. Other n-grams common to writing include *will be*, which requires syntactic knowledge to formulate future time as well as lexical knowledge of non-verbal predicates and *that the*, which requires syntactic knowledge of clausal complements and lexical knowledge about the potential meaning or non-meaning of complementizers (Crossley & Louwerse, 2009).

The remainder of this paper reports on the development of automated indices of n-grams and the assessment of these indices to predict human ratings of essay quality using a corpus of argumentative essays. Our goal is to assess the extent to which automated indices of multiword units are predictive of human judgments of essay quality. In turn, we infer potential cognitive processes associated with assigning ratings to essays the basis of which are linguistic features contained in the text.

Method

Essay Corpus and Essay Scoring

We collected 313 essays written by 313 college freshmen at Mississippi State University (MSU). The essays were based on two SAT writing prompts that asked writers to discuss whether people should admire heroes or celebrities or whether originality is possible. The essays were timed (25 minutes) and no outside referencing was allowed.

Seven expert raters with at least 4 years of experience teaching freshman composition courses at MSU rated the quality of the 313 essays in the corpus. Two raters evaluated each essay based on a commonly used standardized SAT rubric. The rubric generated a holistic quality rating with a minimum score of 1 and a maximum of 6. Raters were informed that the distance between each score was equal. The raters were first trained to use the rubric with 20 similar essays taken from another corpus. A Pearson correlation for each essay evaluation was conducted between the raters. Once the raters reached a correlation of $r = .70$ ($p < .001$), the ratings were

considered reliable. After training, each rater scored a portion of the essays in the corpus such that all essays were scored by at least two raters. The final interrater reliability for all essays in the corpus was $r > .75$. We used the mean score between the raters as the final value for the essay's quality unless the differences between the 2 raters was >= 2, in which case a third expert rater adjudicated the score.

N-gram Indices and Analyses

We developed numerous computational indices to measure n-gram accuracy, n-gram frequency, and n-gram proportions. The purpose of these indices was to evaluate the n-gram production of writers and examine if n-gram production influenced human ratings of essay quality.

Representative corpus. Each of the n-gram indices we developed depended on the co-occurrence of words in the British National Corpus (BNC). We selected three sub-corpora from the BNC to be representative of n-gram occurrences in natural language. These three sub-corpora included a writing, general speaking, and spontaneous speech corpus. For the writing corpus, we selected all written samples from the BNC that were identified as academic, expository, legal, narrative, and newspaper text. In total, these samples comprised over 3000 texts and over 80 million words. For the general speaking sub-corpora, we selected samples from the BNC that were identified as broadcast, interview, lecture, meetings, professional, and spontaneous texts. In total, these samples comprised about 600 texts and totaled about ten million words. For our last sub-corpora, spontaneous speech, we selected only those texts that were identified as spontaneous speech samples. In total, this genre was comprised of about 150 texts that comprised almost four million words. For each of these sub-corpora, we calculated the incidence of bigrams and trigrams that occurred in each sub-corpora at a variety of ranges (e.g., the first 5,000 n-grams, the first 25,000 n-grams, the first 50,000 n-grams). These incidences were then normalized for the length of the sub-corpora. Using these normalized frequency counts from each subcorpora, we developed the indices below.

N-gram accuracy. We developed a set of algorithms to assess the n-gram accuracy of written text by comparing the normalized frequency of n-grams shared in both the reference sub-corpora and the written text of interest. The shared normalized frequencies for each n-gram were placed into two arrays and a correlation was then computed between the two arrays. The *r* value from the correlation was then reported. This correlation represents the similarity between the frequency of occurrences in a representative corpus and a sample text. As such, this index measures the n-gram accuracy of sample text under the expectation that the incidence of n-grams in the representative sub-corpus is typical of natural language use. We predict that higher rated essays will contain n-grams that occur at similar frequencies as in our representative sub-corpora.

N-gram frequency. We also developed a series of algorithms to assess the frequency of n-grams found in a sample text. For these algorithms, a list of all the bigrams and trigrams that occurred in a sample text along with their incidences were computed. Each n-gram was then assigned a frequency score taken from the representative sub-corpora (normed for the incidences of occurrences in the sample text). The frequency score for each n-gram in a text was thus computed and an average score was calculated for all the n-grams in the text. For these frequencies indices, we computed both a raw frequency score and a logarithmic frequency score. These scores represent the frequency of the bigrams in the sample text as reported in the representative sub-corpora. We predict that higher rated essays will contain less frequent n-grams in much the same way the higher rated essays contain less frequent words (McNamara et al., 2009; Crossley et al., 2011).

N-gram proportion. The last algorithms we developed were based on proportion scores. These algorithms report the proportion of n-grams in the sample text that are also found in the representative corpus. We predicted that higher rated essays will contain proportionally fewer n-grams in much the same way that higher quality essays contain a smaller proportion of frequent words.

Analyses. To assess relationships between the n-gram indices and essay quality, correlations were calculated between each index value reported for each individual text and the human ratings assigned to each text. Because these indices were developed to measure similar constructs (i.e., n-gram production), we assumed they would be highly inter-correlated. Issues of multi-collinearity were addressed by selecting only the index that demonstrated the highest correlation with essay quality for each group of indices (e.g., n-gram accuracy, n-gram frequency, n-gram proportions) from each subcorpora (e.g., spoken, spontaneous, and written). We also examined the selected indices for multicollinearity with each other.

Selected indices were entered into a multiple regression to predict the human ratings of essay quality. Because many of the indices were normalized for text length, we hypothesized that the indices would correlate strongly with text length, which is a strong predictor of essay quality (Crossley & McNamara, in press). We thus conducted a second regression analysis that controlled for the length of the essays in the corpus. This analysis allowed us to analyze the unique contributions that the n-gram indices made in explaining human judgments of essay quality beyond text length.

Results

Entire Corpus Analysis

Table 1 presents the n-gram indices from each category with the highest correlations with essay quality. We also include an index of text length. Essays that expert raters

scored as higher quality had fewer frequent bigrams (computed with both logarithmic transformations and raw counts) and lower proportion scores. Higher quality essays also produced shared n-grams at more similar frequency as found in the representative corpus than did lower rated essays. Large effect sizes (defined as $r > .50$) were reported for the three logarithmic frequency indices. A large effect size was also reported for the total number of words in the text. However, the correlation for the total number of words in the text was weaker than that reported by the logarithmic frequency indices.

Table 1
Correlations between indices and holistic essays score

Index: n-gram (size) algorithm (corpus)	*r*
Bigram (25,000) frequency logarithm (spontaneous)	-0.570**
Bigram (50,000) frequency logarithm (all spoken)	-0.567**
Bigram (50,000) frequency logarithm (written)	-0.547**
Total number of words	0.517**
Bigram (25,000) frequency (spontaneous)	-0.505**
Bigram (25,000) frequency (all spoken)	-0.483**
Trigram (50,000) frequency (written)	-0.383**
Bigram (25,000) proportion (all spoken)	-0.236**
Bigram (25,000) proportion (spontaneous)	-0.232**
Bigram (50,000) correlation (all spoken)	0.208**
Bigram (25,000) correlation (spontaneous)	0.200**
Bigram (25,000) correlation (written)	0.189**
Bigram (25,000) proportion (written)	-0.177*

** $p < .001$, * $p < .010$

Before our final selection of variables for the multiple regression, we ensured that no index pair correlated above r => .70 and that each variable passed tolerance tests (i.e., VIF and tolerance values). As predicted, Pearson correlations revealed that many of the indices were highly correlated with one another. In total, four variables did not show multicollinearity with one another. These variables were *bigram (25,000) frequency logarithm (spontaneous), trigram (50,000) frequency (written), bigram (25,000) proportion (all spoken),* and *bigram (50,000) correlation (all spoken)*. VIF and tolerance values were at an acceptable criterion for all these indices (around 1).

A linear regression (step-wise) was conducted including the four variables. Two variables were significant predictors: *bigram (25,000) frequency logarithm (spontaneous*), $t = -11.84$, $p < .001$ *and bigram (25,000) correlation (written)*, $t = 2.262$, $p < .01$. The overall model was significant, $F_{2, 310} = 50.398$, $p < .001$, $r = .580$, $r^2 = .3125$, indicating that the combination of the two variables accounted for 33% of the variance in the human ratings.

Corpus Analysis Controlled for Text Length

We next used the n-gram indices to analyze a corpus of scored essays that were controlled for text length. Text length is a concern for this analysis because many of the n-gram frequency indices we developed showed strong correlations with text length (see Table 2). The corpus used in this study was a subsection of the MSU essay corpus. We selected only the essays that had text lengths between 384 and 500 words. These essays provided us with the greatest number of texts (88) that did not demonstrate significant correlations between text length and the grades assigned by the raters.

Table 2
Correlations between n-gram indices and text length

Index: n-gram (size) algorithm (corpus)	*r*
Bigram (50,000) frequency logarithm (written)	-0.849**
Bigram (50,000) frequency logarithm (all spoken)	-0.826**
Bigram (25,000) frequency logarithm (spontaneous)	-0.790**
Bigram (25,000) frequency (all spoken)	-0.687**
Bigram (25,000) frequency (spontaneous)	-0.681**
Trigram (50,000) frequency (written)	-0.540**
Bigram (50,000) correlation (all spoken)	0.247**
Bigram (25,000) correlation (spontaneous)	0.215**
Bigram (25,000) correlation (written)	0.189*
Bigram (25,000) proportion (spontaneous)	0.089
Bigram (25,000) proportion (written)	0.082
Bigram (25,000) proportion (all spoken)	0.058

** $p < .001$, * $p < .010$

Table 3 (next page) presents the n-gram indices that correlated with human judgments of essay quality for the text length controlled corpus. Essays that expert raters scored as higher quality had lower proportion scores and fewer frequent bigrams (computed with both logarithmic transformations and raw counts). Indices of correlational accuracy were not significantly correlated with essay quality for this corpus. Medium effect sizes (defined as $r > .30$) were reported for all significant correlations.

Three variables did not demonstrate multicollinearity: Bigram (25,000) proportion (all spoken), Bigram (50,000) frequency logarithm (all spoken), and Trigram (50,000) frequency (written). VIF and tolerance values were at an acceptable criterion for all these indices (around 1).

A linear regression analysis was conducted including the three variables from the training set. Only one variable was a significant predictor of essay quality: Bigram (25,000) proportion (all spoken), $t = -3.797$, $p < .001$. The overall regression model was significant, $F_{1,87} = 14.416$, $p < .001$, $r = .377$, $r^2 = .142$, indicating that the one variable accounted for 14% of the variance in the essay ratings.

Table 3
Correlations for restricted range of text length

Index: n-gram (size) algorithm (corpus)	*r*
Bigram (25,000) proportion (all spoken)	-0.377**
Bigram (25,000) proportion (written)	-0.375**
Bigram (50,000) frequency logarithm (all spoken)	-0.372**
Bigram (25,000) frequency logarithm (spontaneous)	-0.360**
Bigram (25,000) proportion (spontaneous)	-0.359**
Trigram (50,000) frequency (written)	-0.331*
Bigram (50,000) frequency logarithm (written)	-0.196
Total number of words (text length)	0.190
Bigram (25,000) frequency (spontaneous)	-0.161
Bigram (25,000) frequency (all spoken)	-0.095
Bigram (25,000) correlation (written)	0.051
Bigram (25,000) correlation (spontaneous)	-0.041
Bigram (50,000) correlation (all spoken)	0.020

** $p < .001$, * $p < .010$

Discussion and Conclusion

The purpose of this study was to explore the extent to which automated indices of n-gram production could explain human ratings of essay quality. A variety of different n-gram indices were developed and tested against a corpus of scored essays. The results demonstrated that n-gram indices that measured bigram frequency and accuracy were strong predictors of human judgments of essay quality surpassing even text length. However, many of the indices developed were highly correlated with text length. For this reason, a second analysis that controlled for text length was conducted. This analysis demonstrated that the bigram index related to proportion was a significant predictor of human judgments of essay quality.

In the first analysis, a number of n-gram indices that measure the frequency of n-grams demonstrated strong correlations with human ratings of essay quality above that yielded by text length. These indices reported negative correlations with human quality scores indicating that essays that contain less frequent n-grams (in this case bigrams) were scored as higher quality. This analysis supports past findings demonstrating that human raters assign higher scores to essays that contain less frequent linguistic items (i.e., essays that are more linguistically sophisticated). Smaller effect sizes were reported for our indices of n-gram proportion scores and n-gram accuracy scores. The n-gram proportion scores were also negatively correlated with human judgments of essay quality demonstrating that essays that contained a lower proportion of n-grams found in our various subcorpora n-gram lists received higher scores. This finding also supports the notion that higher scored essays contain less frequent linguistic features. In contrast, our n-gram accuracy scores were positively correlated with human judgments of essay quality. These correlations indicate that essays that contain the n-grams in our n-gram lists are scored higher if those n-grams are produced at a frequency similar to those found in the representative corpora. A linear regression using frequency, accuracy, proportion scores and text length indices demonstrated that two indices related to n-gram frequency and n-gram accuracy explained 31% of the variance in the human scores. The predictive ability of the n-gram indices is quite high and, in some cases, greater than using a variety of different linguistic indices. For instance McNamara et al. (2009) used a number of Coh-Metrix indices on a similar corpus of human scored essays and were only able to predict 22% of the variance in human ratings using indices of lexical frequency, syntactic complexity, and lexical diversity.

One limitation to our first analysis was the multicollinearity displayed between many of the n-gram frequency indices and text length. All three of the frequency indices that used logarithmic transformations correlated with text length at > .70, while all the other frequency, proportion, and accuracy indices correlated below < .70. To assess the effect of these indices in the absence of text length constraints, we conducted a second analysis that examined only those essays that did not demonstrate significant correlations between the human scores and text length. In this analysis, proportion indices and frequency indices demonstrated significant correlations with human scores of essay quality (although lower than in the first analysis). As in the first analysis, these correlations were negative indicating that essays with less frequent n-grams received higher scores demonstrating that higher scored essays contain fewer commonly used n-grams and lower score essays contain more commonly used n-grams. N-gram accuracy indices did not demonstrate significant correlations with essay quality. A regression analysis showed that one index related to the proportion of bigrams in the essay was the best predictor of essay quality explaining 14% of the variance in human scores. This second analysis demonstrates that the n-gram indices we developed capture unique attribute of language beyond text length and that these attributes are important predictors of essay quality.

Overall, these analyses demonstrate that textual n-grams have an important effect on the judgments made by human raters. Human raters are more likely to judge a text as higher quality if it contains fewer frequent n-grams and a lower proportion of n-grams. Additionally, human raters are likely to score a text as higher quality if the n-grams that are produced occur at a similar frequency as found in the representative corpus. The effects seem to be greatest for bigrams as compared to trigrams and stronger for bigram indices developed from our spontaneous speech

and spoken subcorpora. That bigrams indices were generally more predictive than trigram indices is likely the consequence of sparse data problems in the trigram lists (i.e., there are fewer trigrams per essay than bigrams). Because bigrams are extremely frequent, they have proven effective in many computational applications to include text categorization (Peng, Schuurmans, & Wang, 2003; Jurafsky & Martin, 2000). We also hypothesize that n-gram lists developed from spoken corpora are stronger predictors because they better evince natural language exposure, which would not be reflective of written texts.

One likely reason that n-gram indices are significant predictors of essay quality is that they tap into both the paradigmatic and syntagmatic features of the text. Thus, our n-gram indices likely act as proxies for both the lexical sophistication and the syntactic complexity of a text. One problem with the reported indices is that they do not currently provide explicit information about which n-grams are shared, which n-grams are rare, and which n-grams demonstrate differences in frequency. Having access to such data might allow us to better understand the lexical and syntactic elements within the n-grams that explain the differences in the human ratings. Such information could be used to provide feedback to writers in the classroom or in intelligent tutoring systems like W-Pal.

Acknowledgments

This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES. The authors would like to thank Brad Campbell, Daniel White, Steve Chrestman, Michael Kardos, Becky Hagenston, LaToya Bogards, Ashley Leonard and Marty Price for scoring the corpus of essays.

References

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word combinations. In *Phraseology: Theory, Analysis and Applications*, Anthony Paul Cowie (ed.), pp. 101–122. Oxford: Oxford University Press.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing system. *AI Magazine, 25*, 27-36.

Cowie, A. (1998). Introduction. In *Phraseology: Theory, Analysis, and Applications*, Anthony Cowie (ed.), 1-20. Oxford: Oxford University Press.

Crossley, S. & Louwerse, M. (2007). Multi-dimensional register classification using bigrams. *International Journal of Corpus Linguistics* 12 (4), 453-478.

Crossley, S. A., McNamara, D. S., Weston, J., & McLain Sullivan, S. T. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communications, 28* (3), 282-311.

Crossley, S. A., & McNamara, D. S. (in press). Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. *Journal of Research in Reading.*

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

Dikli, S. (2007). An overview of automated essay scoring of essays. *Journal of Technology, Learning, and Assessment*, 5(1), Retrieved [November 17th, 2011] from http://www.jtla.org

Farghal, M., & Hussein O. (1995). Collocations: A neglected variable in EFL. International Review of Applied Linguistics in Language Teaching 33 (4), 315-331.

Fontenelle, T. (1994). What on the earth are collocations: An assessment of the ways in which certain words co-occur and others do not. *English Today* 10 (4), 42–48.

Hearst, M. (2002). The debate on automated essay grading. *IEEE Intelligent Systems*, *15*, 22-37.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes.* 27, 4–21.

Kellogg, R. & Raulerson, B. (2007). Improving the writing skills of college students. *Psychonomic Bulletin and Review, 14*, 237-242.

Manning, C. D. & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, Massachusetts: MIT Press.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). The linguistic features of quality writing. *Written Communication*, *27*, 57-86.

Nesselhauf, N., & Cornelia T. (2002). Collocations in CALL: An investigation of vocabulary-building software for EFL. *Computer Assisted Language Learning* 15 (3), 251-279.

National Commission on Writing. (2003). *The Neglected "R."* NY: College Entrance Examination Board.

Peng, F., Schuurmans, D. & Wang, S. (2003). Language and task independent text categorization with simple language models. In M. Hearst and M. Ostendorf (Eds.), *HLT-NAACL 2003: Main Proceedings*, pp. 189-196, Edmonton, Alberta.

Rudner, L., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric essay scoring system. *Journal of Technology, Learning, and Assessment*, 4(4). Retrieved [November 17th, 2011] from http://www.jtla.org.

Shermis, M. D. & Burstein, J. (2003). Automated Essay Scoring: A Cross Disciplinary Perspective. Mahwah, NJ: Lawrence Erlbaum Associates.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.