

Lexical richness and success in a standardized academic English writing test

Richard Smith
International University of Japan

Abstract

This article investigates the degree to which raters of essays written for a standardized English academic writing test take into account lexical richness when assigning a quality score. It utilizes a relatively new lexical richness measure, the Lexical Frequency Profile, which has been demonstrated to be stable across compositions on different topics written by the same author provided that there is some control over genre. While previous studies of lexical richness have mostly focused on free writing produced within particular EFL and ESL institutional settings by writers with low to high intermediate levels of linguistic proficiency, this study gathered its data from free writing produced within the public domain of a well-known standardized test, the Analytical Writing Assessment component of the GMAT, by writers whose linguistic proficiencies range from intermediate to native speaker levels. In this study, considerable support was found for Laufer and Nation's (1995) prediction that the Lexical Frequency Profile would be useful for helping to determine the factors that affect judgments of quality in writing. The study's main findings were that (a) at the lower to middle levels, the Profile's lexical richness statistics correlated significantly with holistic score ratings of writing quality and discriminated clearly among writers of differing language proficiency, and (b) at the highest levels of writing quality, these statistics suggested strongly that there exists a threshold level beyond which the trend for lexical sophistication to increase in tandem with writing quality is no longer self-evident.

Key Words: lexical richness, Lexical Frequency Profile

1. INTRODUCTION

There has been a general view that lexis plays a significant role in the creation and in the understanding of meaningful text and especially of meaningful written text (Grabe, 1985). There is evidence that within the professional English L2 writing instruction community, feedback on lexis is regarded as an important part of instructor feedback on L2 learner composition (Cohen & Cavalcanti, 1990). Furthermore, this view is shared in varying degrees by the L2 writers themselves (Leki & Carson, 1994) and by university professors who teach non-native speaking students (Vann, Meyer & Lorenz, 1984; Santos, 1988).

Since the early 1980's, the diffusion of computers and the digitization of text corpora have made it possible to subject this view about the contribution of lexis to written text quality to empirical tests, which typically take the form of analyses of lexical variance within and across texts. A few of these studies have focused on the relation between vocabulary choice in free writing and holistic rating scores of writing quality. An even smaller number have done so in English L2 contexts. These latter attempts to measure the relationship

between the lexical attributes of compositions, commonly referred to as the compositions' *lexical richness*, and holistic ratings of their quality have been inconclusive and even contradictory (Read, 2000: 208-209). While these outcomes perhaps indicate that this area of research will not easily yield tangible results, the smallness of the number of the major studies (three) and the heterogeneity of the conditions under which they were conducted suggest that there is still plenty of scope for investigation. The recent appearance of a new statistical measure of lexical richness in compositions, the Lexical Frequency Profile, which its developers (Laufer & Nation, 1995) have claimed has advantages over the previously used statistical measures, provides grounds for a renewed examination of the relationship between lexical richness in free writing and holistic score ratings of writing quality in an L2 context.

This study will try to build on previous studies in two ways. First, it will examine the three previous major studies on the relation between lexical richness and holistic score ratings of L2 writing quality in a search for ways to improve research design in order to minimize doubts about the reliability of its results. Second, it will look at these studies and at the broader domain of previous studies on L2 lexical richness in free written text and establish the framework within which the questions about the relationship between lexical richness and holistic score ratings of L2 writing quality are most likely to yield useful insights.

Following a review of the previous studies, there will be a discussion of the merits of the Lexical Frequency Profile as a lexical richness measure and of the GMAT Analytical Writing Assessment as a real-world test of writing quality at higher linguistic proficiency levels. This discussion will set the context for the presentation of this study's research aims and questions, the description of its data collection procedure and data treatment, and the presentation and discussion of its results.

2. BACKGROUND

2.1 Previous studies involving holistic score ratings of writing quality

The literature recognizes two major reasons for the inconclusiveness of the three previous major studies of the relationship between the lexical richness of freely written L2 English compositions and holistic ratings of their quality. The first reason is that the previous researchers have at best (Linnarud, 1986; Engber, 1995) identified only moderately significant correlations ($r = .43$ to $.57$) between one of their lexical statistics and holistic ratings of writing quality and at worst (Nihalani, 1981) found no significant relationships between any of their lexical statistics and writing quality. The findings of the first two

researchers need to be further qualified by the observation that, although they both used a similar battery of measures of lexical richness, the significant correlations with holistic ratings of writing quality involved different statistical measures: *lexical individuality* in the case of Linnarud's study and *lexical variation* and *lexical variation minus lexical error* in the case of Engber's study.

The second reason involves doubts about the reliability and validity of the statistical measures and of the data that have been generally used to produce lexical richness statistics. Laufer and Nation (1995) have argued exhaustively that these doubts rest on firm foundations. Firstly, they have shown that there are significant reliability and/or validity weaknesses embedded in each of the seven statistical measures of lexical richness which they found in previous studies. These seven measures include the *lexical individuality* and *lexical variation* measures referred to above. Secondly, they have pointed out that studies of lexical richness in writing need to control for factors other than vocabulary size which might influence the lexical attributes of the writing. These factors include the nature of the writing task and the writer's familiarity with the topic. Most of the attention in the literature has been directed towards the variances in lexical performance induced by the writing task type or genre (Read, 1991; Kroll & Reid, 1994). Reid (1990) has shown, for example, that among topic types found in the TOEFL Test Of Written English (TWE) the comparison/contrast type, which involves taking a position, tends to induce greater lexical density than the description/interpretation of a chart type of topic. This point needs to be taken into account when attempting to collate the results of Linnarud's study, which was based on the analysis of L2 compositions that were responses to pictorial story prompts, with the results of the studies by Nihalani and Engber, which were based on compositions written in response to explanatory topics.

An additional reason for a cautious attitude to the existing research findings relates to questions about the reliability of, or the bias in, the holistic rating scales used in the research. Nihalani (1981) and Linnarud (1986) instructed their composition raters to use an entirely impressionistic rating scale based on single word or two-word descriptors such as "very bad", "bad", "good" and "very good." In neither study did raters undergo any training in the application of these descriptors. The heterogeneity of the raters in Linnarud's study raises an additional reliability question. Engber's (1995) study is superior in most of these respects. It used a slightly adapted form of the six-level TWE rating scale, which has a comprehensive set of descriptors, and its raters were all ESL professionals who had received training in the use of the TWE scale or of an analogous rating scale. The one weakness of the procedure in

Engber's study is that the TWE rating scale includes a few explicit lexical descriptors which may positively bias raters' views of the contribution of lexis to the overall quality of compositions.

The second and third reasons for the existing studies' inconclusiveness highlight a number of design weaknesses which subsequent researchers in this area should try to avoid or minimize. These weaknesses result from problems with (1) the reliability and validity of the lexical richness measures, (2) the type and the consistency of the writing task(s) used to elicit samples of writing and (3) the quality of the rating scales in terms of their provision of clear reference standards for rater agreement on writing quality, the presence in the rating scale of bias towards treatment of lexical richness, and the quality of the raters in terms of their training and/or homogeneity. The first of these problem sets will be considered in the context of a review of the Lexical Frequency Profile (LFP). The second and third problem sets will be discussed in a review of the Analytical Writing Assessment, a standardized writing test which is currently a required component of the Graduate Management Admission Test.

2.2 Studies involving lexical richness in general

Although Read (2000: 208) points out that it is still hard to draw firm conclusions from any of the studies conducted into the lexical richness characteristics of L2 free writing, not all the existing conclusions are equally soft. Most of the studies have focused on subject populations in the intermediate level of English proficiency. This is probably not an accident as it is generally accepted that one of the defining features of the progression through the intermediate levels of proficiency is a "take-off" in the learner's vocabulary size. The studies by Engber (1995), Laufer & Nation (1995), Laufer (1994 & 1995) and Linnarud (1986) all gathered subject populations composed of intermediate learners of English and all of the studies in varying degrees identified variances in lexical richness which were related to variances in language proficiency or in writing quality. The Laufer and Nation study established that there existed a very significant difference between the variances in the LFP measures of the low intermediate subject group of EFL students and the variances in the LFP measures of the high intermediate subject groups.

2.3 A framework for research questions

These initial explorations of the relationship between lexical richness and writing quality point towards two important sets of research questions. The first set of questions concerns the relationship between lexical richness, as measured by the new instrument, the LFP, and

holistic score ratings of writing quality at the intermediate levels of proficiency. The demonstration by Laufer and Nation that there is a relationship between lexical richness and proficiency at these intermediate levels raises the question of why previous studies have failed to capture any truly convincing relationship between lexical richness and measured writing quality at these levels. Although Laufer and Nation have examined and applied the LFP in several ways and predicted in their 1995 study that the LFP would be useful for helping to determine the factors that affect judgments of quality in writing, neither they nor other researchers have yet applied the LFP to an examination of this relationship. The second set of questions should address the same relationship at the higher end of the proficiency spectrum. Ideally, a single study would address both sets of questions, but in practice it is difficult either to assemble a large and diverse enough subject population for such an omnibus study or to find a holistic rating instrument which is sensitive enough to discriminate finely among the several levels of a very wide spectrum of writing proficiencies. In the case of this study, the first constraint alone makes it necessary to focus on the second set of questions.

3. THE LEXICAL FREQUENCY PROFILE

In an attempt to provide a more reliable, valid and detailed alternative to the traditional statistical measures of lexical richness, Laufer and Nation (1995) have developed a new statistical measure, the Lexical Frequency Profile, and have produced arguments and evidence in favor of its reliability and validity. What follows is a summary of the features of the LFP which are salient for this study.

3.1 Differences between the LFP and other lexical richness measures

The LFP differs from the other statistical measures of lexical richness in five significant ways. First, whereas the other measures yield a single, global statistical value as a measurement of lexical richness in a composition, the LFP produces four statistical values for each composition.

Second, and most important, the ratios represented by the statistical values in the other measures are ratios which are calculated internally within each composition or within each cohort of compositions without reference to an external yardstick. The *lexical variation* measure, for example, expresses the ratio in percentage terms of the number of word types (the number of words in a text minus all repetitions and all instances of tense and plural

inflection) to the number of word tokens (the total number of running words) in a text. As Laufer and Nation pointed out in their 1995 study, these internally calculated lexical richness ratios are sensitive to factors other than the writer's vocabulary size, most notably composition length and composition topic. Unlike these internally calculated ratios, the four ratios which are generated by the LFP use as their base reference four vocabulary check lists which represent different frequency levels of vocabulary use among native speakers. These four frequency levels comprise the *1st thousand*, the one thousand most commonly used word families, the *2nd thousand*, the one thousand word families most commonly used after the most common one thousand, the *Academic Word List (AWL)* which comprises the 570 word families which are most commonly found across a wide range of academic texts and which are not represented in the first two levels, and, lastly, the *Not in lists* level, which includes all the word families not found in the other three levels. The LFP measure of any given composition is a measure of the percentages of the word families in the composition which are drawn from each of the four vocabulary levels. A typical result is the sequence, 75-8-10-7, which indicates that 75% of the word families in the composition are drawn from the *1st thousand* level, 8% of the word families are drawn from the *2nd thousand* level, and so on.

The third difference between the LFP and other statistical measures of lexical richness is that the LFP defines knowledge of *words* conservatively in terms of knowledge of entire word families. Laufer and Nation define a word family according to Bauer and Nation's (1993) level 3 on their scale of word knowledge classification, which defines a *word* as the base form plus all inflections and affixes such as *-able*, *-er*, *-ly*, and *un-*. The other statistical measures invariably define knowledge of a word to mean knowledge of the word type (the dictionary word plus tense and plural inflections), with the result that they treat, for example, *happy* and *happily* as separate words and give the same weight to them that they give to completely unrelated words. The LFP calculates its ratios only with reference to instances of words which are not direct derivatives of other words.

The fourth difference lies in the LFP's treatment of lexical error. Several studies, including Engber's 1995 study, have tried to measure lexical error in order to make adjustments to their raw lexical richness statistics. Read (2000: 204-205) has explained that attempts to measure lexical error are problematic in several ways. The most intractable problem of all is the near-impossibility of defining with any precision the boundary between grammatical error and lexical error when the boundaries between grammar and lexis themselves are seen as blurred (Sinclair, 1991). By contrast, Laufer and Nation have adopted a simpler approach to lexical error when making LFP calculations. They dispense with any

attempt to measure lexical error and in their studies have simply omitted any words which struck them as having a clearly incorrect use. They have ignored derivational errors in line with their focus on word families and have corrected minor spelling errors on the grounds that these do not indicate a lack of knowledge of word meaning or word use.

The fifth difference between the LFP and other measures of lexical richness is that Laufer and Nation recommend that all proper nouns should be deleted from the text samples which are analysed.

3.2 Reliability and validity of the LFP

These salient features of the LFP suggest that it should have some robustness and that its results should exhibit some relationship to the English proficiency of its subjects. In their 1995 study, Laufer and Nation presented empirical support for both of these propositions. They showed that LFP values were stable across the two compositions written in response to argumentative prompts by the sixty-five subjects in their study. This feature of the LFP is important for a study which uses data from a standardized writing test in which the writing prompts differ with each test administration. They also showed that the LFP values correlated significantly ($r = .6$ to $.8$) with the subjects' scores on the authors' own productive vocabulary knowledge test, the Active Levels Test. With the exception of the values for the 2nd thousand level, these LFP values also discriminated well among the three proficiency levels of the subjects in the study.

3.3 Current limitations of the LFP

While the LFP has several merits which indicate that it should be a useful tool for lexical richness studies, lexical richness researchers should be aware that it has four limitations. The first limitation is its relative newness. Only three studies (Laufer 1994; Laufer & Nation, 1995; Laufer, 1995) employing the LFP are easily accessible in the literature. None of the three published studies investigated the relationship between the LFP and holistic ratings of writing quality. Furthermore, none of these studies has investigated the issue of the LFP's stability with respect to composition length. In their three studies, Laufer and Nation either obtained compositions of equal length or selected passages of equal length from the compositions.

The second limitation is that the first two LFP levels, the 1st thousand and the 2nd thousand, have been compiled from West's (1953) now aging *General Service List* (Nation & Waring, 1997). There are now in existence several computerized corpora which have been

compiled on more scientific lines and from a more representative corpus of texts than was the case with the *General Service List*. The problem is that these new corpora are not readily accessible to the public and none of their custodians have yet released into the public domain any authoritative frequency lists. In the meantime, therefore, lexical researchers continue to rely on West's list (Read, 2000: 226-227). One of the peculiarities of this list is that, although it was compiled primarily according to *frequency* of use, a secondary criterion was *unavoidability* of use. The 2nd *thousand level* in particular contains a number of words, such as words for household items and for food, whose presence is not justified on the grounds of frequency alone. Many of these words are unlikely to find much representation in academic essays. This peculiarity helps to account for the surprisingly low 2nd *thousand* values which have so far appeared in the published studies. On the other hand, the third LFP level which comprises the *AWL* corpus of word families has been compiled recently according to modern lexicographical procedures (Coxhead, 1998).

The third limitation is that the LFP does not automatically discriminate between homonyms. In their 1995 study Laufer and Nation estimated that their 300 word token text samples contained a mean of between two and three homonyms each. This is a limitation shared by all lexical richness calculations which are not performed manually.

The fourth limitation is that in practice the four statistical values the LFP generates have not been equally significant or useful. The 1st *thousand* vocabulary contains most of the function words and the basic lexical items that any writer must use. The 2nd *thousand* values have proved to be small and rather insignificant. By default, therefore, the *AWL* and *Not in lists* vocabulary represents the more sophisticated lexical choices. In addition, the values generated by these two lists have tended to assume more significance and more power to discriminate when combined together. For these reasons, Laufer (1995) has argued the benefits of treating the combined *Beyond 2000* values as a basic measure of lexical sophistication.

4. THE ANALYTICAL WRITING ASSESSMENT

4.1 General features of the AWA

The Analytical Writing Assessment (AWA) is a relatively new standardized writing test which has been a required component of the Graduate Management Admission Test (GMAT) since October, 1994 (GMAC, 1999a). It is administered by the Graduate Management

Admissions Council (GMAC) on behalf of the Educational Testing Service (ETS), which directly administers the TOEFL. The AWA has two important features in common with the TWE. The AWA rating instrument is a six-point holistic scale which permits half-point graduations and the time allowance for a single AWA essay is thirty minutes. This congruity with a test as well-known to EFL/ESL specialists as the TWE is useful.

4.2 Merits of the AWA for a lexical richness study

The first merit of the AWA is its target candidate population. The designers of the GMAT have explicitly designed the AWA to measure the writing skills of educated native-speakers (GMAC, 2000a). Although the number of GMAT test-takers from countries where English is not the first language has increased throughout the 1990's, test-takers from countries where English is the dominant language still represent a large majority of all GMAT and AWA test-takers. The AWA can therefore lay claim to being a "real-world" test of written English skills which is drawing increasing non-native speaker participation.

Another merit is the AWA's strict control of writing task genre. The AWA requires test-takers to write two essays in one hour. By design, one essay topic is always an "analysis of an argument" topic and the other is always an "analysis of an issue" topic.¹

The third merit of the AWA is the care which its administrators take to ensure a reasonably high degree of scoring uniformity and inter-rater reliability. It provides raters with two parallel holistic rating instruments, one for rating the "analysis of an argument" essay and the other for rating the "analysis of an issue" essay (see Appendix ²). In terms of their design, these are reminiscent of the TWE rating scale, though the specific descriptors they contain are quite different. It should be noted that neither of these rating instruments contains any explicit lexical descriptors. The raters, who are college and university faculty members (GMAC 1999b), undergo training before each and every scoring session. Two raters are assigned to rate each essay, with a third rater assigned if the ratings of the first two raters differ by more than one point. Altogether, four raters are normally assigned to rate the two essays written by each test-taker (GMAC, 1999a).

5. THE STUDY

5.1 Aims

The study aims in three ways at complementing earlier studies of the relationship between lexical richness and free-response writing in English. Firstly, it aims to show whether the

Lexical Frequency Profile can play a useful role in identifying a relationship between lexical richness in free written production and holistic score ratings of the writing quality. Given its subject population and the nature of the writing instrument it employs, the study will confine this examination to the writing produced by writers who represent the upper end of the proficiency spectrum between the intermediate and native speaker levels.

Secondly, if the answer to the first question is positive, the study aims to discover whether the lexical richness measures of the LFP discriminate among writers at each level up the writing quality scale of the AWA or whether they fail to discriminate among the writers at certain levels. The study will also attempt to ascertain whether there exists a threshold level beyond which increments in lexical richness values cease to be associated with increments in ratings of writing quality. While Arnaud (1984), Linnarud (1986) and Waller (1993) showed that there are measurable lexical richness differences between L1 and perceived L1 writing on the one hand and L2 writing on the other hand, it remains an open question whether such perceptions of difference in authorship translate into perceptions of quality difference when these are regulated by a uniform and detailed rating scale.

Thirdly, the study will put these LFP results into some comparative perspective by adding two additional measures of lexical richness: essay word length and *lexical variance*. Apart from Grobe's (1981) study, little research has been conducted on the relationship between essay word length and holistic score ratings of the writing quality. However, it seems intuitively very likely that there will normally be a strong relationship between the two and that this relationship will be closer than those for most other discrete text variables. Grobe (1981) studied writing produced by young L1 writers and found, not surprisingly, that essay length correlated very strongly with ratings of writing quality. Engber's 1995 study used a *lexical variation* measure which was a type-token ratio adjusted for essay length and this showed a moderately significant correlation of .45 with essay quality. This study will use an almost identical instrument to calculate *lexical variation* which is adjusted for essay length. The only difference will be the use of 250 word segments instead of the 126 word segments used in Engber's study.

5.2 Research questions

- a. Will there be a significant relationship between the writers' LFPs and the holistic scores of their essays' writing quality as measured in the AWA component of the GMAT?
- b. If there is a significant relationship, will the relationship express itself in a consistent fashion across all score levels?

- c. Will two other lexical measures of writing quality, essay word length and adjusted type-token ratio, produce results comparable to or stronger than the results produced by the LFP measures?

5.3 Subjects

The subjects were fifty-two candidates for admission to an English-medium MBA School in Japan over the three-year period, 1997-2000. Forty-six of the candidates were admitted and entered the MBA School. The fifty-two subjects represent 20 nationalities: Japan (12), India (6), U.S.A. (6), Indonesia (4), China (3), Philippines (3), Bangladesh (2), Canada (2), Romania (2), Vietnam (2), Guatemala (1), Hong Kong (1), Kenya (1), Malaysia (1), Russia (1), Singapore (1), South Korea (1), Thailand (1), Uzbekistan (1), Western Samoa (1). This nationality distribution was an outcome of the data collection procedure.

5.4 Procedure

5.4.1 Data collection

Two sets of data were collected for this study. The first set of data comprises AWA scores and the copies of the essays which are the bases for these scores. As part of the MBA School's admission requirements, all candidates are required to allow the submission to the School of an official GMAT score report which includes a report of the AWA score and a complete copy of both of the essays written for the AWA.³ The author extracted a sample of AWA scores and essays from the enrolled student collection.⁴ The main constraint on the sampling was the need to gather a collection of essays whose scores were equally distributed across the GMAT rating scale from the minimum score of 0.5 to the maximum score of 6.0 (see Appendix). A second constraint on the sampling was the need to obtain some TOEFL proficiency data about the subjects.

At each half-point level within the 2.5 to 4.5 AWA score range there was no difficulty in identifying among the enrolled student collection at least ten subjects with AWA score reports and TOEFL scores. AWA score reports at the 5.0 level were also plentiful, but for various reasons valid TOEFL scores for these subjects were less plentiful. GMAT's policy of achieving a normal distribution of scores with a fat middle and very lean tails,⁵ also meant that it was difficult to find AWA score reports and essays below the 2.5 score level and above the 5.0 level. An additional problem was that essays rated below 2.0 were usually very short in terms of their word length. In view of these factors, it was decided to keep the sample limited to six subjects chosen at each half-point score level within the score band 2.0-5.5 and

to make up the shortfall at the 2.0 and 5.5 score levels by using a few AWA score reports and essays from outside the enrolled student collection. Six subjects were randomly identified for this purpose and all six granted use permission. In addition, two sets of usable score reports and essays from within the enrolled student collection were located at both the 1.5 and the 6.0 score levels and these were added to the data set.

The second set of data in this study comprises data about English proficiency level. TOEFL proficiency data was available for forty-four of the selected subjects. TOEFL data was obtained from thirty-four of the subjects within two months of the submission of the GMAT score reports. Twenty-six of these TOEFL scores were obtained from traditional paper-based TOEFL administrations and eight were obtained from the recently-introduced computer-based TOEFL administrations. The scores from these eight computer-based score reports were equated to paper-based TOEFL scores by means of an official TOEFL concordance table (ETS, 1998). Ten subjects were high proficiency students who took an Institutional TOEFL on arrival at the institution within six months of submitting GMAT score reports. All ten subjects obtained TOEFL scores above 600. The eight subjects for whom TOEFL scores were not available were all citizens of the U.S.A. and Canada.

5.4.2 Data processing

A peculiarity of the AWA is that it provides a single score rating for the two essays written by test-takers during a single test session. This peculiarity meant that the extraction of representative text samples required some care. In their 1995 study, Nation and Laufer extracted the first 300 word tokens from each essay in order to create text samples of equal length. No reason was given for this procedure, but it can be presumed that it was done to exclude the possibility that variance in composition length might have an effect on the LFP results. In this study the mean average length of the essay sets after deletion of errors and proper nouns was 591 word tokens. Forty-six of the essay sets permitted the extraction of the first 150 word tokens from both essays in the sets. Two of the essay sets were just above 300 word tokens in length and yielded 300 word tokens for analysis with a slightly asymmetrical extraction of 157/143 and 159/141 word tokens from the two essays in the set. Four of the essay sets had a combined length of fewer than 300 word tokens; they contained 240, 242, 260 and 271 word tokens respectively. An ad hoc analysis of the longer essay sets suggested that samples of 250 word tokens produce slightly lower *Beyond 2000* LFP values than the longer samples of 300 word tokens. Since these essay sets all belong to the group of essay sets with lower score ratings, it was decided to keep them in the data base in order to

maintain a balanced subject population.

In accordance with the procedure developed by Laufer and Nation and described above (see Section 3.1), incorrectly used words were omitted from the LFP analysis. The percentage of omissions in relation to the number of word tokens in the extracted samples was never higher than 2% and averaged 0.6%. This low percentage may reflect the relatively high proficiency levels of the writers. Proper nouns were also deleted from the samples. This proved to be an important procedure for the AWA text samples as the AWA writing prompts often include references to the names of imaginary companies, municipalities and products. On average, each essay set sample lost 3.2% of its word tokens to proper noun deletion. Wrong derivatives of words were not considered to be errors and minor spelling errors were corrected. These error and proper noun deletions were performed before counting the number of word tokens in the essays and before extracting the text samples for analysis. The other lexical richness measures, composition length and *lexical variation* adjusted for length, were applied to the full texts. For the *lexical variation* calculations, all derivational errors as well as use errors and proper nouns were omitted from the texts.

Each essay set sample was scanned into a computer text file and the text file was checked and manipulated manually. The samples were then individually analyzed for their LFP characteristics by a Windows compatible program called *VocabProfile* (Heatley, Hwang & Nation, no date). Each essay's adjusted-for-length *lexical variation* characteristics were analyzed by another Windows compatible program called *WordSmith, Version 3* (Scott, 1998).

6. RESULTS AND DISCUSSION

The first two research questions address the relationship between the LFP statistics and the holistic score ratings of the essay sets:

- a. Will there be a significant relationship between the writers' LFPs and the holistic scores of their essay writing quality as measured in the AWA component of the GMAT?
- b. If there is a significant relationship, will the relationship express itself in a consistent fashion across all score levels?

Table 1 presents the correlations between the LFP percentages of the essays and the holistic scores of the essays' writing quality.⁶ In addition to the four percentages calculated for the four LFP levels, a fifth percentage, *Beyond 2000*, which combines the *AWL* and *Not in lists*

Table 1 Correlations between the LFP Measures and Essay Score

| | 1 st 1,000 | 2 nd 1,000 | AWL | Not in lists | Beyond 2000 |
|----------------|-----------------------|-----------------------|--------|--------------|-------------|
| <i>r</i> | .73 | 0.03 | .63 | .62 | .78 |
| <i>p</i> value | <.0001 | .82 | <.0001 | <.0001 | <.0001 |

percentages, is shown. Table 2 shows the means of the same LFP measures for each AWA score level between 2.0 and 5.5. Each score level represents six of the subjects in the study. The means for the score levels 1.5 and 6.0 are not included because each level represents only two subjects. These results show there exist significant relationships between three of the four LFP measures of lexical richness in the essays and holistic score ratings of the essays.

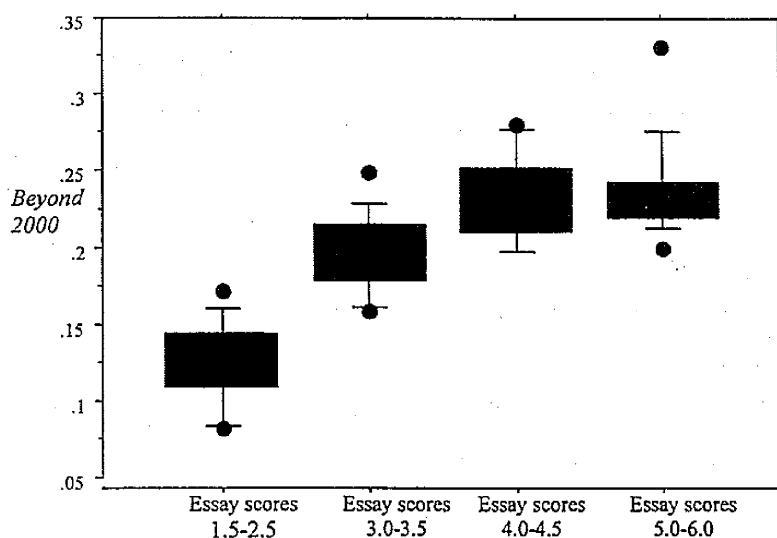
Table 2 Means of LFP Percentages by AWA Score

| Score | 1 st 1,000 | 2 nd 1,000 | AWL | Not in lists | Beyond 2000 |
|-------|-----------------------|-----------------------|------|--------------|-------------|
| 2.0 | 79.6 | 8.9 | 7.4 | 4.1 | 11.5 |
| 2.5 | 78.2 | 8.4 | 8.9 | 4.5 | 13.4 |
| 3.0 | 72.8 | 7.4 | 13.2 | 6.6 | 19.8 |
| 3.5 | 73.4 | 7.8 | 12.1 | 6.7 | 18.9 |
| 4.0 | 69.5 | 8.3 | 15.3 | 6.9 | 22.2 |
| 4.5 | 66.8 | 8.6 | 15.8 | 8.8 | 24.6 |
| 5.0 | 67.8 | 8.6 | 14.8 | 8.8 | 23.6 |
| 5.5 | 67.9 | 8.1 | 14.0 | 9.9 | 23.9 |

The strength of the relationship further increases when the *AWL* and *Not in lists* levels are combined into a single *Beyond 2000* level. Diagram 1 shows the box plots of the *Beyond 2000* value ranges for four sets of holistic score levels: 1.5-2.5 (14 subjects), 3.0-3.5 (12 subjects), 4.0-4.5 (12 subjects), 5.0-6.0 (14 subjects). The score levels are consolidated in this way in order to permit the box plots to exhibit patterns of central tendency and of variance in the value distributions. The plots reveal a clear and stable relationship up to the 4.5 score level between the percentage of sophisticated lexis in the essays and holistic ratings of their quality.

At first glance, the answer to the second research question is a little more complicated. Table 2 shows a secular trend from score level 2.0 to score level 4.5 for lexical choice sophistication to increase as the score level increases. As can be seen, the percentage of *Not in lists* word families in the essays rises in steady increments, while the percentages of both

Diagram 1 Box Plot of Variances in the *Beyond 2000* Values by Essay Score Range



the *AWL* and *Beyond 2000* word families in the essays jump considerably between the 2.5 and 3.0 score levels, but then decline slightly between the 3.0 and 3.5 score levels before moving upwards again at the 4.0 and 4.5 levels. However, this secular trend between the 2.0 and 4.5 score levels becomes an absolute trend if the divisions between the score levels are marked at whole integer, instead of half-point, boundaries. This suggests strongly that the general sophistication of the writers' word choices is a significant discriminator among the essays at the 2, 3 and 4 whole integer score levels. On the other hand, beyond the 4.5 score level this upward trend in the sophistication of lexical choice reverses itself to a slight extent. The only exception to this pattern of reversal is the clear and almost incremental increase in the mean percentages of *Not in lists* word families. One possible explanation for these conflicting trends in lexical sophistication at the highest score levels is that essays above the 4.5 score level are marked not only by a high *ratio* of sophisticated to basic lexical choices, but also by a greater *range* of sophisticated lexical choices than is found in essays at the adjacent lower levels. Another possible and related explanation, which is elaborated in the discussion of the results for research question c, is that at advanced levels of academic writing proficiency writers start giving more emphasis to precision than to experimentation in their lexical choices.

The third research question addresses the comparative significance of the LFP statistics:

- c. Will two other lexical measures of writing quality, essay word length and adjusted type-

token ratio (*lexical variation*), produce results comparable to or stronger than the results produced by the LFP measures?

The correlational results are: $r = .810$ for the essay length/score relationship and $r = .098$ for the adjusted type-token ratio/score relationship. The first result confirms the intuition that this relationship is normally strong. It also suggests that the *Beyond 2000* score relationship produced by this study lies near the upper bounds of the strength of relationship with writing quality we might expect of any lexical measure. The second result is surprisingly low, but it may reveal more about the divergence in the subject populations of this and Engber's 1995 study than about any divergence in the meaning of the studies' results.

Engber's 1995 study, which contained a subject population comprised entirely of learners at the intermediate and high intermediate proficiency levels, represents a useful reference point for a wider discussion of the answers to research questions a & b. Engber's study did not identify any upper quality level "plateau" effects for the values produced by the study's significant lexical richness measures: *lexical variation* and *lexical variation minus error*. Engber hypothesized that such a plateau would be identified if advanced level writers were included in the subject population because, unlike the intermediate writer who is actively experimenting with lexical choice, the advanced writer "may be retrieving items that meet precise specifications from an already adequate lexical base, resulting in less variety but perhaps relatively high quality writing" (151). Although the measurement instruments of this study are not directly comparable with Engber's measurement instruments, the results for research questions a & b lend some support to Engber's view.

Table 3 TOEFL Proficiency Data for the Study's 52 Subjects

| TOEFL Score Band | No. of Subjects |
|------------------|-----------------|
| Native Speaker | 8 |
| 620-677 | 12 |
| 600-617 | 9 |
| 580-597 | 6 |
| 550-577 | 8 |
| Below 550 | 9 |

The same hypothesis may also help to explain the low correlational value obtained for the relationship between essay score and the adjusted type-token ratio (one of the results for research question c), though other factors such as the nature of the AWA prompts and the

expected rhetorical form of the essays may also play a role. Engber noted that her study's subjects were at a stage of English language proficiency where their vocabulary growth had not peaked because they were still engaged in a lot of lexical experimentation, with greater levels of experimentation characteristic of the more proficient students. The present study's subject population, by contrast, spans the levels from intermediate to native speaker. Table 3 shows the TOEFL proficiency data for the forty-four non-native speakers in the study's total subject population of fifty-two. A TOEFL score of 600 represents the 86th percentile of all test-takers who took the paper-based TOEFL between July 1998 and June 1999 (ETS, 1999). If we treat it as the cut-off point that marks the start of "advanced" levels of proficiency, twenty-nine of this study's fifty-two subjects, a majority, belong to the advanced group.

6. CONCLUSIONS AND FUTURE RESEARCH

This study has presented evidence which suggests that the Lexical Frequency Profile can play a useful role in identifying a relationship between lexical richness in free written production and holistic score ratings of writing quality. Furthermore, the study shows that the Profile can identify such relationships even among writers of relatively high linguistic proficiency. At the same time, the data give signs that under the conditions of a standardized essay examination there are some upper limits to this relationship between lexical richness and writing quality. The study also adds incidental confirmation to the finding of Laufer and Nation's 1995 study that the word families which represent the second thousand of the two thousand most common word families do not add much to our understanding of the lexical richness characteristics of essays written in response to argumentative prompts.

Engber (1995) surmised that "... at the intermediate level, lexical variation is a productive strategy for expressing content, whereas linguistically advanced writers may rely more on precision" (150). If this is true, this study suggests that the LFP can reflect this precision at least as well as, and probably better than, other lexical richness measures. Future research could usefully focus on the relationship between the LFP and lexical variation among intermediate level writers as well as on the relationship between the LFP and writing quality at this level. Such research would add another piece to our understanding of the nexus between lexical development, language proficiency and writing success.

Notes

¹ The AWA test instructions and sample essay prompts can be found on pages 28 and 29 in the *GMAT Information Bulletin, 1999-2000*. They are available in downloadable pdf format as part of the document, *The GMAT Analytical Writing Assessment: An Introduction*, on the World Wide Web at: <http://www.gmac.com/publications/radC15C5.pdf> [checked January 23, 2001].

² Because of space constraints, only the rating scale for the "analysis of an issue" essay is reproduced in the appendix. The rating scale for the "analysis of an argument" essay is identical to this rating scale in structure and in most of its content. It differs only in respect of a few specific descriptors. Copies of both rating scales can be found on pages 28 and 29 in the *GMAT Information Bulletin, 1999-2000*. These rating scales are also available in downloadable pdf format as part of the document, *The GMAT Analytical Writing Assessment: An Introduction*, on the World Wide Web at: <http://www.gmac.com/publications/radC15C5.pdf> [checked January 23, 2001].

³ By October, 1997, all GMAT administrations worldwide had become computer-based (GMAC, 2000a). Almost all the GMAT score reports submitted to the host institution during the 1997-2000 period contained clean type-written essays. For this reason, the study targeted the 1997-2000 body of score reports.

⁴ Before drawing a sample of the score reports and their essays, the author contacted the entire MBA student body at the host institution during the 1997-2000 period for general permission to use their GMAT score reports for this study on condition that any published results would be anonymous. Only two students out of the 191 contacted refused permission.

⁵ The figures which follow show the distribution of scores among the AWA test population for the period April 1997 through March 2000. The first figure in each pair of figures shows the AWA score and the second figure shows the percentage of the examinees who scored below that AWA score:

6.0:97; 5.5:91; 5.0:80; 4.5:63; 4.0:43; 3.5:26; 3.0:13; 2.5:6; 2.0:2; 1.0-1.5:1; 0-0.5:0

Mean: 3.9; Standard deviation: 1.0

(Source: GMAC, 2000a: 13)

⁶ Following the practice of previous researchers when presenting similar results, the figures show the simple r correlation, not the r^2 value.

REFERENCES

- Arnaud, P. J. L. (1984). The lexical richness of L2 written productions and the validity of vocabulary tests. In T. Culhane, C. Klein-Braley & D. K. Stevenson (Eds.), *Practice and Problems in Language Testing* (pp. 14-28). Occasional Papers, No. 29. Department of Language and Linguistics, University of Essex.
- Bauer, L. & Nation, I. S. P. (1993). Word Families. *International Journal of Lexicography*, 6, 253-279.
- Cohen, A. D. & Cavalcanti, M. C. (1990). Feedback on compositions: teacher and student verbal reports. In B. Kroll (Ed.), *Second Language Writing: Research Insights for the Classroom* (pp. 155-177). Cambridge: Cambridge University Press.
- Coxhead, A. (1998). *An Academic Word List*. English Language Institute Occasional Publication, No. 18. Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington.
- Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, 4, 139-155.
- ETS (1998). *TOEFL Concordance Tables*. Princeton, N.J.: Educational Testing Service.
- ETS (1999). *TOEFL Test and Score Data Summary, 1999-00 Edition*. Princeton, N.J.: Educational Testing Service.
- GMAC (1999a). *The GMAT Analytical Writing Assessment: An Introduction*. Princeton, N.J.: Graduate Management Admission Council.
- GMAC (1999b). *GMAT Information Bulletin, 1999-2000*. Princeton, N.J.: Graduate Management Admission Council.
- GMAC (2000a). *Guide to the Use of GMAT Scores*. Princeton, N.J.: Graduate Management Admission Council.
- GMAC (2000b). *Profile of Graduate Management Admission Test Candidates: Five-Year Summary 1994-95—1998-99*. Princeton, N.J.: Graduate Management Admission Council.
- Grabe, W. (1985). Written discourse analysis. In R.B. Kaplan, A. d'Anglejan, J.R. Cowan, B. Kachru, G.R. Tucker & H. Widdowson (Eds.), *Annual review of applied linguistics* (Vol.5, pp. 101-123). New York: Cambridge University Press.
- Grobe, C. (1981). Syntactic maturity, mechanics, and vocabulary as predictors of quality ratings. *Research in the Teaching of English*, 15, 75-85.
- Heatley, A., Hwang, K. & Nation, P. (date not given). *VocabProfile* (Computer Program). Wellington, New Zealand: English Language Institute, Victoria University. Available: <http://www.vuw.ac.nz/lals/software.htm> [July 10, 2000].
- Kroll, B. & Reid, J. (1994). Guidelines for writing prompts: clarifications, caveats and cautions. *Journal of Second Language Writing*, 3, 231-255.

- Laufer, B. (1994). The lexical profile of second language writing: does it change over time? *RELC Journal*, 25, 21-33.
- Laufer, B. (1995). Beyond 2000: a measure of productive lexicon in a second language. In L. Eubank, L. Selinker & M. Sharwood (Eds.), *The Current State of Interlanguage*. Philadelphia: John Benjamins Publishing Company.
- Laufer, B. & Nation, P. (1995). Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics*, 19, 255-271.
- Leki, I. & Carson, J. (1994). Students' perceptions of EAP writing instruction and writing needs across the disciplines. *TESOL Quarterly*, 28, 81-101.
- Linnarud, M. (1986). *Lexis in Composition: A Performance Analysis of Swedish Learners' Written English*. Malmö: CWK Gleerup.
- Nation, P. & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: description, acquisition and pedagogy* (pp. 6-19). Cambridge: Cambridge University Press.
- Nihalani, N. K. (1981). The quest for the L2 index of development. *RELC Journal*, 12, 50-56.
- Read, J. (1991). The validity of writing test tasks. In S. Anivan (Ed.), *Current Developments in Language Testing*. Singapore: SEAMEO Regional Language Centre.
- Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.
- Reid, J. (1990). Responding to different topic types: a quantitative analysis from a contrastive rhetoric perspective. In B. Kroll (Ed.), *Second Language Writing: Research Insights for the Classroom* (pp. 191-210). Cambridge: Cambridge University Press.
- Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly*, 22, 69-90.
- Scott, M. (1998). *WordSmith Version 3*. (Computer Program). Oxford: Oxford University Press.
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Vann, R., Meyer, D. & Lorenz, F. (1984). Error Gravity: A Study of Faculty Opinion of ESL Errors. *TESOL Quarterly*, 18, 427-440.
- Waller, T. (1993). Characteristics of near-native proficiency in writing. In H. Ringbom (Ed.), *Near-Native Proficiency in English* (pp. 183-293). Åbo, Finland: English Department, Åbo Akademi University.
- West, M. (1953). *A General Service List of English Words*. London: Longman.

Appendix (from GMAC, 1999b: 28)

Analytical Writing Assessment “Analysis of an Issue” Rating Scale

SCORE

6 OUTSTANDING

A 6 paper presents a cogent, well-articulated analysis of the complexities of the issue and demonstrates mastery of the elements of effective writing.

A typical paper in this category

- explores ideas and develops a position on the issue with insightful reasons and/or persuasive examples
- is clearly well organized
- demonstrates superior control of language, including diction and syntactic variety
- demonstrates superior facility with the conventions (grammar, usage, and mechanics) of standard written English but may have minor flaws

5 STRONG

A 5 paper presents a well-developed analysis of the complexities of the issue and demonstrates a strong control of the elements of effective writing. A typical paper in this category

- develops a position on the issue with well-chosen reasons and/or examples
- is generally well organized
- demonstrates clear control of language, including diction and syntactic variety
- demonstrates facility with the conventions of standard written English but may have minor flaws

4 ADEQUATE

A 4 paper presents a competent analysis of the issue and demonstrates adequate control of the elements of writing.

A typical paper in this category

- develops a position on the issue with relevant reasons and/or examples
- is adequately organized
- demonstrates adequate control of language, including diction and syntax, but may lack syntactic variety
- displays control of the conventions of standard written English but may have some flaws

3 LIMITED

A 3 paper demonstrates some competence in its analysis of the issue and in its control of the elements of writing but is clearly flawed. A typical paper in this category exhibits *one or more* of the following characteristics:

- is vague or limited in developing a position on the issue
- is poorly organized
- is weak in the use of relevant reasons or examples
- uses language imprecisely and/or lacks sentence variety
- contains occasional major errors or frequent minor errors in grammar, usage, and mechanics

2 SERIOUSLY FLAWED

A 2 paper demonstrates serious weaknesses in analytical writing skills. A typical paper in this category exhibits *one or more* of the following characteristics:

- is unclear or seriously limited in presenting or developing a position on the issue
- is disorganized
- provides few, if any, relevant reasons or examples
- has serious and frequent problems in the use of language and in sentence structure
- contains numerous errors in grammar, usage, or mechanics that interfere with meaning

1 FUNDAMENTALLY DEFICIENT

A 1 paper demonstrates fundamental deficiencies in analytical writing skills. A typical paper in this category exhibits *one or more* of the following characteristics:

- provides little evidence of the ability to develop or organize a coherent response to the topic
- has severe and persistent errors in language and sentence structure
- contains a pervasive pattern of errors in grammar, usage, and mechanics which severely interferes with meaning

0 Off-topic, in a foreign language, merely attempts to copy the topic, or consists only of keystroke characters