



Researching Simplicity and Sophistication in Student Writing

RUTH BREEZE*

Universidad de Navarra

ABSTRACT

The programs WordSmith and VocabProfile were used to research lexical differences between essays written in English by Spanish undergraduates and a set of essays independently judged as being of TWE grade 6 standard. The results indicated that writing by this group of students was generally characterised by low lexical variation, a preponderance of high-frequency words and under-use of academic vocabulary in comparison with the target style. Reasons for the apparent lexical simplicity of this sample of student writing are discussed.

KEYWORDS: WordSmith, VocabProfile, L2 writing, lexical variation, lexical range, academic lexis.

**Address for correspondence:* Ruth Breeze. Instituto de Idiomas. Universidad de Navarra. C/Irunlarrea 1. 31080 Pamplona. Tel. 34 948 279172. Fax. 34 948 425742; e-mail: rbreeze@unav.es

I. INTRODUCTION

At university level, the issue of written English style, which often poses special problems for non-native students, takes on considerable importance. Until this stage, it has usually been enough for learners to paraphrase, to simplify, to get the message across. At university, however, the objective is to be exact, to be sophisticated, to express complex ideas in complex sentences, to master the techniques of written cohesion rather than to repeat the same basic words, and to cultivate a high, academic register in both vocabulary and syntax. Since most L2 users entering European English-medium universities have learnt English through standard EFL methodology which places a premium on the language of everyday communication, it would hardly be surprising if their writing were to fall short of this ideal, resembling spoken language and containing more elements of “basic” English than would be deemed appropriate by university teachers (Reid, 1993; Shaw & Liu, 1998; Read, 2000). Vocabulary, in particular, has been identified as a special area of difficulty in L2 academic writing (Leki & Carson, 1994; Muncie, 2002).

In view of this situation, the techniques of corpus linguistics offer the possibility of gaining deeper knowledge of the ways in which L2 writing may be at variance with the target style, providing a tool for quantifying particular features of L2 writers' texts that diverge from reader expectations. Although, as Hinkel points out (2003: 275), “research has not established with certainty what specific lexical and syntactic features, when taken together, can create an impression of a seemingly simplistic or reasonably sophisticated text in written L2 discourse”, it seems likely that choice and variety of vocabulary play a significant role in achieving an appropriate academic style. The present study uses the programs WordSmith and VocabProfile to explore the differences between European L2 writers' texts and the target style in lexical variation and range.

II. METHOD

Researchers have applied a wide range of different measures and criteria to texts written by L2 writers (Polio, 2001). In particular, contrastive studies have brought into focus the mismatch between the target language and L2 writers' productions (Ringbom, 1998; Granger, 1998). As the purpose of the present paper was to obtain information about the way in which students' essays differed from the target style in lexical terms, it was decided that quantitative measures developed by previous authors (Scott & Tribble, 2006) should be used, complemented where appropriate by qualitative analyses.

For the purposes of the present study, a comparison was established between a set of 30 texts written by undergraduates at a Spanish university and a second group of texts that could be taken to represent the target style. This control group consisted of a set of 18 essays from the websites www.wayabroad.com/twe/ and www.testmagic.com.¹ All of these essays had been posted to the websites and independently judged by their editing service as being

worthy of the maximum score on the Test of Written English (TWE). The essays all belonged to the same genre, that of the argumentative essay, in which a statement is presented and the writer is asked whether or not he/she agrees or disagrees, and why. Only essays on general social or business topics were selected, while essays on scientific or technical topics were excluded.

All 30 members of the student group had answered the question: *Businesses should do everything they can to make money. Do you agree?* This question was taken from the list of possible TWE questions for 2003, which was published on the ETS-TOEFL website. The 18 control-group essays were all argumentative in approach, written as answers to an agree/disagree question of this kind, but were slightly more diverse in their subject matter: unfortunately no large corpus of top-grade TWE essays on a single subject could be located at that date. For the purposes of the present study, it is necessary to assume that the samples of language analysed are representative of the way those writers approach the argumentative essay, in terms of register and discourse, and leave aside the possibility that the actual topic of the essay might influence the range of words used, so that essays on business might naturally contain less varied lexis than essays on, say, education or housing. Despite this evident shortcoming, this control group was felt to be sufficiently representative of the target style for comparative purposes.

The main focus of the present study is quantitative, applying various measures that may provide a key to the sophistication or quality of the texts in question. However, where appropriate, the findings of the quantitative studies are illustrated by examples from the texts in question. The purpose of this is to triangulate the data, and to show how the figures point to style-related phenomena which affect the way the text may be evaluated.

III. LEXICAL SIMPLICITY AND SOPHISTICATION

The appropriate use of sophisticated lexis and evidence of a wide range of vocabulary are features that are highly prized on qualitative writing assessment scales. It is highly likely that L2 writers' texts have shortcomings in this area, as their previous instruction has not equipped them with a rich vocabulary. As spoken language has been placed in the foreground, they have probably not been widely exposed to formal written registers, and although spoken language is arguably not "simpler" than written language, some features of it may convey an impression of simplicity when written down. In short, we may surmise that student L2 writers' vocabulary is probably limited in both range (having a lack of synonyms or precise terms) and register (being informal rather than formal and academic).

The crucial question for the design of the present research was that of how the concept of lexical simplicity versus sophistication can be usefully operationalised in the context of L2 writers' texts. The decision was made to investigate the lexical variation present in both sets of texts in terms of the type-to-token ratio, obtained using WordSmith; and to research the lexical range of these texts (percentage of words from first and second thousand most frequent word families in English) and their academic lexical content by

means of VocabProfile. The aim was to compare the language of the two samples, and to determine precisely how the students' essays differed from the target language.

In what follows, brief descriptions are given of the actual measures used for lexical variation, lexical range and academic lexis. The results obtained by comparing the students' essays with the TWE essays are reported and discussed.

III.1. Lexical variation

The term "lexical variation" means the amount of variety among the words used in a particular text. This notion is a useful one, because it is evident that some texts are highly repetitive, reusing the same vocabulary several times, whereas others make use of a more varied range of words. Over the years, applied linguists have used various formulae for calculating lexical variation. The program WordSmith used here employs a formula based on the ratio between the number of word types, that is, *different* words, by the number of tokens, that is, *all* words, calculated as follows:

$$\text{Lexical variation} = \frac{\text{No. of word types}}{\text{No. of word tokens}} \times 100$$

A low index of lexical variation shows that a text has a large number of repeated words, whereas a high index of lexical variation indicates that there is more variety among the vocabulary used, either because the text ranges over a wider variety of subjects, or because the writer has made an effort to use synonymous terms to avoid repetition (Martín, 2003: 166).

One particularly important methodological point to bear in mind when using the index of lexical variation for comparative studies is that there is an in-built bias in the measure itself, because longer texts inevitably repeat more high-frequency running words – prepositions, auxiliaries, pronouns – than shorter texts do. The basic rule is that the shorter the text is, the higher the index of lexical variation. As a general rule, it is often stated that a 1,000-word essay may well have a type-to-token ratio of around 40%, whereas a 100,000-word corpus of essays is likely to have a ratio of about 10% (Meunier, 1998: 32). To avoid this problem researchers generally either use statistics such as WordSmith's STT ratio, or try to ensure that the texts being compared are of similar length (Engber, 1995). In the present study, the latter principle was applied, and comparable samples were prepared by taking the first 180 words of each essay. Indices of lexical variation were then calculated for both sets of essays.

III.2. Lexical variation: results of comparative analysis

The results for the two groups are shown in Table 1.

	Students' essays mean	Students' essays SD	TWE essays mean	TWE essays SD
Type-to-token ratio	58.22	4.83	61.23	3.6
1st 1,000 word families (types)	80.827	5.115	73.283	7.124
2nd 1,000 word families (types)	6.03	2.284	7.072	2.265
Academic words (types)	8.56	3.69	9.49	4.57

Table 1. Comparison between students' essays and TWE essays

Table 1 shows that the mean type-to-token ratio in the texts by the students in this study was lower than the mean in the TWE essays. On the basis of Student's T-test for independent samples, the difference between the two groups was found to be statistically significant ($p < 0.05$). It was noticeable that both the highest and the lowest type-to-token ratios were found in the student group. Student 14, whose text had an abnormally high ratio (70.56), wrote an extremely idiosyncratic essay which read like a collage of disjointed phrases from another text. Student 7's text, which was qualitatively evaluated as one of the poorest essays, had the lowest ratio (50.56).

III.3. Lexical variation: interpretation of results

These results suggest that one shortcoming of the students' texts is that they are lexically less varied. Several reasons may be put forward to account for this phenomenon. One explanation is that the students either did not have, or did not use, a wide vocabulary in English. Moreover, even if students are familiar with a wide range of vocabulary, it has often been noted that student writing is repetitive and lexically unadventurous, probably because L2 writers often prefer to stick to the tried and tested rather than venture into new territory. Whereas good writers make an effort to find synonyms rather than repeat the same words, less proficient writers tend to be satisfied when communication is achieved, and are less concerned with questions of style. A further issue bound up with this question of lexical repetition is the nature of cohesion in L2 writers' texts. It is reasonable to suppose that less proficient L2 writers such as the students in this study are achieving cohesion by repeating

key items of vocabulary, rather than by using a variety of pronouns or synonyms, or by deploying a wider range of syntactic structures (Reynolds, 2001).

Another interpretation of the lack of lexical variation in these L2 writers' texts is that the cognitive overload caused by the complex task of writing a coherent essay in an L2 renders it impossible for the L2 writer to muster up a wide and varied vocabulary. L2 writers are often not sufficiently in control of the language to produce grammatically correct sentences, even though their knowledge of grammar is sound. Similarly, in the lexical area, although they have a wide vocabulary, they are not always able to call it up at the right time, and may not even be aware of the impression they are making in their writing. Principles such as not repeating the same word, which learners are usually aware of when writing in their L1, are not necessarily transferred to the task of writing in an L2.

Finally, despite the relevance of lexical variation in achieving an appropriate style and making a positive effect on readers (Engber, 1995), it is nevertheless true that a lexically varied essay is not necessarily a good one (Meunier, 1998). For in-depth assessment of L2 writers' vocabulary, factors such as appropriate word choice and collocation must be taken into account, which fall beyond the scope of what can be achieved by applying measures of lexical variation.

III.4. Lexical range

It is arguable that indices of lexical variation tell us about the amount of repetition in a text, but say nothing about the quality of the words used, or the relative frequency of simple and more complex words (Laufer & Nation, 1995). Another approach to the issue of lexical simplicity is to examine the level of the vocabulary used by L2 writers in terms of the relative frequency or rarity of that vocabulary in general corpora of English.

The first and second thousand word families in English were established by West in his General Service List (1953). This list contains word families, also sometimes described as headwords or lemmas, and so the actual number of word types (individual forms of the word in question) is much larger, possibly approaching 8,000. By matching L2 writers' texts to the lists of the 1,000 and 2,000 most frequent word families in English, it is possible to gauge the range of the vocabulary used: how many of the tokens, types or word families in the text belong to the 1,000 or 2,000 commonest word families in English. Students with an advanced level of English are generally expected to have acquired a minimum active English vocabulary of 2,000 word families, along with a much larger receptive vocabulary (words understood but not used) (Huntley, 1999; Valcourt & Wells, 1999). Studies of vocabulary reported by Nation (1990) have shown that a basic 2,000-word vocabulary of high-frequency items actually comprises 87% of words in most academic texts, which suggests that mastery of these elements is a crucial first step on the road to becoming an effective writer. The widespread criticism of many L2 writers' texts as being "simple" might be quantifiable in terms of a greater prevalence of words from, say, the first 1,000 most

frequent words in English, and a lower incidence of words from outside this category, than would be found in similar texts by NS or by proficient L2 writers.

In the present study, the VocabProfile program devised and made available by Nation, which operationalises West's General Service List (1953), was used to determine the percentage of words in each text which belonged to the first or second thousand commonest word families in English. It was decided that for the purposes of the present study, it would be most appropriate to consider the percentage of word types belonging to the different levels of frequency, rather than the percentage of word tokens, since the latter would tend to produce an exaggerated result because of the relatively repetitive nature of the vocabulary of many of these texts. The percentage of types was felt to be a more accurate index of the level or richness of the vocabulary in these texts. If a text has a large percentage of word types belonging to the thousand most frequent word families in English, it would tend to indicate that the vocabulary of that text is simple, basic and unadventurous. It might suggest that the writer has a poor vocabulary, or that he/she is transferring habits from everyday spoken language into writing. On the basis of the above, we hypothesised that well-written essays such as those in the TWE group might contain a smaller percentage of word types belonging to the commonest thousand word families than the L2 students essays, and a correspondingly larger percentage of words from the second thousand word families and beyond.

III.5. Lexical range: results of comparative analysis

The percentage of word types belonging to the first and second thousand commonest word families in English was calculated using Nation's VocabProfile program for each of the 180-word samples of 30 students' essays and 18 TWE grade-6 essays. Table 1 shows the results of this comparative study.

Regarding the percentage of word types belonging to the first thousand word families in English, the difference between the two groups was found to be statistically significant when Student's T-test for independent samples was applied ($p < 0.05$). A significantly greater percentage of the word types in the students' essays belonged to the first thousand word families in English than was the case for the TWE essays (80.83% compared to 73.28%). We can regard this as indicative that the students' essays had a poorer or more basic vocabulary than the TWE essays.

As far as the second thousand word families in English were concerned, the difference in means did not attain statistical significance (6.03% compared with 7.07%). On the whole, the student texts had a slightly smaller percentage of word types from base two. None the less, the text with the highest percentage of word types from base two was that of student 28 (11%), while that with the lowest was TWE essay 6 (2.9%).

III.6. Lexical range: interpretation of results

On the basis of the data concerning the first thousand words, it seems clear that the student essays tended to have a higher percentage of words from this very basic list, and accordingly a lower percentage of words from ranges beyond this. One conclusion that can be drawn from this is that the vocabulary in these students' essays tended to be rather simple and lacking in range. These data are consistent with the findings of some other authors (Read, 2005) which indicate that a higher percentage of high-frequency words reflects a lower level of performance.

On the whole, the findings concerning the second thousand word types are somewhat inconclusive. The texts contained many lexical items which belonged neither to the first nor the second thousand word families in English. Some texts contained far more words that did not match with any list, than words that matched with the list for base two. Thus it would be meaningless to say that a lower incidence of words from base two indicates a poorer vocabulary, because it might merely indicate that the student is using a range of words that is far more sophisticated, or perhaps more specialized, than those contained in Nation's second thousand.

In general, the percentage of word types from the first thousand word families could be seen to function as an index of the lexical range of these L2 texts. The students' texts had a higher percentage of words belonging to the first thousand word families than the highly-rated TWE essays did. In quantitative terms, it is possible to confirm that these students' texts made greater use of the simple "core" vocabulary of English: by the objective criterion of the vocabulary profile test, the students' texts were "simpler" than the TWE essays. Most of the highly-rated TWE texts had a lower incidence of "common" words and therefore proportionally more words that came from beyond the central core of English. This test thus goes some way to explaining why the students' texts might be judged "simple" and their TWE counterparts "sophisticated". None the less, it should be remembered that the quantitative measures used are crude, in that they consist of word counts, and no account is taken of the appropriateness or accuracy of the vocabulary that features in the texts. A text with a higher proportion of non-core vocabulary might not be written in better English than a text mainly consisting of words from the core lexicon. As Polio (2001: 100) points out, "risk-takers who use advanced words incorrectly" place the quality of their essay in jeopardy, even though they score higher on a frequency profile.

On the other hand, the results obtained using VocabProfile also tell us little about the nature of the core vocabulary used: simple percentages cannot reveal any varying patterns in the ways the students employed that basic core vocabulary, nor can they tell us about phenomena such as repetition of key words or overuse of simple verbs, which also contribute to the perceived degree of simplicity of a text.

In sum, although this exploration of the lexical range of these sets of texts provides clues to one of the main problems besetting these L2 writers' texts, it also leaves us with various unanswered questions regarding the nature of the vocabulary in these texts.

III.7. Academic lexis

As we have already seen, one of the fundamental features which makes some texts seem sophisticated and others simple is the presence or absence of a wide range of vocabulary. Too great a frequency of very basic vocabulary can make a text appear over-simple, while use of words from beyond the first thousand word families tends to make a favourable impression. However, as we have also seen in the case of the second thousand word families, as we advance beyond the basic thousand-word core it becomes increasingly difficult to draw conclusions about what would be expected.

One approach to this problem is based on the notion of the language domain (sphere of action or area of concern). Any person may have to use language in a variety of different domains, which might include personal (family or social) domains, public and professional domains, and educational domains. The genre of the argumentative essay which is under consideration here belongs to the educational/academic language domain, one which has been under scrutiny from linguists for many years. The existence of specifically "academic" vocabulary was proposed by Xue and Nation (1984), who proposed that an academic wordlist exists consisting of words beyond the basic core of English that are particularly common in academic prose. The 836 word families from Xue and Nation's University Word List (1984) are words which are particularly common in academic as opposed to general texts, because they belong to the language of research, analysis and evaluation. The list thus concentrates on what has sometimes been termed "subtechnical" vocabulary (Miller, 2001). According to Nation (1990), words from the category "academic vocabulary" account for approximately 8% of the running words in any academic text, irrespective of the discipline to which it belongs.³

Recently, an improved academic wordlist has been developed by Coxhead (2000), who took issue with the methodology used to draw up the Xue and Nation list (1984). This list was based on the findings of various studies which had used only small corpora focusing on a rather limited range of academic topics. To improve on this, Coxhead assembled a corpus of around 3.5 million words over a broad range of academic subjects (arts, commerce, law and science each provided approximately one quarter of texts used in the corpus), and applied a criterion of generality across topic, as well as criteria of frequency and non-core status, to select the components of her wordlist. The resulting list contains only 570 word families, which Coxhead found to account for 10% of the total word tokens in her academic corpus. Over 94% of the words in the list were found to occur in 20 or more of the 28 subject sub-areas of the corpus. The Coxhead academic wordlist is now operationalised in the latest version of Nation's VocabProfile program, used here.

The frequency of particular academic words such as those belonging to these academic wordlists makes them important markers of academic or formal-discursive register, and it has been stated that knowledge of this type of vocabulary is an important factor in achieving high scores on the TWE test (Huntley, 1999). Argumentative essays such as those in this study are almost certainly expected to contain a relatively high percentage of words from this academic category. In previous studies, it was found that students whose

writing contained native-like percentages of academic words were more likely to achieve academic success, and that students' writing could be improved efficiently by intensive practice using words from the academic word lists (Huntley, 1999). However, one particular point concerning the sample of students in the present study should be mentioned, which places a question mark over the relevance of these assertions in the present context. A large proportion of academic words derive from Latin, and Spanish-speaking students may have a tendency to use vocabulary of this kind (Sánchez-Hernández & Pérez-Paredes, 2005: 210). In this case, the results of the academic word count would be higher than the students' general level of proficiency might suggest. If anything, we could surmise that these students' English might contain too many words of Latin origin, possibly with long Latinate terms where NS would prefer something that sounds more "English". If this were the case, then mere word counts would be insufficient to analyse the extent of the students' problems with lexical richness and formal academic register: more detailed error analysis would be required to provide insights into this question.

III.8. Academic lexis: results of comparative study

VocabProfile was used to estimate the frequency of academic vocabulary in the sets of texts described above. The results are set out in Table 1.

From the results detailed in Table 1, it is evident that the students used slightly fewer words from this category than the TWE writers (8.56% compared to 9.49%). Scrutiny of the results for individual cases revealed that some of the students used a relatively large percentage of academic words: in four cases, over 14% of the word types used matched with the academic wordlist. In contrast to this, in four essays the percentage of word types from the academic wordlist was less than 5%. The TWE essays also varied considerably in their academic word content: four also had more than 14%, while two had less than 5%.

It has been mentioned above that statistics have been calculated over large corpora which indicate that words from the academic wordlist account for up to 10% of all running words, or word tokens. For this reason, the percentages of academic word tokens in students' essays and TWE essays were also calculated (data not shown). The mean for students' essays was 5.75%, whereas for TWE essays it was 6.59%. It is interesting to note that although the two groups differed slightly, both were below the figures estimated by Nation (1990) and Coxhead (2000) for the usual percentage of academic words in academic texts. Nation (1990) estimated this percentage at around 8% of running words (word tokens) on the basis of the Xue and Nation wordlist. Coxhead states (2000: 226) that her academic wordlist, which is considerably shorter than that of Xue and Nation but which is empirically sounder, accounted for 10% of the total number of word tokens in her 3.5 million-word academic corpus.

III.9. Academic lexis: interpretation of results

The data described above point to slight differences between the students' essays and the TWE essays in the mean use of both word types and word tokens from the academic word list, but these differences did not attain statistical significance. It is also evident that the Spanish-speakers' predicted bias towards Latinate words did not lead the students in this sample to make as much use of academic vocabulary as the successful TWE writers did.

It is interesting to note that neither the TWE essays nor the students' texts, taken as groups, contained as high a percentage of academic words as would be usual for academic texts. It is possible to speculate as to whether the length of the texts had any bearing on this result. First of all, it might seem that the shortness of these texts might provide some clue as to why these texts had a lower incidence of academic words. However, it is equally arguable that the reverse might be expected to occur: short texts of the rather condensed kind required by the TWE essay could actually be expected to have a higher incidence of academic words than the lengthy texts used in Coxhead's academic corpus.

Other authors (Read, 2005) have argued that quantitative data on lexical features such as the percentage of academic words should be complemented by qualitative studies which investigate the way these words are being handled in context. In view of the potential importance of academic vocabulary as a marker of a sophisticated style, the students' essays were examined in more detail to determine exactly how the frequency of academic words may relate to the quality of the text. Presented below are two examples from the students' essays, which serve to illustrate the findings in a more meaningful way. The words that have been underlined in these texts are from Coxhead's academic wordlist.

Essay 1: A student essay in which 2.8% of the word tokens belong to the academic wordlist. First of all I think it is useless to promote any business unable to make any profit. In the other hand, every business should get earnings but not any prices, I mean not every way to get a profit is fear for workers or is not allowed by law. Some international enterprises search to turn out with lowest cost by hiring children in far east - as India, Indonesia - making football shoes and balls by Nike; or cheap workforce as immigrants in some European countries. It is a great way to obtain biggest profits from the managers point of view. But from the human rights point of view to explode children as their receive a poor wage for a hard work. The same thing happens with illegal immigrants while they don't get any education. The shareholders earn big profits with business, next time they want to earn more money. And if you have big profits you pay big taxes too and fiscal fraud is an easy way, not to pay all the taxes that you should, then you have more money to divide to the shareholders. Fiscal fraud happens probable in very big business. All business try to increase these profits but usually hiring children and illegal immigrants happens in big business. The ridiculous thing is that business could make a loss, although their targets are profits.

Essay 10: An example from a student essay in which 9.3% of word tokens belonged to the academic wordlist. This statement is wrong, because in a business involves not just profit, also it involve people. A ethical choice are based on the personal moral philosophy of the decision maker, this philosophy is learned through the process of socialization with

friends, family and by formal education. Its also influenced by the social, business and corporate culture in which a person finds him. Also in every business have a profit responsibility, that is maximize profits for their owners, but each company have to evaluate the situation to make a decision because like this profit responsibility, exists the social responsibility that means that organizations are part of a larger society are accountable to society for their actions, this is very difficult because the diversity of values in the different social, business and corporate culture. This decision depends on the person, because in some cases, the society don't have it in mind, so all the consequences are not expected and the problem could grow. For example a consumer confusion over which products are environmentally safe is also apparent, because you really don't know.

In these examples, taken from essays 1 (2.8%) and 10 (9.3%), the academic vocabulary used in each case is underlined. Judged subjectively, essay 10 (9.3% of word tokens belonged to the academic wordlist) has a richer and more mature range of vocabulary than essay 1 (2.8%), and the use of the VocabProfile program helps to confirm this subjective judgement. Similarly, essay 1 (2.8%) seems to have a particularly impoverished lexis, and the VocabProfile program provides quantitative support for this impression. On the other hand, these examples also point to some of the shortcomings of this kind of experiment. Firstly, although essays 9 and 10 use more academic words, this could partly be due to lexical repetition. Essay 10 uses the phrase "corporate culture" twice in the excerpt tested, so this one phrase accounts for four of the instances of academic vocabulary. Secondly, even if academic words are used, this may partly be due to the Spanish speaker's preference for Latinate vocabulary. Thus the writer of essay 9 may use the phrase "to acquire a skill" not in order to achieve the sense of a higher register ("acquire" rather than "learn"), but because a Spanish speaker might automatically opt for "acquire" because it is similar to "adquirir". Thirdly, the academic words that are used (particularly those borrowed from Spanish) may be used inappropriately, and so far from being evidence of a superior command of English, are in fact a sign that the student has a limited vocabulary. One example of this in essay 1 is the phrase "to promote any business": "promote" belongs to the academic wordlist, and therefore adds to the proportion of academic words in the text, but the term is used here in a way that is not idiomatic, because "set up" or "establish" would be more usual collocates for "business".

Despite these problems, the data gathered concerning the academic words in this set of student essays and TWE essays point clearly to an under-use of academic vocabulary when compared with large-scale studies of academic prose, which was especially acute in the student essays. This experiment indicates that vocabulary may be one of the key problems in student writing: texts with a low frequency of academic words, coupled with a high frequency of words from the basic 1,000-word core of English, may give an impression of over-simplicity or even lexical impoverishment. This is clearly one of the key areas in which these students need assistance.

IV. DISCUSSION

To offer a satisfactory conclusion, it is necessary to comment both on the methods applied, and on the results obtained. Regarding the methods, it is clear that WordSmith and VocabProfile permitted a more detailed and objective study of lexical aspects of the students' writing than would have been possible by traditional qualitative assessment methods. The type-to-token ratio provides a solid basis for gauging repetitiveness, while the percentages of words from the first thousand word families and the academic wordlist, in particular, serve as indicators of lexical simplicity or sophistication. On this issue, our results appear to contradict the findings of a recent large-scale study (Shaw & Weir, 2007) which detected no significant differences in the lexical range of student writing over a broad spectrum of levels. In our view, the lack of variation in the latter study is likely to have been influenced by the types of writing task used, which encompassed a range of social genres, including only a small proportion of what might be regarded as academic text types. In these circumstances, these authors' conclusion that "a lexical profile analysis may be too crude a measure to differentiate between better and poorer performances" (Shaw & Weir, 2007: 104) would seem to be overstated. Although measures of variation and range evidently cannot be used in isolation, the present study suggests that quantitative tools do provide evidence that could be used in combination with other factors to build a profile of student texts. Moreover, for practising teachers of writing, they afford useful insights into the ways in which student work may differ from the target style.

As far as the actual results of this study are concerned, the foregoing studies have suggested that writing by this group of students tended to be characterised by rather low lexical variation, a preponderance of high-frequency words and under-use of academic vocabulary in comparison with top-grade TWE essays. Moreover, some students in the group produced writing that diverged greatly from the target in several of these categories. Overall, it seems that lexical simplicity may be a significant problem in these students' writing. These students need further help and guidance in this specific area, so that they can learn ways of developing their vocabulary and using it appropriately in writing.

The problem that could be raised here is that this whole phenomenon might simply be framed as a problem of "general language competence", and of a limited lexical repertoire. It is certainly extremely important for L2 writers to extend their linguistic resources in general and their knowledge of vocabulary in particular, particularly if the goal is to project "sophistication" rather than "simplicity", but this is a long-term project that might not be attainable within the context of university English courses.

Against this argument, we could maintain that since writing is not a skill that requires spontaneous output in real time, and texts can be produced slowly with appropriate use of reference material, it would be inaccurate to say that the problem is ever simply one of language proficiency. Students who produce lexically limited texts can be assumed to have low awareness of the impression caused by using simplistic vocabulary, since they have not made use of the opportunities they have to remedy the situation. All students, even those with limited linguistic resources, should be able to improve their writing skills with

appropriate guidance, indication of what is expected, and help as to how to enrich and check their texts. The students in this group had obviously not been sensitised to these issues. There are various other reasons why "general language competence" probably does not offer a satisfying explanation for the phenomenon under discussion in this chapter. These students' general command of English had been assessed as B2 to C1, which would generally be taken to indicate that they were able to use a wide range of structures and vocabulary in different registers, and at least had the potential to use an appropriate style. If they under-perform in writing, the reason may lie in factors other than their general command of English.

An alternative explanation for the "simplicity" of many of the texts studied here is that many of the "simple" features of these students' texts could well be a result of transferring habits from speech to writing (Hinkel, 2003). This suggests that rather than lacking linguistic resources, many students may simply not be sensitised to the problem of acquiring a formal written style. They write as they speak, with unfortunate consequences for the lexical content of their written productions, as well as implications for the syntax, cohesion, coherence and genre of their texts.

A related but slightly different question of motivation behind limited vocabulary use is that of the "safety-first" approach to writing. It has been surmised that the reason for poor lexical range displayed in learners' texts may not be an actual lack of vocabulary, so much as a preference for the tried and tested. Vocabulary is of paramount importance in the whole issue of simplicity and sophistication, and the learners' limited lexical range and variety may be the main reason why their writing seems "dull, repetitive and unimaginative, with many undeveloped themes" (Ringbom, 1998: 50). But the limited lexicon is perhaps more a matter of habit or choice than of actual linguistic impoverishment. Previous authors have observed that in the stressful situation of having to organise thoughts and manage discourse in a foreign language, many L2 writers prefer to stick to familiar words, to play safe rather than run risks (Hasselgren, 1994). In such cases, further sensitisation and training are needed in order to alert students to the importance of taking risks with words, and to teach them techniques (such as appropriate dictionary use) for taking the risk out of lexical adventurousness.

To conclude, it is sufficient to say that these comparisons between two groups of texts using WordSmith and VocabProfile have shed light on the relative lexical simplicity of the students' essays. This simplicity is itself far from simple, and may have various causes. But it is also a phenomenon which should not be regarded as immutable, since it may well arise out of some basic misunderstandings of the nature of writing and the type of written product that is required.

NOTES

¹Available on those sites on 20 January 2004.

²The Test of Written English was an optional component of the TOEFL paper-based test, which is now an obligatory part of the writing section of TOEFL iBT, the most common English language proficiency test used to assess language competence for study purposes. It consists of an opinion essay written in 30 minutes, which was formerly scored on a scale from 1 to 6, and is now awarded points from 1 to 5.

³Read (2005) found that the percentage of academic words in IELTS candidates' spoken performance was roughly proportional to their general level of performance: students who obtained grades 6 to 8 had around 10% academic words, while students with lower levels of performance (around CEF B1) had only 5.9%.

REFERENCES

- Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34:2, 213-238.
- Engber, C. (1995). The relationship of lexical proficiency to the quality of ESL composition. *Journal of Second Language Writing*, 4, 139-155.
- Granger, S. (1998). The computer learner corpus: a versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on Computer*. Harlow: Addison, Wesley, Longman, 3-18.
- Hasselgren, A. (1994). Lexical Teddy Bears and Advanced Learners: a Study into the Ways Norwegian Students Cope with English Vocabulary. *International Journal of Applied Linguistics*, 4:2, 237-260.
- Hinkel, E. (2003). Simplicity without elegance: features of sentences in L1 and L2 academic texts. *TESOL Quarterly*, 37:2, 275-301.
- Huntley, H. (1999). Master: a university word list reader. *TESL-EJ*, 4:2.
- Laufer, B. & Nation, I. S. P. (1995). Vocabulary size and use: Lexical richness in L2 written productions. *Applied Linguistics*, 16:3, 307-322.
- Leki, I. & Carson, J. (1994). Students' perceptions of EAP writing instruction and writing needs across the disciplines, *TESOL Quarterly* 28, 81-101.
- Martín Martín, P. (2003). Análisis contrastivo de los componentes estructurales y gramaticales de los resúmenes de los artículos científicos. *Revista Española de Lingüística*, 33:1, 153-183.
- Meunier, F. (1998). Computer tools for the analysis of learner corpora. In S. Granger (Ed.), *Learner English on Computer*. Harlow: Addison, Wesley, Longman, 19-37.

- Miller, P. (2001). VocabProfile. *Vocabulary Acquisition Research Group Encyclopedia*. University of Wales. (available at www.swan.ac.uk/cals/calsres/encyclopedia/vp.htm 18 February 2005).
- Muncie, J. (2002). Process writing and vocabulary development: comparing lexical frequency profiles across drafts, *System*, 30, 225-235.
- Nation, I. S. P. (1990). *Teaching and Learning Vocabulary*. New York: Newbury House.
- Pearson, J. & Bowker, L. (2002). *Working with Specialised Language: A Practical Guide to Using Corpora*. London: Routledge.
- Polio, C. (2001). Research methodology in second language writing research: the case of text-based studies. In T. Silva, and P. K. Matsuda (Eds.), *On Second Language Writing*. Mahwah, NJ: Lawrence Erlbaum, 91-115.
- Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.
- Read, J. (2005). Applying lexical statistics to the IELTS speaking test. *Cambridge ESOL Research Notes*, 20, 12-16.
- Reid, J. (1993). *Teaching ESL Writing*. Eaglewood Cliffs, NJ: Prentice Hall.
- Reynolds, D. (2001). Language in the balance: Lexical repetition as a function of topic, cultural background, and writing development. *Language Learning*, 51, 437-476.
- Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: a cross-linguistic approach. In S. Granger (Ed.), *Learner English on Computer*. Harlow: Addison, Wesley, Longman, 41-52.
- Sánchez-Hernández, P. & Pérez-Paredes, P. (2005). Examining English for Academic purposes students' vocabulary output: corpus-aided analysis and learner corpora, *Revista Española de Lingüística Aplicada*, 1, 201-212.
- Scott, M. & Tribble, C. (2006). *Textual Patterns. Keywords and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.
- Shaw, P. & Liu, E. (1998). What develops in the development of second language writing. *Applied Linguistics* 19, 225-254.
- Shaw, S. & Weir, C. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge: Cambridge University Press.
- Virtanen, T. (1998). Direct questions in argumentative student writing. In S. Granger (Ed.), *Learner English on Computer*. Harlow: Addison, Wesley, Longman, 94-106.
- West, M. (1953). *A General Service List of English Words*. London: Longman.
- Xue, G. & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication*, 3, 215-229.