Indexing Second Language Vocabulary in the Intermediate GEPT

Hintat Cheung
National Taiwan University
hintat@ntu.edu.tw

Siaw-Fong Chung
National Chengchi University
sfchung@nccu.edu.tw

Sophia Skoufaki[1]
National Taiwan University
sophiaskoufaki@ntu.edu.tw

Abstract

The Language Training and Testing Center (LTTC) and the Graduate Institute of Linguistics at National Taiwan University have been collaborating since 2007 on the construction of the LTTC English Learner Corpus. The corpus will consist of language samples produced by Taiwanese learners of English who have sat the General English Proficiency Test (GEPT), a language proficiency examination administered by the LTTC. This paper will first give an overview of the corpus content and the procedures involved in its compilation. Then, it will provide further insight into vocabulary use by English learners in Taiwan by examining the role of lexical variation in judgements about the writing quality of GEPT texts, as these judgements are reflected in text grades. The data are the 2,000 writing samples from the Intermediate Level General English Proficiency Test (Intermediate GEPT) which have been processed so far. In the Intermediate GEPT, writing ability is partly assessed through the production of 120-word written texts. These texts are graded holistically. Apart from describing vocabulary use in these texts, the present study also aims to probe into the lexical dimension of the holistic grading system used by the LTTC for Intermediate GEPT marking.

Keywords: Intermediate GEPT, lexical variation, vocabulary use, writing assessment, holistic grading.

Introduction

The Language Training and Testing Center (LTTC) and the Graduate Institute of Linguistics (GIL) at National Taiwan University (NTU) have been collaborating since 2007 on the construction of the LTTC English Learner Corpus[2]. The corpus will consist of language samples by Taiwanese learners of English who have sat the General English Proficiency Test (GEPT), a language proficiency examination administered by the LTTC. This paper will first provide an overview of the corpus and the procedures involved in its compilation. Then it will use corpus data to examine an issue which has yielded conflicting results in previous studies, namely, the relationship between lexical variation (otherwise called "lexical diversity" and "lexical variability") and holistic essay grading. Lexical variation will be examined through two different lexical variation measures and in relation to two English word lists created while keeping in mind the needs of Taiwanese learners of English, the English Reference Wordlist for the College Entrance Examination in Taiwan (Jeng, 2002) and LTTC's word list for the GEPT. The data are the 2,000 writing samples from the Intermediate Level General English Proficiency Test (Intermediate GEPT) which have been processed so far.

[1] Author names are presented alphabetically. Skoufaki is the corresponding author.

[2] This project is directed by Prof. Hintat Cheung of GIL. The project co-directors are Dr Zhao-Ming Gao from the Department of Foreign Languages and Literatures at NTU and Dr Siaw-Fong Chung from the Department of English at National Chengchi University. The project members are the postdoctoral research associate, Dr Sophia Skoufaki, the research assistant and administrator Ms Sumei Chen, and two PhD students, Ms Sally Chen and Ms Claire Chiyi Wu. Ms Shanju Lin was the research assistant and administrator in the academic year 2008-9.

LTTC English Learner Corpus: Content

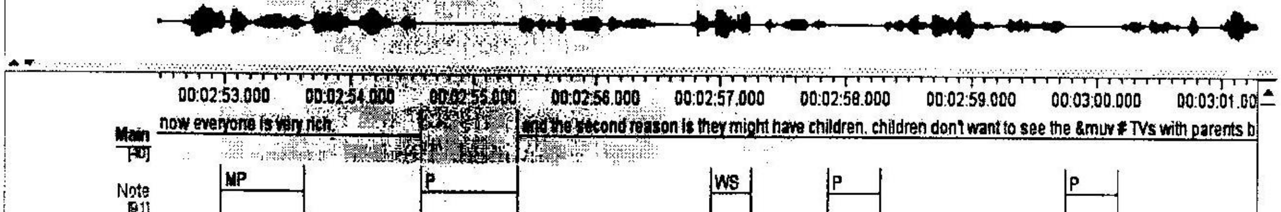
In the current, first phase of corpus construction, 2,000 written-production and 400 oral-production samples from the Intermediate GEPT examination have been processed. The corpus is currently being extended through the addition of 2,000 more writing samples from the Intermediate GEPT examinations and 1,000 such samples from the High-Intermediate GEPT examinations.

The written samples which have been processed are short paragraphs on three topics. Two of these topics are questions about personal preferences (favourite idol and exotic food) and a third topic asks test-takers to explain why many elementary-school children in Taiwan are nearsighted and to propose effective ways of preventing nearsightedness. The oral samples contain answers to ten questions and the description of a picture. Three sets of answers and image descriptions have been processed.

LTTC English Learner Corpus: Design and Compilation

Metadata about the performance and characteristics of each test-taker are available for each sample. These metadata are the score that was assigned by test-graders to each sample, the broad region of Taiwan (e.g., North, East) where the test was taken, the age, gender, education level of the test-taker, his or her major if the test-taker was a college graduate, whether the test-taker was a student or not, and whether (s)he had lived in an English-speaking country, and if so, for how long.

The digitization of the written and oral samples relied on the principle of having each sample transcribed by two persons, comparing the two transcriptions through a computer program and then making revisions manually based on the differences between the transcripts. In terms of the written samples, the 2,000 hand-written samples were initially scanned and then each sample was typed in the computer independently by two student helpers. The two versions of each sample were compared by a computer program and revisions were done manually. The oral samples were initially recorded on cassette tapes, then digitized and finally student helpers were trained to transcribe them using the software ELAN (EUDICO Linguistic Annotator) (Hellwig, van Uytvanck, & Hulsbosch, 2008) and add tags using the CHAT (CHILDES) format (MacWhinney, 2008). Tags were added inside the body of the transcriptions for repetitions, self-corrections, incomprehensible sounds, lengthened vowels and other characteristics. Tags for filled (FP) and unfilled (P) pauses, mispronunciations (MS) and word stress (WS) errors were added in the tier below that of the main transcription. Figure 1 below shows the soundwave and the transcription of part of an ELAN file. The transcription appears in two tiers. The first tier contains the text and CHAT symbols and the second codes for unfilled pauses, mispronunciations, and word stress errors.



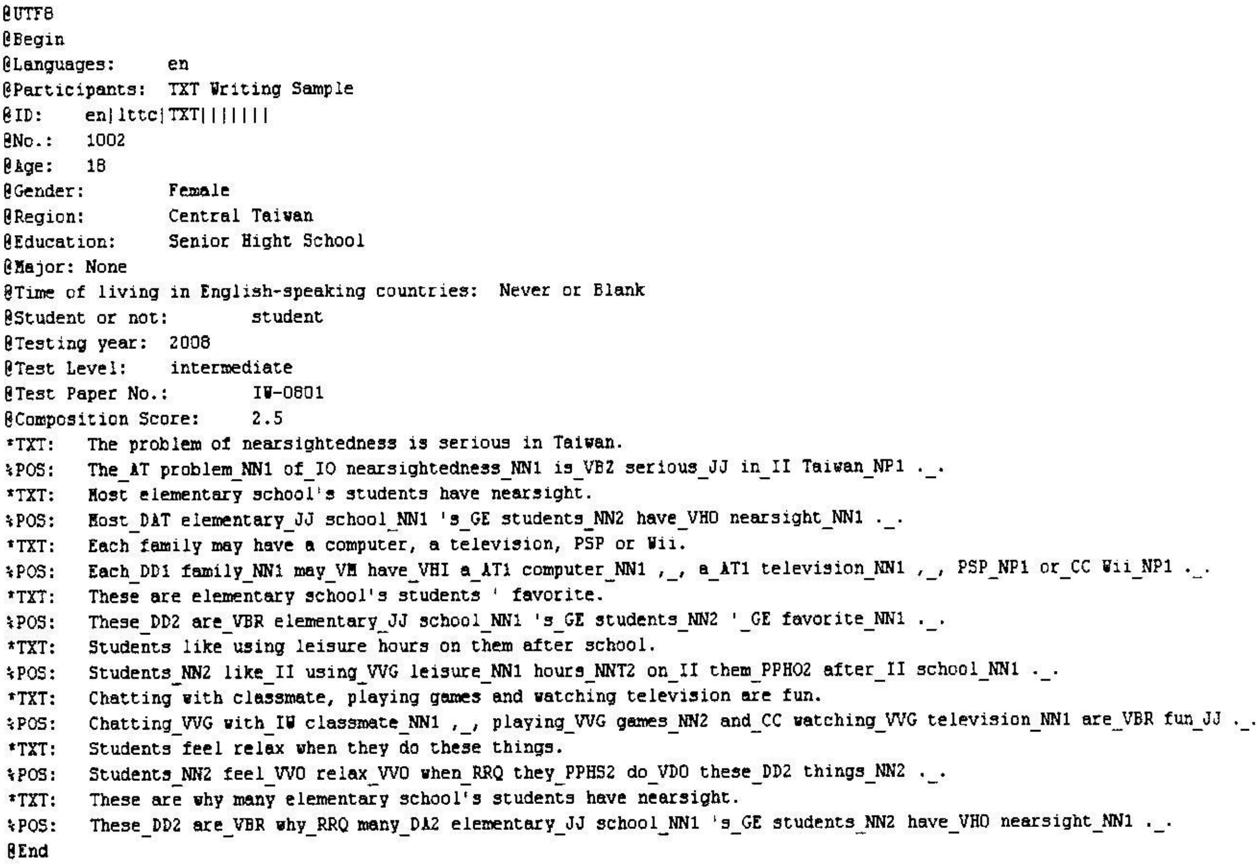
now everyone is very rich.

and the second reason is they might have children. children don't want to see the &muv # TVs with parents b

Figure 1. Soundwave and transcription segment from an ELAN oral file

Two student helpers transcribed each oral sample and these two versions were compared through a computer program. The samples were then revised manually.

The last stage in the processing of the sample files is part-of-speech tagging. So far, all the written samples and half of the oral samples have been part-of-speech tagged with the CLAWS4 tagger (Garside & Smith, 1997). Following part-of-speech tagging, the aforementioned metadata were added to the samples through a software program so that each sample has the structure of a CHAT file. Figure 2 below shows a writing sample in CHAT format.



```
@UTF8
@Begin
@Languages:     en
@Participants:  TXT Writing Sample
@ID:    en|lttc|TXT|||||||
@No.:   1002
@Age:   18
@Gender:        Female
@Region:        Central Taiwan
@Education:     Senior Hight School
@Major: None
@Time of living in English-speaking countries:  Never or Blank
@Student or not:        student
@Testing year:  2008
@Test Level:    intermediate
@Test Paper No.:        IW-0801
@Composition Score:     2.5
*TXT:   The problem of nearsightedness is serious in Taiwan.
%POS:   The_AT problem_NN1 of_IO nearsightedness_NN1 is_VBZ serious_JJ in_II Taiwan_NP1 ._.
*TXT:   Most elementary school's students have nearsight.
%POS:   Most_DAT elementary_JJ school_NN1 's_GE students_NN2 have_VH0 nearsight_NN1 ._.
*TXT:   Each family may have a computer, a television, PSP or Wii.
%POS:   Each_DD1 family_NN1 may_VM have_VHI a_AT1 computer_NN1 ,_, a_AT1 television_NN1 ,_, PSP_NP1 or_CC Wii_NP1 ._.
*TXT:   These are elementary school's students ' favorite.
%POS:   These_DD2 are_VBR elementary_JJ school_NN1 's_GE students_NN2 '_GE favorite_NN1 ._.
*TXT:   Students like using leisure hours on them after school.
%POS:   Students_NN2 like_II using_VVG leisure_NN1 hours_NNT2 on_II them_PPHO2 after_II school_NN1 ._.
*TXT:   Chatting with classmate, playing games and watching television are fun.
%POS:   Chatting_VVG with_IW classmate_NN1 ,_, playing_VVG games_NN2 and_CC watching_VVG television_NN1 are_VBR fun_JJ ._.
*TXT:   Students feel relax when they do these things.
%POS:   Students_NN2 feel_VV0 relax_VV0 when_RRQ they_PPHS2 do_VD0 these_DD2 things_NN2 ._.
*TXT:   These are why many elementary school's students have nearsight.
%POS:   These_DD2 are_VBR why_RRQ many_DA2 elementary_JJ school_NN1 's_GE students_NN2 have_VH0 nearsight_NN1 ._.
@End
```

Figure 2. A written sample in CHAT format

In these files the metadata appear as the header of each file, followed by the writing sample. Every other line in the sample has the same content as the line above it but it also includes CLAWS part-of-speech tags.

Lexical Variation and Holistic L2 Production Scoring

Many current EFL writing performance examinations use holistic grading of texts written by examinees. Research has shown that in holistic grading judges are affected by the text as a whole rather than by individual elements (Gunning, 2006; Wolcott & Legg, 1998). However, the role that individual linguistic characteristics of L2 production play on judgements about writing quality has also been examined. Lexical variation is among the lexical characteristics which have been examined in this respect.

Before reviewing the relevant literature, it is necessary to explain more what "lexical variation" means in these studies through considering how it is measured. The various ways in which 'lexical variation' has been measured indicate various understandings of the term, but a broad definition is the one by Malvern and Richards (2002: 87): 'the variety of active vocabulary deployed by a speaker or writer'. The various measures used indicate that different researchers define this variation in vocabulary used differently. The most well-known measure of lexical variation is the Type-Token ratio (TTR), which divides the word types by the word tokens of a piece of linguistic output. What counts as a word type differs among studies. Some consider them to be word types per se, that is, each word form is counted separately, without including in its group any inflected or derived forms related to it. Others consider them to be word lemmas, that is, the root word forms and their most frequent inflected forms. Yet others define word types as word families, that is, groups of word forms which do not only include the root and the most frequent inflected forms but also the derived forms with the derivational affixes up to a certain level in Bauer and Nation's (1995) English affix frequency list. In this study, Bauer and Nation counted the frequency of English affixes in corpora and organised them in nine frequency bands. Level 3 word families included the inflected forms of words and also word forms with frequent derivational suffixes. Many researchers use the affixes included in levels up to Level 3 of this list when their data is from elementary to intermediate learners (Milton, 2009: 10). TTR has been found to be unreliable because it systematically decreases as text length increases (e.g., Richards, 1987). This fact means that it is methodologically wrong to measure lexical variation through TTR in studies which use texts of different length.

A series of measures which are algebraic transformations of TTR have been proposed to help avoid this problem, namely, the Guiraud index, Herdan's index, and Uber index. These measures purport to compensate for text length differences, however, they have all been found to be unstable (see Tweedie & Bayen, 1998 for a thorough evaluation of these measures' reliability). Measures such as D (Malvern & Richards, 2002), which use a curve-fitting approach that takes into account the position and shape of the entire TTR curve seem to be more reliable, but even they do not agree among themselves in their classifications of texts in terms of lexical variation (Jarvis, 2002).

More recently, some researchers have stressed the need for a combination of quantitative measures and qualitative examinations of the lexical differences among pieces of language output from high- and low-proficiency speakers (e.g., Daller, van Hout, & Treffers-Daller, 2003; Meara & Bell, 2005). One of the reasons for this change in research method is that different texts can have the same type/token or other ratio indicating lexical variation, but they may differ significantly in terms of the frequency of 'difficult' words in them. This difficulty is usually defined in terms of frequency of words in native English (the more infrequent a word, the more difficult it is considered) and word lists are formed according to this frequency criterion (e.g., the word lists by Paul Nation which come with the software RANGE). The issue of what constitutes difficult vocabulary for language learners is also controversial, but we will examine it later in this literature review and will, for the time being, focus on this last kind of lexical variation measures. The Lexical Frequency Profile (LFP) by Laufer and Nation (Laufer and Nation 1995: 311) "shows the percentage of words a learner uses at different vocabulary frequency levels in her writing – or, put differently, the relative proportion of words from different frequency levels". Based on this list of data, Laufer (1995) proposed a way to locate the proportion of the advanced vocabulary in a text. She considered the word families which are beyond the first 2,000 word families in the word list used in the LFP as advanced vocabulary, since they occur less frequently in native-speaker corpora than those in the first 2,000 word families. She used the percentage of these word families as an indication of lexical sophistication. Meara and Bell (2005: 8-9) have criticised the use of the first 2,000 word families as 'basic vocabulary' because, according to them, "in a typical text" the percentage of words levels above the first 2,000 word families.

This approach has inspired other researchers to improve on it and propose similar measures. Meara and Bell (2005) proposed P_Lex. It is a measure which "looks at the distribution of difficult words in a text, and returns a simple index that tells us how likely the occurrence of these words is." (Meara & Bell, 2005: 9). They claim that it is more reliable with texts varying in length that LFP. Daller, van Hout and Treffers-Daller (2003) modified the Guiraud index, so that its formula has advanced word types rather than all the word types in a text as its numerator ('advanced Guiraud index').

Coming back to our review of studies which examine whether lexical variation plays an important role in holistic grading, we can see that results vary greatly across studies. In terms of the holistic grades given to L2 oral output, some studies indicate that, when measured through measures which take qualitative vocabulary elements into account, lexical variation correlates significantly with holistic grading (Daller & Phelan, 2007: 234). In terms of L2 written output, Engber (1995) asked students from a variety of language backgrounds who had taken an intensive English course at a US university to write an essay on the same topic. She found a substantial correlation between lexical variation and holistic grades without having deleted the lexical errors from the L2 essays she examined and an even higher correlation after deleting lexical errors. Daller and Phelan (2007) correlated a variety of linguistic characteristics in essays with the holistic grades assigned to the essays according to IELTS criteria. They used various measures of lexical variation. Whereas the correlations with the purely quantitative lexical variation measures were not significant, the highest correlations among all linguistic characteristics were achieved by the lexical variation measures which combined both quantitative and qualitative elements, such as the advanced Guiraud index. By contrast, in Lorenzo-Dus and Meara (2005) lexical diversity measured with D did not differ significantly across holistic grades of L2 essays. Baba (2009) found that lexical diversity measured with D did not correlate with holistic scores given to summaries written by Japanese advanced English language learners. Schmitt (2005), as summarised in Shaw and Weir (2007: 103-105), compared lexical variation across Cambridge ESOL examinations of different language proficiency levels. He used STTR, a quantitative measure provided as part of WordSmith Tools software. It is a variation of TTR which controls for text length. He found an increase in lexical variation between the Key English Test (KET) and the Preliminary English Test (PET) but no significant increase between PET and the First Certificate in English (FCE). He also compared performance according to Laufer and Nation's LFP, which he considers a measure of lexical sophistication rather than a measure of lexical variation, but he did not find any increase in LFP across the aforementioned English proficiency examinations.

A few studies examined the effect that individual learner differences have on the relationship between lexical variation and holistic grades. Jarvis (2002) tested, among other things, the relationship between lexical variation and holistic grades through a variety of lexical variation measures in short narratives produced by Swedish and Finnish learners at various levels of English language proficiency and by English native speakers. He found that D was the most stable measure and that there was a significant correlation between the two best measures, D and U, and the holistic grades in their totality and also by each participant group. He also examined the relationship between L1 and level of lexical variation, but concluded that this relationship is not straightforward and is probably affected by other factors such as age and years of English instruction. Yu (forthcoming) found that, after outlier removal, lexical variation as measured by D could account for 11% of the variance of the holistic grades given to writing samples from past Michigan Examination Language Assessment Battery (MELAB) tests. However, he found that certain learner characteristics played an important role on the significance or magnitude of this relationship. The relation between D and holistic grades was stronger for men than for women. For the Chinese L1 learners, D did not correlate significantly with holistic grades.

This brief literature review points to several reasons for the conflicting results among studies. An important reason has to be the use of different lexical variation measures across studies. Most of the studies which found a significant relationship between lexical variation and holistic grading used measures which assessed both lexical variation both quantitatively and qualitatively, such as the advanced Guiraud index. Moreover, lexical measures vary in terms of their reliability when they are used with texts of varying length and this fact holds even for the measures which are considered the most reliable (Jarvis, 2002). Finally, when measures which take lexical difficulty into account are used, they use different word lists to distinguish between easy and difficult vocabulary (e.g., LFP uses a different word list from P_Lex).

Differences in the definition of 'word type' probably also yield different results. Depending on the researcher's assumptions about the organisation of a typical participant's mental lexicon, 'word type' can vary in terms of its size. Some researchers consider each inflected and derived form of a root as distinct word types (e.g., Yu, forthcoming). Others define it as the root form and all its frequently inflected forms, that is, according to its lemma (e.g., Engber, 1995). Some researchers who work with learners with English proficiency from the elementary to the upper-intermediate level consider as word types word families which include word root forms and their combinations with the affixes in level 3 of Bauer and Nation's (1993) frequency list. Such different definitions of 'word type' mean that studies vary in the number of lexical items which fall under the same word type, thus affecting lexical measure calculations.

The differences in data cleaning procedures form another likely reason for the conflicting results. Some researchers exclude lexical errors from the analysis while others include only lexical errors they are uncertain about (e.g., Laufer & Nation, 1995), correct the spelling errors (e.g., Morris & Cobb, 2004; Laufer & Nation, 1995), or correct only those which seem to be genuine spelling mistakes (e.g., Yu, forthcoming). This fact can affect the degree of relationship found between lexical variation and holistic grading, as Engber (1995) indicates. Some researchers exclude proper nouns from the analysis (e.g., Laufer & Nation, 1995), others include them in the lowest level of word lists involved in the calculation of lexical variation measures (e.g., Morris & Cobb, 2004), and others include them in the analysis without any pre-processing (e.g., Yu, forthcoming). These various approaches affect the number of types and tokens included in the lexical variation measures' calculations.

As indicated above, individual learner differences can play an important mediating role in the relationship between lexical variation and holistic grading. Most studies differ in terms of the L1 and years of L2 instruction of their participants, so findings among studies are not comparable.

Finally, studies differ in terms of task characteristics. At a broad level, most studies ask participants to write paragraphs or essays on specific topics, but there are also some which ask them to describe a picture (Daller, van Hout, & Treffers-Daller, 2003) or film (Jarvis, 2002), or to summarise a text (Baba, 2009). At a lower level, tasks have been shown to yield differences in lexical variation and its relationship to holistic grading because of the participants' familiarity with an essay topic or because of differences in topic phrasing (Carlson et al., 1985), so such variation in essay prompts may affect the relationship under consideration.

The aforementioned factors probably affect estimates of the relationship between lexical variation and holistic grading. Therefore, these findings cannot be generalised to different learner populations. This conclusion makes the question of whether there is a significant relationship between lexical variation and holistic grading in the writing samples which have been processed in the LTTC English Learner Corpus so far worth investigating. This research question is interesting also because (a) most lexical variation studies have been conducted with advanced proficiency learners and (b) very few studies include Chinese L1 learners (i.e., Engber, 1995; Yu, forthcoming). The learners who have produced the writing samples used in this study are all of low-intermediate English proficiency and have Mandarin Chinese as their first language. Therefore, this study aims to give more information about the relationship between lexical variation and holistic grading in this under-represented learner population.

The second research question addressed by this study relates to the methodological issue of how lexical variation should be assessed. As mentioned above, most studies which have found a strong relationship between lexical variation and holistic grading have used lexical variation measures which are not only quantitative but also include a qualitative element about the level of word difficulty in a text. This study will therefore use both simple and advanced Guiraud index to measure lexical variation. It is expected that this study will replicate most previous studies in that a significant relationship between lexical variation and holistic grading will be found only when lexical variation is measured through the *advanced* Guiraud index.

The present study is also motivated by the fact that different studies use different word lists as the basis of lexical variation measures with a qualitative element. Moreover, most of these studies use word lists which are based on an understanding of word difficulty as low word frequency in a native speaker corpus. However, as Nation (2004: 12) has pointed out, perhaps word frequencies in native-speaker corpora are not a good indication of word difficulty for learners because especially foreign language learners do not have the same level of access to linguistic input equivalent to that of the native speakers. This claim is further bolstered by research findings summarised in Milton (2009). Milton notes that although applied linguistic theory advises textbook writers to introduce new word in terms of their frequency in native-speaker corpora (that is, from the most to the least frequent words), this is not feasible because "the content and sequencing of textbooks is likely to be driven more by the practical concerns of structuring a workable text and less by the theory as it exists." (Milton, 2009: 196). These considerations suggest that using word lists which have been constructed with Taiwanese learners in mind is more appropriate to our study. Therefore, in this study two word lists made in Taiwan have been used to calculate the lexical variation measures. One is the English Reference Wordlist for the College Entrance Examination in Taiwan (Jeng, 2002) (CEEW) and the other is LTTC's word list for the GEPT (GEPT list). Further explanation about why these two word lists have been chosen will be provided in the next section of this paper. The use of two word lists, rather than one, is considered beneficial for data analysis since lexical variation measures based on different word lists may reveal different findings about lexical variation. Using more than one word list to calculate lexical measures is also an innovatory element of the study, since all other such studies have used only one word list.

In sum, the current study addresses the following research questions:

1. Which is the relationship between lexical variation and holistic grading in 2,000 Intermediate GEPT writing samples?
2. How do the simple and advanced Guiraud index compare as indices of lexical variation in these writing samples?
3. How does the use of two different word lists affect the findings for research questions (a) and (b)?

Method

Rationale

Since the data in this study are available from the LTTC English Learner Corpus, the main part of the method consists of the statistical procedures involved in data analysis. The first main step in data analysis is the calculation of simple and advanced Guiraud indices based on the lemmatisation of the writing-sample files via the two aforementioned Taiwanese learner word lists, CEEW and the GEPT list. The simple Guiraud index is calculated by dividing the total number of word types with the square root of the total number of tokens in a text. The advanced Guiraud index is calculated by dividing the 'advanced' word types by the square root of the total number of tokens in a text. The second step is the comparison of the scores from all the Guiraud indices across the holistic grades given to the writing samples. The final step is the correlation between the holistic grades and the indices which differed significantly across holistic grades to assess the strength of the relationship between holistic grades and each lexical measure.

The procedures followed to obtain the word types used in the advanced Guiraud calculations will be summarised in the 'Data preparation' section below. Before summarising these procedures, it is necessary to offer the motivation for important aspects of the method of this study, namely,

1. the use of the simple and advanced Guiraud indices rather than other lexical variation indices
2. the definition of 'word type' as lemma
3. the selection of CEEW and the GEPT list as the bases for the Guiraud indices, rather than other word lists
4. the selection of the word level(s) in the CEEW and the GEPT list which would be considered as 'basic vocabulary' in the advanced Guiraud indices
5. the inclusion of words apart from those belonging in the levels in CEEW and the GEPT list considered to constitute 'basic vocabulary' in the advanced Guiraud indices.

As mentioned in the section of the literature review which discussed the reliability of lexical variation measures, all lexical variation measures are unreliable but to various degrees. The Guiraud index is not as unreliable as TTR but it is less reliable than D, for example. However, even very advanced lexical variation measures require a minimum text length for reliable performance. For example, Meara and Bell (2005) after testing the reliability of their lexical variation measure P_Lex, conclude that it is stable for texts of at least 150 words. Similarly, D needs to be applied to texts at least 50 words long to be reliable (Yu, forthcoming: 10). In this exploratory study, the Guiraud index was used because it was decided that a measure which can be easily computed would be appropriate. More computationally demanding measures, such as those proposed by Jarvis (2002), can be used in follow-up studies.

The selection of Guiraud and advanced Guiraud was made also in order to replicate Daller, van Hout, and Treffers-Daller (2003). In that study, lexical variation in the speech of German-dominant and Turkish-dominant German-Turkish bilinguals was compared through the type/token ratio, the Guiraud index, and versions of TTR and the Guiraud test which take lexical difficulty into account (what they call 'advanced TTR' and 'advanced Guiraud'). They found that the advanced Guiraud distinguished between the two speaker groups better than the other measures. By doing the same kind of data analysis with Taiwanese low-intermediate learners of English, the current study tests whether advanced Guiraud is useful for learners of lower proficiency and different L1 and L2 than those in Daller, van Hout, and Treffers-Daller (2003).

Finally, the selection of advanced Guiraud rather than only the Guiraud index was made in order to compare the relationship between a purely quantitative lexical variation measure (Guiraud) and holistic grading with the relationship between a lexical variation measure which is both quantitative and qualitative (advanced Guiraud) and holistic grading. As mentioned above, most studies which have found a significant effect of lexical variation on holistic grading had used lexical variation measures which are both quantitative and qualitative, so it would be interesting to see if this pattern of results would be replicated here.

As mentioned above different researchers operationalise 'word type' differently. In this study 'word type' means the root word form and its most frequent inflected forms; in other words, it equals more or less the meaning of 'lemma'. We decided not to define 'word type' as a word family with family members ending with the suffixes included in levels 1-3 of Bauer and Nation's (1995) list because we do not know yet whether we are warranted to make the assumption that word families are organised in that way in the mental lexicon of the learners who produced the writing samples. Moreover, the affix frequency list is meant to reflect receptive vocabulary knowledge, which can facilitate word recognition in reading (Bauer & Nation, 1993: 268). Therefore, it is unclear whether using this notion of word family is appropriate for our writing samples, which required productive vocabulary knowledge to be constructed. 'Word type' was not defined as a word type per se. For example, the verb form 'read' and 'reads' would be considered different word types in that analysis whereas in ours they are exponents of the same word type. This decision was made because it was felt that at low-intermediate level learners have mastered the basic inflectional morphology, so word families in their mental lexicon include inflected forms, at least.

The word lists used in this study to calculate the lexical variation measures were constructed with Taiwanese learners of English in mind. CEEW was constructed in order to determine the words that Taiwanese high school students are expected to use to pass different levels of college entrance examinations. It consists of six levels. The first four level of the CEEW are expected to be used in the general college entrance examination and the rest are expected to be used in subject-specific examinations.

Each level has 1080 headwords, so the whole list includes 6,480 headwords in total. 35 sources were used to construct the list, including

1. Junior and Senior high school English text books in Taiwan
2. USA primary school text books
3. Other English textbooks used in UK, USA, Canada, Japan, China and Taiwan.

The GEPT word list includes words which can be used in the questions posed to GEPT test-takers. There is a separate word list for each level of the GEPT examination, that is, an Elementary, Intermediate, and a High-Intermediate level. It includes 8,239 headwords in total. The Elementary level includes 2,263 headwords, the Intermediate level 2,682, and the High-Intermediate level 3,294. This word list was constructed in a similar way to the CEEW list. Table 1, used with permission of the LTTC, lists the sources for each level of the list.

Table 1

The reference sources of the GEPT word list

Elementary	**Intermediate**	**High-intermediate**
Collins Cobuild English Dictionary (bands 4-5)	Collins Cobuild English Dictionary (bands 3-5)	Collins Cobuild English Dictionary (bands 2-5)
English Reference Wordlist levels 1-2 compiled by **College Entrance Examination Center** (CEEC)	**English Reference Wordlist** levels 1-4 compiled by **College Entrance Examination Center** (CEEC)	**English Reference Wordlist** levels 1-6 compiled by **College Entrance Examination Center** (CEEC)
Taiwan MOE's 2000 words for Grades 1-9	China's High School English Vocabulary	China's English vocabulary for College English 4 and 6

As one can expect, both word lists are different from the ones often used in lexical variation studies in that they are not based on frequencies in native corpora. In the CEEW, all words from the reference sources were listed and those which appeared more than twice in the reference sources were kept as candidates for levels 1 to 4 of the word list. From these words, those which appeared more than five times were selected and those which would appear in the first four levels were selected by English language teachers. The same procedure but for words appearing three times led to the creation of the other two word list levels. A similar procedure, based on frequency in reference sources and expert judgement was followed in the compilation of the GEPT word list. This learner-oriented approach to the formation of the word lists means that these word lists stand a good chance of overlapping with the vocabulary used by GEPT test-takers. This rationale underlies the selection of these word lists rather than native-corpora based word lists in this study.

Both these lists rather than other Taiwanese English learner word lists were selected because of their breadth. They are the two Taiwanese English learner word lists which include the largest number of headwords. Using extensive word lists was thought to yield as few spuriously 'unfound' words as possible in the writing samples. They were also selected because of their recency as compared to other Taiwanese English learner word lists (for a comparison of many Taiwanese English learner word lists seen Su, 2006: 4). An important reason why these word lists were used to calculate different versions of the Guiraud index rather than using only one of them was that their levels partly overlap and they also partly differ in the distribution of headwords across levels, as Table 2 indicates. It was thought that using two word lists of big and similar size but which do not completely overlap in total headwords as well as level by level headwords might lead to interesting comparative results about the relationship between lexical variation and holistic grading.

Table 2
Percentage of headwords from the LTTC's wordlists for GEPT examinations (GEPT list) which also exist in the English reference wordlist for the college entrance examination in Taiwan (CEEW)

CEEW	*GEPT list*			Total percentage
	Elementary level	Intermediate level	High-intermediate level	
Level 1	12.16%	0.78%	0.15%	13.09%
Level 2	10.94%	1.88%	0.37%	13.19%
Level 3	1.67%	11.16%	0.54%	13.37%
Level 4	0.56%	11.55%	0.93%	13.04%
Level 5	0.1%	2.27%	10.92%	13.29%
Level 6	0.06%	1.37%	11.71%	13.14%
Not in CEEW	1.45%	3.8%	15.63%	20.88%
Total percentage	26.94%	32.81%	40.25%	100%

The decision about which level or levels of each word list would be considered basic and which advanced for the calculation of lexical variation measures with a qualitative element was difficult to make because no vocabulary knowledge test has been administered to the writers of the GEPT writing samples used in this study. In terms of the GEPT word list, since it includes one level per GEPT examination, we considered the level below the one of the writing samples, that is, the Elementary one, as the basic vocabulary and the compilation of the Intermediate and the High-Intermediate levels as the advanced vocabulary. In CEEW, each level contains fewer headwords than in the GEPT word list, therefore a decision had to be made about which levels would constitute, jointly or not, the basic vocabulary. Level 1 has a baseline equivalent to Mogilner (1992) children's writer's word book (Jeng, 2002). Level 2 is in line with the ministry of education's 9 year education English list, less the ones already appearing in level 1. Level 1 and 2 together contain 2,160 headwords, which is a number close to the one of the Elementary level in the GEPT word list. Therefore, it was decided that one computation of the advanced Guiraud measures would be done with Level 1 as basic vocabulary and another with Level 1 and L2 together as basic vocabulary.

As indicated by findings in Chung and Wu (2009), an earlier study on the lexical variation of the same data, the writing samples on the topic about the writer's favourite exotic food contained food terms (e.g., 'sushi', 'kimchi') which do not exist in CEEW and the GEPT list. Moreover, the writing samples on the topic about the writer's favourite idol included many proper nouns (e.g., 'Jay', 'US'). Finally, an inspection of some of the writing samples by the third author indicated that some files included numerals and combinations of numerals and letters or punctuation marks (e.g., '50','PS2','7-11'), which do not appear in the word lists and, so, would be grouped under 'unfound words' unless processed separately. To avoid creating a wrong impression about the proportion that words found in each word list level, these words were located and were added to the word types which fell under the first level of each word list after word types in the data has been categorised among each word list's levels.

Data preparation

As mentioned earlier, all corpus samples were tagged using the CLAWS tagger (Garside & Smith, 1997). Figure 3 below gives an example of such tagged texts.

There_EX were_VBDR many_DA2 kind_NN1 of_IO food_NN1 in_II the_AT world_NN1 ,_, most_DAT of_IO them_PPHO2 were_VBDR delicious_JJ ,_, and_CC fast_JJ food_NN1 was_VBDZ my_APPGE favorite_NN1 ._.

Figure 3. The first sentence of a tagged writing sample

The tagged samples were processed through software in two stages. Figure 4 below gives an overview of these procedures.

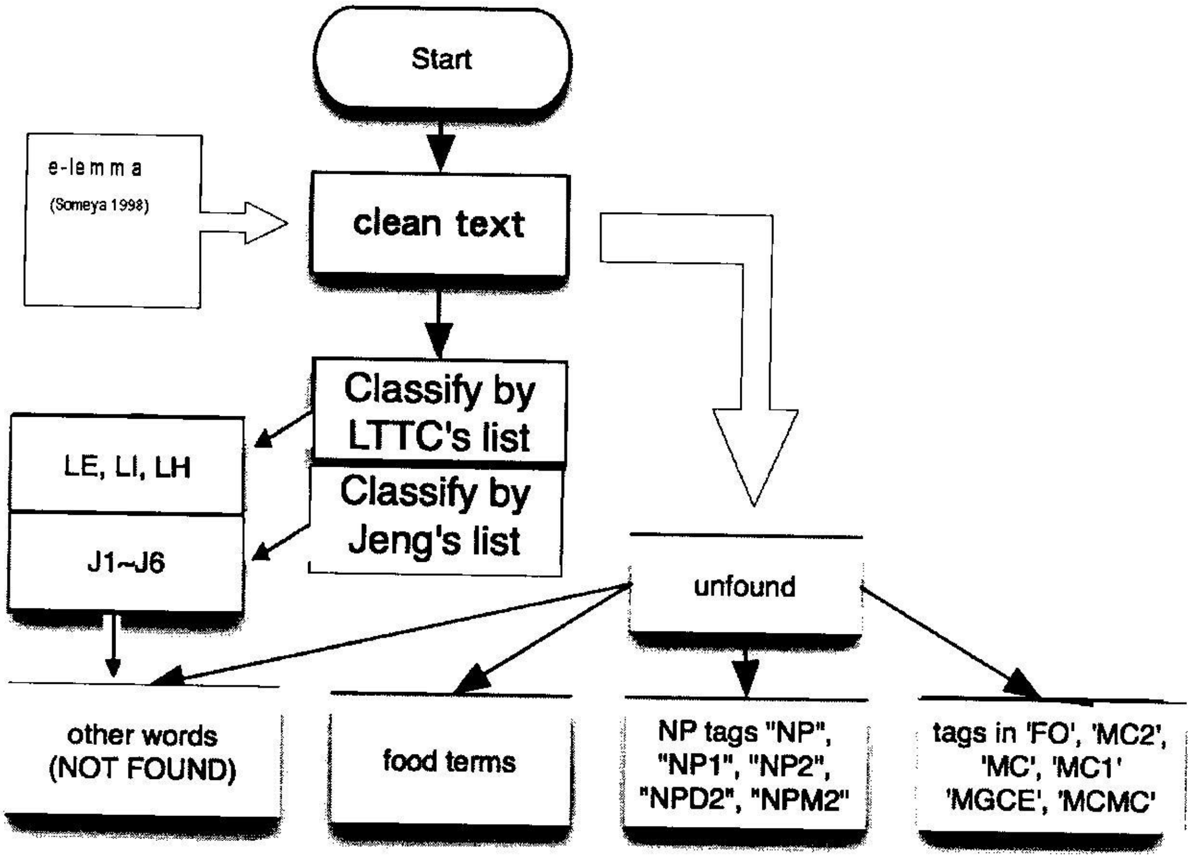
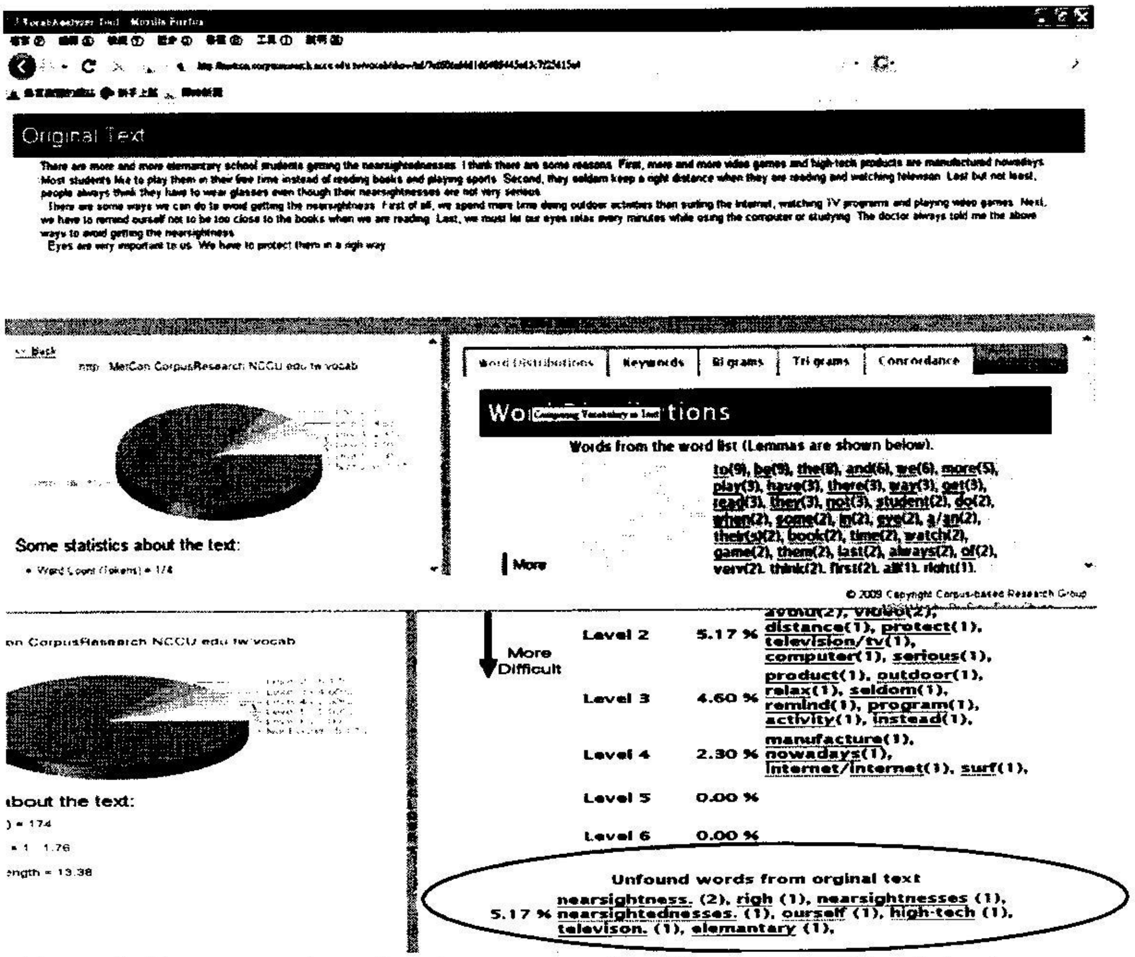


Figure 4. Software procedures conducted on the writing samples

In the first stage, before their tags were removed to be further classified according to the two word lists, three groups of data were first extracted, namely the food terms, numerals and proper nouns.

In the second stage, all tags were cleaned and the raw data were then lemmatized using e-lemma (Someya, 1998). The words in the cleaned writing samples were then classified according to each word list via the software VocabAnalyser (Chung, Chao, & Hsieh 2009). The two screenshots in Figure 5 show the output of this software for a sample writing sample.



There are more and more elementary school students getting the nearsightednesses. I think there are some reasons. First, more and more video games and high-tech products are manufactured nowadays. Most students like to play them in their free time instead of reading books and playing sports. Second, they seldom keep a right distance when they are reading and watching televison. Last but not least, people always think they have to wear glasses even though their nearsightnesses are not very serious.

There are some ways we can do to avoid getting the nearsightness. First of all, we spend more time doing outdoor activities than surfing the Internet, watching TV programs and playing video games. Next, we have to remind ourself not to be too close to the books when we are reading. Last, we must let our eyes relax every minutes while using the computer or studying. The doctor always told me the above ways to avoid getting the nearsightness.

Eyes are very important to us. We have to protect them in a righ way.

Some statistics about the text:

- Word Count (Tokens) = 174

Words from the word list (Lemmas are shown below).

to(9), be(9), the(8), and(6), we(6), more(5), play(3), have(3), there(3), way(3), get(3), read(3), they(3), not(3), student(2), do(2), when(2), some(2), in(2), eye(2), a/an(2), their(s)(2), book(2), time(2), watch(2), game(2), them(2), last(2), always(2), of(2), very(2), think(2), first(2), all(1), right(1).

© 2009 Copyright Corpus-based Research Group

on CorpusResearch.NCCU.edu.tw/vocab

about the text:
) = 174
= 1 1.76
ength = 13.38

		avoid(2), video(2),
Level 2	5.17 %	distance(1), protect(1), television/tv(1), computer(1), serious(1),
Level 3	4.60 %	product(1), outdoor(1), relax(1), seldom(1), remind(1), program(1), activity(1), instead(1),
Level 4	2.30 %	manufacture(1), nowadays(1), Internet/internet(1), surf(1),
Level 5	0.00 %	
Level 6	0.00 %	

Unfound words from orginal text

5.17 % nearsightness. (2), righ (1), nearsightnesses (1), nearsightednesses. (1), ourself (1), high-tech (1), televison. (1), elemantary (1),

Figure 5. Two screenshots showing an example of the output of VocabAnalyzer

When words were categorised according to CEEW, they were distributed across the six levels of the CEEW. When words were categorised according to the GEPT list, they were distributed across the Elementary, Intermediate and High-Intermediate levels of the list. In word categorisation according to each word list, words not found in a list were placed in a separate category.

The lists of the food terms, proper nouns, and numerals were proofread by the third author for two reasons. First, proper nouns and numerals were located using the CLAWS4 tags, but part-of-speech tagging through CLAWS is not 100% correct in the parsing of language learner data. Therefore, it was necessary to eliminate from the lists of proper nouns and numerals the words which were wrongly tagged as such. Second, a few of the food terms were already in the CEEW or GEPT list, so these words were eliminated from the food word list. Before calculating the Guiraud indices, these words were added to the first level of each of the two words lists, so that the number of words spuriously unfound in the lists would be diminished.

Results and Discussion

As mentioned above, the data were the 2,000 Intermediate GEPT writing samples which have been processed so far for the LTTC Learner Corpus. Nine of these files were empty, so the final number of files in the study is 1,991.

The writing samples had 264,285 tokens in total. The shortest text had 14 tokens and the longest 288. The mean number of tokens was 132.74 (*s.d.* 38.01). The number of word types differed depending on which word list was used to count the word types. According to the CEEW, total types were 150,682. Their mean was 75.68 (*s.d.* 17.68). According to the GEPT list, total types were 165,782. Their mean was 83.27 (*s.d.* 19.13).

Writing scores spanned from 1 to 5 and increased by 0.5 increments. The mean writing score was 3.391 (*s.d.* 0.89). Because there was only one case of score 1 (file 0519) and only two cases of score 1.5 (files 0981, 0991), scores were compiled into low (scores 1, 1.5, 2), mid (scores 2.5, 3, 2.5), and high (scores 4, 4.5, 5).

Table 3 shows the mean and standard deviation for all the versions of the Guiraud index used in this study.

Table 3
Mean and SD for Guiraud and advanced Guiraud indices

Lexical variation measure	*Mean*	*SD*
Guiraud with CEEW list	6.09	0.73
Guiraud with GEPT list	6.04	0.75
Advanced Guiraud with CEEW L1 as basic vocabulary	1.41	0.45
Advanced Guiraud with CEEW L1+L2 as basic vocabulary	0.84	0.34
Advanced Guiraud with GEPT list's LE as basic vocabulary	1.71	0.51

Table 4 shows the mean and standard deviation for all the versions of the Guiraud index used in this study per score band.

Table 4
Mean and SD for Guiraud and advanced Guiraud indices per score band

Lexical diversity measure	*Writing score bands*					
	Low		*Mid*		*High*	
	Mean	SD	Mean	SD	Mean	SD
Guiraud with CEEW list	5.50	0.70	5.99	0.67	6.44	0.63
Guiraud with GEPT list	5.37	0.72	5.92	0.69	6.40	0.64
Advanced Guiraud with CEEW L1 as basic vocabulary	0.31	0.11	0.34	0.10	0.38	0.11
Advanced Guiraud with CEEW L1+L2 as basic vocabulary	0.53	0.19	0.59	0.17	0.63	0.18
Advanced Guiraud with GEPT list's LE as basic vocabulary	0.84	0.24	0.87	0.20	0.86	0.21

As can be seen from Table 4, the means of all Guiraud indices increase as score bands increase.

The research questions of this study were all examined through the same statistical analyses. Therefore, all the statistical analyses will be presented together.

As a first step towards addressing the research questions, a one-way ANOVA was conducted where the data from all the Guiraud indices formed the list of dependent variables and the score band was the independent variable. Significant differences between pairs of levels of the score bands for at least one of the Guiraud indices, taken together with the rising trends in Guiraud index scores across score bands, would mean that it is likely that lexical variation contributes to holistic grading. The second part of the statistical analysis will correlate the scores in each of the Guiraud indices shown to differ significantly in the ANOVA analysis with the writing score bands. This analysis will show us how strongly lexical variation indices are associated with the writing score bands.

The distributions of the data from both simple Guiraud measures across writing score bands were normal, but those of all the advanced Guiraud measures were positively skewed. The latter were normalised through the log transformation so that they would constitute appropriate dependent variables at the ANOVA analysis and appropriate predictor variables in the correlation which would follow.

Table 5 shows that all Guiraud measures differed significantly across score bands except the log of the advanced Guiraud index based on the GEPT list. This agreement among nearly all the measures supports the validity of the finding, although, of course, using other kinds of lexical variation measures would help us make this inference with more confidence.

Table 5
F value and significance level for each of the Guiraud indices when compared by writing score bands in a one-way ANOVA analysis

Lexical variation measure	*F**	*Sig.*
Guiraud with CEEW list	215.51	0.000
Guiraud with GEPT list	240.07	0.000
Log of advanced Guiraud with CEEW L1 as basic vocabulary	51.78	0.000
Log of advanced Guiraud with CEEW L1+L2 as basic vocabulary	38.07	0.000
Log of advanced Guiraud with GEPT list's LE as basic vocabulary	2.13	0.119

Degrees of freedom were 2 between and 1988 within groups.

The post-hoc tests indicate that all the Guiraud indices which differ significantly across score bands also differ significantly across all pairings of the score bands. As the requirement for homogeneity of variance was met for the simple Guiraud indices, the LSD and Bonferroni tests were computed for them. All paired comparisons were significant at $p<0.0001$ for both post-hoc tests. In terms of the post-hoc tests in the ANOVA analyses with the advanced Guiraud indices, the assumption of homogeneity of variance was not met for the one based on CEEW's level 1, but was met for the other two. However, there were also large differences in group sizes (group 1: 245 cells, group 2: 985 cells, group 3: 761 cells). Since post-hoc tests which do not require homogeneity of variances are also recommended when there are large group size differences, tests which do not assume equal variances were also conducted for all Guiraud measures (both simple and advanced). Tamhane's T2 and Games-Howell tests were both significant for all paired comparisons at p<0.0001 except when advanced Guiraud was computed on the basis of the GEPT word list.

It is interesting to consider why the advanced Guiraud index based on the GEPT measure did not differ significantly across score bands. This finding may be related to the fact that while levels in CEEW are evenly distributed – each level containing 1,080 headwords – headword distribution across the three levels of the GEPT word list is not even. The Elementary level includes a smaller proportion of the GEPT list's headwords than the other two levels of the list, as shown in the last row of Table 2. Moreover, out of the 20.8% of GEPT headwords which were not found in CEEW, most of them are at the High-Intermediate level. If these words were removed from the GEPT list, perhaps the advanced Guiraud measure based on this list would behave similarly to the ones based on the CEEW. In general, the large proportion of advanced words in the GEPT word list may mean that it is more appropriate as the basis for qualitative lexical variation measures of advanced rather than low proficiency language samples.

Based on the findings from the ANOVA analysis, the second part of the statistical analysis correlated the scores from all Guiraud indices except the advanced GEPT Guiraud index. The correlation was a non-parametric one, Kendall's tau, because the writing score bands, a categorical variable, are the predicted variable. Table 6 shows the correlation coefficient and its significance for each of these Guiraud indices.

Table 6
Correlation of predictor Guiraud indices with writing score bands

Lexical variation measure	*r*
Guiraud with CEEW list	0.33**
Guiraud with GEPT list	0.34**
Log of Advanced Guiraud with CEEW L1 as basic vocabulary	0.17**
Log of Advanced Guiraud with CEEW L1+L2 as basic vocabulary	0.14**

**p < 0.0001

Table 6 indicates that the Guiraud indices which differ significantly among score bands in the ANOVA analysis account for small portions of the variance of the writing score bands. The highest correlations were those of the two simple Guiraud measures, but they were only moderate. The results of this correlation analysis suggest that lexical variation, at least as it is measured by these variants of the Guiraud index, does not play an important role in the holistic grading of these Intermediate GEPT writing samples.

Conclusion

Before summarising the conclusions which may be drawn from this study, it is necessary to point out the limitations of the study. First, many factors except lexical variation affect writing quality and, therefore, essay marking. Such factors are L2 proficiency (Sasaki and Hirose, 1996), L1 vocabulary knowledge (Schoonen et al., 2002, 2003), and L1 writing proficiency (Schoonen et al., 2003). This study does not control for them, so any possible overlap or interaction between lexical variation and these factors was not examined. Second, because of the large number of samples it was impossible to exclude lexical errors from the analysis. As indicated in Engber (1995), where lexical variation was calculated both before and after excluding lexical errors, perhaps if errors had been excluded, the correlations between the lexical variation indices and holistic grading would be higher. Finally, as other algebraic transformations of TTR, the Guiraud index has been shown to overcompensate for longer texts (Jarvis, 2002); therefore, it may have provided a wrong profile of lexical variation for longer texts. In later analyses of the data more reliable lexical variation measures will be used. Of course, this study also has the shortcomings that all the lexical variation studies which also examine lexical sophistication have. The meanings of polysemous words are ignored in the analysis and formulaic expressions are also ignored (Shaw and Weir, 2007: 55).

The results of the study indicate a positive but weak relationship between lexical variation as measured by simple and advanced Guiraud measures based on two Taiwanese English learner word lists. The fact that the correlations between the advanced Guiraud indices and the writing score bands were lower than those between the simple Guiraud indices and the writing score bands is difficult to interpret because of the unreliability of the Guiraud index across texts of different length. However, a plausible explanation may be that at a low-intermediate level of language proficiency as the one in the Intermediate GEPT examination, the use of rare words occurs rarely, so it is not something that distinguishes the high- from the low-performing learners in holistic grading. This explanation fits in well with the finding of Daller, van Hout and Treffers-Daller (2003). They compared the performance of simple and advanced Guiraud indices as a way to distinguish between two groups of fluent German speakers, German-dominant and Turkish-dominant German-Turkish bilinguals. In other words, if a lexical variation measure which takes into account how much advanced vocabulary is included in a text helps to distinguish between subgroups of highly proficient language speakers, then it is unlikely to be useful in distinguishing among levels of low language proficiency.

From a methodological point of view, the use of two word lists to calculate lexical variation measures and of two versions of the Guiraud index indicates that which word list is used in combination with which lexical variation measure may affect the results. Contrary to all results from other word list and Guiraud index combinations, the advanced Guiraud index based on the GEPT list did not differ significantly across writing score bands.

Finally, the use of both a purely quantitative and a qualitative-cum-quantitative version of the same lexical variation index in this study indicates that, contrary to expectation, both kinds of measures vary across writing score bands. Further research with different lexical variation measures needs to be done in order to examine whether these findings will be replicated.

Acknowledgements

We would like to thank the LTTC for giving us permission to use the GEPT word list in this study. We would also like to thank Mr August Chao for creating the software used in the data preparation procedures and Dr Zhao-Ming Gao for discussing aspects of the study with us. All errors remain our own.

References

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. *Journal of Second Language Writing*, 18, 191-208.

Bauer, L. and Nation, P. (1993). Word families. *International Journal of Lexicography*, 6(4), 253-279.

Carlson, S., Bridgeman, B., Camp, R. and Waanders, J. (1985). *Relationship of admission test scores to writing performance of native and nonnative speakers of English.* TOEFL Research Report No. 19. Princeton, NJ: Educational Testing Service.

Chung, S.-F, Chao F.Y. A. & Hsieh, Y.-C. (2009). VocabAnalyzer: A Referred Word List Analyzing Tool with Keyword, Concordancing and N-gram Functions. Poster presented at the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 23). Also appeared in Kwong Olivia (Ed.), *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation.* Hong Kong: City University of Hong Kong Press (pp. 638-645). December 3-5.

Chung, S.-F. & Wu, C.-Y. 2009. Effects of Topic Familiarity on Writing Performance: A Study based on GEPT Intermediate Test Materials. Presented at *The 2009 LTTC International Conference on English Language Teaching and Testing.* National Taiwan University, Taiwan. March 6-7.

Daller, H. & Phelan, D. 2007. What's in a teacher's mind? Teacher ratings of EFL essays and different aspects of lexical richness. In H. Daller, J. Milton, and Treffers-Daller, J. (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 234-44). Cambridge: Cambridge University Press.

Daller, H., van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics,* 24, 197-222.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing* 4: 139-155.

Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech, and A. McEnery (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 102-121). Longman, London.

Gunning, T. G. (2006). *Assessing and correcting reading and writing difficulties* (3th ed.). Boston: Pearson Education Inc.

Hellwig, B., van Uytvanck, D., & Hulsbosch, M. (2008). *EUDICO Linguistic Annotator (ELAN) version 3.6 manual.* Downloaded on 13 January 2009 from http://www.lat-mpi.eu/tools/elan/

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, 19, 57-84.

鄭恆雄，張郇慧，程玉秀，顧英秀。2002。《大學入學考試中心高中英文參考詞彙》。台北：財團法人大學入學考試中心基金會。[Jeng, H, Chang, H, Chen, Y.H., Gu, Y.H. (2002). *English Reference Wordlist for the College Entrance Examination in Taiwan.* Taipei: College Entrance Examination Center.]

Laufer, B. (1995). Beyond 2000 - a measure of productive lexicon in a second language. In L.Eubank, M. Sharwood-Smith, & L.Selinker. (Eds.), *The Current State of Interlanguage* (pp. 265-272). Amsterdam: Benjamins.

Laufer, B. and Nation, P. (1995). Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics,* 16, 307-322.

Lorenzo-Dus, N. & Meara, P. (2005). Examiner support strategies and test-taker vocabulary. *International Review of Applied Linguistics in Language Teaching,* 43(3), 239-258.

MacWhinney, B. (2008). *The CHILDES project. Tools for analyzing talk – Electronic version. Part 1: The CHAT transcription format.* Downloaded on 13 January 2009 from http://childes.psy.cmu.edu/manuals/chat.pdf

Malvern, D. & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity, *Language Testing*, 19(1), 85–104.

Meara, P. & Bell, H. (2001). P_lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect,* 16: 5-19.

Milton, J. (2009). *Measuring second language vocabulary acquisition.* London: Multilingual Matters.

Mogilner, A. (1992). *Children's Writer's Word Book.* Ohio: Writer's Digest.

Morris, L. and Cobb, T. 2004. Vocabulary profiles as predictors of the academic performance of Teaching English as a Second Language trainees. *System,* 32, 75-87.

Nation, P. (2004). A study of the most frequent word families in the British National Corpus. In P.Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 1-13). Amsterdam: John Benjamins.

Richards, B.J. (1987). Type/token ratios: what do they really tell us? *Journal of Child Language,* 14, 201-209.

Sasaki, M., & Hirose, K. (1996). Explanatory variables for EFL students' expository writing. *Language Learning,* 46, 137–174.

Schmitt, N. (2005). *Lexical resources in Main Suite examinations.* Cambridge: UCLES internal report.

Shaw, S. & Weir, C.J. (2007). *Examining Writing in a Second Language.* Studies in Language Testing 26. Cambridge University Press and Cambridge ESOL.

Schoonen, R., van Gelderen, A., de Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language Learning,* 53, 165–202.

Schoonen, R., van Gelderen, A., de Glopper, K., Hulstijn, J., Snellings, P., Simis, A., & Stevenson, M. (2002). Linguistic knowledge, metacognitive knowledge and retrieval speed in L1, L2, and EFL writing. In S. Ransdell & M.-L. Barbier (Eds.), *New directions for research in L2 writing* (pp. 101–122). Dordrecht, Netherlands: Kluwer Academic.

Someya, Y. (1998). E-lemma list. (free online software available at http://www.lexically.net/downloads/version4/downloading%20BNC.htm)

Su, C.-C. (2006). A Preliminary Study of the 2000 Basic English Word List in Taiwan. *The Proceedings of the 23rd International Conference on English Teaching and Learning in the Republic of China* (pp. 1002-1016). Taipei: Kaun Tang.

Tweedie, F.J. & Baayen, R.H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and theHumanities,* 32, 323–352.

Wolcott, W. & Legg, S. M. (1998). *An overview of writing assessment: Theory, research and practice.* Urbana: National Council of Teachers of English.

Yu, G. (forthcoming) Lexical diversity in writing and speaking task performances. *Applied Linguistics* (advanced online access on June 4 2009). doi:10.1093/applin/amp024