An Experiment in the Ability of Raters to Evaluate Lexis in Writing

Rachael Ruegg

Akita International University, Japan

Abstract

For many years, people have considered lexis and grammar separately in the context of teaching and learning English. In the assessment of second language writing, lexis and grammar continue to be considered separately. However, recent corpus studies have questioned this approach and argued that lexis and grammar are fundamentally inseparable. While the assessment of lexis and grammar as two distinct qualities lends face validity to assessment criteria, the corpus literature suggests that raters may not be able to accurately distinguish the two. The current study examines the ability of raters to separate lexis and grammar when using an analytic rating scale to assess timed essays. In this experiment, the lexical content of 27 essays was manipulated before rating in order to determine the effect of lexical accuracy, lexical variation and lexical richness on lexis and grammar scores. From the results, it seems that raters are sensitive to lexical accuracy, but not lexical variation nor lexical richness. In addition, the manipulation of lexical qualities had a significant effect on grammar scores but not on lexis scores, supporting the idea that raters find it challenging to distinguish lexis from grammar.

1. Introduction

For many years, people have considered lexis and grammar separately in the context of the teaching and learning of English. However, recent corpus studies (see Biber, Conrad, & Cortes 2004; Gries & Stefanowitsch, 2004; Hoey, 2005; Hoey & O'Donnell, 2008; Hunston, 2008; Römer, 2009; Sinclair, 2004) have questioned this traditional approach and argued that lexis and grammar are fundamentally inseparable.

Within the context of language assessment, in the assessment of writing, lexis and grammar continue to be considered separately. Many assessment criteria, both those used in tests of English writing and those developed for use in classroom assessment mention lexis and grammar as two separate qualities to evaluate. (e.g. IELTS (n.d.), The ESL composition profile (Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughey, 1981), The ELTT rating scale for writing (Austrian University ELTT Group, n.d.), TOEFL iBT (ETS., n.d.)

Discussion with a large number of colleagues suggests that classroom instructors also tend to consider lexis and grammar separately. While the assessment of lexis and grammar as two distinct qualities lends face validity to assessment criteria, the corpus literature (e.g. Hoey, 2005; Römer, 2009) suggests that raters may not be able to accurately distinguish the two. Therefore, continuing to assess lexis and grammar separately may decrease the reliability of assessment instruments. Very little research has been published attempting to ascertain whether raters or classroom instructors are able to distinguish between lexis and grammar. Indeed, little has even been published discussing where this distinction might lie.

In a previous study (Ruegg, Fritz & Holland, 2011), it was observed that raters had difficulty separating lexis and grammar when using an analytic rating scale to assess timed essays. The current study further examines this issue by using an experimental design. It was considered by the researchers that an experimental design would provide stronger results than the observational study because of the ability to control the content of the essays rated. The purpose of this experimental study is to confirm the results of the previous observational study. The research questions for the current study are:

1. When rating timed essays that vary greatly in lexical quality are raters more sensitive to accuracy, variation, or richness of words used when rating essays for lexis and grammar?
2. Do variations in lexical accuracy, variation and richness have a stronger effect on lexis scores or grammar scores?

2. Review of Literature

2.1 Defining lexis

Although there has been an abundance of research on lexis, few studies have attempted to define lexis. A prime example of this is in a study by Bardovi-Harlig and Bofman (1989) which investigated the relationship between syntactic complexity and overall accuracy in writing. They identified three different categories of errors: syntactic (e.g. word order), morphological (e.g. preposition errors), and lexical-idiomatic (i.e. vocabulary). However, although they defined the first two categories, they failed to explain what was included in the third category.

Several studies have attempted to define vocabulary. For the purpose of his study, Chastain (1990) defined vocabulary errors as “[a] a wrong word; [b] a missing word; or [c] an extra word” (p. 11). All of these errors could be classified as word choice errors, whereas none of them pertain to word formation. Kobayashi and Rinnert (1992) also took into account word choice and not word formation in their definition of lexis. They identified three kinds of vocabulary errors: word choice, awkward formation (considered at the sentence level rather than at the word level) and transitional problems.

For the purpose of this study, lexis constituted three separate qualities: lexical accuracy (number of content words used accurately), lexical variation (number of types used) and lexical richness (the average frequency level of words used).

2.2 Assessing Lexis

Although assessing both lexis and grammar separately is widespread, there are few studies investigating the validity or reliability of such assessments. A number of studies

have examined the impact vocabulary plays in raters' assessments of writing (see Engber, 1995; Laufer & Nation, 1995; Linnarud, 1986). However, Engber (1995) and Linnarud (1986) both looked at holistic ratings rather than at specific ratings for lexis. Santos (1988) found that university instructors rated lexical errors as the most serious type of errors and suggested that more focus should be placed on word choice within instruction of English as a Second Language.

The present study aims to ascertain what effect variation in lexical accuracy, lexical variation and lexical richness has on ratings of timed writing using analytic rating scales for lexis and grammar (see appendix A).

2.3 Distinguishing lexis and grammar

Distinguishing between lexis and grammar is notoriously difficult. Some items appear quite clearly describable as belonging to the lexis category, such as spelling and word choice. Some seem to be equally clearly describable as belonging to the grammar category, such as article usage and sentence structure. On the other hand, a vast majority of items could be classified into either category, or both, depending to a large extent on who was doing the categorising. Few studies have attempted to draw a distinction between lexis and grammar, possibly because finding agreement between sources on such a distinction is very challenging. In reality, it appears that there is not one clear distinction to be drawn between lexis and grammar but rather that some items are more lexical in nature, while others are more grammatical. Thinking about lexis from this perspective, Halliday (2004) describes his notion of a "lexico-grammar cline". Halliday (2004: 64) describes lexis at one end of the cline as 'open sets, specific in meaning' and grammar at the other end as 'closed systems, general in meaning'. In between the two extremes fall a number of items such as prepositions, adverbs and conjunctions. Whether these are more grammatical or lexical in nature depends on who is doing the classifying.

Recent corpus linguistics studies suggest that lexis and grammar are fundamentally inseparable. Hoey (2005) proposes his theory of the lexicon, called 'lexical priming', which is based on corpus findings about the way words interact with each other as well as common patterns of use. He states in his book that "the theory reverses the roles of lexis and grammar, arguing that lexis is complexly and systematically structured and that grammar is an outcome of this lexical structure" (p. 1). Similarly, Sinclair (1991) states that the way words can be used is more fixed than the way grammar can be used. Likewise, Gries and Stefanowitsch (2004) investigated like lexical constructions such as "the university's budget" and "the budget of the university" and found that sometimes "each of the two members...is a construction in its own right with its own meaning" (p.124). Römer (2009) concludes that "studies, based on large collections of authentic text from a range of different sources, have provided massive evidence for the interdependence of lexis and grammar" (p.140).

In a previous study, Ruegg, Fritz and Holland (2011) investigated raters' sensitivity to different lexical qualities when rating timed essays using the same analytic scale used in the present study. They found that although lexical accuracy did significantly predict lexis scores, grammar scores predicted lexis scores much more strongly than any other variable. Although two separate analytic rating scales were used to assess lexis and grammar, in 69.37% of ratings the same score was given for both scales. This begs the question; are raters able to distinguish lexis from grammar? The purpose of the current study is to ascertain the ability of raters to distinguish vocabulary from grammar when rating writing for

both lexis and grammar using analytic rating scales when the writing varies greatly in lexical quality.

2.4 Experimental studies on the assessment of writing

A study by Freedman (1979) used an experimental design to ascertain the extent to which raters take content, organisation, sentence structure and mechanics into account when rating writing. The essays in her study were manipulated to be stronger or weaker in terms of these four categories. Following this, raters were asked to rate the essays in terms of the same four categories. It was found that raters commented the most on sentence structure and mechanics, whereas content and organisation affected the scores they assigned more.

A more recent study by Fritz and Ruegg (2013) also used an experimental design, manipulating the vocabulary of essays before asking raters to assess their lexical quality. Their study looked specifically at the influence of different lexical qualities on raw scores assigned for lexis when rated using an analytic rating scale. A similar research method will be employed in the present study, manipulating essays before asking raters to evaluate them. However, this study investigates the interplay between lexis and grammar which was identified by Ruegg, Fritz and Holland (2011) in their observational study, using an experimental research design.

3. Method

This study was conducted at a foreign language university in Eastern Japan. Every year the members of the in-house English proficiency test collaborative research group decide on the writing prompt for the test, which is taken by all students in their first and second year of study in all but one department within the university. In June of 2009 the following prompt was chosen for the writing section of the test:

"Eating meat is bad for your health and raising animals just to eat causes them to suffer. Sheep, pigs and chickens spend their lives in terrible conditions until they are killed for the supermarket shelves. You can get everything you need for a healthy diet without meat. Therefore, more people should become vegetarian." **Give your reaction to the above statement and support your answer with specific reasons and examples.**

After the prompt had been decided upon, a student at the university was asked to write a 30 minute essay based on the prompt. A second year student from the department in which students do not take the test was selected to write the sample essay. Students in the different departments within the university are at similar proficiency levels, therefore it was assumed that this essay would be comparable to actual test essays, which were written by students at the end of their first and second years of study.

According to Gao (2013: 73), function words "play a grammatical role in a sentence, such as conjunctions, prepositions and adverbs". Nation (2001: 430-431) offers a list of function words of the English language. For the purpose of investigating the lexical content of the sample essay, first, all function words (defined as those appearing in Nation's (2001) list) were removed. Words which appeared in the prompt were also removed because in the test booklet it is stated that examinees will not be given credit for the use of words which appear in the prompt and during the rater training raters are also trained not to give credit for those words. After both function words and the words that appeared in the prompt had been removed, the sample essay contained thirty-two content words; these are the words that were

manipulated in order to create the essays for this study. The original essay written for this research can be seen in appendix B, with the 32 content words underlined.

In order to create essays of varying lexical quality, three different types of manipulations were performed: manipulations of lexical accuracy, lexical variation and lexical richness. Low, medium and high levels were determined for all three lexical qualities and an equal number of essays were created at each level for each lexical quality. An essay with every possible combination of lexical qualities was used when creating the essays ranging from low accuracy, low variation and low richness, to high accuracy, high variation and high richness. There are 27 different possible combinations of lexical qualities and there were therefore 27 manipulated essays in the study.

Ruegg, Fritz and Holland (2011) suggest four different types of lexical errors. For the purpose of this research, two of the types of errors identified by Ruegg Fritz and Holland (2011) were selected for this study in order to create essays of varying lexical accuracy: 1) using the word out of context; and, 2) using the wrong part of speech. When manipulating the accuracy of lexis, errors related to either word choice or to part of speech. Many would argue with the classification of the wrong part of speech as a lexical error, rather than a grammatical one. However, Nation (2001) states that there are three concepts involved in knowing a word: word form, word meaning and word use. Both of the error types used in this study fall within Nation's (2001) definition of word meaning.

The following is an example of a word choice error (using a word out of context):
Original Sentence: First I think there are trouble people who are keeping animals.
Manipulated Sentence: First I think there are trouble people who are managing animals.

An example of a part of speech error follows:
Original Sentence: First I think there are trouble people who are keeping animals.
Manipulated Sentence: First I think there are trouble people who are keeper animals.

Essays with low accuracy consisted of 32 inaccurate content words. Essays with medium accuracy consisted of 16 inaccurate content words and 16 accurate content words. Essays with high accuracy consisted of 32 accurate content words. The essay with the lowest overall lexical quality (low variation, low richness, low accuracy) can be seen in appendix C. The essay with the highest overall lexical quality (high variation, high richness, high accuracy) can be seen in appendix D. A different student was employed to handwrite the 27 manipulated versions of the essay on official test paper.

In creating essays of varying lexical variation, the 32 content words were considered in terms of how many were the same or similar in meaning. Out of the 32 words, 18 different meanings were expressed, while the other 14 were the same or similar in meaning to those 18. For example; different variations of the words 'believe', 'know' and 'think' made up seven of the 32 content words. Without changing the meaning of the essay considerably, it would be possible to replace these seven words with just one word. In this way, it was found that 18 different words were needed to maintain the original meaning of the essay. On this basis, essays with low lexical variation had 18 different content words. Essays with high lexical variation had 32 different content words. Essays with medium lexical variation each had 25 different content words, in order to place them halfway between the lexical variation of the low essays and that the high essays. In order to increase lexical variation, an online thesaurus (www.thesaurus.com) was used to find synonyms, or near synonyms, for the 32

content words and ones that would not seem out of context or change the original meaning of the essay was chosen.

In creating essays of varying lexical richness, RANGE software (Nation, 2005) was used to analyse all 32 content words and their synonyms. Words at the 1,000 word level, the 3,000 word level and above the 3,000 word level (listed as ‘not in the lists’ in the RANGE output) were chosen to replace the content words. An essay with low lexical richness contained 32 content words which all came from the 1,000 word level. An essay with medium lexical richness contained 32 content words which all came from the 3,000 word level. An essay with high lexical richness contained 32 content words which all came from above the 3,000 word level.

Raters of the writing section of the test are assigned a random selection of essay numbers to rate for each test administration. For the January 2010 administration, the essays created for this study were assigned nonexistent essay numbers and then also randomly assigned. In this way, the raters were not made aware of the fact that these were not real student essays. Each rater was assigned 39 real test essays along with three of the essays created for this study to rate. To ensure that two of the essays created for this study would not be rated sequentially, the essays for this study were placed at the beginning, in the middle and at the end of the list of assigned essays.

In total there were 28 raters of writing for the January 2010 administration of the test, however, one of the raters was the researcher conducting the present study. For this reason, 27 raters rated the essays for this study. The 27 raters were all EFL instructors at the university where the research was conducted with at least a Master’s degree in TESOL, Linguistics or a related field. Almost 60% of the raters (16) were males and just over 40% (11) were females. They were from the Australia, Bulgaria, Canada, England, Ireland, Jamaica, Japan, New Zealand, Scotland and the United States. A rater norming session was conducted within a week prior to the test administration and all raters were required to attend the session. The main purpose of the rater norming session was to increase interrater reliability. At the discussion-based session, each analytic rating scale was explained before sample essays were rated and discussed. For example, the analytic rating scale for lexis asks raters to consider variety and control, so during the rater norming session raters were told what variety and control mean in the context of vocabulary.

Immediately after the test administration the essays created for this study were combined with the actual test essays for rating. There were 895 real test essays. All 922 essays were rated using the KEPT analytic rating scales. The real test essays were double rated, the data was cleaned and scaled using many-facet Rasch modelling in order to provide scores to students. Following this, the data from the 27 essays created for this study was cleaned. Although it was intended for each essay to be rated 3 times, the final number of ratings for each essay after cleaning ranged from 1 to 4 because of rater error and some questionable ratings being excluded. In total there were 72 ratings used to calculate the 27 essays’ scores. The data was then scaled together with the 895 test essays using many-facet Rasch modelling. Rasch modelling adjusts scores taking into consideration the comparative strictness or leniency of each rater for each rating scale. This increases both the inter-rater and the intra-rater reliability of the test.

After the essay data was scaled, multivariate analysis of variance (MANOVA) was employed with the Rasch adjusted lexis and grammar scores as the dependent variables and

lexical variation, lexical richness and lexical accuracy as the independent variables. The lexical variation scores represented the number of types in each essay, so essays with low lexical variation had 65 different types, essays with medium lexical variation had 71 types and those with high lexical variation had 78 types. The lexical richness scores represented the average frequency in the English language of the words used in each essay. Essays with low lexical richness had a lexical richness score of 1, indicating that all of the words in the essay came from the 1,000 word level. Essays with medium lexical richness had a richness score of 1.54, because the score included function words (which came from the 1,000 word level) in addition to 32 content words from the 3,000 word level. Essays with high lexical richness had a lexical richness score of 3.17, which demonstrates the large number of low frequency words those essays contained. The lexical accuracy scores represented the number of content words used accurately in each essay. Therefore, essays with low lexical accuracy had a lexical accuracy score of 0, essays with medium lexical accuracy had a score of 16 and those with high lexical accuracy had a score of 32, indicating that all content words were used correctly in terms of the correct part of speech being used in the correct context. Since all of the manipulations were to the lexis rather than the grammar of the essay, the analytic rating scale suggests that we should expect no significant relationship between the grammar scores and the lexical variation, lexical richness or lexical accuracy.

4. Results and Discussion

The descriptive statistics for each variable can be seen in Table 1. The descriptive statistics include the number of cases, as well as the mean, standard deviation, skewness values and kurtosis values. The possible lexis scores which could be assigned according to the analytic rating scale for lexis range from 0 to 4 points. The lexis scores assigned to the manipulated essays ranged from 1.08 to 3.21 and the standard deviation was 0.6048. The possible grammar scores which could be assigned according to the analytic rating scale for grammar also ranged from 0 to 4. The grammar scores assigned to the manipulated essays ranged from 0.71 to 2.99 and the standard deviation was 0.5967. Skewness and kurtosis values were measured in order to determine whether the data met the assumption of normal distribution. The skewness values ranged from -0.015 to 0.561 and the kurtosis values ranged from -1.560 to -0.377. According to George and Mallery (2010), Skewness and Kurtosis values between -2 to +2 indicate acceptable measures.

Variable	Mean	SD	N	Skewness	Kurtosis
Lexis	2.1237	0.60480	27	0.234	-0.525
Grammar	1.8500	0.59669	27	-0.015	-0.377
Accuracy	16.0000	13.31280	27	0.000	-1.560
Variation	71.3333	5.41366	27	0.099	-1.560
Richness	1.9033	0.93997	27	0.561	-1.560

Table 1. Descriptive Statistics

The results of the MANOVA show that lexical accuracy had a significant effect on the dependent variables at the 0.05 level; Wilks' Lambda = 0.650, $F(2, 22) = 5.932$, $p = 0.009$. On the other hand, lexical variation (Wilks' Lambda = 0.913, $F(2, 22) = 1.045$, $p = 0.369$) and lexical richness (Wilks' Lambda = 0.877, $F(2, 22) = 1.539$, $p = 0.237$) had no significant effect on the dependent variables. This shows that despite accuracy and range

both being mentioned on the analytic rating scale for lexis and during rater training, raters are affected to a significantly larger extent by accuracy of lexis than they are by variation or richness of words used.

Independent Variable	df	F	p	η^2
Accuracy	2	4.262*	.018	.111
Richness	2	1.680	.194	.047
Variation	2	.069	.933	.002

* $p < .05$

Table 2. Between-Subjects Factors ANOVA (Lexis scores)

Furthermore, the tests of between subjects effects found that the effect of lexical accuracy on grammar scores was strongly significant; $F(1) = 12.256$, $p = 0.002$, whereas the effect of lexical accuracy on lexis scores did not reach the level of significance; $F(1) = 4.233$, $p = 0.051$. All the manipulations that were carried out were changes in lexical quality rather than grammatical changes. Considering that in this study lexis was manipulated, a wider range would be expected in the lexis scores than in the grammar scores. Moreover, these changes should lead to more variety in lexis scores and have a lesser effect on grammar scores. The results of the tests of between subjects effects showed, however, that the opposite was true. It is clear from these results that while lexical accuracy does have a significant effect on scores assigned for lexis and grammar, it has a significantly stronger effect on grammar scores than on lexis scores.

5. Limitations and suggestions for further research

Although statistically strong results were found, suggesting that asking raters to assess lexis and grammar separately may be problematic, there are several limitations with the current study that need to be taken into consideration. In addition, further research would be beneficial.

If raters have any concerns about the essays they are rating, they are instructed to first of all rate them as they usually would and then let the rating room supervisor know of their concerns. Some of the raters became suspicious that cheating might have occurred when rating the essays created for this study. There are several possible reasons for this suspicion. It was decided for each rater to rate three different essays created for this research in order to increase the total number of ratings. However, as raters were only required to rate 39 real test essays which only took around half a day, three similar essays was probably too many to read in a limited period of time and may have aroused suspicion. In future research it is suggested that the total number of real test essays to be rated be carefully considered before deciding how many research essays each rater should rate. Although it is ideal to have three separate ratings for each essay for research purposes, in this case two ratings per essay may have avoided this suspicion and prevented problems.

Another possible reason for the raters' misgivings was that in January 2010 the test was administered differently from previous administrations, because of this some real examinees cheated on the writing section of the test by plagiarising the writing of the person

sitting next to them. Finding that some examinees had cheated may have made raters more alert to possible cheating while they were rating.

In addition to this, it is possible that the manipulations made to the lexical qualities of the essays were too advanced for the test population. On the other hand, the grammar in the essay was of a fairly low level. The disparity between the high lexical quality of some essays and the low grammatical quality may have aroused suspicion in some raters.

The raters' misgivings are not considered to have affected the ratings in this study as the few ratings that seemed dubious were deleted from the sample. However, in order to prevent raters becoming suspicious of essays during this kind of research, effort should be made to limit the number of manipulated essays which need to be rated by each rater, ensure that real examinees do not have opportunities to cheat during the exam and not to manipulate certain qualities in essays too extremely.

Word choice may be considered to be close to the lexis end of the lexico-grammar cline, on the other hand, word formation may be further towards the grammar end. In this study, the characteristic associated with word formation was part of speech. It is possible that some raters consider part of speech to be grammatical rather than lexical knowledge and this may be one reason that errors relating to lexical accuracy affected grammar scores more than lexis scores. However, Nation (2001) and others in the field of vocabulary consider that knowing the different parts of speech is one component of vocabulary knowledge.

In this study, raters did not receive any specific training in relation to the distinction between lexis and grammar. It would be beneficial to train raters specifically in the distinction between lexis and grammar during rater training sessions. Specifically, terms such as 'word formation' should be explained in detail to ensure that raters go into the rating process with the same concept in mind. Similar research to this, or an observation such as the one carried out by Ruegg, Fritz and Holland (2011), which was carried out after such a detailed training session would be a valuable addition to the field.

6. Conclusions

Although in the rating scales it is stated that raters should consider lexical variation (referred to as 'variety' in the rating scales) and lexical accuracy (referred to as 'control' in the rating scales) when rating writing for lexis, neither lexical variation nor lexical accuracy significantly predicted the scores given for lexis in this study. In addition to this, it was considered by the researchers that raters might be more sensitive to the frequency level of words used (lexical richness). However, it is clear that raters in this study were not sensitive to any of these lexical qualities. The strong extent to which lexical accuracy affected grammar scores begs the question whether raters are able to distinguish between lexis and grammar when rating writing.

The fact that lexical accuracy affected grammar scores so strongly suggests that it might be beneficial to collapse the two rating scales into a single lexico-grammar scale. The correlation between lexical accuracy and grammar scores was much stronger than that between lexical accuracy and lexis scores. Furthermore, lexical variation and lexical richness did not have a significant effect on lexis scores. In addition to this, the grammar of the essays in this study was not manipulated in any way, but the range of grammar scores was slightly wider than that of lexis scores. Under these circumstances, it seems that raters are unable to distinguish lexical quality from grammatical quality.

The purpose of this experiment was to confirm the findings of a previous observational study by Ruegg, Fritz and Holland (2011). The results of the current study seem to confirm the finding that it is challenging for raters to distinguish lexis from grammar. In writing tests such as this one, as well as in classroom assessment, ongoing research is required to verify whether the ratings given to language learners are valid and assessment practices need to be constantly evaluated in view of such research to improve measurement validity.

References

- Austrian University ELTT Group. (n.d.). *Austrian University ELTT rating scale for writing*. Retrieved July 19, 2007, from http://www.uni-klu.ac.at/ltc/downloads/ELTT_Writing_Scale.pdf
- Bardovi-Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition*, 11(1), 17-34.
- Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 2, 371-405.
- Chastain, K. (1990). Characteristics of graded and ungraded compositions. *The Modern Language Journal*, 74, 10-14.
- Engber, C. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, 4(2), 139-155.
- ETS. (n.d.). *For test-takers: TOEFL Paper-based Test (PBT): Writing score guide*. Retrieved July 19, 2007, from <http://www.ets.org>
- Freedman, S. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology*, 71 (3), 328-338.
- Fritz, E. & Ruegg, R. (2013). Rater sensitivity to lexical accuracy, sophistication and range when assessing writing. *Assessing Writing*, 18, 173-181.
- Gao, J. (2013). Basic cognitive experiences and definitions in the Longman Dictionary of Contemporary English. *International Journal of Lexicography*, 26(1), 58-89.
- George, D. & Mallery, P. (2014). *IBM SPSS statistics 21 step by step: A simple guide and reference*. New Jersey: Pearson Education.
- Gries, S. & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, 9(1), 97-129.

- Halliday, M. (2004). *An introduction to functional grammar* (3rd ed.). London: Hodder Arnold.
- Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.
- Hoey, M. & O'Donnell, M. (2008). Lexicography, grammar, and textual position. *International Journal of Lexicography*, 21(3), 293-309.
- Hunston, S. (2008). Starting with the small words: Patterns, lexis, and semantic sequences. *International Journal of Corpus Linguistics*, 13(3), 271-295.
- IELTS. (n.d.). *An overview of IELTS academic writing*. Retrieved July 19, 2007, from http://www.cambridgeesol.org/teach/ielts/academic_writing/aboutthepaper/overview.htm
- Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfiel, V., & Hughey, J. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.
- Kobayashi, H., & Rinnert, C. (1992). Effects of first language on second language writing: Translation versus direct composition. *Language Learning*, 42(2), 183-215.
- Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307-321.
- Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. Malmo, Sweden: Liber Forlag Malmo.
- Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
- Nation, I. S. P. (2005). Range and frequency: Programs for Windows based PCs. [Computer software and manual]. Retrieved June 3, 2008, from <http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx>
- Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, 7(1), 141-163.
- Ruegg, R., Fritz, E. & Holland, J. (2011) Rater sensitivity to qualities of lexis in writing. *TESOL Quarterly*, 45(1), 63-80.
- Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly*, 22(1), 69-90.
- Sinclair, J. (1991). *Corpus concordance collocation*. Oxford: Oxford University Press.
- Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.

Appendices

Appendix A

KEPT Essay Rating Scales

	Organization Think about: Coherence Structure	Lexis Think about: Variety Control	Grammar Think about: Range Accuracy	Content Think about: Relevancy to the main idea Ideas that are supported and developed
0	No coherence or organisation, unconnected sentences which communicate little.	Demonstrates minimal word knowledge.	Phrases or sentences produced, but many inaccuracies make message/writing difficult to understand	A list of sentences with no logical connection and/or are irrelevant.
1	Some attempts to organise information but with little connection between ideas apparent	A limited variety of vocabulary with little control.	Inadequate range of grammar used repetitively or inaccurately.	Ideas are connected but not relevant, developed or supported
2	Obvious attempts to organize information though sometimes the lack of coherence creates ambiguity.	Uses an adequate variety of vocabulary with moderate control.	An adequate range of grammar used, with inaccuracies that impede the understanding of sentences.	Ideas are connected, relevant, but not supported or developed.
3	The writing displays an organizational structure which enables the message to be followed although sometimes the lack of coherence might create ambiguity.	Uses a wide variety of vocabulary but there are inaccuracies in word choice and formation.	An adequate range of grammar but occasionally accuracy affects the understanding of sentences.	Ideas are connected and relevant. They are supported but the main idea is not developed.
4	The writing displays a coherent organizational structure which enables the message to be followed effortlessly.	Uses a wide variety of vocabulary, accurately and with control.	A wide range of grammar used accurately.	The ideas are relevant, well supported and developed.

Appendix B

Original essay

I think we should not become vegetarian. I have two reasons. First, I think there are trouble people who are keeping animals to eat for us if I stop eating meats. Their working is to grow them, sell them and get money. If I stop to eat meats their work must damage. Second reason is about my health. I have known that meats need our human body because meats make our blood and support our body, so only vegetable is not good for our health. These reasons support my opinion which I think we should eat also meats. I know that animals to eat by human have not been kept carefully. I think the keeper should give more love for them, and also many people should think about them when they eat them by having thankful mind.

Appendix C

Essay with the lowest lexical quality

I think we should not become vegetarian. I have two reasons. First, I think there are trouble people who are keeper animals to eat for us if I stop eating meats. Their working is to grow them, sale them and get money. If I stop to eat meats their working must struggling. Second reason is about my health. I have thought that meats need our humane bodily because meats make our bloody and support our bodily, so only greens is not well for our health. These reasons support my think which I think we should eat also meats. I think that animals to eat by humane have not been kept careful. I think the keeping should give more lovely for them, and also many people should think about them when they eat them by having please minding.

Appendix D

Essay with the highest lexical quality

I reckon we should not become vegetarian. I have two reasons. First, I surmise there are impaired people who are nurturing animals to eat for us if I refrain from eating meats. Their endeavor is to nourish them, merchandise them and get cash. If I desist to eat meats their toil must ravaged. Second reason is about my health. I have comprehended that meats need our mortal torso because meats fortify our hemoglobin and support our anatomy, so only vegetable is not advantageous for our health. These reasons support my conviction which I deem we should eat also meats. I envisage that animals to eat by hominid have not been reared compassionately. I foresee the breeder should give more affection for them, and also many people should envision about them when they eat them by having gratified disposition.