A Summary of ‘A Comparative Statistical Assessment of Different Types of Writing by Japanese EFL College Students’

Brian MCNEILL

1.0 Introduction

This paper represents only a brief summary of the contents of my PhD thesis (McNeill, 2006). Future papers will develop these results in more detail, and I hope to elaborate much more on the findings.

The approach of the thesis was to investigate a dichotomy which appears to exist in the literature on corpus research using statistical relationships of text features: while one branch of this research says different types of text are ‘different’, particularly narrative and non-narrative texts, and much has been written on the nature of these differences (e.g. see Biber, 1989 or Grabe, 1987); when other research has been done on statistical relationships of text features this ‘difference’ has been ignored and texts of different types of writing have been combined together in the same target corpus. The main examples given of this problem were taken from prominent sources: Polio (1997) (from *Language Learning*), Ortega (2003) (from *Applied Linguistics*), and Wolfe-Quintero, Inagaki and Kim (1998) (from *University of Hawai’i Press*). All three are investigations into statistical relationships of text features, and the corpus used in each contain a range of text types. Thus the overall approach of this thesis was to investigate the validity of this practice, and determine the nature of statistical relationships of text features in texts of different types.

The aim of the thesis was stated as: this thesis seeks to determine, in a small corpus of Japanese EFL college student writing, what relationships exist between holistic score and specific count/ratio measures in a range of text types done under two writing conditions. A set of research questions gave more focus:

(1) Can holistic score be measured with sufficient reliability, when the people carrying out the scoring are novice essay readers?

(2) What relationships exist between the individual text features and the holistic scoring for the set as a whole?

(3) What relationships exist between the individual text features and the holistic scoring for the different types of essays?

(4) What types of differences are there between the various types of essays?

(5) What types of differences are there between the test essays and the term essays?

(6) Do all these measures combine into a smaller set of factors, each of which separately influences the essay reader's scoring?

The corpus used in this study was comparatively small, about 275,000 words[1], and therefore more suited to the examination of specific features of texts rather than generalizations about language (McCarthy & Carter, 1997). The corpus was composed of both term and test writing in seven essay types: Self-introduction, Descriptive, Compare/Contrast, Process, Persuasive, Narrative, and Report. It therefore has nine sub-corpora, seven for each of the essay types and two for test essays whose prompts come from the seven term essays. It is these sub-corpora which are compared, with each other and between the term and test writing of each essay type.

The main comparisons made in the thesis involve holistic score and a set of count/ratio measures. The comparisons are made for the corpus as a whole, for each sub-corpora, between sub-corpora, and between term and test writing within each sub-corpora. These comparisons serve to highlight what differences exist between text types.

Regarding the count/ratio measures used in the thesis, they can be divided into three groups: count measures, formula measures, and calculated measures. A list of measures with each group is given in Figure 1. Though there are a great many measures used, only the most significant results will be discussed below.

Figure 1: Measures used in the thesis.

Count measures	Formula measures	Calculated measures
Words - tokens, types, families Sentences T-units Errors Error-free-T-units (EFT) Function words - tokens, types, families Lexical Frequency Profile - token, type and family counts of: - first 1000 word level - second 1000 word level - Academic word list - 'not in the lists' Lexical sophistication - counts of words with two letters, three letters, … up to thirteen-or-more letters	Flesch Score Grade Level	Words per sentence Words per T-unit T-units per sentence EFT per T-units Errors per T-unit Errors per Sentence Errors per 100 words Type/Token ratio (TTR) Guiraud's G (Corrected TTR) Average word length 'Standardized' TTR (100 words) Percents of function word tokens, types and families Lexical Density calculated using word tokens, types and families Percents of tokens, types and families beyond the 1000 word level

		Percents of tokens, types and families beyond the 2000 word level Other TTRs - type/token, family/token, family/type ratios - Guiraud's G calculated for type/token, family/token, family/type Percent of words with… - two or more letters - three or more letters etc., up to… - thirteen or more letters

Definitions for the measures in this thesis were carefully researched. Word-token, word-type and word-family definitions followed Bauer and Nation (1993) and Coxhead (1999); Errors, T-units and Error-free T-units followed Polio (1997), Kroll (1990), and of course Hunt (1965); the list of function word tokens, types and families is a rather extensive list complied from many sources (but primarily Bauer & Nation (1993) and lists given in the software packages used (*Vocab Profile:* Nation (2003) and *Wordsmith Tools* (1996)); the Lexical Frequency Profile measures follow Laufer (1997); and finally Guiraud's G was selected as the type/token ratio of choice (also referred to as Root TTR) following examination of Malvern and Richards (1997) and Tweedie and Baayen (1998). Note these are all primary references for the measures and the original thesis should be consulted for other references which contributed to the choice of that measure.

2.0 Main findings of the thesis

The following subsections will give brief details about the main findings of the thesis.

2.1 Satisfactory inter-rater reliability of the holistic scoring session was achieved

Holistic score was the cornerstone for the project, the measure that would be used in almost all comparisons. It was therefore important that a satisfactory level of inter-rater reliability be achieved in the essay reading to substantiate all other work. The difficulty was that novice essay readers were employed to do the scoring, and hence Research Question #1 and whether they would be able to achieve sufficient reliability.

Overall reliability for triplets of readers for this project were in a range of Chronbach's Alpha values of .7162 to .7609. This level of alpha is considered to be sufficiently high for general comparisons of group behavior, but not sufficiently high for decisions related to individual cases (Glass & Hopkins, 1996). It was fortunate that each essay was scored by three readers because the alpha values for pairs was in a much lower range, .5847 to .7929. We can have much more confidence in the results with the average score from three readers. I suggest that the alpha values for pairs were low due to the following reasons: for each essay set a range of topics was allowed (not one specified topic), and seven of nine sub-corpora were un-timed writing. These two criteria have been noted to add difficulty to holistic scoring (Shermis & Daniels, 2003). For much more detailed conclusions about specific essays, a complete re-read of the essay set with experienced essay readers would be required.

2.2 The essay types are different

The nine sub-corpora were generally found to group together into four sets: a base set composed of Self-introduction, Descriptive, Compare/Contrast, Process, and Persuasive; two single-essay sets of Narrative and Report, and a fourth set composed of the two test essays. Narrative and Report essay types were most different, especially from each other, with the base set of ‘Exposition essays’ falling between them, and the set of two test essays falling outside all others. While each member of the base set had some individual feature that was significantly different from other members of the base set, when the whole set of measures is considered they are found to not be that different. That Narrative and Report text types were so different from each other is of great interest, and future research will delve into the nature of these differences and how teachers can be better informed.

2.3 Word count correlates most strongly with holistic score

Of all the measures used in the thesis, word-token count was found to correlate most strongly with holistic score. The more a student wrote, the more likely they were to achieve a higher holistic score. It must be stated, however, that the correlation values were only moderately high (Spearman’s R less than 0.6), so length is not the only factor that influences holistic score (as can be seen in the factor analysis of the data, explained below in Section 2.6). Also it seemed that, on average, the more the students wrote, the lower the correlation value between word count and holistic score was, with the two ‘longest’ essay types having non-significant relationships between word count and holistic score. It was suggested that there may be a ‘threshold’ of about 400 words where an essay reader places less emphasis on length and more emphasis on other text features. This important item will be a focus of future research, as if it can be more clearly demonstrated it would clearly be valuable information toward the design of future text research of this kind.

2.4 Errors have no effect on holistic score

Both error counts and Error-free T-unit counts were found to function in the same way as the other length counts, and were of no use in comparison of the essay types. Error rates (errors per T-unit, errors per sentence, errors per 100 word) and Error-free T-unit rates (EFT/T) were found to not correlate with holistic score at all. For the target student group, Japanese EFL college students, though errors were present in their writing, they did not appear to have an influence on the holistic score awarded to an essay. Note that while seven of nine sub-corpora represent term writing that had undergone re-writing and editing, the same was true for the two test essays. Errors did not appear to influence the holistic score.

This result supports recent recommendations for EFL/ESL classroom methodology where it is suggested that teachers should encourage students not to focus too much on errors in their writing but rather focus on getting their ideas down on paper.

2.5 The importance of lexical range and lexical variation in essay writing

Lexical range was measured as the counts of word-types and word-families, and this contributed to the measures of Percent of Function Words (calculated for tokens, types and families) and Lexical Density (calculated for tokens, types and families). Lexical variation was measured as type/token ratio (TTR), and as length was not controlled for in this investigation a corrected TTR was required, hence the choice of Guiraud's G. TTRs were calculated in three ways: type/token ratio, family/token ratio, and family/type ratio.

Results for the measures of lexical range and lexical variation proved to be important, and also opened a new branch of investigation: while these measures did not produce results of interest when the base of the measure was word-tokens, these measures did produce significant results when the base of the measure was word-type. In particular, the measure of Percent of Function Word-types gave significant results as to its ability to predict holistic score, and this investigation appears to be the first to use a measure of this kind. I call for more investigation of this measure where statistical relationships of text features are the focus of study.

Corrected TTR also gave significant results for the two new methods of calculation, corrected family/token and corrected family/type ratios. To my knowledge these two methods of calculating a TTR have never been investigated, and in this thesis they proved to be valuable measures in both their correlation with holistic score as well as in their ability to distinguish between term and test writing.

2.6 Factor analysis: factors influencing the essay readers in their scoring of essays

All the statistical measures were included in a factor analysis of the data (more accurately, an exploratory Principle Component Analysis). Remembering that interpretations of factor analyses are rather subjective, the position chosen for this investigation was examining what factors influence the essay readers' awarding of holistic score. In a factor analysis, an 'ideal' solution contains a small number of factors each with a small set of variables that load on them, creating a distinct solution with easily labeled factors[2] (Garson, 2004). In the case this investigation, nine to eleven factors emerged depending on the essay type, demonstrating the complexity of the holistic scoring process.

The most important result that emerged from the factor analysis in the comparison of the different types of writing was that, generally speaking, identical factors emerged in each of the nine essay types, but their rank order was different. Factor #1, labeled ‘Length count’ due to the fact that in was composed of all the length related count variables, was the dominant factor in all essay types indicating the importance of the relationship of length to holistic score. This factor was found to explain between 22% and 31% of the variance in holistic score. That this percentage is not large (e.g. 75%) indicates that essay reading/scoring is not simple, and many factors influence essay readers in their awarding of holistic scores. Holistic scoring is a complex process.

For each of the remaining factors, their rank order was different depending on the essay type. This was interpreted as demonstrating that not only are the essay types different, but also the manner in which the essay readers interpret the essays is different in their awarding of holistic score.

Other results of note involved errors, lexical sophistication, and type/token ratios. While error rates were shown not to have a correlation with holistic score, a factor with all the error rate variables loading on it (thus titled ‘Error rate’) emerged in each essay type, not in the same rank position but rather of the same magnitude (and in different rank position relative to the magnitude of the other factors in each essay type). In this interpretation of the factor analysis, it would seem that the presence of errors does have an effect on the essay readers’ awarding of holistic score, and that effect is constant across all essay types.

As for lexical sophistication, while no correlation was found between holistic score and any level of word length or average word length, a factor which could be labeled ‘Lexical Sophistication’ emerged in each essay type. What was interesting was that the range of word lengths that loaded on the factor was different for the various essay types. This was interpreted as indicating that a general presence of a proportion of words in certain word-length ranges had an effect on the essay readers' awarding of holistic score, but this interpretation is vague and in serious need of qualification as to which specific lexical sets contribute to the factor.

Finally, as with error rates and lexical sophistication above, while no correlation with holistic score was evident, a factor emerged demonstrating the importance of lexical range and variation, which was labeled 'Type/Token ratio'. It was suggested in the thesis that this is evidence of the complexity of the essay reading/scoring task, and while simple mathematical models like correlations cannot completely account for the influence of lexical range and variation, the human mind, being a superior calculator, can interpret these variations in word patterning and word use and put this information to use in the awarding of holistic score.

2.7 Term writing versus test writing: it seems to be the same

Two results were found that may have impact, (1) that generally the means of measures from the term and test essays for any given essay type were not significantly different, and (2) the means of holistic score were not significantly different for two essay types, Compare/Contrast and Report. The first result implies that the specific statistical features of any given essay type do not vary significantly between the two conditions of writing, term (un-timed writing) and test (timed writing), and this lends credence to the idea that the essay types have unique properties and are therefore different from each other. When writing in any particular essay type, the characteristics of that essay type result in statistical features specific to that text type, and different to other essay types. Qualification of these differences in future work may serve classroom teachers well.

The second result justifies what is common practice in some areas of testing. In the two text types of Compare/Contrast and Report, the condition of writing does not appear to have an influence on the holistic score. Thus writing done in timed-test conditions is found to be similar to un-timed term writing, and therefore best represents what a student can do in term report writing. These two text types are better suited than the other text types to the testing situation, for example placement testing and proficiency testing, as the writing done under test conditions was found to be of similar quality to that which could be done in non-test conditions. This gives credence to ETS's traditional choice of compare/contrast prompts in the Test of Written English (TOEFL-TWE), as well as their plans to move toward summarization-style prompts in the next updating of the TWE (Cumming, et al., 2006). School staff who conduct placement testing should be aware that these two text types are ones that best demonstrate what a student can do in term writing conditions.

3.0 Future work

This article is only a brief summary of the findings of my PhD thesis (McNeill, 2006), and the thesis itself is only a brief overview of what can be learned by examining the statistical relationships of different types of writing. There are a great many avenues of research which are open to investigation, some of which are as follows:

(1) Greater detailing of the holistic score. Conduct a re-read of the essay set with experienced essay readers to investigate what different levels of inter-rater reliability could be achieved.

(2) Examine methods of computing inter-rater reliability to determine if the current levels are accurate/appropriate (as suggested in Johnson, Penny & Gordon, 2001; and Copeland, 1996).

(3) Examine the proposal of the '400 word threshold' as a level at which length becomes less important than other text features in influencing

essay readers' awarding of holistic score.

(4) Examine the proposal that word-type counts are a more dominant feature, particularly regarding function word type counts and ratios (rather that word-token counts), focusing on the idea that lexical range is an important text feature in influencing essay readers' awarding of holistic score.

(5) Focus on the specific differences that exist between Narrative texts and Report texts (narrative being simple narrative with dialog, report being a news event summary/opinion). In particular, does the style of narrative (e.g. with dialog, without dialog, first-person narrative, third-person narrative) affect the nature of the relationships of the statistical features of the text.

(6) While T-units are an interesting text measure, clause counts can give more detail (Wolfe-Quintero, Inagaki & Kim (1998)) and a deeper examination of clause types (e.g. subordination, embedding, and nominalization) and their distribution within and across text types will help inform writing teachers and their instruction of students.

(7) Perhaps most important for classroom teachers, a detailing of how all the statistical features of the texts are realized lexically, that is, what actual differences in word use occur in the different text types, and a detailing of how these differences can be taught to students in writing courses.

Then, there is the question as to whether the results obtained from the relatively small corpus used in the thesis can generalize to a larger corpus of the same types of writing thereby substantiating the claims made in the thesis, and indeed generalize to other groups of writers (L1 writers, and L2 writers of varying nationalities and proficiency levels). The creation of and comparison with a similar corpus of much greater size is needed.

This will be formidable work indeed.

And finally, the question as to the relevance of these statistical findings to actual classroom practice. Only when these statistical items are qualified and specific lexical differences between text types clarified will teachers actually be able to improve instruction to their writing students. This is perhaps the most important avenue of research in need of pursuit, and the one that offers the greatest rewards to the classroom.

4.0 Conclusion

This article represents only a brief summary of the findings of the thesis. For greater clarification, one is recommended to refer to the thesis itself. While the thesis was a huge project, it represents only an overview of what can be learned from the examination of the statistical relationships of individual text features in different types of writing

Footnotes

[1] 'Small' when compared to major corpora such as the Bank of English (Collins-Cobuild, 2002), currently at more than 300 million words.

[2] A factor analysis groups variables together that display similar patterns of correlation. Inspection of these different factors (or the groups of variables) leads the researcher to create a title for each factor based on the nature of the variables in each group.

References

Bauer, L. & Nation, P. (1993). Word families. *International Journal of Lexicography*, Vol. 6, No. 4, pp. 253–279.

Biber, D. (1989). A typology of English texts. *Linguistics*, Vol. 27, No. 1, pp. 3–43.

Collins-Cobuild (2002), *Bank of English* – English language corpus (Birmingham, UK: Collins Cobuild).

Copeland, H.L. (1996). *Consistency of seven different methods of estimating reliability for a writing performance assessment exam*. University of Virginia: PhD thesis (UMI 9724643).

Coxhead, A. (1999). A new academic word list. *TESOL Quarterly*, Vol. 34, pp. 213–238.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., and James, M. (2006). *Analysis of Discourse Features and Verification of Scoring Levels for Independent and Integrated Prototype Written Tasks for the New TOEFL*, TOEFL Monograph Report No. MS-30. Princeton, NJ: Educational Testing Service.

Garson, D. (2001, 2004). *Factor Analysis* (web page). Accessed 2005/08/27, <http://www2.chass.ncsu.edu/garson/pa765/factor.htm>.

Glass, G.V. & Hopkins, K.D. (1996*). Statistical Methods in Education and Measurement, 3rd Edition*. Boston: Allyn & Bacon.

Grabe, W. (1987). Contrastive rhetoric and text-type research. In U. Connor & R. B. Kaplan (Eds.), *Writing Across Languages: analysis of L2 text*. Reading, MA: Addison-Wesley.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels*, National Council of Teachers of English Research Report No. 3. Champaign, IL: NCTE.

Johnson, R.L., Penny, J. & Gordon, B. (2001). Score resolution and the interrater reliability of holistic scores in rating essays. *Written Communication*, Vol. 18, No. 2, pp. 229–249.

Kroll, B. (1990). What does time buy? ESL student performance on home versus class compositions. In B. Kroll (Ed.), *Second Language Writing: research insights from the classroom*. Cambridge: Cambridge University Press.

Laufer, B. (1997). Beyond 2000: a measure of productive lexicon in a second language. In L. Eubank, L. Selinker, and M. Sharwood-Smith (Eds.), *The Current State of Interlanguage*. Amsterdam: John Benjamins.

Malvern, D. & Richards, B. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving Models of Language*. Clevedon: Multilingual Matters.

McCarthy, M. & Carter, R. (1997). Written and spoken vocabulary. In N. Schmitt (Ed.), *Vocabulary: description, acquisition, and pedagogy*. Cambridge: Cambridge University Press.

McNeill, B.R. (2006). *A comparative statistical assessment of different types of writing by Japanese EFL College Students*. University of Birmingham, UK: unpublished PhD thesis.

Nation, I.S.P. (2001, 2003), *Vocab Profile* (software downloaded 4 May 2001), <http://www.vuow.ac.nz/nation_p/ software.download>.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: a research synthesis of college-level L2 writing. *Applied Linguistics*, Vol. 24, No. 4, pp. 492–518.

Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning*, Vol. 47, No. 1, pp. 101–143.

Shermis, M. & Daniels, K. (2003). Norming and scaling for automated essay scoring. In M. Shermis & J. Burstein (Eds.), *Automated Essay Scoring*. London: Lawrence Erlbaum.

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

Wolf-Quintero, K., Inagaki, S. & Kim, H-Y. (1998). *Second Language Development in Writing: measures of fluency, accuracy and complexity*. Honolulu: Second language teaching and curriculum center, University of Hawai'i.

WordSmith Tools (1996). Oxford: Oxford University Press.