Tracking Lexical Development in the EDC Classroom: RED Corpus Study #2

Timothy A. Opitz

ABSTRACT

The current study utilizes Lauffer and Nation's Lexical Frequency Profiler in an exploratory analysis of a learner spoken English corpus in order to contribute to the greater knowledge and discussions about oral proficiency assessment and inform SLA pedagogy. Lexical sophistication has been identified in current literature as having strong correlations with proficiency and increasing interest is being given to including multi-party discourse skills in the assessment. A sample of learner multi-party academic discussions was analyzed in terms of lexical proficiency and compared to a native speaker sample. Overall, the Lexical Frequency Profiler did not show clear longitudinal changes and it appears that topic knowledge supersedes any other variable when evaluating lexical features. An analysis of the frequency of Offlist words provides some potential for future research concerning using lexical features that are indicative of multi-party discourse skills and facilitate lay perceptions of proficiency.

INTRODUCTION

The present study is intended to further our understanding of speaking skills assessment and contribute to discussions about informed SLA pedagogy by examining a learner-learner spoken English corpus with Lauffer and Nation's Lexical Frequency Profiler (LFP). As ESL/EFL classrooms have gravitated towards communicative based approaches, assessment of oral skills has likewise received more attention (Poonpon, 2010). The most recognized and reputable English language proficiency testing services have incorporated a speaking skills assessment section. The TOEFL iBT, EIKEN, CEFR, and IELTS testing services have all adopted an assessment method of speaking skills based on extremely similar features that require a trained rater to make a judgement about the candidates' abilities based on holistic descriptors within broader analytical dimensions. For example, all of these major testing services include "language use" as an analytical dimension. Band descriptors to distinguish proficiency include language such as: "fairly automatic and effective use of grammar and vocabulary" (TOEFL iBT), "Can interact with a degree of fluency and spontaneity the makes regular interaction with NS without strain for either party" (CEFR) and "uses a wide vocabulary resource" (IELTS). While training raters to understand the nuances of these band descriptors has resulted in a some degree of inter-rater reliability, there is a distinct lack of consensus about defining terms such as "fluency" (McCarthy 2010) or even "proficiency" (Iwashita, Brown, McNamara & O'Hagan, 2008) Additionally, the distinct lack of empirical underpinnings (McCarthy 2010) for band descriptors has motivated numerous researchers to test empirically valid features of spoken English for assessment purposes.

While researchers are quick to point out that there are key differences between written and spoken English (Biber & Gray, 2013; Tian, 2013) and that no single feature in isolation should be considered a valid measure (Biber & Gray, 2013; Lu, 2012; Osborne, 2007), a survey of the literature reveals that lexical sophistication and lexical density are frequently identified as strongly correlating with oral proficiency. In a seminal study, Lauffer and Nation (1995) work from the assumption that "richer vocabulary is characteristic of better language knowledge" (p.316) and conclude that learners' productive use of language can be measured by a vocabulary test. Similarly, Wolfe-Quintero, Inagaki and Kim (1998) claim "lexical richness is manifest in second language (L2) use in terms of the sophistication and range of an L2 learner's productive vocabulary." (p. 1) Subsequently, Lauffer and Nation's LFP has been used a common reference point for empirically

evaluating lexical sophistication as one indicator of second language proficiency. Lenko (2000) used PELCRA (Polish and English Language Corpora for Research and Applications) to look at three measures of lexical richness (TTR, mean TTR, and LFP) concluding that the LFP is "the most reliable instrument to distinguish between learners at different proficiency levels." (p. 105) Using the TOEFL iBT pilot oral tests as data, Iwashita et al. (2008) attempted to identify features that correlate to rater perceptions concluding that raters tend to overvalue temporal production features such as pronunciation, hesitation phenomena, and grammatical accuracy while the role of vocabulary is undervalued. To the contrary, using the Spoken English Corpora of Chinese Learners, Lu (2012) looked at 26 distinct empirical measures of lexical richness finding no correlation between rater perceptions and lexical density or lexical proficiency. Lu (2012) cited discrepancies between written and spoken English, and instead suggested some of the transformed measures of lexical variation (e.g. Malvern and Richards' D measure) as having the strongest correlation. In a recent analysis of task type effect on proficiency measures, Biber and Gray (2013) used the LFP on both written and oral learner texts holding to the principle that "lexical frequency profiles discriminate between proficiency levels and correlate well with other measures of vocabulary size with lower proficiency learners using higher proportion of high frequency words and higher proficiency learners using more words from the less frequent or Offlist words." (p.4)

Still, a faction of researchers remain discontent with the current model of proficiency assessment because it lacks an element of interaction. The current assessment practices usually involve a brief introduction, a prepared oral monologue and a brief question and answer session, usually 1-2 minutes, with the rater asking questions and the candidate answering. McCarthy (2010) makes a strong argument for assessing an additional "interactive" dimension arguing that lay perceptions of fluency and proficiency include some element of discourse level skills. Furthermore, spontaneous speech places the highest demands on automaticity (Dornyei, 2009) and the competition for turn-taking during multi-party discussions leaves little time for planning (Bialystock, 1982). Therefore, it stands to reason that spontaneous multi-party discussions would provide a more accurate representation of a learner's control over language by placing strenuous cognitive demands on speed of recall. It is in this spirit that the present study aims to fill a gap in the literature and apply the LFP to learner-learner multiparty discussions in an exploratory study. Three basic research questions were attempted to be addressed:

1. Which features of the Lexical Frequency Profile are noticeably different between three proficiency levels of EDC learners and native speakers during multi-party discussions?
2. Do learners show longitudinal development of lexical proficiency in multi-party discussions?
3. Which, if any, lexical features can be empirically identified as indicative of discourse level proficiency in multi-party discussions?

METHOD

As a sample of learner-learner multi-party discussions, the RED Corpus (Rikkyo English Discussion Corpus) was utilized. The RED Corpus transcribed a total of sixty distinct multi-party discussions over the course of an academic year at Rikkyo University. Out of necessity to make the data compatible with the LFP software, the RED Corpus texts were de-annotated with restarts, reformulations, and repetitions retained. While the transcription conventions of the RED Corpus itself could affect the quality and representativeness of the texts (Buck, 2017; Opitz, 2016), the texts should functionally represent authentic samples of typical learner behavior. Additionally, two roughly comparable multi-party oral texts were found on MICASE (Michigan Corpus of Academic Spoken English) for comparison. The RED Corpus and MICASE texts have a number

of contextual similarities. They are both multi-party discussions in an academic context, both occurring as in-class discussions, and both discussions occur between university students of roughly the same age. In order to heighten the comparability, the relevance of topic was considered. The RED Corpus does represent six distinct topics (communication, social issues, happiness, English in Japan, Japanese culture, gender equality) and MICASE discussions about culture and gender were chosen to be comparable. The MICASE transcripts were similarly de-annotated for use in the LFP. The resulting text files were analyzed with the LFP on the LexTutor website (http://www.lextutor.ca/cgi-bin/vp/eng/output.pl).

RESULTS

The Web VP Classic v. 4 version of the LFP was used and seven categories of data are presented in Table 1 and Table 2. The K1 category represents the 1000 most frequent words in English. Within the K1 category, the percentage of function words and content words is also given. The K2 category is the next 1000 most frequently used English words. The AWL category is the most frequently appearing academic words list and Offlist category includes any words that did not fall into any other category. Lastly, lexical density as the ratio of content words to total words was included to supplement the lexical profile snapshot.

Table 1: LFP analysis of RED Corpus by proficiency level compared to MICASE

	RED L4	RED L3	RED L2	MICASE
K1	82.41%	84.89%	85.52%	82.22%
function	42.82%	47.36%	48.30%	52.75%
content	39.60%	37.53%	37.22%	29.47%
K2	4.21%	4.02%	3.53%	2.45%
AWL	2.16%	2.37%	1.98%	3.24%
Off list	11.21%	8.71%	8.96%	12.09%
Lexical density (content/total)	0.57	0.53	0.52	0.47

Table 1 displays the LFP data by the pre-designated learner proficiency levels upon entering university along with the NS MICASE data as a baseline for comparison.

Research Question 1: Which features of the Lexical Frequency Profile are noticeably different between three proficiency levels of EDC learners and native speakers during multi-party discussions?

The percentage of K1 words appears to be roughly equivalent to the MICASE sample. However, contrary to Lauffer and Nation's (1998) assertion that lower proficiency learners rely more heavily on K1 tier vocabulary, the percentage of K1 words actually increases in sync with learner proficiency in the RED Corpus data. The increase in K1 words is speculatively due to the quantitative increase in the use of function words as the higher proficiency learners become more adept at quickly recalling more complicated grammatical structures which require the use of linking words such as prepositions for dependent clauses. Some evidence of this can be seen in the Appendix. Closely related, there is also a linear progression of increasing percentage of K1 function words by proficiency although the percentages are significantly lower than the MICASE data. Surprisingly, the highest percentage of K2 words were manifested by the lowest proficiency learner and the NS MICASE sample had the lowest percent. The quantity of Offlist words tends

to vary, especially when viewed longitudinally in Table 2. Nonetheless, the NS MICASE speakers used the largest percentage of off list vocabulary. The most interesting aspect of the Offlist words is not so much in the quantity but the quality. There is significant deviation of word type between the NS MICASE speakers and the RED Corpus learners. This idea will be addressed more fully when discussing RQ3. The lexical density decreases in a linear fashion from lowest proficiency to NS. The learners are may be over-dependent on vocabulary and unable to quickly parse out more complex grammatical structures. Also, there could be some negative L1 transfer occurring as the RED Corpus learners are all Japanese L1 speakers and have yet to distinguish the difference between Japanese being a topic-comment structured language and English necessitating a subject even in daily spoken conversations. Some evidence to support this is seen in the Appendix when looking at the frequency of place holding and non-referential pronouns "it" and "that".

Research Question 2: Do learners show longitudinal development of lexical proficiency in multi-party discussions?

Table 2 presents a compilation of all RED corpus proficiency levels tracked longitudinally at six points throughout the course of the one year discussion program.

Table 2: LFP analysis of RED Corpus multi-party lexical sophistication (longitudinal)

Spring	Week 2	Week 8	Week 12	Fall	Week 2	Week 8	Week 12
K1	88.0%	85.7%	84.57%		87.49%	77.15%	85.86%
function	49.95%	50.01%	50.10%		45.96%	40.93%	44.94%
content	38.04%	35.69%	34.47%		41.53%	36.22%	40.92%
K2	2.64%	5.12%	4.54%		3.38%	4.99%	2.78%
AWL	1.82%	1.60%	1.79%		1.28%	3.13%	3.60%
Off list	7.55%	7.58%	9.1%		7.85%	14.73%	7.76%
Lexical density (content/total)	0.50	0.50	0.50		0.54	0.59	0.55

In short, none of the features examined with the LFP show development over time. If anything, the data tends to show a counter-intuitive regression between the Spring and Fall semesters with the percentage of K1 function words and lexical density in the Fall semester trendy away from the MICASE sample in Table 1. It is unlikely to assume that the RED Corpus learners' lexical proficiency decreased over the course of the discussion program. The most plausible explanation for this discrepancy is likely the confounding variable of topic familiarity. Support for this argument can be found when looking at the MICASE samples when separated by topic.

Table 3: LFP analysis of MICASE discussions by topic

	MICASE culture	MICASE gender
K1	86%	80.09%
function	55.37%	51.28%
content	30.63%	28.81%
K2	2.32%	2.53%
AWL	3.02%	3.36%
Off list	8.66%	14.02%

Lexical density (content/total)	0.45	0.49

Even in the NS speaker MICASE data, there is significant deviation of lexical sophistication between the culture discussion and the gender discussion.

Research Question 3: Which, if any, lexical features can be identified as indicative of discourse level proficiency in multi-party discussions?

Because the LFP was not designed with the intent of analyzing multi-party discourse, the greatest utility of using the LFP to analyze learner-learner discussions does not lie so much in what the LFP was able to categorize in terms of K1, K2 or AWL tiers but in the quality of what was dumped into the Offlist. The topic likely has a large impact on the quality and types of vocabulary that appear in the Offlist. Additionally, some common words that occur in multi-party interaction which are not associated with professional interpretations of fluency appear in the Offlist. Browsing over the Offlist tokens, three clear trends can be observed. First, the frequency of "yeah" and "okay" as well as what McCarthy (2010) would call turn-opener tokens such as "mhm" or "erm" shows a clear difference. An example of a poverty of variety in discourse level confluence tokens can be seen by comparing the frequency of "yeah" and "okay" tokens in the RED Offlist to the MICASE Offlist.

Table 4: Percent of "yeah" and "okay" tokens by RED Corpus proficiency groups compared to MICASE

	RED L4	RED L3	RED L2	MICASE
Yeah and okay tokens	6.15%	5.2%	5.51%	1.83%

The MICASE data shows a much larger variety of turn-opening tokens categorized as Offlist such as "alright", "um", "mhm", and "uhhuh" as a percent of total tokens. Second, the use of individual first names is significantly greater in the MICASE samples. A quick browse of token frequency lists revealed the RED L4 learners used proper first names 12 times, RED L3 count was 31, RED L2 35, and the NS MICASE speakers used proper first names 90+ times. The third readily noticeable feature was the frequency of casual speech reductions in the MICASE data. Some of the more frequently occurring tokens were "cuz" and "wanna". Transcribing casual speech reductions was pre-meditated decision when compiling the RED Corpus. However, there were only two or three instances in the entire 48,000+ tokens of the RED Corpus.

DISCUSSION

The present study has several implications for assessment purposes. First, the discrepancy between the percentage of K1 function words between the RED Corpus learners and NS MICASE speakers in Table 1 appears to indicate that spontaneously recalling grammatical forms requires a high degree of automaticity. The assertion that raters tend to over-value grammatical accuracy (Iwashita et al., 2008) does not appear to be justified. If anything, a correlation between proficiency and frequency of function word tokens is evident. While the use of function words did not exhibit longitudinal development, as might have been expected throughout the two semesters of the EDC program, the over-riding importance of topic knowledge is introduced as a consideration. As evidenced by the disparity of K1 function word tokens by NS participants in Table 3, even NS will exhibit a strong variance of lexical sophistication when discussing different topics. The significance for assessment purposes is that learners may demonstrate noticeably different lexical sophistication based on the topic. The current oral proficiency assessment format,

which tends to utilize a brief introduction, a prepared monologue and a brief question and answer session, may be adequate to broadly distinguish between proficiency levels. However, learners may be at either a severe advantage or disadvantage depending on their specific topic knowledge of the prepared monologue or the question and answer portion of an assessment test. Candidates and assessors alike may be better served by having a candidate perform the same task, say for example a prepared monologue, three times on three distinct topics rather than just once. Of course, the assessment procedure itself would become more unwieldly and labor intensive but, the benefits may outweigh the costs. Additionally, the qualitative analysis of the Offlist words lends support for the voices calling for oral proficiency assessment to include a more interactional element (McCarthy, 2010). The current interaction during assessments is typically limited to a dyad of rater and candidate with the rater predominantly asking the questions. If oral proficiency assessment is to capture a full range of linguistic competence, recourse for consideration of uninitiated lay perceptions of proficiency should be considered. Possible features can be found in the Offlist tokens. A considerably wider variety of confluence tokens are exhibited by NS during multiparty discussions and the frequency of casual speech reductions, even in this academic discussion context, and the use of first names have potential to be developed as features for analysis. Provided that satisfactory justification for including these features was provided, adding this analytical dimension to the band descriptors and providing rater training would be an achievable task.

The corresponding implications for SLA pedagogy are fairly straight forward. Namely, even though pedagogy in general has been moving towards communicative approaches, form focused instruction still has a place. As a simple example of the need for declarative knowledge of forms, consider the frequency of the articles "a" and "the" in Appendix A. "The" is the second most frequently used word in the CANCODE sub-corpus of spoken English social discussions and the fourth most frequent in the samples of MICASE NS academic discussions. However, "the" does not appear in the top 20 most frequently used words by the lowest proficiency (L4) RED Corpus learners or the L3 learners and peaks at the twelfth most frequent among the highest proficiency level L2 learners. Article usage is taught early in most students English education history and most students have extensive declarative knowledge of the form. However, automatizing use in spontaneous oral production is still problematic. Also, the findings and speculations of the present study tend to support including some aspects of multi-party discussion management and casual speech reductions into the curriculums in order satisfy lay perceptions of proficiency.

CONCLUSION

The validity of the current study is clearly limited in a number of ways. Foremost, the sample size is rather small and may not be large enough to mitigate the confounding variable of individual idiosyncrasies. (Opitz, 2016) Secondly, the study is limited by the range of contexts that are being analyzed. The RED Corpus data comes nearly exclusively from native Japanese speaking English learners in a university academic discussion context and the MICASE data is taken from only two distinct university academic discussions by American university students. Clearly, more contexts would need to be evaluated before any of the speculations presented here could be extrapolated into general conclusions. However, the exploratory nature of the study does open up possibilities for future research.

REFERENCES

Bialystock, E. (1982). On the relationship between knowing and using linguistic forms. *Applied Linguistics,3*(3), 181-206.

Biber, D., and Gray, B. (2013). Discourse Characteristics of Writing and Speaking Task Types on the TOEFL iBT Test: A Lexico-Grammatical Analysis. Retrieved from: https://www.ets.org/Media/Research/pdf/RR-13-04.pdf.

Buck, J. (2017). Developing the Rikkyo English Discussion Corpus. The 2016 PanSIG Proceedings, Okinawa. (in press)

Dornyei, Z. (2009). *The Psychology of Second Language Acquisition*. Oxford: Oxford University Press.

Iwashita, N., Brown, A., McNamara, T., and O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24-49.

Lenko-Szymanska, A. (2000). How to trace the growth in learners' active vocabulary? A corpus based study. TALC Papers. Retrieved from: http://www-gewi.kfunigraz.ac.at/talc2000/Htm/index1.htm.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' narratives. *The Modern Language Journal*, *96*(2), 190-208.

McCarthy, M. (2010). Spoken fluency revisited. *English Profile Journal*, *1*(4), doi:10.1017/S2041536210000012.

Opitz, T. (2016). The efficacy of drilling backchannels in a CLT context: A corpus analysis of Japanese ESL earners. *New Directions in Teaching and Learning English Discussion*, *3*(1), 200-210.

Osborne, J. (2007). Investigating L2 fluency through oral learner corpora. *Spoken Corpora in Applied Linguistics*, M.C. Campoy & M.J. Luzón (eds.). Frankfurt: Peter Lang, 181-197.

Poonpon, K. (2010). Expanding a second language speaking rating scale for instructional and assessment purposes. Spaan Fellow Working Papers in Second or Foreign Language Assessment, vol. 8, 69-94.

Tian, X., (2013). Distinguishing spoken English from written English: Rich feature analysis. *English Language Teaching*, *6*, 72-78.

Wolfe-Quintero, K., Inagaki, S., & Kim H. (1998). Second language development in writing: Measures of fluency, accuracy, and complexity. *Studies in Second Language Acquisition,23*(3), 423-425.

APPENDIX – 20 most common words in RED Corpus by proficiency level compared to MICASE and CANCODE

Frequency Rank	RED L4	RED L3	RED L2	MICASE	CANCODE
1	I	I	I	like	I
2	yes	ok	yes	I	the
3	you	you	yeah	that	and
4	ok	is	ok	the	you
5	yeah	yes	is	and	it
6	is	think	you	you	yeah
7	do	so	think	of	a
8	think	to	to	to	to
9	to	yeah	so	it	that
10	no	do	and	um	was
11	so	and	do	of	a
12	and	a	the	mhm	in
13	what	can	not	what	oh
14	understand	in	but	is	it's
15	Japanese	not	no	in	know
16	a	but	a	think	no
17	not	my	can	so	mm
18	why	what	see	it's	like
19	in	don't	have	no	but
20	can	very	in	ok	he