A Multivariate Analysis of Interlanguage Differences between Learner Levels

Peter Longcope

Since Selinker (1972) first introduced it, the concept of interlanguage has played an important role in helping us understand second language acquisition. More recently, (Larsen-Freeman, 2006) researchers have begun to view interlanguage as a system composed of a number of sub-systems, including, but not limited to, syntactic complexity, grammatical accuracy, fluency, and lexical variation. Over the years, research has been done on these different sub-systems to clarify what dimensions are important in interlanguage development.

Syntactic Complexity

With respect to syntactic complexity in a second language, Monroe (1975) found that six measures of syntactic complexity in the writing of university students studying French increased systematically according to the year in school students were. More specifically, freshmen had the lowest scores for all measures; the scores increased for sophomores, and then for juniors and seniors, and finally for graduate students. While the students' scores were not necessarily statistically different from the scores of the students just before or just after them, the overall growth for all measures was found to be statistically significant. Included in these measures were the measures of mean clauses per T-Unit and mean words per T-Unit.

More recently, Ortega (2003) performed a meta-analysis on studies looking at development in syntactic complexity in second language writing. She found that learners of different levels differed by as many as 2 words per T-Unit and as much as 0.2 clauses per T-Unit. She also found that the rate of development was slower for learners in an EFL context than for learners in an ESL context and that learners' ultimate level of attainment was lower for learners in an EFL context than for learners in an ESL context.

Grammatical Accuracy

In looking for an index of ESL development, Larsen-Freeman (1978, 1983) and Larsen-Freeman and Strom (1977) looked at a few global measures of grammatical

accuracy. They found that percentage of error-free T-Units and mean words per error-free T-Unit both increased fairly consistently with learner level. They did find, however, that the two measures were not always able to distinguish between two adjacent levels.

Fluency

Much of the research into the fluency developments that second language learners make has been done by comparing computational differences in interlanguage samples with assessments of those samples made by native speakers in order to find what fluency measures play the most important part in fluency development. Lennon (1995) took this approach and found that change in dysfluency markers, such as self-corrections and repetitions, did not help explain native speaker ratings of learner fluency. Derwing et al (2004), using a similar methodology, also found that self-repetitions did not correlate well with native speaker ratings of fluency; however, they did find that, especially for lower-level learners, temporal measures, including speech rate correlated well with them.

Other research has looked at how learners of different levels differ from one another with respect to fluency measures. In one such study, Riggenbach (1991) also found that measures of self-repair do not contribute much in distinguishing one level of learner from another. Furthermore, she found, like Derwing at al (2004) that rate of speech can help in distinguishing one level of learner from another.

Lexical Variation

With respect to the final interlanguage sub-system to be discussed, Wolfe-Quintero, Inagaki, and Kim (1998) reviewed a large number of studies investigating second language learner development in writing. Regarding lexical variation, they found that one of the most important measures was a modified form of type-token ratio, the modification being that instead of dividing the number of lexical types by the total number of lexical tokens, the total number of lexical types is divided by the square root of twice the number of lexical tokens. This modification is done in order not to bias the measure against longer samples, which, due to the nature of language, will include a larger number of repeated lexical items than shorter samples.

Conclusion

Most of the studies discussed above, with the exception of the studies on fluency, have investigated interlanguage development by focusing on written development rather than oral development. Furthermore, the studies generally focus on the given sub-system in isolation; in other words, in discussing the findings of the study, the

researchers look at only one sub-system at a time. In order to truly understand interlanguage development as well as the interlanguage differences between disparate levels of learners, the discussion of change within any one of the sub-systems needs to be informed by change that may be occurring in other sub-systems at the same time.

Another problem with many of the studies discussed above is that they have only looked to find whether or not disparate levels of learners differ with respect to a given measure. Ziliak and McCloskey (2008) point out that such analysis is lacking because there usually is some difference between different groups; however, the difference might not be noted if the sample sizes are not large enough. They argue that the more important question that should be asked is, "How big is the difference?" They say that rather than simply looking to see if differences are statistically significant, researchers should focus on the size of the differences and then turn to the phenomenon under investigation to discuss whether or not differences of that size are scientifically important.

Therefore, the research questions for this study are the following:

1) What combination of sub-systems can account for interlanguage differences in the oral production of second language learners at disparate levels?
2) In each of the sub-systems, how large are the differences between learners?

Methodology

Subjects

The subjects for this study were 25 first year university students at a prestigious private university in western Japan. All subjects were native Japanese speakers who had been studying English for at least six years. All the subjects came from three of seven English classes in the same faculty at the university. Students who volunteered to participate in this study were exempted from having to take a conversation test in their English classes at the end of the first semester.

Classes

At the beginning of the school year, all the students in the faculty took the G-TELP. Based on their scores on the test, students were placed into one of seven classes in an attempt to keep learners of the same level together. Using the students' G-TELP scores ANOVAs were done on the different classes. The three classes that were chosen for this study were classes for which the G-TELP reading scores, listening scores, and grammar scores were each determined to be statistically significantly

different from each other. In other words, the reading scores, listening scores, and grammar scores for the intermediate level group were statistically significantly different from the reading scores, listening scores, and grammar scores of the higher level group as well as the lower level group.

Units of Analysis

For the measures of complexity and accuracy used in this study, the main unit of analysis is the AS-Unit, defined in Foster, Tonkyn, and Wigglesworth (2000) as "a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either" (p. 365). In determining what would or would not be included as an AS-Unit, for the purposes of this study, Foster, et al.'s second level was used (p.370); that is, all one-word utterances and echoic responses were excluded from analysis. This decision was made because it was felt that inclusion of such utterances would misrepresent learners' interlanguages.

Data Collection Instrument

All data for this study were collected by means of guided interviews between the individual subjects and the researcher. Interviews lasted anywhere between five and fifteen minutes and covered general personal information. Topics that were discussed during the interviews included the learner's hometown, personal interests and hobbies, travel experience (both domestic and foreign), and plans for the upcoming summer vacation.

Measures

The measures used in this study are the following: for syntactic complexity, words per AS-Unit and clauses per AS-Unit; for grammatical accuracy, percentage of error-free AS-Units; for fluency, words per minute; and for lexical variation, the modified type-token ratio discussed above.

Statistical Analysis

The first analysis done is a multivariate analysis of variance (MANOVA); however, prior to doing that, correlation coefficients were determined between each of the measures in order to establish which measures could be grouped together for multivariate analysis. After performing the multivariate analysis, the effect size f was calculated for each measure. The effect size f can be understood as the standard deviation of the means of the groups divided by the standard deviation within the populations (see Cohen, 1988). Considering the effect size will allow a discussion of the size of the difference that exists for each of the measures.

Results

The group means and standard deviations for each of the measures used in this study can be found in table 1, and the correlation coefficients for each measure with every other measure are given in table 2.. As can be seen from table 2, all but two of the measures used in this study correlated with each other at a statistically significant level. The two measures that did not correlate with each other for these learners were the measures of lexical variation and percentage of error-free AS-Units. Therefore, MANOVAs were run on two different groupings of measures: (1) words per minute, lexical variation, words per AS-Unit, and clauses per AS-Unit and (2) words per minute, words per AS-Unit, clauses per AS-Unit, and percentage of error-free AS-Units.

Table 1: Group means and standard deviations for each variable

	LEX		WPM		W / ASU		S / ASU		EFASU	
	Mean	ST Dev	Mean	ST Dev	Mean	ST Dev	Mean	ST Dev	Mean	ST Dev
Upper	5.4552	0.5617	70.1144	16.9808	5.8529	1.2087	0.1975	0.1131	0.3586	0.1402
Middle	5.0655	0.4846	51.6755	15.3951	5.1178	1.0310	0.1229	0.0662	0.3247	0.0668
Lower	4.4577	0.3957	36.9107	7.6132	4.2276	0.8726	0.1104	0.0752	0.2688	0.0975

Table 2: Correlation coefficients of measures

	LEX	WPM	WASU	SASU	EFASU
LEX	--				
WPM	0.5797 *	--			
W / ASU	0.6646 *	0.7381 *	--		
S / ASU	0.3508 *	0.4221 *	0.6307 *	--	
EFASU	0.2902	0.5015 *	0.6629 *	0.6592 *	--

N = 25

* $p < 0.05$

The results from the two MANOVAS are given in table 3. As can be seen, the differences between the different groups of learners with respect to the groupings of measures reached statistical significance for both groupings. Therefore, it can be concluded that the different levels of learners were statistically different from each other when lexical variation, words per minute, words per AS-Unit, and clauses

per AS-Unit were considered together, and when words per minute, words per AS-Unit, clauses per AS-Unit, and percentage of error-free AS-Units were considered together.

Table 3: Results of MANOVAs (Wilks' Lambda)

	Value	F	df	Error df	Sig.	Eta Squared	Observed Power
LEX, WPM, W/ASU, S/ASU	0.373	3.029	8	38	0.01	0.389	0.914
WPM, W/ASU, S/ASU **EFASU**	0.445	2.367	8	38	0.036	0.333	0.817

Finally, the effect size f for each of the measures is given in table 4. The size of each measure is given overall, as well as the size between the individual groups. The effect sizes are interpreted here in accordance with Cohen (1988). As can be seen, the effect size for each of the measures except percentage of error-free AS-Units is large, that is greater than 0.40, when the three groups are taken together. More specifically, lexical variation has an effect size of 0.844 when all groups are considered together; words per minute has an effect size of 0.9646; words per AS-Unit has an effect size of 0.636; and clauses per AS-Unit has an effect size of 0.4653. As for percentage of error-free AS-Units, this measure has a medium effect size of 0.364.

Table 4: Effect sizes (f) for all measures

	LEX	WPM	W / ASU	S / ASU	EFASU
Overall	0.844 ***	0.9646 ***	0.636 ***	0.4653 ***	0.364 **
Upper - Middle	0.376 **	1.1461 ***	0.3315 **	0.4301 ***	0.1715 *
Middle - Lower	0.6768 ***	1.2021 ***	0.46 ***	0.0894	0.353 **
Upper - Lower	1.0282 ***	2.6233 ***	0.7713 ***	0.4554 ***	0.3723 **

*** Large f > 0.40
** Medium f > 0.25
* Small f > 0.10

When considering the size of the differences for each measure between the individual groups, the largest differences are for the measure words per minute. As can be seen in table 4, the effect size for this measure is 1.1461 between the upper and middle groups, 1.2021 between the middle and lower groups, and 2.6233 between the upper and lower groups. All of these are more than twice the necessary 0.40 necessary to be considered large.

The effect sizes between the individual groups for lexical variation are also fairly high. The effect size between the upper and middle groups is 0.376, which is not quite high enough to be considered large, but is close. The effect size for lexical variation between the middle and lower groups is 0.6768, and between the upper and lower groups is 1.0282, both well above the 0.40 necessary to be considered large.

Similar to lexical variation, the effect sizes between the individual groups for words per AS-Unit are also consistently somewhat large. The size of the effect for this measure between the upper and middle groups is 0.3315 and so is considered a medium-sized effect. The size of the effect for this measure between the middle and lower groups us 0.46 and so is considered a large effect. Finally, the size of the effect between the upper and lower groups is 0.7713.

The effect sizes for the measure of percentage of error-free AS-Units follows a similar pattern to the pattern established by the measures of lexical variation and words per AS-Unit but is different in degree. The effect size for percentage of error-free AS-Units between the upper and middle groups is 0.1715. This is considered a somewhat small effect size. The effect size between the middle and lower groups is 0.353, a medium-sized effect, and the effect size for the measure between the upper and lower groups is 0.3723, again, a medium-sized effect.

Finally, the effect size pattern for clauses per AS-Unit is quite different from the others. The effect size between the upper and middle groups is a large 0.4301. The effect size between the middle and lower groups, however, is only 0.0894. This effect size is below the 0.1 that would be necessary to consider the effect small. Finally, the effect size between the upper and lower groups is a large 0.4554.

Discussion

The first research question for the study was what combination of sub-systems can account for interlanguage differences in the oral production of second language learners at disparate levels. Based on the MANOVAs presented above, there appear

to be two separate combinations of sub-systems that can describe the differences between the disparate groups in this study. The first of those combinations is lexical variation, fluency, and syntactic complexity, while the second is fluency, syntactic complexity, and grammatical accuracy. So for these learners, the construct of interlanguage development can be defined as development in fluency and syntactic complexity along with development in either lexical variation or grammatical accuracy. What is unclear, however, is why lexical variation and grammatical accuracy failed to correlate at a statistically significant level with each other. On the one hand, where the correlation between the two measures is approaching significance, it is possible that the sample sizes in this study were too small. On the other hand, it is possible that the construct of interlanguage development is not monolithic and so these two sub-systems contribute to different dimensions of it—dimensions that are not wholly separate as they overlap at the points of syntactic complexity and fluency. More research needs to be done to clarify this.

With respect to the effect sizes of the different measures, it may be helpful to look at a graph of the two combinations of measures. In these graphs, the raw scores of each measure for each group have been converted to z-scores, or standardized scores, so that they can be placed on the same graph. When looking at figures 1 and 2, it is clear that the three groups have interlanguages that are very different from one another. What is more, in figure 1, with the exception of the measure of clauses per AS-Unit, the differences between each of the three groups is quite stark: the upper group is much higher than the other two groups on all three measures and the lower group is much lower than the other two groups on all three measures. In figure 2, the picture is similar if somewhat muted. In this case, the distinctions between the three groups with respect to grammatical accuracy is less obvious than with the other measures.

Conclusion

The purpose of this study was to look at the interlanguage differences between groups of second language learners at disparate levels. What was found was that these learners differed in two ways, with respect to lexical variation, fluency, and syntactic complexity, and with respect to fluency, syntactic complexity, and grammatical accuracy. Furthermore, it was found that while the effect sizes for the different measures between groups was mostly large, the effect sizes were not uniform.

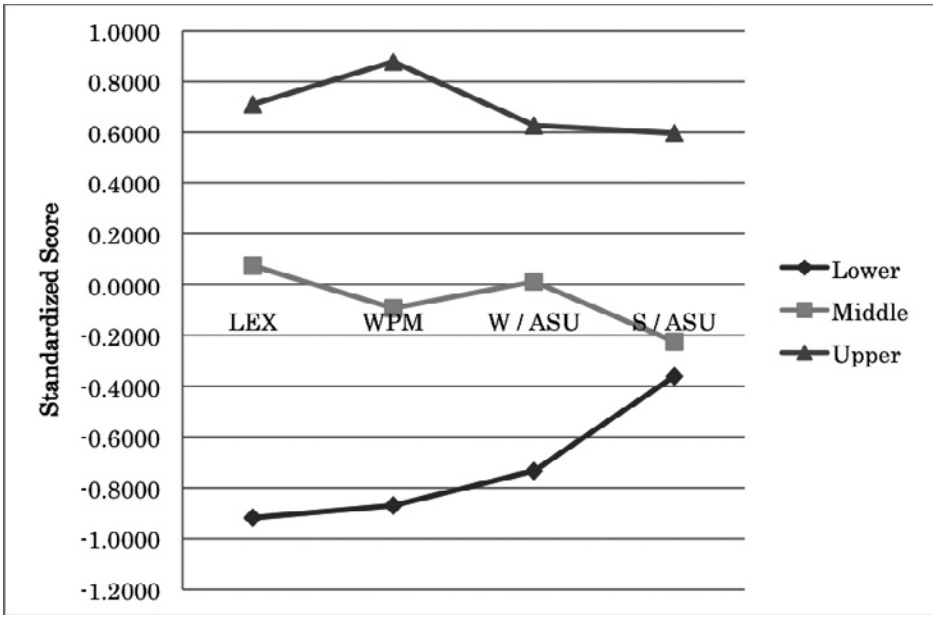

Figure 1: Group means of subjects' standardized scores on lexical variation, fluency, and syntactic complexity

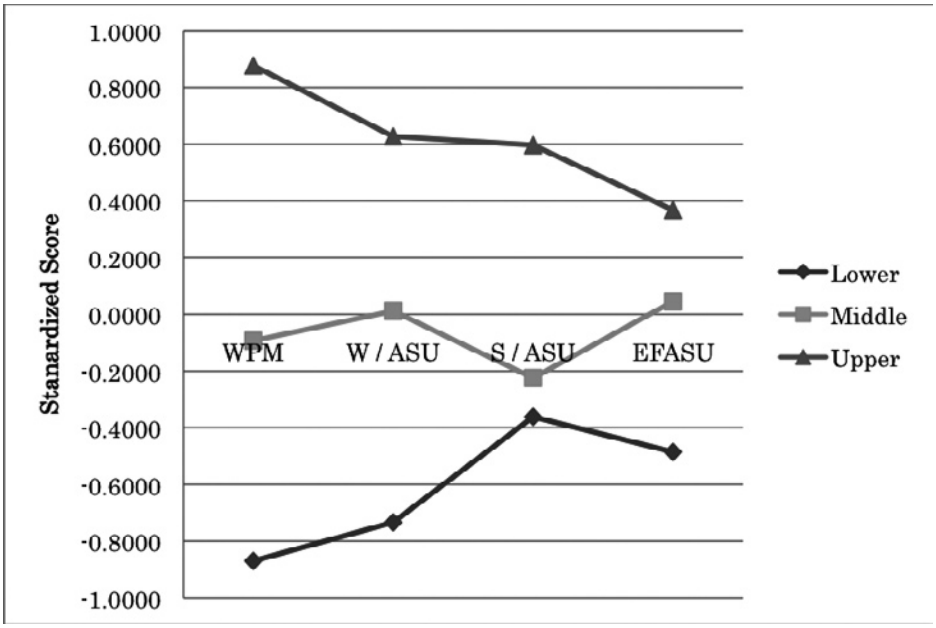

Figure 2: Group means of subjects' standardized scores on fluency, syntactic complexity, and grammatical accuracy

References

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Derwing, T., Rossiter, M., Munro, M., and Thomson, R. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*, 655-679.

Foster, P., Tonkyn, A., and Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*, 354-375.

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly*, *12*, 439-448.

Larsen-Freeman, D. (1983). Assessing global second language proficiency. In H. Seliger and M. Long (Eds.), *Classroom Oriented Research* (pp. 287-304). Rowley, MA: Newbury House.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Lingsuitcs*, *27*, 590-619.

Larsen-Freeman, D. and Strom, V. (1977). The construction of a second language acquisition index of development. *Language Learning*, *27*, 123-134.

Lennon, P. (1995). Assessing short term change in advanced oral proficiency: Problems of reliability and validity in four cases. *ITL Review of Applied Linguistics*, *109-110*, 75-109.

Monroe, J. (1975). Measuring and enhancing syntactic fluency in French. *French Review*, *48*, 1023-1031.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*, 492-518.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, *14*, 423-441.

Selinker, L. (1972). Interlanguage. *IRAL*, *10*, 209-231.

Wolfe-Quintero, K., Inagaki, S., and Kim, H-Y. (1998). *Second Language Development ion*

Writing: Measures of Fluency, Accuracy, and Complexity. Technical Report No. 17. Honolulu, HI: University of Hawai'i, Second Language Teaching and Curriculum Center.

Ziliak, S. and McCloskey, D. (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor, MI: The University of Michigan Press.