Examining GRE Writing Predictive Validity:

Comparing GRE Response Samples with MICUSP

Introduction

As one test that most universities in American require taking when students want to be admitted to graduate school, GRE plays a pivotal role in college life. ETS claims that the question types in the test "closely reflect the kind of thinking … in graduate or business school." (ETS, n.d.)

Specifically in writing, in one of ETS research reports, Breland et al (1999b) scrutinize two studies (Breland et al, 1999a; Powers et al, 1996), and conclude that GRE, having two essay writing in writing section, has relatively reliable predictive validity, meaning that the test scores are correlated with students' later writing performance in graduate school. This paper aims to examine part of the predictive validity of GRE's writing section.

Methodology

This paper aims to answer two questions:

(1) Does getting a high score in GRE writing necessarily entail writing a grade-A paper in graduate school?

(2) To what extent to the linguistic features (lexical and syntactical complexity, propositional idea density and coherence) in GRE writing resemble papers written in graduate school?

To answer the two questions, I plan to examine the two corpora on lexical, syntactic,

propositional idea density and coherence level.

Data collection

Two corpora are built for this preliminary research. One includes 15 essay responses from "GRE CAT: Answers to the Real Essay Questions" (Stewart, 2003). All the responses in the book are scored at least five (six is the highest score). However, only responses for analyzing issues are collected for the fact that responses for analyzing arguments are highly dependent on the argument topics, and therefore language use may be restricted. The goal of analyzing issues in GRE test is to critically elaborate the complex nature of issues, take a side and argue with supporting reasons and examples. This genre resembles argumentative essays. An example for analyzing an issue can be "*It is often necessary, even desirable, for political leaders to withhold information from the public.*" (Stewart, 2003) The fields of these sample issues are controlled within education, politics and culture for the sake of comparison with the second corpus.

The second corpus comprises of ten papers from Michigan Corpus of Upper-Level Students Paper (MICUSP). The number of papers is reduced from fifteen (in the first corpus) to ten due to the fact that papers in MICUSP are generally much longer than timed essays in GRE responses. All of the writers are first or second year graduate students (not restricted to native or non-native speakers of English). The genre of all the papers is argumentative paper in order to compare with the first corpus. The writers major in education, political science or sociology in agreement with the first corpus.

Here is a summary of the data collected:

Corpora	GRE Response	MICUSP
Number of texts	15	10
Writer	GRE test takers	First/second year graduate students
Genre	Analyzing an issue	Argumentative Essay
Field of study	Education, politics, culture	Education, political science, sociology

Annotation and Analytical Procedures

In order to see how the GRE writing can reflect future writing in graduate schools, one aspect is to see if the linguistic features in two corpora have significant differences. To achieve that, the two corpora are POS-tagged by using Stanford POS Tagger (Klein & Manning, 2003), lemmatized by using Morpha (Minnen et al, 2001) and analyzed by using Lexical Complexity Analyzer (Lu, 2012). Subsequently, the corpora are parsed and analyzed by using L2 Syntactic Complexity Analyzer (Lu, 2010). Last, I use CPIDR (Brown et al, 2008) to examine the propositional idea density and Latent Semantic Analysis (Foltz et al, 1998) to examine the sentential coherence of the corpora. All the results are loaded in IBM SPSS for further statistic analysis.

First, I run an independent-samples t test to examine the difference of means of the 25 indices retrieved from Lexical Complexity Analyzer (LCA) between two corpora. The total numbers of sentences, word types, and lexical types, etc., are excluded from this part of the analysis as they are not comparable.

Second, the same procedure applies to the results retrieved from L2 Syntactic

Complexity Analyzer (L2SCA). Independent-samples t test is run on the means of fourteen indices between two corpora. (Same, the total numbers of sentences, clauses, etc., are excluded).

Last, an independent-samples t test is run on the means of propositional idea density and sentential coherence between the two corpora. Results will be shown and discussed in the following section.

Results

Lexical Difference

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
ld	Equal variances assumed	.327	.573	5.838	23	.000	.05100	.00874	.03293	.06907
ls1	Equal variances assumed	.002	.963	.989	23	.333	.01700	.01719	-.01856	.05256
ls2	Equal variances assumed	1.125	.300	-2.746	23	.011	-.04867	.01772	-.08532	-.01201
vs1	Equal variances assumed	.001	.974	5.498	23	.000	.10467	.01904	.06529	.14405
vs2	Equal variances not assumed	8.560	0.008	-2.616	10.445	.025	-4.03000	1.54057	-7.44289	-.61711
cvs1	Equal variances not assumed	5.117	.033	-2.201	10.773	.050	-.47100	.21397	-.94316	.00116
ndw	Equal variances not assumed	14.028	0.001	-7.103	9.501	.000	-479.43333	67.50177	-630.91484	-327.95183
ndwz	Equal variances assumed	.288	.597	-.835	23	.412	-1.00000	1.19782	-3.47789	1.47789
ndwerz	Equal variances assumed	2.324	.141	.625	23	.538	.35000	.55966	-.80775	1.50775
ndwesz	Equal variances assumed	.020	.888	2.852	23	.009	1.23000	.43132	.33775	2.12225

ttr	Equal variances not assumed	6.101	.021	9.093	12.898	.000	.20000	.02199	.15245	.24755
msttr	Equal variances assumed	.366	.551	2.557	23	.018	.02567	.01004	.00490	.04643
cttr	Equal variances assumed	.933	.344	-4.777	23	.000	-1.49767	.31349	-2.14617	-.84916
rttr	Equal variances assumed	.922	.347	-4.776	23	.000	-2.12033	.44391	-3.03863	-1.20203
logttr	Equal variances not assumed	5.184	0.032	7.161	12.731	.000	.05300	.00740	.03698	.06902
uber	Equal variances assumed	.089	.768	3.460	23	.002	2.51300	.72624	1.01066	4.01534
lv	Equal variances not assumed	11.519	0.002	6.872	12.337	.000	.22200	.03230	.15183	.29217
vv1	Equal variances assumed	.738	.399	5.929	23	.000	.20600	.03475	.13412	.27788
svv1	Equal variances not assumed	6.433	0.018	-6.746	11.594	.000	-46.7036	6.92266	-61.84562	-31.56171
cvv1	Equal variances assumed	1.111	.303	-8.012	23	.000	-2.14067	.26718	-2.69336	-1.58797
vv2	Equal variances assumed	.014	.906	6.143	23	.000	.04400	.00716	.02918	.05882
nv	Equal variances assumed	2.161	.155	7.125	23	.000	.21600	.03032	.15328	.27872
adjv	Equal variances assumed	.691	.414	5.721	23	.000	.05267	.00921	.03362	.07171
advv	Equal variances assumed	.254	.619	5.069	23	.000	.02733	.00539	.01618	.03849
modv	Equal variances not assumed	7.513	0.012	6.888	12.177	.000	.07967	.01157	.05451	.10483

As shown above, only three (highlighted) out of twenty-five indices are not significantly different, which are ls1 (lexical sophistication), ndwz and ndwerz (number of different random words). Intuitively, we would think that lexical items in papers in MICUSP should be more diverse because unlike GRE responses, they are not timed; writers would have more time to polish their writing and be careful with their word choices. However, if we look at the lower section of the table, ttr, lv, vv1, vv2, nv,adjv, advv, and modv (type-token ratio,

variations in lexicals, verbs-I, verbs-II, nouns, adjectives, adverbs, and modifiers) in GRE responses are all significantly higher than papers in MICUSP. Only if we use corrected number of type-token ration, lexical and verb variation will we get the result of MICUSP higher than GRE responses.

Syntactical Difference

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
MLS	Equal variances assumed	3.850	.062	-3.246	23	.004	-6.48154	1.99667	-10.61197	-2.35111
MLT	Equal variances assumed	.233	.634	-2.043	23	.053	-3.25248	1.59164	-6.54503	.04007
MLC	Equal variances assumed	1.574	.222	2.124	23	.045	1.08702	.51176	.02838	2.14567
C/S	Equal variances not assumed	7.351	0.012	-3.385	12.357	.005	-.76056	.22469	-1.24856	-.27256
VP/T	Equal variances assumed	.012	.913	-1.551	23	.135	-.32856	.21184	-.76679	.10967
C/T	Equal variances assumed	.659	.425	-3.180	23	.004	-.46237	.14540	-.76315	-.16159
DC/C	Equal variances assumed	3.616	.070	-.752	23	.460	-.03008	.04002	-.11288	.05271
DC/T	Equal variances assumed	.053	.820	-1.812	23	.083	-.24402	.13466	-.52259	.03454
T/S	Equal variances assumed	.407	.530	-1.666	23	.109	-.09836	.05904	-.22049	.02376
CT/T	Equal variances assumed	1.211	.283	-1.140	23	.266	-.06770	.05939	-.19055	.05516
CP/T	Equal variances assumed	.009	.927	-.617	23	.543	-.05735	.09299	-.24971	.13501

CP/C	Equal variances assumed	.688	.415	1.209	23	.239	.06205	.05132	-.04412	.16822
CN/T	Equal variances assumed	1.327	.261	-2.022	23	.055	-.59407	.29381	-1.20186	.01373
CN/C	Equal variances not assumed	6.110	.021	.897	20.878	.380	.08260	.09212	-.10904	.27424

The table above shows the syntactic difference between the two corpora. We can see only five out of fourteen indices are significantly different, which are MLS (mean length of sentence), MLC (mean length of clause), C/S (clauses per sentence), C/T (clauses per T-unit), and CP/C (Coordinate phrases per clause). In detail, papers in MICUSP have longer sentences, more clauses per sentence and per T-unit, while GRE responses have longer clauses and more coordinate phrases per clause. In terms of mean length of T-unit, clauses per sentence, verb phrases per T-unit, dependent clauses per clause, dependent clauses per T-unit, T-units per sentence, complex T-unit ratio, coordinate phrases per T-unit, complex nominals per T-unit and complex nominals per clause, there are no significant differences between the two corpora.

Propositional Idea Density

	Texttype	N	Mean	Std. Deviation	Std. Error Mean
IdeaDensity	GRE	15	.56027	.017621	.004550
	MICUSP	10	.53080	.013571	.004292

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Idea Density	Equal variances assumed	.960	.337	4.467	23	.000	.029467	.006596	.015821	.043112

Propositional idea, derived from psycholinguistic research, is "taken to be the basic unit involved in the understanding and retention of the text" (Covington, 2008). The higher the propositional idea density is, the more difficult it is to understand a text. The way of calculating it is to divide the total number of propositions by the total number of words.

From the table shown above, there are no significant difference between the variance of the two corpora ($p > 0.05$). We can see GRE responses (mean 0.56) have significantly higher propositional idea density than do papers in MICUSP (mean 0.53). The significance level is smaller than 0.001.

Coherence

Group Statistics

	Texttype	N	Mean	Std. Deviation	Std. Error Mean
Coherence	GRE	15	.27867	.052897	.013658
	MICUSP	10	.26500	.063814	.020180

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Coherence	Equal variances assumed	.447	.511	.583	23	.566	.013667	.023440	-.034823	.062156

LSA is used in this process. It provides the degree of semantic relatedness between the two adjoining segments. Folze et al (1998) state that this approach is reliable in predicting the coherence of the texts.

As shown in the table, the mean of coherence in GRE responses is 0.27, and in

MUCUSP is 0.265. The result shows there is no significant between the two corpora's means ($p > 0.05$).

Discussion

From the results I retrieved from SPSS after comparison between the two groups, we can conclude that from the surface level, GRE responses have more diverse lexical items than MICUSP, and more ideas in the text, while both genre share a lot of common characteristics on syntactic level and coherence level.

However, the result on lexical is somewhat limited due to the fact that every text in GRE responses is tremendously shorter than a real graduate level paper, which leads to more lexical variation. Affected by the size of the text, graduate papers are meant to contain more function words, such as articles, pronouns, auxiliary verbs, etc. If we use the corrected model of indices, we can actually see some of the lexical variations (such as corrected verb variation) in GRE are lower than MICUSP. Plus, the total number of sophisticated lexicals in MICUSP is twice as many as in GRE responses. We have no confidence to say that the writers in GRE responses corpus would be able to produce the similar amount of sophisticated lexicals if they write longer. Therefore, overall on lexical level, I would argue that the result could only be valid if a writer who gets high-score GRE responses continues his/her lexical development and language use in the future.

On syntactic level, if a student who writes a good GRE response wants to write a grade-A paper, s/he needs to write longer sentences with more clauses, in a word, more

syntactically complex. The reason why GRE responses have more coordination phrases appears to be that in such a short text, one needs those to express the logical transitions between sentences and paragraphs, while in longer papers, the logical transitions are more accomplished by sentence meanings.

On the propositional idea density level, intuitively, we would think as a graduate paper, it should contain more ideas comparing to a timed essay. Surprisingly, GRE responses contain more ideas in sentences. Since sentences in graduate papers are significantly longer than GRE responses, the possible reason causing graduate papers' lower density would be sentences containing more function words. Furthermore, although the difference is significant in statistics, (mean difference 0.029), in reality, it is not so conspicuous in a sense. The result indicates that if both writers of the two corpora write a 100-word long sentence together, the sentence in GRE response will contain mere two to three more propositional idea words than in graduate paper. Let alone if a one writes a sentence shorter then 50 words, the difference will be minute. Hence, I argue that if a writer can write a good GRE response on the propositional idea density level, s/he can perform well in a graduate paper.

On coherence level, the corpora don't show significant differences, which means that the two genre samples are more or less equally coherent. Nevertheless, there is some limitation using this method. LSA only provides coherence of the texts according to the semantic relations within inter- and intra- sentential level. It doesn't carry out a comprehensive analysis comparing to Coh-Metrix (Graduate-level paper is too long for Coh-Metrix to analyze). Therefore, GRE responses show similar coherence level to MICUSP

graduate papers to some extent.

For the two research questions, with the limitation in the research methods, it still appears that if one can write a good response in the GRE test, there should be little difficulty writing a graduate paper. And regarding to linguistic features in writing, GRE responses very similar, even slightly superior to graduate papers.

Pedagogical Implication

These two corpora can also be used for other purposes besides answering the two research questions. Because this is only a preliminary research, the corpora are not big enough. If we have more comprehensive corpora in different fields, we can extract the frequently used academic words (function words excluded), to form academic word lists for every field by using AntConc (Anthony, 2014). Also, we can collect texts to form corpora based on the genre. For example, if the corpus of graduate argumentative papers is big enough, we can use AntConc to generate most frequently used verbs to for the purpose of arguing, and use those as a list to teach students. AntConc can also be used to query most frequent N-grams to teach pragmatic phrases. For example, in these two corpora, some frequently used phrases are “in terms of”, “in relation to”, “and so forth”, etc. These can be taught to students to diversify their vocabulary.

Conclusion

After several means comparisons between the two corpora on lexical, syntactic,

propositional idea density and coherence level, the results show that overall, the linguistic features in GRE responses resemble those in MICUSP papers. Specifically, GRE responses have higher lexical variation than MICUSP, while the sample size difference limits the argument. Only if GRE writers can keep their pace on the development of lexical units can they write a grade-A graduate paper. Meanwhile GRE writers need to write longer sentences and clauses to better explore their syntactic complexity. Fortunately, they are already good enough in propositional idea density and coherence.

In a nutshell, ETS GRE does have a predictive validity in terms of writers' performance due to the similarity after comparisons of the two corpora.

Reference

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1999a). Writing assessment in admission to higher education: Review and framework. *College Board Report* No. 99-3. New York: College Entrance Examination Board.

Breland, H. M., Kubota, M. Y., & Bonner, M. W. (1999b). The performance assessment study in writing: Analysis of the SAT II: Writing Subject Test. *College Board Report* No. 99-4. New York: College Entrance Examination Board.

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., & Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior research methods*, 40(2), 540-545.

Covington, M. (2009, March). Idea density–A potentially informative characteristic of retrieved documents. In *IEEE SoutheastCon* (pp. 5-8).

ETS. (n.d.). Retrieved from http://www.ets.org/gre/revised_general/about/?WT.ac=grehome_greabout_b_150213

Foltz, P. W., Kintsch, W.,& Landauer, T. K. (1998). The measurement of textual Coherence with Latent Semantic Analysis. *Discourse Processes*, 25, 285-307.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474-496.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral

narratives. *The Modern Language Journal*, 96(2), 190-208.

Minnen, G., J. Carroll, and D. Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering* 7:207–223.

Powers, D. E., Fowles, M. E., & Boyles, K. (1996). Validating a GRE writing test (GRE No. 93-26B; ETS RR No. 96-27). Princeton, N]: *Educational Testing Service*.

Stewart, M. A. (2003). *GRE CAT: Answers to the Real Essay Questions*. Thomson.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003, May). Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1 (pp. 173-180). *Association for Computational Linguistics*.