A Case Study of Text and Discourse Based on Juku, Coh-Metrix, and Linguistic Inquiry Word Count (LIWC)

Jia Liu
Foreign Languages College, Guangxi University
Nanning, Guangxi, China;
1522796414@qq.com

Ying Shen*
Foreign Languages College, Guangxi University
Nanning, Guangxi, China;
Institute of Intelligent Systems, The University of Memphis
Memphis, Tennessee, USA
shenying388@gmail.com; sying@memphis.edu

***Abstract*—This paper aims to investigate the main writing features in the English time-limited compositions by third-year English major students at Guangxi University in China. Three automated assessing tools were adopted: Juku automated scoring system designed in China; Coh-Metrix automated computational evaluation tool constructed by researchers in the USA; and Linguistic Inquiry Word Count (LIWC). The results showed that there are great differences in five dimensions in the writings of high and low score compositions based on the calculation of Juku and Coh-Metrix, and there exists a strong correlation between Chinese English majors' argumentative writings and formal texts standardized by LIWC.**

Keywords-Juku automated scoring system; Coh-Metrix automated computational evaluation tool; Linguistic Inquiry Word Count (LIWC)

Ⅰ. INTRODUCTION

Juku is one of the automated scoring tools used by many universities in China. It was designed by Beijing Speech network technology. It can analyze vocabulary, grammar, collocations, and give an overall score to a text or discourse based on the technology of natural language processing, that of analysis of corpus, and that of education assessing. It is easy and free to use.

In China, many researchers have investigated how to use Juku to serve for the English writing teaching and learning. For example, Gu (2012) conducted an experiment to identify whether or not Juku could help students develop their writing abilities. The result of his experiment shows that the writing qualities in the experiment group improved more than the writing quality in the control group. Jiang & Ma (2013) also conducted an experiment to test the functions of Juku that may account for the improvements in English writings. The finding was that Juku can determine the correctness of vocabulary to some extent based on the stored corpus in Juku, but it cannot judge the appropriateness of the content or semantic meanings that are off the topic. Sometimes some evident language mistakes cannot be recognized by Juku (Chen, 2011). He (2013) found that the scores given by Juku are apparently higher than those given by veteran English writing teachers.

Coh-Metrix is a computational tool used to evaluate the linguistic features of a text or discourse. It was designed by the Institute of Intelligent Systems, the University of Memphis in the USA. This tool has been developed from version 1.0 to the present version 3.0. Coh-Metrix 3.0 can be used to confirm 108 indices, which can represent the cohesion of the explicit text and coherence of the mental representation of the text (McNamara & Graesser, 2014).

Graesser et al. discuss five major factors that account for most of the variance in texts across grade levels and text categories: word concreteness, syntactic simplicity, referential cohesion, causal cohesion, and narrativity. They consider the importance of both quantitative and qualitative characteristics of texts for assigning the right text to the right student at the right time.(Graesser, McNamara, & Kulikowich 2011)

Coh-Metrix has been used to analyze text by many researchers in China. For instance, Du & Cai (2013) found that easability, and frequency of words, referential cohesion, and sentence length have a close relationship to the quality of writings by English major students in China. Based on the analysis of the data, they constructed a formula that can interpret 44% of highly scored English writings. This formula is: final score of a writing=118.633−(0.778×easability)+(0.062×word numbers)−(0.15 ×the minimum concreteness of content words).

Linguistic Inquiry Word Count (LIWC) is a text analysis software program designed to calculate the degree to which people use different categories of words across a wide array of texts (Pennebaker, Booth, and Francis, 2007). Because LIWC can demonstrate the features of vocabulary use by the calculation of the mean use of individual variables (including self-references, social words, positive emotions, negative emotions, overall cognitive words, articles and big words), it has been applied in much research concerning analyzing the characteristics of vocabulary use. (Duan, et al. 2014)

In this paper, the following research questions will be addressed:

• Are there any relationships between high and low score compositions in relation to linguistic features in the selected argumentative writings according to three

automated scoring systems (Juku, Coh-Metrix, and LIWC)?

• What can we learn from the statistic analysis of this paper?

Ⅱ. RESEARCH METHODOLOGY

A. Subjects, topics, and time

150 English compositions from 45 third-year English major students at Guangxi University were analyzed. The selected topics are confined to argumentative writings because in China argumentative writings are the main genres represented in TEM-8 (Test for English Majors Band 8). Selected subjects are students who chose the author's English academic writing class as a selective lesson during the first semester of 2014.

B. Software

Juku Automated Scoring System, Coh-Metrix Automated Computational Evaluation Tool, and Linguistic Inquiry Word Count (LIWC).

C. Procedure

First, 150 compositions were input into the Juku automated scoring system after the author corrected all the spelling mistakes in all the compositions; then final scores of each composition were drawn from the automated scoring system Juku.

Second, the 150 English compositions were divided into three groups according to the data from Juku: high score group (scores higher than 90), intermediate score group (scores higher than 80), and low score group. (scores higher than 70)

Third, 50 English compositions in each group were copied into Coh-Metrix automated computational evaluation tool and then 108 indices of the linguistic and discourse representations of subjects' English compositions were automatically calculated.

Fourth, 50 English compositions in each group were pasted into LIWC to draw out the data of seven variables of vocabulary use.

Ⅲ. RESULTS AND DISCUSSIONS

• Noun overlap, adjacent sentences (binary and mean); argument overlap, adjacent sentences (binary and mean); stem overlap, adjacent sentences (binary and mean); noun overlap, all sentences (binary and mean); and argument overlap, all sentences (binary and mean) present a high relationship with high score compositions. (refer to Table 1, items 28 through 33)

• Lexical diversity demonstrated a significant relationship with the high score compositions. In other words, type-token ratio, content word lemmas; type-token ratio (all words); MTLD (all words) and VOCD (all words) affect much the quality of argumentative writings. (refer to Table 1, items 46 through 49)

• Two indices of syntactic complexity are highly related with the high score compositions. Put in other words, left embeddedness, words before main verb, and the mean number of modifiers per noun phrase are more commonly seen in high score compositions than those in low score ones. (refer to Table 1, items 67 and 68)

• Syntactic pattern density of preposition phrase density, agentless passive voice density, and gerund density are much higher in the the high score English writings than those in low score ones. In other words, preposition phrase density (incidence), agentless passive voice density (incidence), and gerund density (incidence) have the strong influence on the quality of argumentative compositions. (refer to Table 1, items 77 and 78)

• Word information of noun, verb and adjective incidences in highly scored English writings are evidently existing than in lowly scored writings. Put in other words, the number of left embeddedness, words before main verb (mean) and the number of modifiers per noun phrase (mean) are significantly more than those in low score compositions. In addition, word information of concreteness for content words (mean), imagability for content words (mean), hypernymy for verbs (mean) and hypernymy for nouns and verbs (mean) have a high relationship with high score compositions. (refer to Table 1, items 97 and 98)

• Based on Table 2, it is clearly seen that the lexical use tendency of seven variables [Self-references (I, me, my), Social words, Positive emotions, Negative emotions, Overall cognitive words, Articles (a, an, the), Big words (> 6 letters)] in the high score groups through low score groups' argumentative compositions is highly correlated with that of formal texts rather than personal texts. From this it can be considered that Chinese English majors have gained the ability to write their argumentative compositions from at least seven variables as a whole even though there exist obvious differences in using each individual variable.

Table Ⅰ: Consecutive Outcomes of 150 English Argumentative Compositions Based on Juku and Coh-Metrix

Descriptive	DESPC	DESSC	DESWC	DESPL	DESPLd	DESSL	DESSLd	DESWLsy	DESWLsyd	DESWLlt	DESWLltd
Low score (70)	4.88	19.78	340.18	4.189	1.88076	18.4959	8.78346	1.52176	0.85498	4.57282	2.53514
Middle score (80)	4.9	21.08	356.28	4.34734	2.00152	17.66262	8.58336	1.594	0.94948	4.78736	2.65818
High score (90)	4.66	13.02	346.14	2.67532	0.94328	37.49478	21.26744	1.64864	0.97412	4.99136	2.81772
Items	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Text Easability Principal Component Scores	PCNARz	PCNARp	PCSYNz	PCSYNp	PCCNCz	PCCNCp	PCREFz	PCREFp			
Low score (70)	0.2596	59.3226	-0.07386	48.9744	-0.4475	38.3288	0.40012	60.7826			
Middle score (80)	0.00548	50.0182	0.09414	53.5716	-0.42168	38.2992	-0.43346	36.3766			
High score (90)	-0.10438	46.3418	-1.81284	23.4434	0.40104	59.0682	0.89752	58.9794			
Items	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**			
Text Easability Principal Component Scores	PCDCz	PCDCp	PCVERBz	PCVERBp	PCCONNz	PCCONNp	PCTEMPz	PCTEMPp			
Low score (70)	1.0718	75.3082	0.39092	62.7254	-2.64246	3.6054	-0.60204	31.8482			
Middle score (80)	0.95942	75.1628	0.2611	56.2724	-3.05508	2.101	-0.71636	31.0958			
High score (90)	1.25038	79.99	0.3311	57.275	-2.71406	3.9846	-0.9264	27.251			
Items	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**			
Reference Cohesion	CRFNO1	CRFAO1	CRFSO1	CRFNOa	CRFAOa	CRFSOa	CRFCWO1	CRFCWO1d	CRFCWOa		
Low score (70)	0.44704	0.63264	0.53636	0.3801	0.53694	0.4647	0.13518	0.12146	0.11306		
Middle score (80)	0.33192	0.50606	0.40042	0.27318	0.42452	0.3264	0.10542	0.11158	0.0839		
High score (90)	0.6246	0.7253	0.72226	0.56772	0.67136	0.65746	0.1163	0.07668	0.11152		
Items	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**		
Reference Cohesion	CRFCWOad	LSASS1	LSASS1d	LSASSp	LSASSpd	LSAPP1	LSAPP1d	LSAGN	LSAGNd		
Low score (70)	0.11466	0.2425	0.16666	0.2303	0.16452	0.46732	0.0955	0.33338	0.13336		
Middle score (80)	0.1004	0.18794	0.15038	0.17386	0.1503	0.37746	0.10314	0.29898	0.12354		
High score (90)	0.08028	0.28814	0.1178	0.18938	0.09828	0.40032	0.09568	0.29802	0.15002		
Items	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**		
Lexical Diversity	LDTTRc	LDTTRa	LDMTLD	LDVOCD							
Low score (70)	0.6431	0.46552	69.87366	73.17234							
Middle score (80)	0.70644	0.50592	84.34416	88.00562							
High score (90)	0.7492	0.5424	93.49182	93.55942							
Items	**46**	**47**	**48**	**49**							
Connectives	CNCAll	CNCCaus	CNCLogic	CNCADC	CNCTemp	CNCTempx	CNCAdd	CNCPos	CNCNeg		
Low score (70)	98.70362	32.079844	48.86584	19.18672	18.49786	14.4845	52.32358	0	0		
Middle score (80)	107.02248	34.09912	43.59986	18.27312	17.26696	12.22482	59.07066	0	0		
High score (90)	101.72216	32.120116	44.93724	19.51774	18.99014	13.41554	54.60536	0	0		
Items	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**		
Situation Model	SMCAUSv	SMCAUSvp	SMINTEp	SMCAUSr	SMINTEr	SMCAUSlsa	SMCAUSwn	SMTEMP			
Low (70)	29.23322	45.655268	16.65386	0.58046	1.91618	0.10742	0.51292	0.7816			
Middle score (80)	31.32654	50.4151	18.13498	0.59004	1.40858	0.10214	0.51138	0.76756			
High score (90)	18.73192	35.9958	9.12538	1.05852	2.34734	0.08248	0.47634	0.75362			
Items	**59**	**60**	**61**	**62**	**63**	**64**	**65**	**66**			
Syntactic Complexity	SYNLE	SYNNP	SYNMEDpos	SYNMEDwrd	SYNMEDlem	SYNSTRUTa	SYNSTRUTt				
Low score (70)	4.82836	0.8252	0.64022	0.87858	0.85724	0.10886	0.10422				
Middle score (80)	4.72646	0.78488	0.65426	0.89042	0.86982	0.11386	0.10898				
High score (90)	8.06788	0.85896	0.49186	0.67754	0.66382	0.07328	0.07454				
Items	**67**	**68**	**69**	**70**	**71**	**72**	**73**				
Syntactic Pattern Density	DRNP	DRVP	DRAP	DRPP	DRPVAL	DRNEG	DRGERUND	DRINF			
Low score (70)	345.14418	227.02258	36.21434	97.03494	6.819	13.15948	16.3945	21.70734			
Middle score (80)	357.30	223.239	37.68576	100.0499	7.1775	11.5737	14.79602	17.68458			

	83	9									
High score (90)	361.6887	212.81996	34.22756	113.13512	7.82886	10.58786	21.25382	20.07952			
Items	**74**	**75**	**76**	**77**	**78**	**79**	**80**	**81**			
Word Information	WRDNOUN	WRDVERB	WRDADJ	WRDADV	WRDPRO	WRDPRP1s	WRDPRP1p				
Low score (70)	239.37358	114.27688	100.43978	63.21856	74.1534	7.4206	31.77006				
Middle score (80)	244.995	116.04968	102.2869	61.16122	79.15904	6.3816	37.24956				
High score (90)	259.0718	119.68128	109.037	59.45116	57.60416	5.7428	21.73056				
Items	**82**	**83**	**84**	**85**	**86**	**87**	**88**				
Word Information	WRDPRP2	WRDPRP3s	WRDPRP3p	WRDFRQc	WRDFRQa	WRDFRQmc	WRDAOAc				
Low score (70)	5.67096	3.10182	15.57704	2.39748	3.05936	1.24938	356.99378				
Middle score (80)	5.0146	2.92876	13.68572	2.29058	2.98986	1.17566	366.0049				
High score (90)	1.55306	1.80408	15.54126	2.26938	2.97554	0.82034	373.16064				
Items	**89**	**90**	**91**	**92**	**93**	**94**	**95**				
Word Information	WRDFAMc	WRDCNCc	WRDIMGc	WRDMEAc	WRDPOLc	WRDHYPn	WRDHYPv	WRDHYPnv			
Low score (70)	579.99111	357.5257	393.690902	429.96194	4.19216	6.84896	1.55956	1.78254			
Middle score (80)	575.7212	353.15152	390.246884	431.24542	4.18514	6.69792	1.56302	1.77528			
High score (90)	575.1455	360.44532	397.424006	432.91318	4.03698	6.53606	1.5828	1.8359			
Items	**96**	**97**	**98**	**99**	**100**	**101**	**102**	**103**			
Readability	RDFRE	RDFKGL	RDL2								
Low score (70)	59.32084	9.58022	21.93748								
Middle score (80)	54.05508	10.10768	18.31662								
High score (90)	31.29726	18.487	15.92714								
Items	**104**	**105**	**106**								

Table Ⅱ: Correlations between three groups in seven variables [Self-references (I, me, my), Social words, Positive emotions, Negative emotions, Overall cognitive words, Articles (a, an, the), Big words (> 6 letters] based on LIWC

		High score group	Middle score group	Low score group	Personal texts	Formal texts
Middle score group	Pearson Correlation	1				
	Sig. (2-tailed)					
	variables	7				
Middle score group	Pearson Correlation	.992**	1			
	Sig. (2-tailed)	.000				
	variables	7	7			
Low score group	Pearson Correlation	.982**	.989**	1	.	
	Sig. (2-tailed)	.000	.000			
	variables	7	7	7		
Personal texts	Pearson Correlation	.646	.721	.709	1	.
	Sig. (2-tailed)	.117	.068	.074		.
	variables	7	7	7	7	
Formal texts	Pearson Correlation	.989**	.995**	.993**	.718	1
	Sig. (2-tailed)	.000	.000	.000	.069	
	variables	7	7	7	7	7

**. Correlation is significant at the 0.01 level (2-tailed).

Ⅳ. CONCLUSIONS

First, from the data analysis five dimensions (referential cohesion, lexical diversity, syntactic complexity, syntactic pattern density, and word formation) in a text or discourse are confirmed to exert the important role of getting a high score on argumentative compositions.

Second, during the copying of low score compositions, many spelling mistakes were obviously observed in low score compositions while this was seldom the case in high score compositions. However, because each spelling mistake was corrected before they were copied into Coh-Metrix, it can be inferred that the main reason of low score compositions is not necessarily vocabulary misspellings, many other reasons can be predicted.

Third, based on the LIWC outcomes, it can be concluded that English majors in China can to a great extent write formal compositions required by some necessary standards. Especially, the use of seven variables [Self-references (I, me, my), Social words, Positive emotions, Negative emotions, Overall cognitive words, Articles (a, an, the), Big words (> 6 letters)] in the argumentative compositions written by Chinese English majors seems equal to that of formal texts calculated by LIWC.

ACKNOWLEDGMENT

This research is financially supported by the core research project (XBS100008) of Guangxi University in 2010 and Innovative Research Foundation for National University Students in 2014 (NO141059349).

REFERENCES

[1] A. C. Graesser, D. S. McNamara, and J.M. Kulikowich., Coh-Metrix: Providing Multilevel Analyses of Text Characteristics, *Educational Researcher*, vol.40, pp.223-234, 2011.

[2] C. H. Gu & L. Wang, An empirical study of college English writing based on Juku. *Journal of Yangzhou University*, vol.16, no.4, pp.92-96, 2012.

[3] D.S. McNamara, A.C. Graesser, P.M. McCarthy, & Z. Cai, *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge Press, 2014.

[4] H. Y. Du & J. T. Cai, Expected model of accessing writing quality of Chinese English learners based on Coh-Metrix. *Modern Foreign Languages*,vol.3, pp.34-39, 2013.

[5] J. W. Pennebaker, R.J. Booth, and M.E. Francis. *Linguistic Inquiry and Word Count: LIWC,* Austin, TX: LIWC. net. 2007.

[6] X. L. He, The study of validity and reliability of Juku for the assessing of English compositions. *Modern Education Technology*, vol.5, pp.55-59, 2013.

[7]Y. Duan et al. Linguistic style and social historical context: an automated linguistic analysis of Mao Zeddong's speeches, *Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference*, pp.43-46, 2014.

[8] Y. Chen, Investigation and analysis on the application of automated essay scoring system of college English writing teaching, *Chinese Journal*, vol.11, pp.70-72, 2011.

[9] Y. Jiang & W.L. Ma, Achievements or challenges? Inquiring the intelligent tutor's functions on Chinese English writings. *Electronic Education Research*, vol.7, pp.35-38, 2013.