



Automating and Evaluating Measures of L2 Lexical Complexity

Xiaofei Lu

Department of Applied Linguistics

Pennsylvania State University

XXL13@psu.edu

CALICO-09, Tempe, AZ



Outline

- Introduction
- Measures of L2 lexical complexity
- Automating measures of lexical complexity
- Evaluating measures of lexical complexity
- Conclusion



Introduction

- Conceptualizing lexical complexity
 - A multidimensional feature of learner language use encompassing lexical density, sophistication, and variation (Wolfe-Quintero et al. 1998; Read 2000)
 - Does not focus on error
- Research goals
 - Design a computational tool to automate lexical complexity analysis using 25 measures
 - Evaluate the relationship of these measures plus the D measure to L2 oral proficiency



Introduction (cont.)

○ Motivation

- Lexical complexity an important construct in L2 teaching and research
- Relationship between lexical complexity and proficiency claimed in many test rating scales



Lexical complexity measures

- L2 complexity measures proposed in SLA studies and reviewed in
 - Wolfe-Quintero et al. (1998)
 - Read (2000)
 - Malvern et al. (2004)
- Measures of the following three dimensions
 - Lexical density
 - Lexical sophistication
 - Lexical variation



Lexical density

- Proportion of lexical words (N_{lw} / N) (Ure 1971)
- Previous findings
 - Lower in spoken than written texts (Halliday 1985)
 - Affected by various sources (O'Loughlin 1995)
 - Relation to L2 writing non-significant (Engber 1995)
- Inconsistent definition of lexical words
 - All nouns and adjectives
 - Adverbs with adjective base
 - Full verbs (excluding modal/auxiliary verbs)



Lexical sophistication

- Five measures examined

- LS1: N_{slw} / N
- LS2: T_s / T
- VS1: T_{sv} / N_v
- CVS1: $T_{sv} / \text{sqrt}(2N_v)$
- VS2: T_{sv}^2 / N_v



Lexical sophistication (cont.)

- Previous findings
 - LS1: NS-NNS dif sig (Linnarud 86); non-sig (Hyltenstam 88)
 - LS2: sig pre-and post-essay dif (Laufer 04)
 - VS1: sig NS-NNS dif (Harley & King 89)
- Varying definitions of sophistication
 - 2000-word BNC frequency list (Leech et al. 01)



Lexical variation

- 20 measures examined
- 4 based on NDW
 - NDW: Number of different words
 - NDW-50: NDW in first 50 words of sample
 - NDW-ER50: mean NDW of 10 random 50-word subsamples
 - NDW-ES50: mean NDW of 10 random n-word sequences



Lexical variation (cont.)

- 7 based on TTR for total vocabulary
 - Type token ratio (TTR)
 - Mean TTR of all 50-word segments (MSTTR)
 - LogTTR, Corrected TTR, Root TTR, Uber
 - The D measure (McKee et al. 2000)
- 9 based on TTR for word classes
 - $T_{\{LW, V, N, Adj, Adv, Mod\}} / N_{lw}$
 - $T_v / N_v, T_v^2 / N_v, T_v / \text{sqrt}(2N_v)$



Lexical variation (cont.)

- Previous findings
 - NDW and TTR useful, but affected by sample size
 - Transformations of NDW/TTR not equally useful
 - D claimed superior; results mixed (Jarvis 02; Yu 07)
 - Mixed results for word class TTR measures
 - No consensus on a single best measure



Research questions

- How does LD relate to quality of L2 oral narratives?
- How do the LS measures compare with and relate to each other as indices of quality of L2 oral narratives?
- How do the LV measures compare with and relate to each other as indices of quality of L2 oral narratives?
- How do LD, LS and LV compare with and relate to each other as indices of quality of L2 oral narratives?



Data selection

- Spoken English Corpus of Chinese Learners (Wen et al. 05)
 - Transcripts of TEM-4 Spoken Test data in 96-02
 - Students ranked within groups of 32-35
 - Task two data used – 3-minute oral narratives
 - 12 groups of data used (99-02; N=32-35 each)
- Example topic (2001)
 - Describe a teachers of yours whom you found unusual



Data cleaning

- Each file contains
 - A header with various information
 - A transcript of a student's performance in Task 2
 - Erroneous form enclosed in brackets and preceded by corrected form
- Files cleaned as follows
 - Headers and tagged erroneous words removed
 - Fillers removed (e.g., *ah, eh, er, mm, oh, um*)
 - Extra elliptical marks removed



Computing the D measure

- Cleaned text files converted to CHAT format
 - The textin utility in CLAN (MacWhinney et al. 2000)
 - CHAT: Codes for the Human Analyses of Transcripts
- Analysis of CHAT files
 - The vocd utility in CLAN
 - Average of 3 measures taken for each sample



Computing other measures

- Part-of-speech tagging (Stanford tagger)
 - It_PRP happened_VBD one_CD year_NN ago_RB ._.
- Lemmatization (Morpha)
 - It_PRP happen_VBD one_CD year_NN ago_RB ._.



Computing other measures (cont.)

- Type and token counting
 - **Types:** w, sw, lw, slw, v, sv, n, adj, adv
 - **Tokens:** w, lw, slw, v
 - {sw, slw, sv} = {w, lw, v} not in the 2,000-word BNC frequency list (Leech et al. 2001)
 - Punctuation marks not counted as type or token
- Computing ratios of 25 measures other than D



Analysis

- Spearman's rho computed for each group
 - X: test takes' rankings within the group
 - Y: Values of each of the 26 measures
- Meta-analysis to combine results
 - Combined effect size (rho) = weighted average
 - Weights computed using inverse variance method
 - Samples with larger N, smaller SE weighted more
- Pearson's correlation between the measures



Results

	P<.01	P<.05	P<.10	Avg rho	Sig.
LD	0	1	0	.011	.836
LS1	0	0	0	.048	.355
LS2	0	0	0	.050	.336
VS1	0	0	2	.133	.010
CVS1	0	1	2	.166	.001
VS2	0	1	2	.165	.001



Results (cont.)

	LS1	LS2	VS1	CVS1	VS2
LS1	1				
LS2	.637**	1			
VS1	.456**	.391**	1		
CVS1	.414**	.382**	.966**	1	
VS2	.381**	.359**	.909**	.935**	1



Results (cont.)

	P<.01	P<.05	P<.10	Avg rho	Sig.
NDW	10	1	0	.526	.000
NDW-50	1	0	2	.176	.001
NDW-ER	2	3	1	.282	.000
NDW-ES	3	2	3	.319	.000



Results (cont.)

	NDW	NDW-50	NDW-ER50
NDW-50	.274**		
NDW-ER50	.579**	.448**	
NDW-ES50	.478**	.448**	.731**



Results (cont.)

	P<.01	P<.05	P<.10	Avg rho	Sig.
TTR	0	0	0	-.038	.459
MSTTR	2	2	3	.332	.000
CTTR	7	2	2	.430	.000
RTTR	7	2	2	.429	.000
LogTTR	0	0	0	.082	.113
Uber	0	3	2	.204	.000
D	4	3	1	.352	.000



Results (cont.)

	TTR	MSTTR	CTTR	RTTR	LogTTR	Uber
MSTTR	.592**					
CTTR	.525**	.737**				
RTTR	.526**	.737**	1.000			
LogTTR	.953**	.697**	.719**	.719**		
Uber	.882**	.746**	.845**	.845**	.949**	
D	.528**	.741**	.805**	.806**	.626**	.762**



Results (cont.)

	P<.01	P<.05	P<.10	Avg rho	Sig.
LV	1	0	0	.063	.225
VV1	0	1	0	.047	.363
SVV1	2	3	1	.291	.000
CVV1	2	3	1	.290	.000
VV2	0	0	0	-.091	.080
NV	0	1	0	.015	.778
AdjV	0	1	0	.074	.155
AdvV	0	1	3	.203	.000
ModV	0	0	3	.173	.001

Results (cont.)

	LV	VV1	SVV1	CVV1	VV2	NV	AdjV	AdvV
VV1	.786**							
SVV1	.452**	.639**						
CVV1	.457**	.649**	.994**					
VV2	.524**	.443**	.530**	.535**				
NV	.714**	.283**	.120*	.119*	.324**			
AdjV	.511**	.443**	.103*	.105*	-.089	.321**		
AdvV	.208**	.153**	.212**	.215**	.173**	.137**	-.038	
ModV	.558**	.475**	.192**	.196**	-.002	.354**	.886**	.414**



Results (cont.)

- Results for Research Question 4 being analyzed
 - Relationship between density, sophistication and variation



Conclusion

- Use of NLP technology to automate analysis of lexical complexity using 26 measures
- Corpus-based evaluation of how these measures relate to and compare with each as indices of L2 proficiency
- Potential applications by learner, teacher and researcher