



TRINITY
COLLEGE LONDON



Measuring lexical complexity in L2 spoken production: Evidence from the Trinity Lancaster Corpus

Raffaella Bottini

r.bottini@lancaster.ac.uk

 [@RaffaBottini](https://twitter.com/RaffaBottini)

Lexical complexity: the construct

Lexical density

the proportion
between content
and function
words

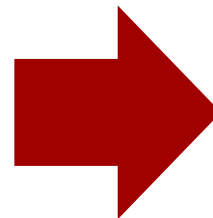
Lexical diversity

the use of a
range of different
words

Lexical sophistication

the use of
advanced words

Text complexity



Lexical items:

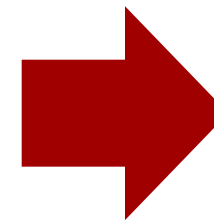
- Wide variety
- Low frequency

Lexical complexity in L2 English

Much discussed topic with experimental studies on

Methodological aspects:

- validation of measures (e.g. McCarthy & Jarvis, 2013)
- automatic tools to measure lexical complexity:
 - TAALES and TAALED* (Kyle & Crossley, 2015)
 - Lexical Complexity Analyzer* (Lu, 2012)
 - Coh-Metrix* (Graesser et al, 2004)



- none of these tools compute **all** existing indices of lexical complexity
- not flexible
- mainly written **reference corpora**

Learner language:

- lexical choices across proficiency levels (e.g. Kim et al., 2018)

Lexical complexity in L2 English speech

- Few studies
- Limited research focus in terms of components of lexical complexity
- Small corpora (e.g. 244 texts in Kyle & Crossley, 2015)
- Insufficient metadata: learners' characteristics (e.g. age, L1)
- Proficiency rated using different scales
- Tasks: mainly monologues, no topic choice, analyses on combinations of tasks
- Using different lexical measures + mainly written reference corpora

Lexical complexity in L2 English speech

- Few studies on lexical complexity in **L2 speech**
- Limited research focus in terms of **components of lexical complexity**
- Small corpora (e.g. 244 texts in Kyle & Crossley, 2015)
- Insufficient metadata: **learners' characteristics** (e.g. age, L1)
- Proficiency rated using different scales
- **Tasks**: mainly monologues, no topic choice, analyses on combinations of tasks
- Using a variety of lexical measures + mainly written reference corpora

How can lexical complexity be measured in L2 speech?

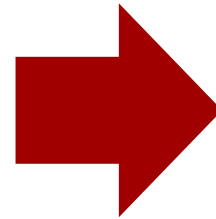
Validating lexical complexity indices

Validation involves “accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations” (AERA, 2014: 11)

Existing validating studies:

- only on a group of lexical diversity measures
- sensitivity to variation of text length
- parallel sampling method:
 score on a whole text VS
 average score on sections of a text
- mainly on L1 written language
- two studies on L2 English speech

(Lu, 2012; Koizumi & In'nami, 2012)



In this study:

- All indices of lexical complexity
- Learner and task-related variables
- Correlations with text length using full texts
- Large dataset of L2 English speech

Research question

How reliable is the performance of lexical complexity measures on L2 spoken production?

Reliability is an “independent characteristic of a test score” based on its generalizability and “consistency [...] across instances of the testing procedure” (AERA 2014: 33–34).

e.g. sensitivity of lexical complexity measures to variations of text length

Trinity Lancaster corpus



Size

4.2 million words - 2,053 speakers

Language

Graded Examination in Spoken English (GESE)
Trinity College London

Proficiency

B1, B2, C1/C2 levels

L1

variety of backgrounds (e.g. Argentina, China, India, Italy, Mexico, Spain, ...)

Ages

8 to 72 years old

Trinity General English Spoken Exams(GESE)



		Proficiency			Topic choice (familiarity)	Interlocutors' roles	Interactivity
		B1	B2	C1/2			
TASKS	Presentation			✓	candidate	candidate-led	monologic
	Discussion	✓	✓	✓	candidate	jointly-led	dialogic
	Interaction		✓	✓	examiner	candidate-led	dialogic
	Conversation	✓	✓	✓	examiner	jointly-led	dialogic

Trinity General English Spoken Exams(GESE)



	Proficiency			Topic choice (familiarity)	Interlocutors' roles	Interactivity
	B1	B2	C1/2			
Presentation			✓	candidate	candidate-led	monologic
Discussion	✓	✓	✓	candidate	jointly-led	dialogic
Interaction		✓	✓	examiner	candidate-led	dialogic
Conversation	✓	✓	✓	examiner	jointly-led	dialogic

Dataset

A subset of the Trinity Lancaster Corpus

		No. of learners	tokens			
			total	mean (SD)	min	max
proficiency level	B1	933	651,018	697.77 (190.28)	199	1,591
	B2	805	727,591	903.84 (215.36)	393	1,694
	C1/2	315	345,479	1,096.76 (258.98)	335	1,917
	All levels	2,053	1,724,088	839.79 (256.51)	199	1,917

Methodology

- Creating a wordlist from the Spoken BNC2014 (Love et al., 2017) based on ARF (Brezina & Gablasova, 2015)
- Creating ***Lex Complexity Tool*** (Bottini, under review)
existing and new complexity indices + Spoken BNC2014 wordlist
- Measuring lexical complexity
- Statistical analysis: Pearson's correlations with text length
linear regression analysis (AIC)

Lexical complexity: indices

Lexical density

Lexical diversity

Lexical sophistication

type-token ratio

segmental values

probability of different words

TTR
CTTR
RTTR
LogTTR
Maas
Uber

MSTTR
MTLD
MATTR

(Voc-D)
HD-D

Lexical diversity and text length

var. 1	var. 2	<i>r</i>	95% CI
No. of tokens	ttr	-.66	[-.68, -.64]
	cttr	.38	[.34, .42]
	rtrr	.38	[.34, .42]
	logttr	-.36	[-.39, -.32]
	maas	-.15	[-.19, -.10]
	uber	.14	[.10, .19]
	msttr	.31	[.27, .34]
	mtld	.31	[.27, .35]
	mattr	.30	[.26, .34]
	hd-d	.34	[.30, .38]

All $p < .001$

Inter-index correlations with TTR:

Maas $r = -.53$ CI [-.56, -.50]

Uber $r = .61$ CI [.59, .64]

	small	medium	large
Correlation (<i>r</i>)	.25	.4	.6

(Plonsky & Oswald, 2014)

(cf. McCarthy & Jarvis, 2007, 2010)

Lexical diversity and text length

var. 1	var. 2	<i>r</i>	95% CI
No. of tokens	ttr	-.66	[-.68, -.64]
	cttr	.38	[.34, .42]
	rttr	.38	[.34, .42]
	logttr	-.36	[-.39, -.32]
	maas	-.15	[-.19, -.10]
	uber	.14	[.10, .19]
	msttr	.31	[.27, .34]
	mtld	.31	[.27, .35]
	mattr	.30	[.26, .34]
	hd-d	.34	[.30, .38]

All $p < .001$

Inter-index correlations with TTR:

MTLD $r = .23$ CI [.19, .27]

MATTR $r = .28$ CI [.24, .32]

HD-D $r = .26$ [.21, .33]

	small	medium	large
Correlation (r)	.25	.4	.6

(Plonsky & Oswald, 2014)

(cf. McCarthy & Jarvis 2013)

Lexical diversity and text length

var. 1	var. 2	<i>r</i>	95% CI
No. of tokens	ttr	-.66	[-.68, -.64]
	cttr	.38	[.34, .42]
	rttr	.38	[.34, .42]
	logttr	-.36	[-.39, -.32]
	maas	-.15	[-.19, -.10]
	uber	.14	[.10, .19]
	msttr	.31	[.27, .34]
	mtld	.31	 [.27, .35]
	mattr	.30	[.26, .34]
	hd-d	.34	[.30, .38]

All $p < .001$

	small	medium	large
Correlation (<i>r</i>)	.25	.4	.6
Effect size (r^2)	.01	.09	.25

(Plonsky & Oswald, 2014)

Regression analysis: lexical diversity

Outcome	Predictors	Estimate	SE	<i>p</i>	F (Df)	Adj. <i>r</i> ²	<i>p</i>
MTLD	tokens	.01	.00	***	51.38 (6, 2989)	.09	***
	task disc.	4.27	.46	***			
	C1/C2 level	4.04	.49	***			
	B2 level	1.98	.38	***			
	L1 Spanish	.05	.30	.87			
	L1 Chinese	2.46	.37	***			
MATTR	tokens	.00	.00	***	65.19 (6, 2989)	.11	***
	task disc.	.02	.00	***			
	C1/C2 level	.03	.00	***			
	B2 level	.01	.00	***			
	L1 Spanish	.00	.00	*			
	L1 Chinese	.01	.00	***			
HD-D	tokens	.00	.00	***	109.10 (6, 2989)	.18	***
	task disc.	.01	.00	***			
	C1/C2 level	.02	.00	***			
	B2 level	.01	.00	***			
	L1 Spanish	-.01	.00	***			
	L1 Chinese	.01	.00	***			

- Outcomes:
- lexical indices
- Predictors:
- no. of tokens
 - task type
 - proficiency level
 - L1

- Baseline values:
- Conversation task
 - B1 proficiency level
 - L1 Italian

Lexical complexity: indices

Lexical sophistication

1. lexical unit

lemmas

$$vs3 = \frac{\text{s. verb types}}{\text{total verb types}}$$

ns

adjs

adv

3. type of indices

frequency bands

$$ls1 = \frac{\text{s. lex. tokens}}{\text{total lex. tokens}}$$

$$ls2 = \frac{\text{s. types}}{\text{total types}}$$

vs1

vs2

cv

mean frequency

$$\log AW = \log (\text{mean freq. of all words})$$

logCW

logFW

2. reference corpus

Spoken BNC2014

Lexical sophistication and text length

var. 1	var. 2	<i>r</i>	95% CI
No. of tokens	ls1	-.02 (<i>p</i> =.41)	[-.06, .03]
	ls2	.32	[.28, .36]
	vs1	-.01 (<i>p</i> =.62)	[-.05, .03]
	vs2	.29	[.25, .33]
	vs3	.19	[.15, .23]
	cvs1	.31	[.27, .34]
	ns	.18	[.14, .22]
	adjs	.20	[.16, .24]
	advs	.12	[.08, .16]
	logAW	.03 (<i>p</i> =.21)	[-.02, .07]
logCW	.11	[.07, .16]	
logFW	-.01 (<i>p</i> =.74)	[-.05, .04]	

All *p* < .001 except where otherwise specified.

	small	medium	large
Correlation (<i>r</i>)	.25	.4	.6

(Plonsky & Oswald, 2014)

Lexical sophistication and text length

var. 1	var. 2	<i>r</i>	95% CI
No. of tokens	ls1	-.02 (<i>p</i> =.41)	[-.06, .03]
	ls2	.32	[.28, .36]
	vs1	-.01 (<i>p</i> =.62)	[-.05, .03]
	vs2	.29	[.25, .33]
	vs3	.19	[.15, .23]
	cvs1	.31	[.27, .34]
	ns	.18	[.14, .22]
	adjs	.20	[.16, .24]
	advs	.12	[.08, .16]
	logAW	.03 (<i>p</i> =.21)	[-.02, .07]
	logCW	.11	[.07, .16]
logFW	-.01 (<i>p</i> =.74)	[-.05, .04]	

All *p* < .001 except where otherwise specified.

	small	medium	large
Correlation (<i>r</i>)	.25	.4	.6
Effect size (<i>r</i> ²)	.01	.09	.25

(Plonsky & Oswald, 2014)

Regression analysis: frequency bands

Outcome	Predictors	Estimate	SE	<i>p</i>	F (Df)	Adj. <i>r</i> ²	<i>p</i>
ls2	tokens	.00	.00	***	84.89 (6, 2989)	.14	***
	task disc.	.03	.00	***			
	C1/C2 level	.01	.00	*			
	B2 level	.00	.00	.70			
	L1 Spanish	-.01	.00	***			
	L1 Chinese	-.03	.00	***			
adjs	tokens	.00	.00	***	26.50 (6, 2989)	.05	***
	task disc.	.07	.01	***			
	C1/C2 level	.04	.01	***			
	B2 level	.00	.01	.72			
	L1 Spanish	-.03	.01	***			
	L1 Chinese	-.05	.01	***			
vs3	tokens	.00	.00	***	35.72 (6, 2989)	.07	***
	task disc.	.05	.01	***			
	C1/C2 level	.03	.01	***			
	B2 level	.01	.00	.25			
	L1 Spanish	-.02	.00	***			
	L1 Chinese	-.03	.00	***			

Outcomes:

- lexical indices

Predictors:

- no. of tokens
- task type
- proficiency level
- L1

Baseline values:

- Conversation task
- B1 proficiency level
- L1 Italian

Regression analysis: mean frequency

Outcome	Predictors	Estimate	SE	<i>p</i>	F (Df)	Adj. <i>r</i> ²	<i>p</i>
logAW	tokens	.00	.00	**	33.25 (6, 2989)	.06	***
	task disc.	-.01	.00	*			
	C1/C2 level	-.01	.00	*			
	B2 level	.00	.00	.56			
	L1 Spanish	.02	.00	***			
	L1 Chinese	-.01	.00	***			
logCW	tokens	.00	.00	**	96.20 (6, 2989)	.16	***
	task disc.	-.03	.01	**			
	C1/C2 level	.10	.01	***			
	B2 level	.00	.01	.78			
	L1 Spanish	.09	.01	***			
	L1 Chinese	.08	.01	***			
logFW	tokens	.00	.00	.45	8.63 (6, 2989)	.02	***
	task disc.	-.01	.00	**			
	C1/C2 level	-.02	.00	***			
	B2 level	.00	.00	.16			
	L1 Spanish	.00	.00	*			
	L1 Chinese	-.01	.00	**			

Outcomes:

- lexical indices

Predictors:

- no. of tokens
- task type
- proficiency level
- L1

Baseline values:

- Conversation task
- B1 proficiency level
- L1 Italian

Regression analysis: comparison

Outcome	Predictors	Estimate	SE	<i>p</i>	F (Df)	Adj. <i>r</i> ²	<i>p</i>
adjs	tokens	.00	.00	***	26.50 (6, 2989)	.05	***
	task disc.	.07	.01	***			
	C1/C2 level	.04	.01	***			
	B2 level	.00	.01	.72			
	L1 Spanish	-.03	.01	***			
	L1 Chinese	-.05	.01	***			
logCW	tokens	.00	.00	**	96.20 (6, 2989)	.16	***
	task disc.	-.03	.01	**			
	C1/C2 level	.10	.01	***			
	B2 level	.00	.01	.78			
	L1 Spanish	.09	.01	***			
	L1 Chinese	.08	.01	***			

Regression analysis: mean frequency

Outcome	Predictors	Estimate	SE	<i>p</i>	F (Df)	Adj. <i>r</i> ²	<i>p</i>
logAW	tokens	.00	.00	**	33.25 (6, 2989)	.06	***
	task disc.	-.01	.00	*			
	C1/C2 level	-.01	.00	*			
	B2 level	.00	.00	.56			
	L1 Spanish	.02	.00	***			
	L1 Chinese	-.01	.00	***			
logCW	tokens	.00	.00	**	96.20 (6, 2989)	.16	***
	task disc.	-.03	.01	**			
	C1/C2 level	.10	.01	***			
	B2 level	.00	.01	.78			
	L1 Spanish	.09	.01	***			
	L1 Chinese	.08	.01	***			
logFW	tokens	.00	.00	.45	8.63 (6, 2989)	.02	***
	task disc.	-.01	.00	**			
	C1/C2 level	-.02	.00	***			
	B2 level	.00	.00	.16			
	L1 Spanish	.00	.00	*			
	L1 Chinese	-.01	.00	**			

Summary of results

Correlations with text length: selection of indices which are independent from text length

- MATTR, HD-D, MTLD
- ls2, vs3, ns, adjs, advs + logAW, logCW, logFW

Regression analysis: associations with learner and task-related features

→ selection of indices tailored for research on learner language

- MTLD
- sophistication indices based on word classes (content words)

Lex Complexity Tool:

- flexible
- spoken BNC2014 wordlist
- all existing + new indices

Limitations

- only two tasks examined
- only three L1s in the regression analysis
- A1 and A2 proficiency levels not included
- multi-word lexical items not considered

Conclusions and future directions

This study:

Methodological aspects of lexical complexity



Next steps:

Closer focus on:

- effect of learners' individual characteristics (proficiency, age, L1)
- effect of task related features (interactivity and topic familiarity)

References




- AERA. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.
- Bottini, R. (under review). Measures of lexical diversity and sophistication in L2 speech: A validation study based on the Trinity Lancaster Corpus.
- Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics*, 36(1), 1–22.
- Eguchi, M., & Kyle, K. (2020). Continuing to explore the multidimensional nature of lexical sophistication: The case of oral proficiency interviews. *The Modern Language Journal*, 104(2), 381–400.
- Gablasova, D., Brezina, V., & McEnery, T. (2019). The Trinity Lancaster Corpus: Development, description and application. *International Journal of Learner Corpus Research*, 5(2), 126-158.
- Graesser, A., McNamara, C., Louwse, D., & Cai, S. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36(2), 193-202.
- Kim, M., Crossley, S., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *Modern Language Journal*, 102(1), 120-141.
- Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, 40(4), 554–564.
- Kyle, K. (2019) Measuring Lexical Richness. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 454-476). Routledge.
- Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4), 757-786.
- Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3), 319-344.
- Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *Modern Language Journal*, 96(2), 190-208.
- McCarthy, P. M., & Jarvis, S. (2013). From intrinsic to extrinsic issues of lexical diversity assessment: an ecological validation study. In S. Jarvis, & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 45-78). John Benjamins.
- McCarthy, P. M., & Jarvis, S. (2010). MTL, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42, 381–392.
- McCarthy, P. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4), 459–488.
- Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research: Effect sizes in L2 research. *Language Learning*, 64(4), 878–912.



TRINITY
COLLEGE LONDON



Thank you

r.bottini@lancaster.ac.uk
 @RaffaBottini