Measuring speaking proficiency using features of lexical and lexicogrammatical use

Kristopher Kyle (University of Oregon)

UNIVERSITY OF OREGON

Colleagues involved in this and related projects:

Masaki Eguchi
UO Linguistics

Cindy Berger
(Duolingo)

Scott Crossley
(GSU)

Danielle
McNamara (ASU)

And MANY Others!

Overview of talk

- Overview of (some) features of productive proficiency
- Importance of multivariate models
- Current study

Productive lexical proficiency: Words

What linguistic features affect readers’ and interlocutors’ perceptions of language proficiency?

- Historically has focused on characteristics of word use
  - Engber (1995)
  - Laufer & Nation (1995)
  - Meara & Bell (2001)
  - Crossley & Cobb (2014)
  - etc.
- More proficient writers tend to use:
  - less frequent (more sophisticated) lexical items
  - a wider variety of lexical items (given a particular task)

Productive lexical proficiency: Words

- Other word level sophistication features have also been used:
  - concreteness (as a proxy for salience)
    - *apple* is highly concrete, while *empathy* is less concrete
  - contextual diversity
    - the number of lexical and or semantic contexts in which a word is used
    - *food* is used in a wider range of contexts than *blender*
  - etc.
- Lexical diversity
  - see Jarvis (2013a,b; 2017) for a multidimensional take on lexical diversity
  - Many measures still used vary with text length (incl. Guiraud/Root TTR!)
  - Some (MATTR, MTLD) are stable across text lengths (see, e.g., Zenker & Kyle 2021)

Productive lexical proficiency: Bigrams

- However, proficient word use extends beyond the word-level (Nation, 2001; Römer, 2009; Sinclair, 1991)
  - using “sophisticated” words in inappropriate grammatical and/or semantic contexts does not represent proficient lexical use
- Recent research has demonstrated that more proficient speakers and writers tend to use more strongly associated (contiguous) bigrams
  - Bestgen & Granger (2014)
  - Eguchi & Kyle (2020)
  - Garner, Crossley, & Kyle (2020)
  - etc.

Productive lexical proficiency: Dependency bigrams

- (contiguous) Bigrams have at least two important drawbacks:
  - capture grammatical “errors” as well as less conventional word choices (see Polio & Yoon, 2021)
  - do not capture relationships between words that are not contiguous
    - e.g., They **kicked** the **ball** to their teammate. (verb-direct object)
- One solution: Dependency bigrams
  - Captures strength of association between words in a particular grammatical relationship, regardless of location in utterance
  - Can be accurately extracted (Kyle & Eguchi, 2021 report annotation accuracy around 95%)

Productive lexical proficiency: Dependency bigrams

- Recent research has demonstrated that dependency bigrams are meaningfully related to writing proficiency:
  - Paquot (2018, 2019)
  - Kyle & Eguchi (2021)
- Kyle & Eguchi (2021)
  - TOEFL independent essays
  - word, contiguous bigram, dependency bigram indices
  - small to moderate correlations
  - multivariate model explained ~23% of variance in TOEFL writing scores

Table 10
The final predictor model

Predictors	*B*	*SE*	*p*	β	95% CI for β *LL-UL*	R^2
(Intercept)	3.427	.035	< .001			
McD	1.338	.373	< .001	.177	.080 – .274	.067
Noun–Amod (MI)	.206	.055	< .001	.154	.073 – .234	.038
USF	-.018	.005	.001	-.165	-.259 – -.070	.057
Verb–Advmod Delta P Strongest	7.681	1.686	< .001	.187	.106 – .267	.046
Verb–Dobj (MI)	.273	.070	< .001	.159	.079 – .239	.034
Observations	480					
R^2 / R^2 adjusted	0.242 / 0.234					
BIC	1156.138					

Note. *B* = beta weight; β = standardized beta weight; CI = confidence interval; *LL* = lower limit; *UL* = upper limit; SE = standard error.

Current study

No research I am aware of has examined:

- speaking proficiency + dependency bigrams
- speaking proficiency +:
  - word indices
  - contiguous bigram indices
  - dependency bigram indices

Research Questions

1. What is the relationship between OPI scores and word, contiguous bigram, and dependency bigram indices?
2. What is the relationship between OPI scores and an optimal model including a combination of word, contiguous bigram, and dependency bigram indices?

Method: Learner Corpus

- National Institute of Information and Communications Technology Japanese Learner English (JLE) corpus
- learner utterances from 1,281 oral proficiency interviews
- Proficiency scores (based on a revised ACTFL OPI rubric) ranged from 1-9

Descriptive statistics for SST JLE Corpus (n = 1281)

	Mean	Standard Deviation	Median
SST Scores	4.664	1.574	4.000
Number of words	886.283	340.079	849.000

Method: Linguistic Analysis

- Corpus-based indices were extracted from the spoken portion of COCA (Davies, 2010) using Spacy (version 2.1.8) and in-house Python scripts.
- Other indices (e.g., concreteness) were derived from relevant databases
- NOTE: As we discuss each linguistic index, we will also look at the results for RQ1

Word-level indices

- Lexical diversity
- Word frequency (log transformed)
  - adjectives
  - adverbs
  - nouns
  - verbs
  - all content words
- Concreteness
- Contextual diversity

Lexical diversity

- Moving average type-token ratio (MATTR)
  - All lemmas
  - Content lemmas
- 50-word moving window
- Text length stable (Covington & McFall, 2010; Zenker & Kyle, 2021)
- Correlates well with direct judgements of lexical diversity (Kyle, Crossley, & Jarvis, 2021)

Lexical diversity

Correlations between score and lexical diversity indices

	Score	mattr50_aw	mattr50_cw
Score	1		
mattr50_aw	0.477	1	
mattr50_cw	0.265	0.792	1

Word Frequency

- Word frequency (log transformed)
- Previous writing research has found a negative relationship between corpus frequency and proficiency/writing quality
  - Laufer & Nation (1995); Crossley & Cobb (2014); Kyle et al. (2018), etc.
- Previous speaking research has found a POSITIVE relationship between frequency and proficiency/spoken production quality
  - Eguchi & Kyle (2020) **[OPIs]**
  - Kyle & Crossley (2015) **[TOEFL iBT Speaking]**
  - Berger, Crossley, & Kyle (2019) **[informal conversations]**

Word Frequency

Examples of high and low frequency words

Word classification	High Frequency (> 100 per million)	Low Frequency (< 20 per million)
adjective modifier	*big, good*, *new, other*	*hardy*, *metallic*, *rusty*, *strained*
adverbial modifier	*actually, always, now*, *really*	*incidentally*, *saintly*, *tersely*, *unsightly*
lexical main verb	*have*, *go*, *say*, *tell*	*codify*, *encroach*, *patronize*, *rebuke*
noun	*family*, *people*, *story, time*	*camper*, *dragon*, *evasion*, *libertarian*
content words	*good*, *have*, *work*, *year*	*drip*, *mortality*, *occasional*, *terribly*

Word Frequency

Correlations between score and word frequency indices

	Score	adj	verb	noun	cw
Score	1				
adjective modifier frequency (log)	0.111	1			
lexical main verb frequency (log)	0.209	0.137	1		
noun frequency (log)	0.449	0.252	0.295	1	
content word frequency (log)	0.461	0.353	0.670	0.801	1

Concreteness (as a proxy for salience)

- Concreteness (based on norms from Brysbaert et al., 2014)
- Words that are more concrete are theorized to be easier to learn (Paivio, 1971; Schwanenflugel et al., 1988) than words that are less concrete, likely because they are more salient (Crossley et al., 2016).
  - more concrete: *apple*, *bellybutton*, and *cookie*
  - less concrete: *doubt*, *pride*, and *rarely*
- More proficient users expected to use less concrete words (on average) than less proficient users.
  - citations

Concreteness

Correlation between score and concreteness

	Score	concreteness
Score	1	
concreteness	-0.609	1

Contextual distinctiveness

- Contextual distinctiveness refers to the number of lexical and/or semantic contexts in which a word typically occurs.
- Based on word associate tasks (e.g., USF norms, Nelson et al., 2004)
    - i.e., the number of stimuli words that elicited a particular word.
    - Words that are elicited by many stimuli (e.g., *car*, *food*, and *music*) are less contextually distinct than those elicited by fewer stimuli (e.g., *blender*, *giver*, and *flirt*).
- Based on corpus co-occurrence (e.g., McDonald & Shillcock, 2001)
    - i.e., the predictability of a word's lexical context (a 5-word window) based on relative entropy
    - word that occur in a wider range of contexts (e.g., *close*, *good*, and *visit*) are more predictable than those with restricted use (e.g., *allegedly, kennel*, and *postpone*)

Contextual distinctiveness

Correlations between score and contextual diversity indices

	Score	McD	USF
Score	1		
McD (corpus based)	-0.237	1	
USF (WAT based)	-0.274	-0.196	1

Multiword indices

- Contiguous bigram strength of association
- Dependency bigram bigram strength of association
  - noun-adjective
  - verb-adverb
  - verb-direct object
  - verb-subject

Strength of association

- Six strength of association measures were calculated for bigram indices, including T, MI, MI^2, delta p (LR), delta p (RL), and delta p (max).

Association strength formulas used with bigrams

Index	Formula
T	$\frac{observed - expected}{\sqrt{observed}}$
Mutual information (MI)	$\log\left(\frac{observed}{expected}\right)$
Mutual information squared (MI^2)	$\log\left(\frac{observed^2}{expected}\right)$
Delta-p (left to right)	$P(Word2 \mid Word1) - P(Word2 \mid -Word1)$
Delta-p (right to left)	$P(Word1 \mid Word2) - P(Word1 \mid -Word2)$
Delta-p (max)	$maximum\_score(deltap\_LR \mid deltap\_RL)$

Contiguous bigram indices

- Six strength of association measures were calculated for contiguous bigram indices, including T, MI, MI^2, delta p (LR), delta p (RL), and delta p (max).

Examples of strongly and weakly associated bigrams

Dependency relationship	Strongly associated (MI > 7)	Weakly associated (MI < 3)
lemmatized bigrams	*licensing requirement, lone holdout, rear projection, super delegates*	*big credit, campaign handle, empty out, level management*

Contiguous bigram indices

Correlations between score and contiguous bigram indices

	Score	T	MI	MI^2	deltap w1 cue	deltap w2 cue	deltap strgst
Score	1						
bigram SOA (T)	0.323	1					
bigram SOA (MI)	0.198	0.604	1				
bigram SOA (MI^2)	0.545	0.674	0.535	1			
bigram SOA (deltap w1 cue)	-0.037	0.495	0.426	0.458	1		
bigram SOA (deltap w2 cue)	0.365	0.474	0.525	0.619	0.389	1	
bigram SOA (deltap strgst)	0.191	0.535	0.527	0.615	0.842	0.803	1

Dependency bigram indices

- Dependency strength of association measures were calculated for lemmas in five dependency relationships

Examples of strongly and weakly associated dependency bigrams (lemma forms)

Dependency relationship	Strongly associated (MI > 7)	Weakly associated (MI < 3)
noun – adjective	*rehearsal hall, sippy cup, square peg, tidal surge*	*fine product, great service, musical education, national concern*
verb – adverb	*classically train, loosely affiliate, mortally wound, partially submerge*	*do anyway, immediately assume, quickly seem, resolve somehow*
verb – direct object	*flee persecution, heal rift, spew hatred, twiddle thumb*	*assure voter, enter number, maintain degree, use money*
verb – subject	*river crest, gallon leak, Greece default, militia disband*	*advocate call, batman be, he offend, they encounter*

Dependency bigram indices

Correlations between score and dependency bigram indices

	Score	noun – adjective	verb – adverb	verb – direct object	verb – subject
Score	1				
noun – adjective (deltap depcue)	0.116	1			
verb – adverb (deltap depcue)	-0.189	-0.059	1		
verb – direct object (deltap strgst)	-0.384	-0.024	0.286	1	
verb – subject (deltap govcue)	0.461	0.039	-0.1	-0.224	1

Summary of correlation analysis (RQ1)

Correlation matrix

	Score	mattr50	adj freq	verb freq	noun freq	concrete ness	USF	bigram SOA	noun – adj SOA	verb – adv SOA	verb – dobj SOA	verb – nsubj SOA
Score	1											
mattr50 aw	0.477	1										
adjective modifier frequency (log)	0.111	0.052	1									
lexical main verb frequency (log)	0.209	-0.019	0.137	1								
noun frequency (log)	0.449	0.284	0.252	0.295	1							
concreteness	-0.609	-0.338	-0.104	-0.295	-0.407	1						
USF (WAT based)	-0.274	-0.233	0.173	0.048	0.11	0.361	1					
bigram SOA (MI^2)	0.545	0.218	0.154	0.449	0.379	-0.473	-0.076	1				
noun – adjective (deltap depcue)	0.116	0.043	-0.111	0.037	0.16	-0.064	-0.042	0.143	1			
verb – adverb (deltap depcue)	-0.189	-0.107	-0.071	-0.107	-0.101	0.158	0.034	-0.085	-0.059	1		
verb – direct object (deltap strgst)	-0.384	-0.107	-0.068	-0.17	-0.237	0.365	0.124	-0.203	-0.024	0.286	1	
verb – subject (deltap govcue)	0.461	0.27	0.052	0.225	0.223	-0.483	-0.114	0.398	0.039	-0.1	-0.224	1

Summary of correlation analysis (RQ1)

Correlations between indices and score (sorted)

Index	Score
concreteness	-0.609
bigram SOA (MI^2)	0.545
mattr50_aw	0.477
verb – subject (deltap govcue)	0.461
noun frequency (log)	0.449
verb – direct object (deltap strgst)	-0.384
USF (WAT based)	-0.274
lexical main verb frequency (log)	0.209
verb – adverb (deltap depcue)	-0.189
noun – adjective (deltap depcue)	0.116
adjective modifier frequency (log)	0.111

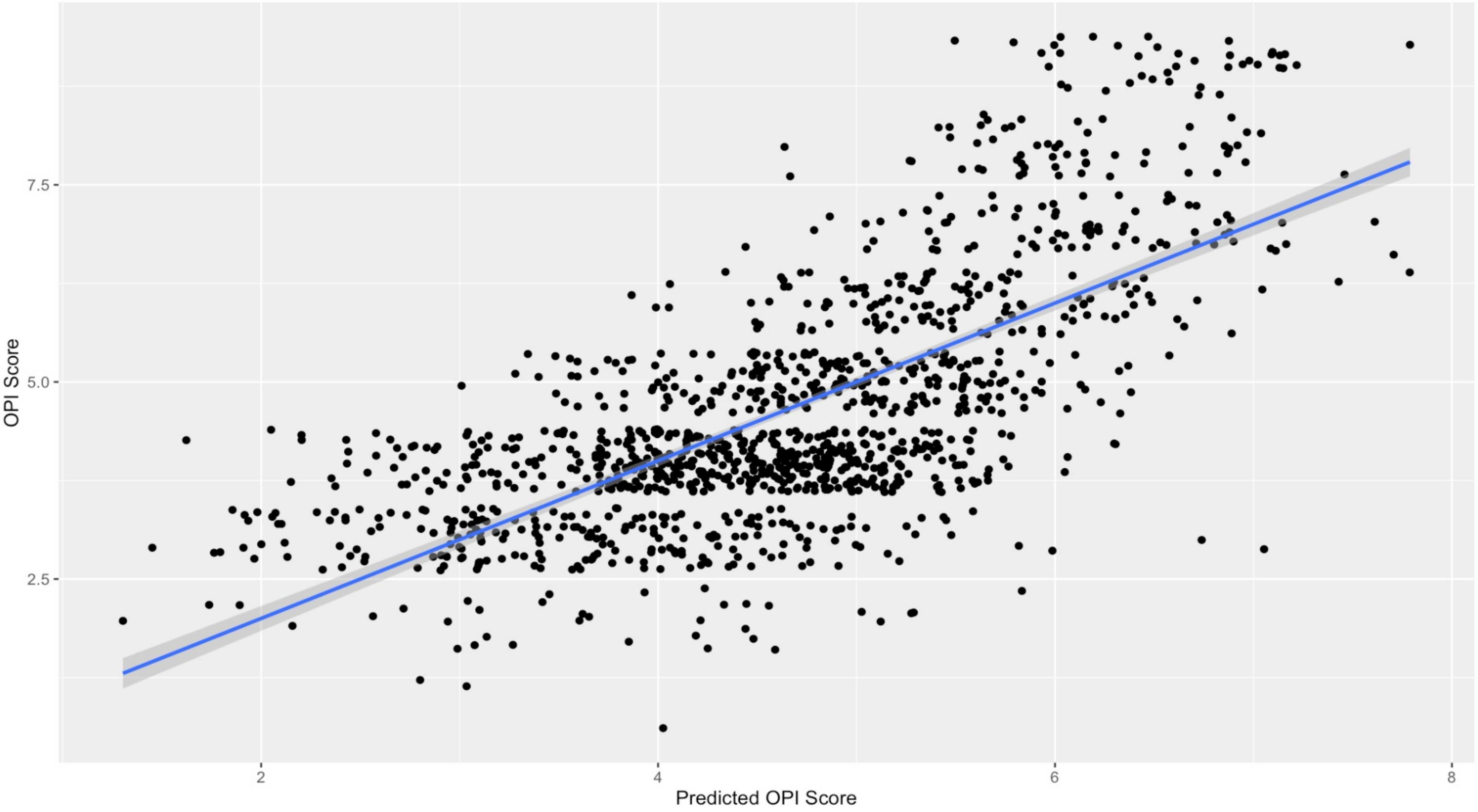
Multivariate analysis (RQ2)

- A multivariate multiple regression was conducted to determine the degree to which indices of lexicogrammatical sophistication could explain the variance in OPI scores.

Results

Final regression model

	relative importance	Estimate	Std. Error	t value	p
(Intercept)		-14.895	0.947	-15.723	< 0.001
adjective modifier frequency (log)	0.008	0.129	0.049	2.641	0.008
bigram SOA (MI^2)	0.186	1.565	0.095	16.525	< 0.001
mattr50_aw	0.135	10.131	0.740	13.691	< 0.001
noun – adjective (deltap depcue)	0.006	1.614	0.886	1.823	0.069
USF (WAT based)	0.044	-0.045	0.006	-7.728	< 0.001
verb – adverb (deltap depcue)	0.018	-4.708	0.990	-4.758	< 0.001
verb – subject (deltap govcue)	0.107	11.110	1.200	9.255	< 0.001



7.5
5.0
2.5
OPI Score
2
4
6
8
Predicted OPI Score

Discussion/Summary: RQ1

- Strong indices in each level:
  - lexical diversity
    - *r* = .477
  - words
    - concreteness *r = -*.607
    - noun frequency *r* = .449
  - contiguous bigram SOA
    - *r* = .545
  - dependency bigram SOA
    - verb-subject *r* = 0.461

Discussion/Summary: RQ2

- ~50% of the variance in OPI scores explained by the model
  - Lexical diversity: 13.5%
  - Word-level indices: 5%
  - Bigram SOA: 19%
  - Dependency bigram SOA: 13%
- Each index type/lexicogrammatical level contributed to the model
- Demonstrates the complexity of our representations of speaking proficiency
- Suggests that including indices from each level is useful when modeling productive (lexical) proficiency

Future Directions

- Explore verb-adverb and verb-direct object relationships in other contexts (negative correlations with score)
- include verb-VAC indices (Kyle & Crossley, 2017)
- systematically investigate the effects of:
  - mode * task type * prompt * L1
- Investigate in other languages

Thanks!

Tools - NLP TOOLS FOR THE S

linguisticanalysistools.org/tools.html

NLP TOOLS FOR THE SOCIAL SCIENCES

HOME BIOS TOOLS USEFUL LINKS

SiNLP

to the number of words, number of types, TTR, letters per word, number of paragraphs, number of sentences, and number of words per sentence for each text. In addition, users can analyze texts with regard to their own custom dictionaries. Click here to learn more

TAACO

TAACO is an easy to use tool that calculates 150 indices of both local and global cohesion, including a number of type-token ratio indices, adjacent overlap indices, and connectives indices. The tool also measures text overlap between two texts (intertextual cohesion). (TAACO 2.0 now available!) Click here to learn more

TAALED

TAALED is an analysis tool designed to calculate a wide variety of lexical diversity indices. Homographs are disambiguated using part of speech tags, and indices are calculated using lemma forms. Indices can also be calculated using all lemmas, content lemmas, or function lemmas. Click here to learn more

TAALES

TAALES is a tool that measures over 400 classic and new indices of lexical sophistication, and includes indices related to a wide range of sub-constructs. Included are indices for both single words and n-grams. Starting with version 2.2, TAALES also provides comprehensive index diagnostics. (TAALES 2.2 now available!) Click here to learn more

TAASSC

TAASSC is an advanced syntactic analysis tool that measures fine-grained indices of clausal and phrasal complexity, classic indices of syntactic complexity, and frequency-based verb argument construction indices. Click here to learn more