

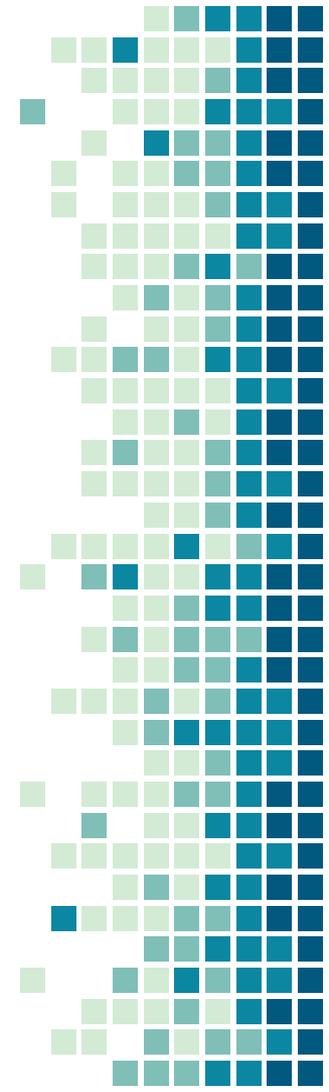
*Lexical Sophistication
Measurements: Applications in
Teaching and Assessment*

Ben Naismith
University of Pittsburgh
TESOL 2019 International Convention
March 13th, 2019



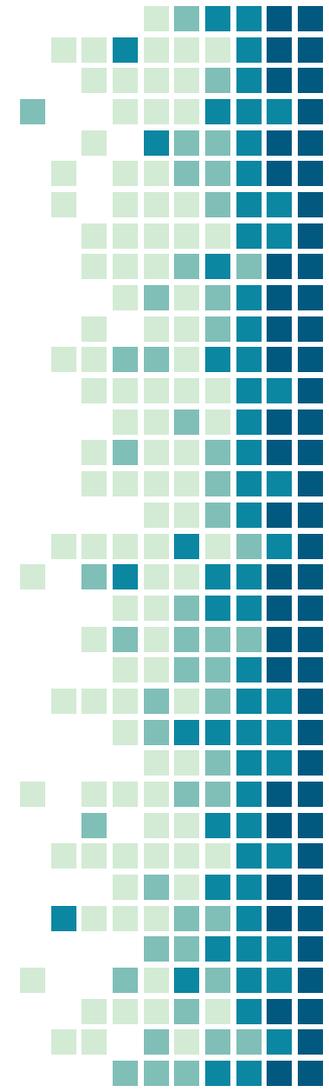
Overview

1. What is lexical sophistication?
2. Measuring lexical sophistication
3. Lexical sophistication tools
4. Practical uses for lexical sophistication data
5. Questions



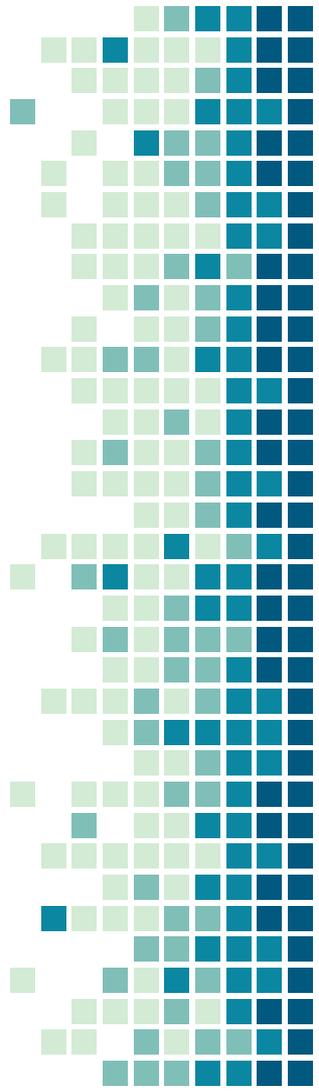
A little math to start...

- How many words are there in English?
 - at least 250,000 distinct English words (OED)
 - How many words do you typically teach in a one-hour class?
 - 6-12?
 - How many classes would it take to teach them all?
 - $250,000 / 10 = 25,000!$ (one class per day for 68 years)
- Essential to prioritize and select vocab in a principled way



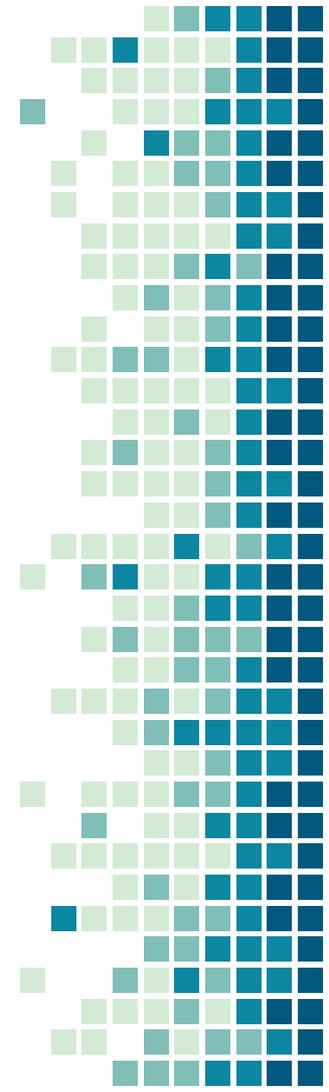
Lexical richness

- Importance of vocabulary widely accepted (Cobb & Horst, 2015)
- Different aspects of 'lexical richness':
 - *Lexical depth* – knowledge and accuracy of usage
 - *Lexical diversity* – range of different words used
 - *Lexical sophistication* – range of advanced words used



Lexical sophistication

- Compare: "*I am very tired.*" vs. "*I am utterly exhausted.*"
→ "*utterly/exhausted*" are more *sophisticated* words
- Teacher perceptions of proficiency correlate with measures of lexical sophistication (Daller & Phelan, 2007)
- Key descriptor for assessing lexical proficiency, e.g. IELTS band descriptors (British Council, 2018)



Lexical sophistication

- Compare: "utterly/exhaustively"

→ "utterly/exhaustively"

- Teacher performance measures

- Key descriptive band descriptors

7	<ul style="list-style-type: none">▪ uses a sufficient range of vocabulary to allow some flexibility and precision▪ uses less common lexical items with some awareness of style and collocation▪ may produce occasional errors in word choice, spelling and/or word formation
6	<ul style="list-style-type: none">▪ uses an adequate range of vocabulary for the task▪ attempts to use less common vocabulary but with some inaccuracy▪ makes some errors in spelling and/or word formation, but they do not impede communication

Measuring Lexical Sophistication

- How to determine 'advanced words'?
→ *Frequency bands* compiled from corpora, e.g.

BNC/COCA: frequency bands (Davies, 2008-)

1000: the, and, language, show, study

2000: accurate, focus, reaction, variable

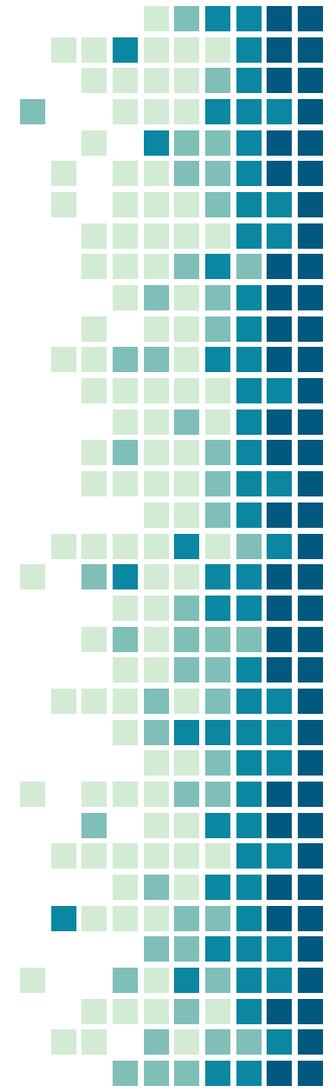
3000: compromise, grammar, mask, visual

5000: inhibitory, retrieve, verb, parallel

10000: decompose, lexicon, neural, paradigm, prefix

13-20k: impervious, orthography, neuroscience, connectionist

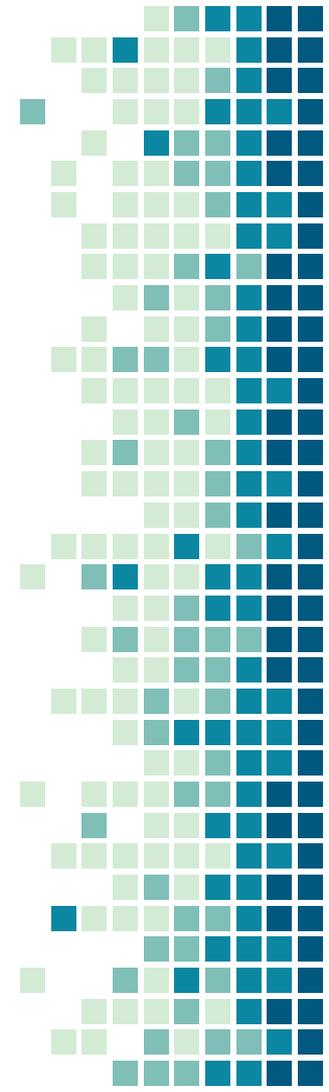
- 3000-9000 bands considered to be critical for instruction (Schmitt & Schmitt, 2014)



Advanced Guiraud (AG)

- Type-Token Ratio (TTR) based on 'advanced' types only
- Omits high-frequency types below 2000 band (though any list of 'advanced types' can be used)
- Effective in distinguishing between groups/levels, e.g. (Daller & Xue 2007; Juffs 2019)

$$AG = \frac{\textit{Advanced Types}}{\sqrt{\textit{Tokens}}}$$



Text 1

I met my friend Nife while I was studying in a middle school. I was happy when I met him because he was a good student in our school. We continued the middle and high school together in the same school. We were studying in the different classes in the middle school; however, in the high school we were studying in the same class. We went to many places in the free time while we were studying in the high school. When we finished from the high school, I went to K.S. University and he went to I.M. University. While we were enjoying in academic life, we made many achievement in these universities. I graduated when Nife was studying in the last semester in the university. After that, I got a job. Fortunately, it was nearby my home. I worked two years then I got scholarship from ministry of high education in my country. When I came here to U.S., my friend Nife arrange some documents to study at grad school in Malaysia.

Text 2

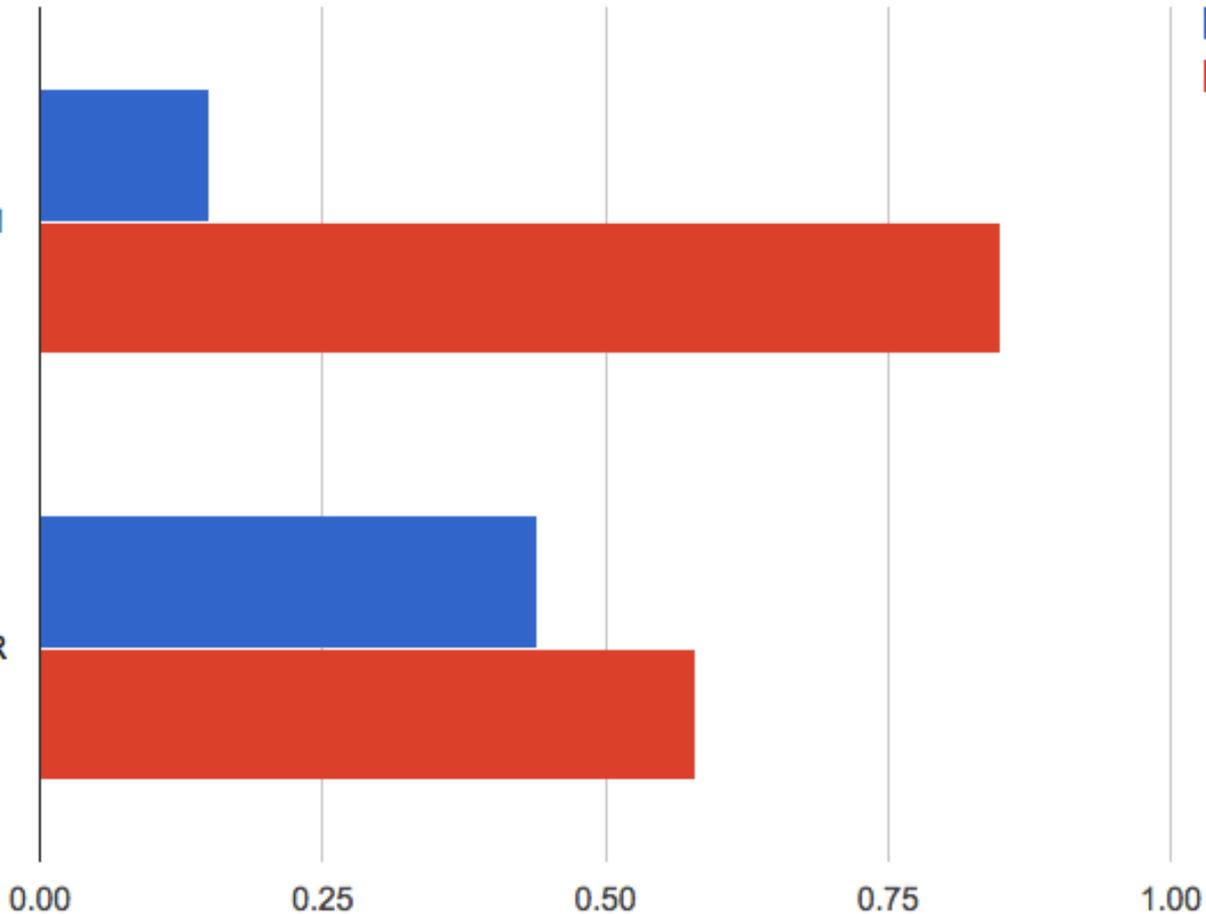
I first met my friend Nife while I was studying in middle school. I recall being delighted when I met him because he was an outstanding student. We then progressed throughout middle and high school together at the same institution. Initially, we were studying in different classes in middle school; however, in high school we were reunited in the same cohort. I remember in our free time, we used to explore the surrounding areas. Upon completing high school, I enrolled in K.S. University, and he opted for I.M. University. While we were enjoying academia, we accomplished many achievements at our respective universities. Eventually, I graduated while Nife was still completing his last semester. Subsequently, I got a job, which fortunately was nearby my home. There, I worked for two years before obtaining a scholarship from the ministry of higher education in my country. When I came here to U.S., my friend Nife completed the necessary arrangements to attend grad school in Malaysia.

Lexical Sophistication

CVS1

Lexical Diversity

TTR



Tools for measuring lexical sophistication

Option 1: Lots of data

- e.g. TAALES (Kyle & Crossley, 2015)
- 400 indices of lexical sophistication
- free downloadable program
- + huge amount of potentially useful data
- - difficult to quickly interpret

Tool for the Automatic Analysis of Lexical Sophistication

Instructions

Options

MRC Frequency
 BNC Word Frequencies
 BNC Ngram Frequencies
 SUBTLEXus Frequencies
 AWL
 AWL Sublists
 AFL
 AOE
 ELP Word Information
 ELP Response Time Norms
 Contextual Diversity
 LSA

 Psycholinguistic Norms
 Hypernymy & Polysemy

COCA Options

Word Frequency and Range

 academic
 fiction
 magazine
 news
 spoken

Bigram Frequency, Range, and Association Strength

 academic
 fiction
 magazine
 news
 spoken

Trigram Frequency, Range, and Association Strength

 academic
 fiction
 magazine
 news
 spoken

Data Input

Your selected input folder:

(No Folder Chosen)

Your selected output filename:

(No Output Filename Chosen)

Run Program

Program Status

...Waiting for Data to Process

al sophistication

ossley, 2015)

sophistication

erpret

	A	B	C	D	E	F	G	H	I
1	Filename	Word Count	KF_Freq_AW	KF_Freq_AW	KF_Ncats_AV	KF_Nsamp_A	TL_Freq_AW	TL_Freq_AW	Brown_Freq
2	SLA final pap	4494	9728.02799	2.8292545	12.5434402	260.662974	38988.483	3.45675699	1462.34493

Tools for measuring lexical sophistication

Option 2: Do it yourself

- e.g. Python
- + you can choose which 'basic' words to ignore
- - it requires basic coding literacy
- Toolkit for measuring AG to be released soon

```
56 def adv_guiraud(text, freq_list='NGSL', custom_list=None,
57                 spellcheck=True, supplementary=True):
58     """
59     Calculates advanced guiraud: advanced types / sqrt(number of tokens)
60     By default, uses NGSL top 2k words as frequency list
61     custom_list is a custom list of common types for frequency list
62
63     Args:
64         text: Input string to calculate AG for
65         freq_list: string specifying which freq list to use. Must be one
66                   of {'NGSL', 'PET', 'PELIC', 'SUPP'}
67         custom_list: if not None, used instead of freq_list (can pass own list
68                     of strings containing common types to ignore for AG
69         spellcheck: Boolean flag to ignore misspelled words (rough spellcheck
70                     with wordnet.synsets())
71         supplementary: Include NGSL supplementary vocabulary in addition to
72                       specified list
73
74     Returns:
75         Calculated AG
76     """
```

Tools for measuring lexical sophistication

Option 3: Web-based program

- aihaiyang.com/software/lca/single/ (Ai & Lu, 2010)
- Input text(s), choose indices, get results
- 'Corrected VS1 (CVS1)' closest to Advanced Guiraud
- + easy to use and read results
- - visualizations don't always work, can't choose the 'basic words'

Step 1: Enter text #1

Accurate Measurement of Lexical Sophistication with Reference to ESL Learner Data Ben Naismith University of Pittsburgh Department of Linguistics 4200 Fifth Ave, Pittsburgh, PA 15260 1-412-624-5900 bnaismith@pitt.edu Brianna Hill University of Pittsburgh School of Computing and Information 4200 Fifth Ave, Pittsburgh, PA 15260 1-412-624-5900 blh82@pitt.edu Na-Rae Han University of Pittsburgh Department of Linguistics 4200 Fifth Ave, Pittsburgh, PA 15260 1-412-624-5900 naraehan@pitt.edu Alan Juffs University of Pittsburgh Department of Linguistics 4200 Fifth Ave, Pittsburgh, PA 15260 1-412-624-5900 juffs@pitt.edu Daniel Zheng University of Pittsburgh Department of Electrical and Computer Engineering 4200 Fifth Ave, Pittsburgh, PA 15260 1-412-624-5900 daniel.zheng@pitt.edu

ABSTRACT One commonly used measure of lexical sophistication is the Advanced Guiraud (AG; [9]), whose formula requires frequency band counts (e.g., COCA; [13]). However, the accuracy of this measure is affected by the particular 2000-word frequency list selected as the basis for its calculations [27]. For example, possible issues arise when frequency lists that are based solely on native speaker corpora are used as a target for second language (L2) learners (e.g., [8]) because the exposure frequencies for L2 learners may vary from that of native speakers. Such L2 variation from comparable native speakers may be due to first language (L1) culture, home country teaching materials, or the text types which L2 learners commonly encounter. This paper addresses the aforementioned problem through an English as a Second Language (ESL) frequency list

Step 2: Select indice(s)

Advanced Guiraud

Option 3: Web-based pr

- aihaiyang.com/soft
- Input text(s), choose
- 'Corrected VS1 (CVS1'
- + easy to use and re
- - visualizations don
- 'basic words'

Lexical density (LD)

Lexical Sophistication

- Lexical sophistication-I (LS1)
- Lexical sophistication-II (LS2)
- Verb sophistication-I (VS1)
- Verb sophistication-II (VS2)
- Corrected VS1 (CVS1)

Lexical Variation

NDW

- Number of different words (NDW)
- NDW (first 50 words) (NDWZ-50)
- NDW (expected random 50) (NDW-ER50)
- NDW (expected sequence 50) (NDW-ES50)

TTR

- Type/Token ratio (TTR)
- Mean Segmental TTR (50) (MSTTR-50)
- Corrected TTR (CTTR)
- Root TTR (RTTR)
- Bilogarithmic TTR (logTTR)
- Uber Index (Uber)

Verb diversity

- Verb variation-I (VV1)
- Squared VV1 (SVV1)
- Corrected VV1 (CVV1)

Lexical word diversity

- Lexical word variation (LV)
- Verb variation-II (VV2)
- Noun variation (NV)
- Adjective variation (AdjV)
- Adverb variation (AdvV)
- Modifier variation (ModV)

Tools for measuring lexical sophistication

Advanced Guiraud

Option 2: Web-based program

Numeric Results

The following tabular results can be copied and pasted into a plain text file, and subsequently imported into spreadsheet or database software for further processing and analysis.

Text #1

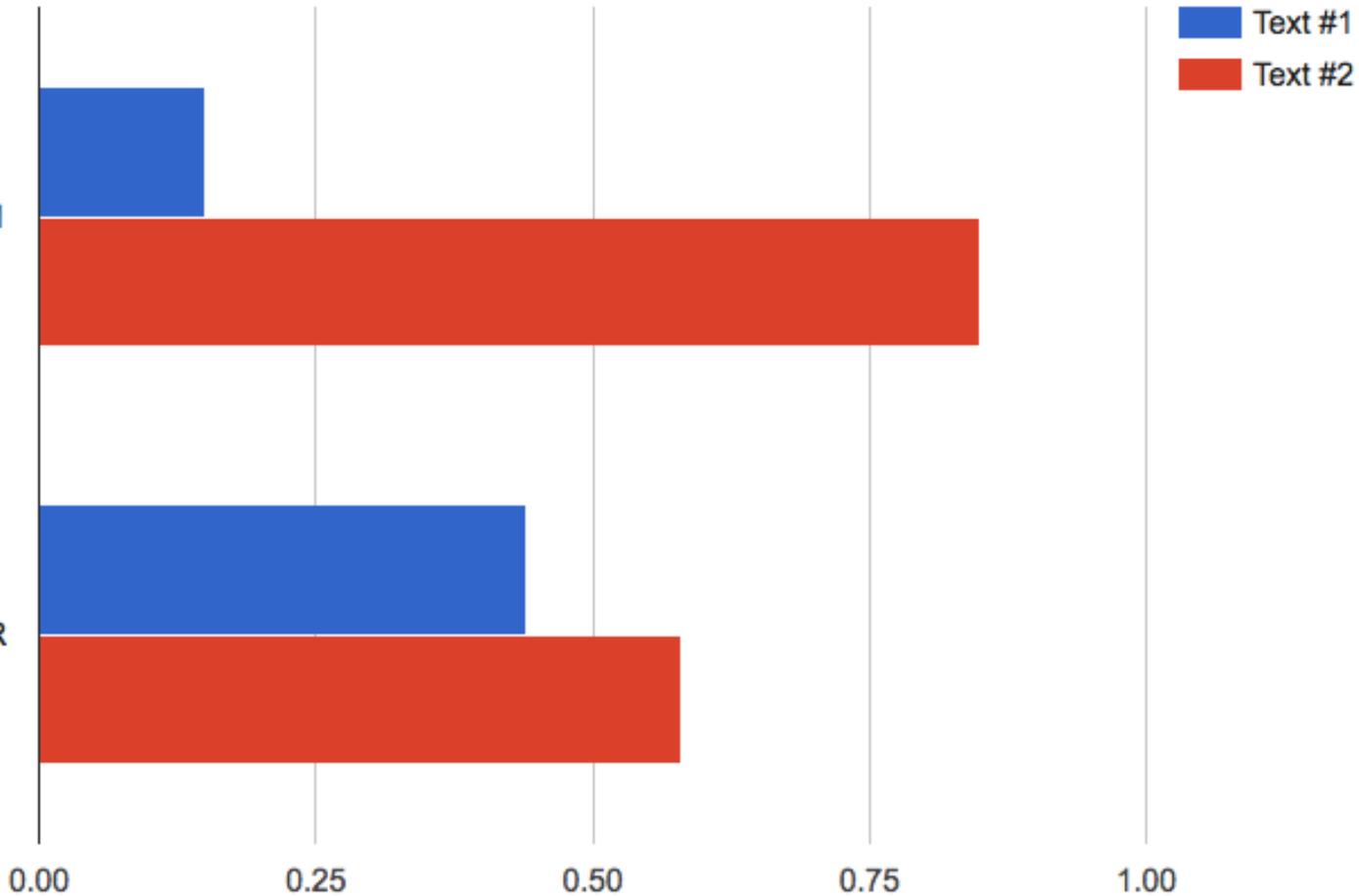
LD, LS1, LS2, VS1, VS2, CVS1, NDW, NDWZ, NDWERZ, NDWESZ, TTR, MSTTR, CTTR, RTTR, LOGTTR, UBER, VV1, SVV1, CVV1, LV, VV2, NV, ADJV, ADVV, MODV
0.52, 0.43, 0.43, 0.12, 7.95, 1.99, 1347, 33, 42.30, 42.10, 0.21, 0.78, 11.92, 16.86, 0.82, 21.43, 0.42, 92.28, 6.79, 0.31, 0.07, 0.27, 0.07, 0.03, 0.09

Lexical Sophistication

CVS1

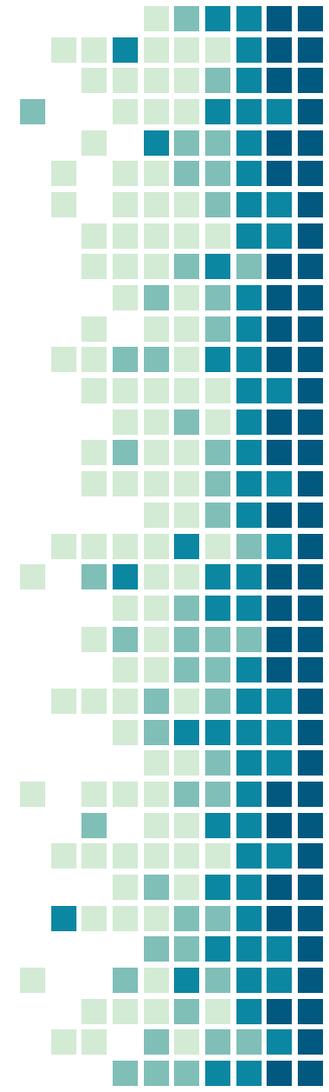
Lexical Diversity

TTR



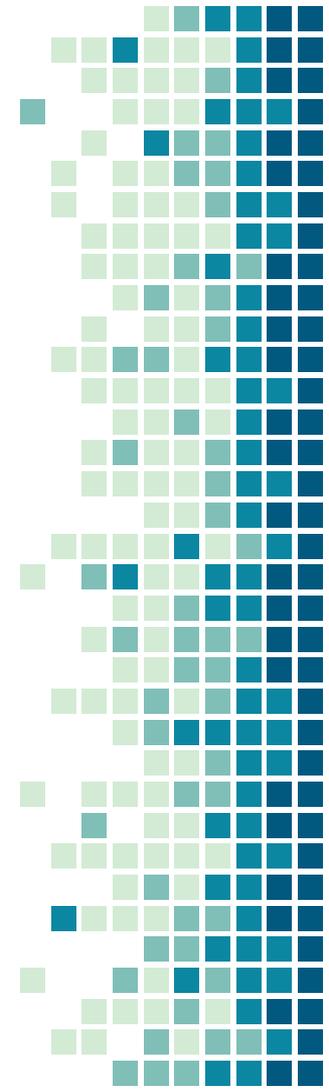
Practical uses – Assessment

- Diagnostically (e.g. program entrance tests)
 - *Lexical proficiency good predictor of overall proficiency (Crossley et al. 2012)*
- Summatively (e.g. end of level)
 - *One measure of achievement tests*
 - *Especially if texts are completed electronically*
- Objective (automated) measure in addition to other subjective measures



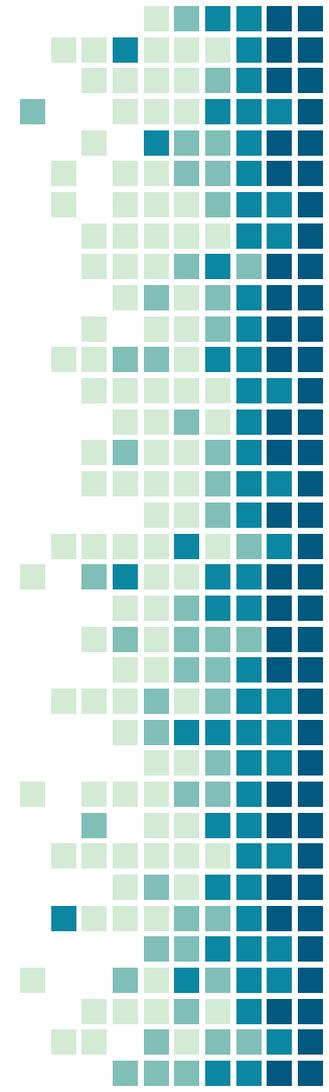
Practical uses – Assessment benchmarks

- Based on Naismith et al. (2018) data, AG could be a useful metric for measuring intermediate to advanced gains as students prepare for academic or professional study, e.g.
 - Low intermediate < .30
 - Intermediate < .50
 - Academically 'ready?' > 1.0
- Vocabulary size and speed have been shown to correlate with measures of academic writing and GPA, i.e. they are predictors of academic achievement. (Roche & Harrington 2013)



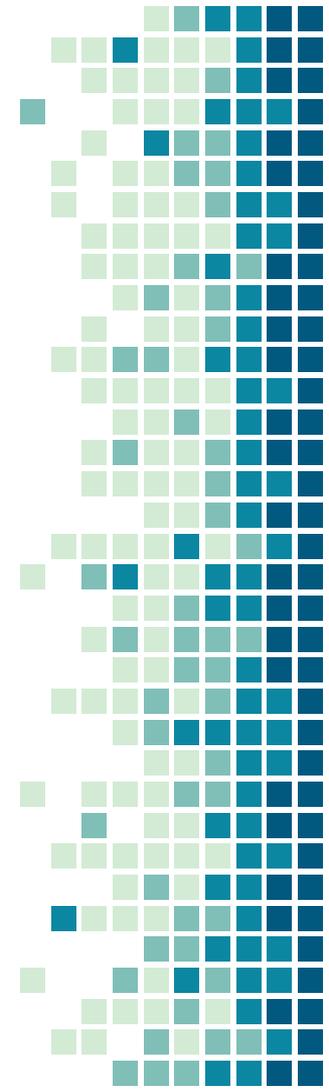
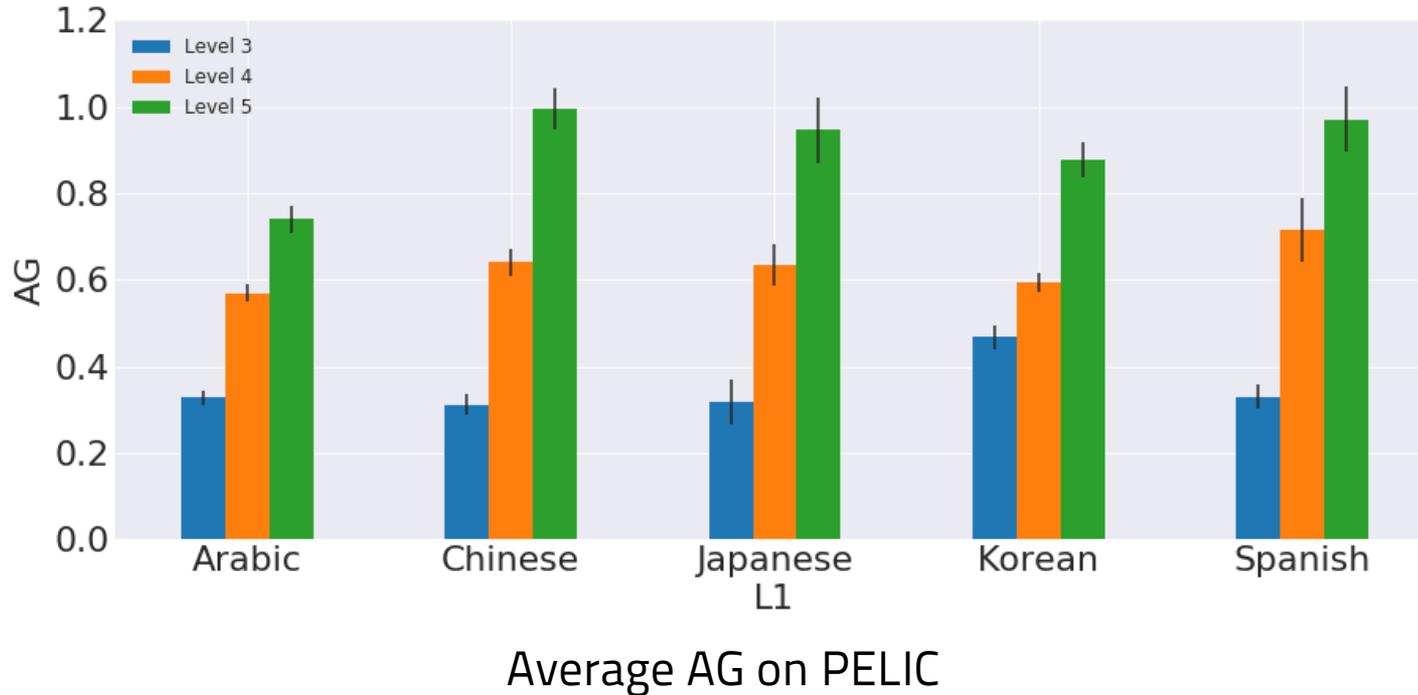
Practical uses – Tracking development

- Formative assessment
 - *From the start to the end of the semester or year*
 - *As part of ongoing learning portfolios
(quantitative/tangible measure of progress)*
- Student-led
 - *Students check own work (1st draft, 2nd draft, etc.)*
 - *As part of in-class activities*



Practical uses – Tracking development

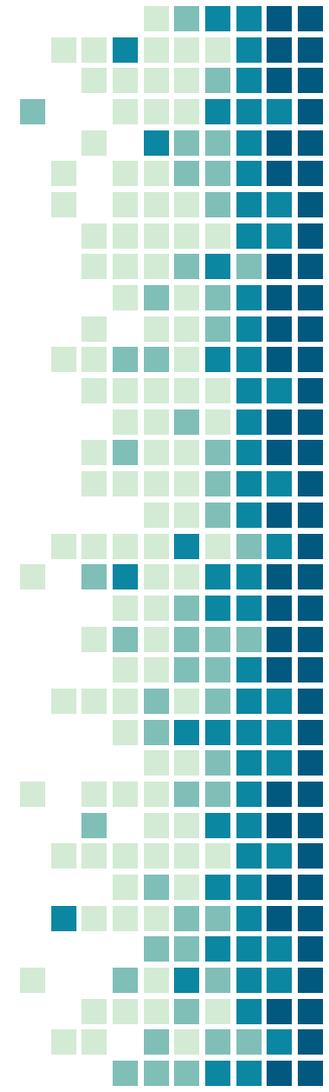
Students at the University of Pittsburgh ELI



Practical uses – Vocabulary lists

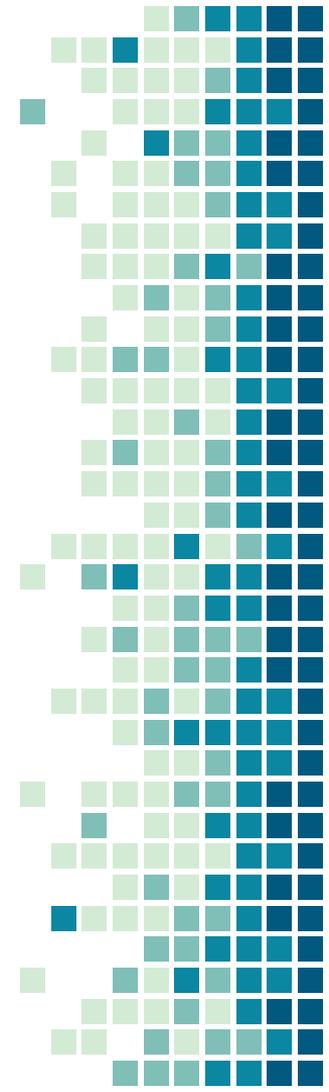
(Principled) vocabulary lists can be used to...

- standardize what learners should know upon reaching a certain level / entering a program
→ e.g. PET list at B1 on CEFR
- push learners to use more sophisticated lexis
→ lexis in next 1000 frequency band (lextutor.ca)
- calculate AG scores sensitive to learners' context
→ e.g. PSL-3 for Pittsburgh IEPs



Practical uses – Selecting lexis

- Selecting from specific frequency bands
 - E.g. COCA 100k:
(https://corpus.byu.edu/coca/files/100k_samples.txt)
- Checking lexis in texts to use in class (whether from coursebook or authentic material)
 - Lex tutor VP Compleat – word family frequencies
(<https://lextutor.ca/cgi-bin/vp/comp/>)
 - English profile – CEFR bands
(<http://www.englishprofile.org/wordlists/text-inspector>)



rank	freq*	texts#	word	lemma	PoS (part of speech; see http://ucrel.lancs.ac.uk/claws7tags.html)	
-----	-----	-----	-----	-----	-----	
20	2095853	178290	have	have	vh0	
40	1057682	148994	would	would	vm	
60	769838	107674	your	your	appge	
80	530956	140946	first	first	md	
100	415055	126347	even	even	rr	
120	327849	115203	in	in	rr	
140	269151	106081	while	while	cs	
160	226115	83873	each	each	dd	
180	199507	92450	before	before	ii	
200	177630	51374	men	man	nn2	
220	165431	75812	around	around	ii	
240	157480	71742	point	point	nn1	
260	145415	72889	almost	almost	rr	
280	140818	45743	program	program	nn1	
300	132652	71346	least	least	rr	
320	125207	53910	wanted	want	vvd	
340	120337	59495	everything		everything	pn
360	114168	51904	used	use	vvn	
380	106086	32416	looked	look	vvd	
400	99926	26397	police	police	nn2	
420	95253	55975	comes	come	vvz	
440	91819	29313	policy	policy	nn1	
460	87347	48189	everyone		everyone	pn
480	83736	30986	show	show	nn1	
500	79205	44851	quite	quite	rr	

Profile summary (token %)/cumul.	
K-1 (88.1)	88.1
K-2 (3.4)	91.5
K-3 (2.5)	94.0
K-4 (1.3)	95.3
K-5 (0.4)	95.7
K-6 (0.2)	95.9
K-7 (0.2)	96.1
K-9 (0.2)	96.3
K-10 (0.4)	96.7
K-12 (1.5)	98.2
OFF (1.89)	≈100%

REPEATED 2+ K-Families
to remove with care
when simplifying
to lower K or C level (?)

2_primeval
8_boa
2_elephant
2_swallow
2_geography
2_digest
2_magnificent
3_constrictor
3_constrictors

once when i was six years old i saw a magnificent picture in a book called true stories from nature about the primeval forest it was a picture of a boa constrictor in the act of swallowing an animal here is a copy of the drawing boa in the book it said boa constrictors swallow their prey whole without chewing it after that they are not able to move and they sleep through the six months that they need for digestion i pondered deeply then over the adventures of the jungle and after some work with a colored pencil i succeeded in making my first drawing my drawing number one it looked something like this hat i showed my masterpiece to the grown ups and asked them whether the drawing frightened them but they answered frighten why should any one be frightened by a hat my drawing was not a picture of a hat it was a picture of a boa constrictor digesting an elephant but since the grown ups were not able to understand it i made another drawing i drew the inside of a boa constrictor so that the grown ups could see it clearly they always need to have things explained my drawing number two looked like this elephant inside the boa

A1112 types / 333 tokens
51.61% / 70.70%[\(Show words\)](#)**A2**50 types / 67 tokens
23.04% / 14.23%[\(Show words\)](#)**B1**19 types / 22 tokens
8.76% / 4.67%[\(Show words\)](#)**B2**15 types / 15 tokens
6.91% / 3.18%[\(Show words\)](#)**C1**2 types / 2 tokens
0.92% / 0.42%[\(Hide words\)](#)

digesting (1)

digestion (1)

C24 types / 4 tokens
1.84% / 0.85%[\(Hide words\)](#)

masterpiece (1)

pondered (1)

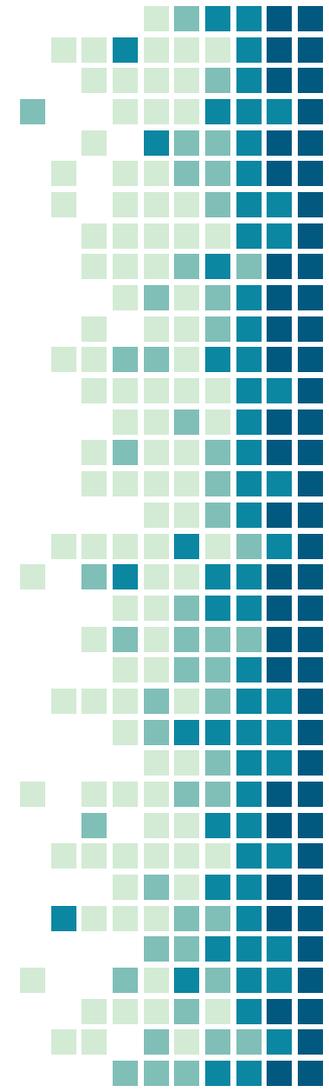
prey (1)

tiresome (1)

Unlisted15 types / 28 tokens
6.91% / 5.94%[\(Show words\)](#)

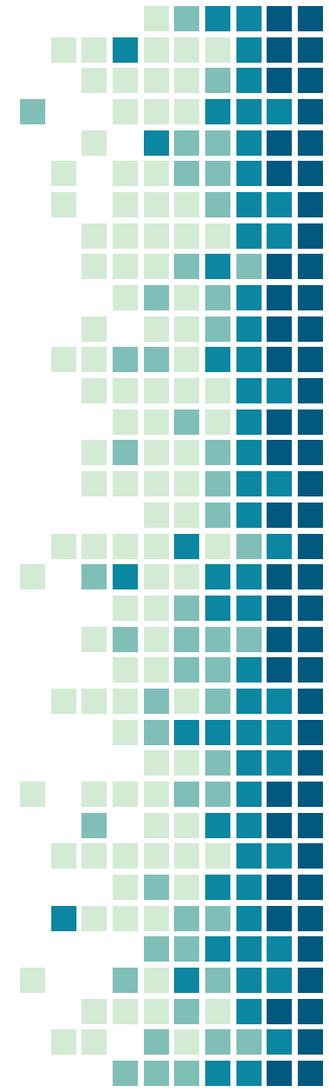
Wish list / Future research

- A more transparent, easy-to-use tool for measuring the AG of texts
- AG benchmarks used by different programs
- Effects of using different 'basic words' lists for calculating AG



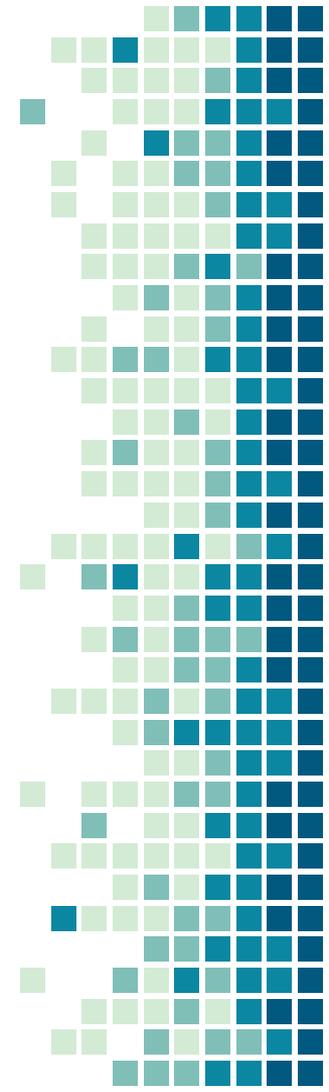
Resources

- Lexical sophistication tools
 - TAALES (Kyle & Crossley, 2015) <https://www.linguisticanalysistools.org/taales.html>
 - Lexical Complexity Analyzer (Ai & Lu, 2010) <https://aihaiyang.com/software/lca/>
 - Pitt Lexical Toolkit (forthcoming) <https://github.com/ELI-Data-Mining-Group/Pitt-ELI-Corpus>
- Public lists
 - NGSL (Browne, 2014) <http://www.newgeneralservicelist.org>
 - PET (Cambridge English, 2012) www.cambridgeenglish.org/Images/84669-pet-vocabulary-list.pdf
- Frequency bands
 - Lex tutor (Cobb, N.D.) www.lextutor.ca/vp/comp
 - COCA 100k (Davies, 2008-) <https://www.wordfrequency.info/purchase.asp>



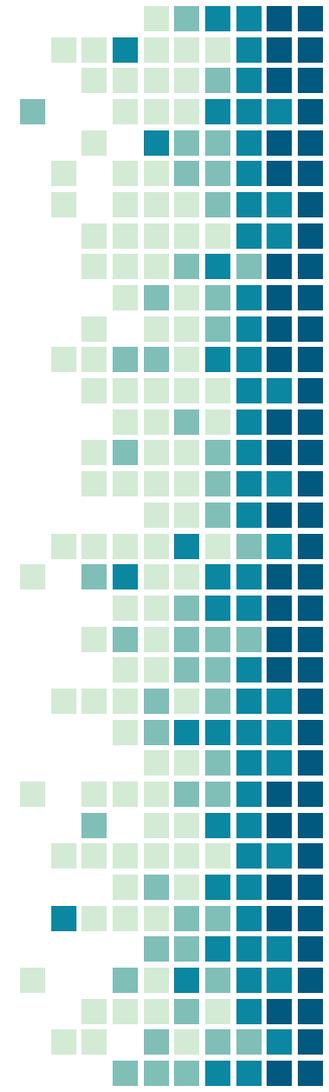
References

- Ai, Haiyang & Lu, Xiaofei. 2010. A web-based system for automatic measurement of lexical complexity. Paper presented at the 27th Annual Symposium of the Computer-Assisted Language Consortium (CALICO-10). Amherst, MA. June 8-12.
- British Council. 2018. Take IELTS with British Council: Public band descriptors. Retrieved Nov 2018 from https://takeielts.britishcouncil.org/sites/default/files/IELTS_task_2_Writing_band_descriptors.pdf
- Browne, C. 2014. A New General Service List: The Better Mousetrap We've Been Looking For? *Vocabulary Learning and Instruction*, 3(2), 1-10.
- Cambridge English: Preliminary and Preliminary for Schools Vocabulary List. 2012. Retrieved from <http://www.cambridgeenglish.org/images/84669-pet-vocabulary-list.pdf>
- Cobb, T. *Compleat Web VP* [computer program]. Accessed 29 Dec 2018 at <https://www.lex tutor.ca/vp/comp/>
- Cobb, T. & Horst, M. 2015. "Learner corpora and lexis". In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 185-206). Cambridge: Cambridge University Press.
- Crossley, S. A., Salsbury, T., & McNamara, D. S. 2012. Predicting the proficiency level of language learners using lexical indices. *Language Testing*, 29(2), 243-263.

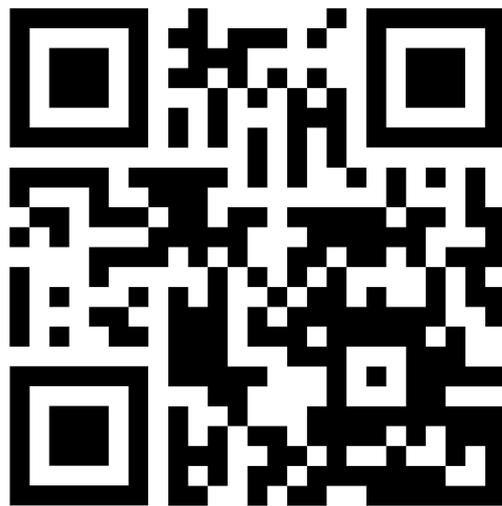


References (cont.)

- Daller, H. & Phelan, D. 2007. "What is in a teachers' mind?", in Daller, Milton, Treffers-Daller (Eds.) *Modelling and Assessing Vocabulary Knowledge*, (234–244). Cambridge: Cambridge University Press.
- Daller, H. & Xue, H. 2007. "Lexical richness and the oral proficiency of Chinese EFL students". In H. Daller, J. Milton, and J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (150-164). New York: Cambridge University Press.
- Davies, Mark. 2008-. *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. Available online at <https://corpus.byu.edu/coca/>.
- Juffs, A. 2019. "Lexical development in the writing of English Language Program Students". In R. M. DeKeyser and G.P. Botana (Eds.), *Reconciling pedagogical demands with pedagogical applicability*. Amsterdam: John Benjamins.
- Kyle, K. & Crossley, S. A. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly* 49(4), pp. 757-786. doi: 10.1002/tesq.194
- Naismith, B., Han, N.-R., Juffs, A., Hill, B. L., & Zheng, D. 2018. "Accurate Measurement of Lexical Sophistication with Reference to ESL Learner Data", conference paper in *Educational Data Mining 2018*, Buffalo, NY.
- Oxford dictionaries. N.D. How many words are there in the English language? [Blog post]. Retrieved January 2019 from <https://en.oxforddictionaries.com/explore/how-many-words-are-there-in-the-english-language/>
- Roche, T. and Harrington, M. 2013. *Language Testing in Asia*, 3:12, 1-13. doi: 10.1186/2229-0443-3-12



Questions or
comments?



View PDF

Thank you for attending.

- <https://github.com/ELI-Data-Mining-Group/Pitt-ELI-Corpus>
- www.pitt.edu/~ben25
- bnaismith@pitt.edu
- [@bennaismithelt](#)