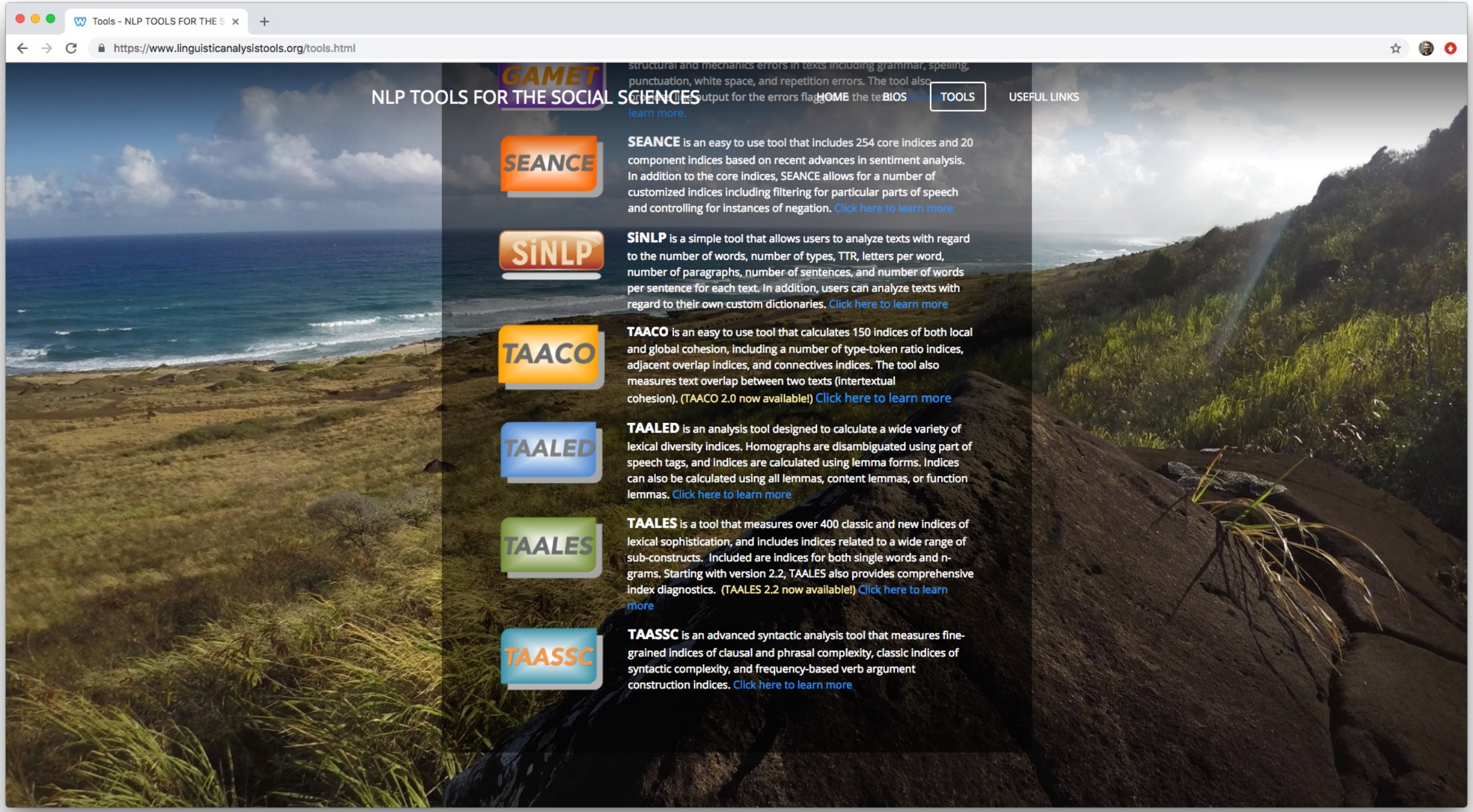
More than you ever wanted to know about: Lexical diversity

Day 1

Kristopher Kyle, University of Oregon

https://github.com/LCR-ADS-Lab/

UNIVERSITY OF OREGON


Tools - NLP TOOLS FOR THE S
https://www.linguisticanalysistools.org/tools.html
NLP TOOLS FOR THE SOCIAL SCIENCES
HOME
BIOS
TOOLS
USEFUL LINKS
GAMET
structural and mechanics errors in texts including grammar, spelling, punctuation, white space, and repetition errors. The tool also
learn more.
SEANCE
SEANCE is an easy to use tool that includes 254 core indices and 20 component indices based on recent advances in sentiment analysis. In addition to the core indices, SEANCE allows for a number of customized indices including filtering for particular parts of speech and controlling for instances of negation. Click here to learn more
SiNLP
SiNLP is a simple tool that allows users to analyze texts with regard 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

Thank you!

Colleagues involved in related projects

Hakyung Sung
UO + RIT

Masaki Eguchi
UO + Waseda

Fred Zenker
TU Braunschweig

Scott Crossley
Vanderbilt

Scott Jarvis
NAU

And MANY Others!

Learner Corpus Research and Applied Data Science Lab (LCR-ADS)

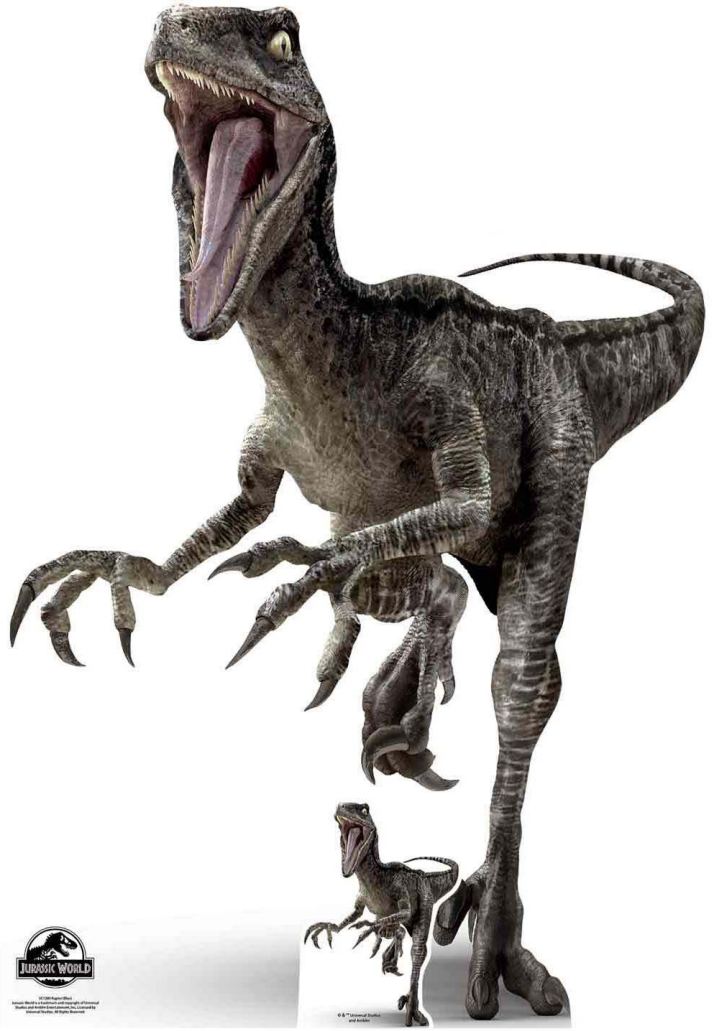
- We are interested in **characterizing** features of productive language use
  - across contexts of use/target language use domains (e.g., modes, registers, etc.)
  - across proficiency levels
  - across periods of time (e.g., development)
- We **evaluate** and **refine** methods of characterizing features of language use
  - cohesion
  - lexical diversity
  - lexical and lexicogrammatical sophistication
  - syntactic complexity
- We **develop** automated methods of characterizing features of language use
  - Stance-taking features (Eguchi, 2022; Eguchi & Kyle 2023 2024)
  - Characteristics of verb argument construction use (Kyle & Crossley, 2017; Kyle et al., 2021; Kyle & Sung, 2024; Sung & Kyle 2024a, under review).

Two important aspects of measuring a construct with an index

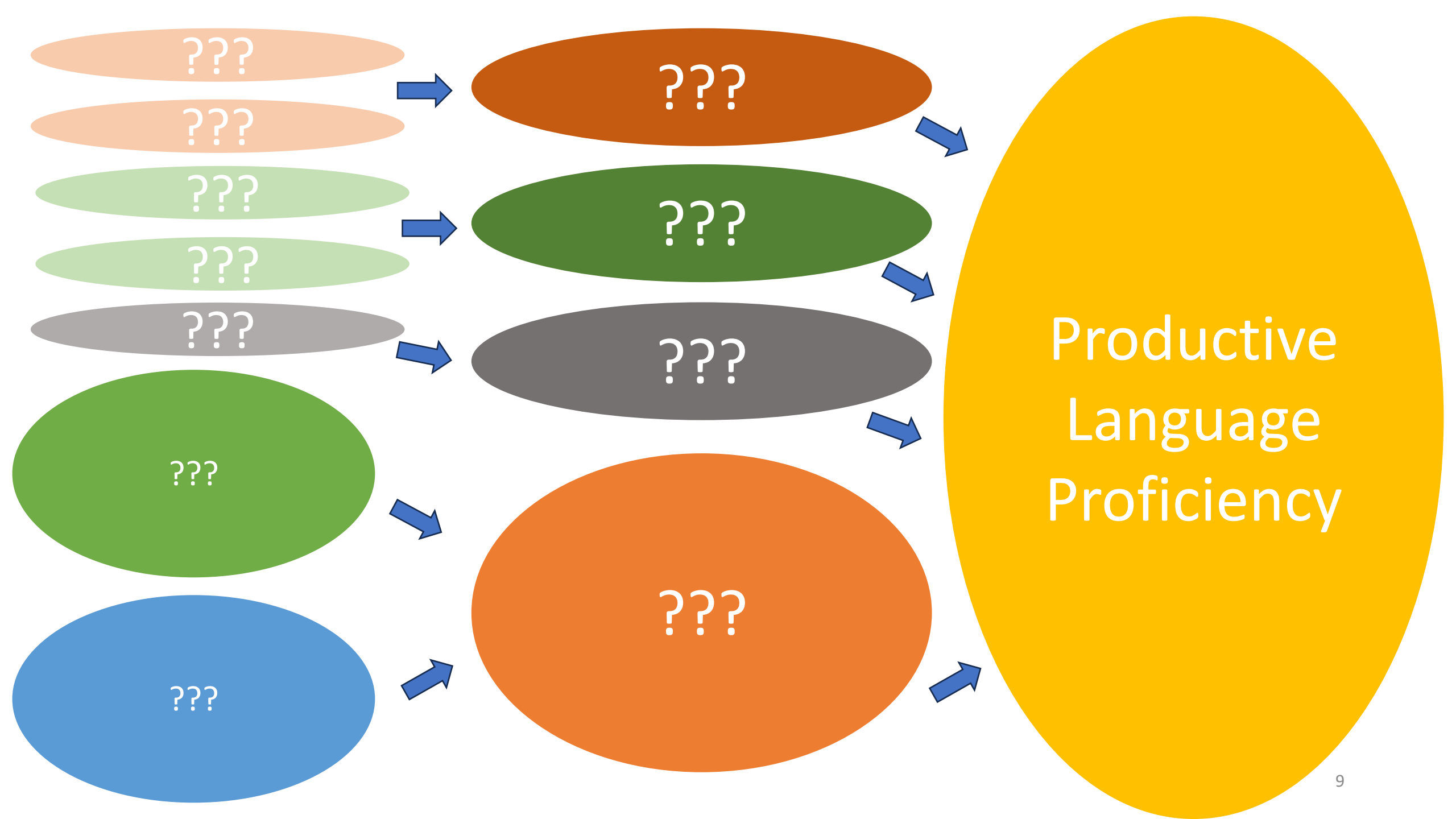
- Reliability
  - Consistent results
- Validity
  - Index measures what it is intended to measure

Defining Constructs

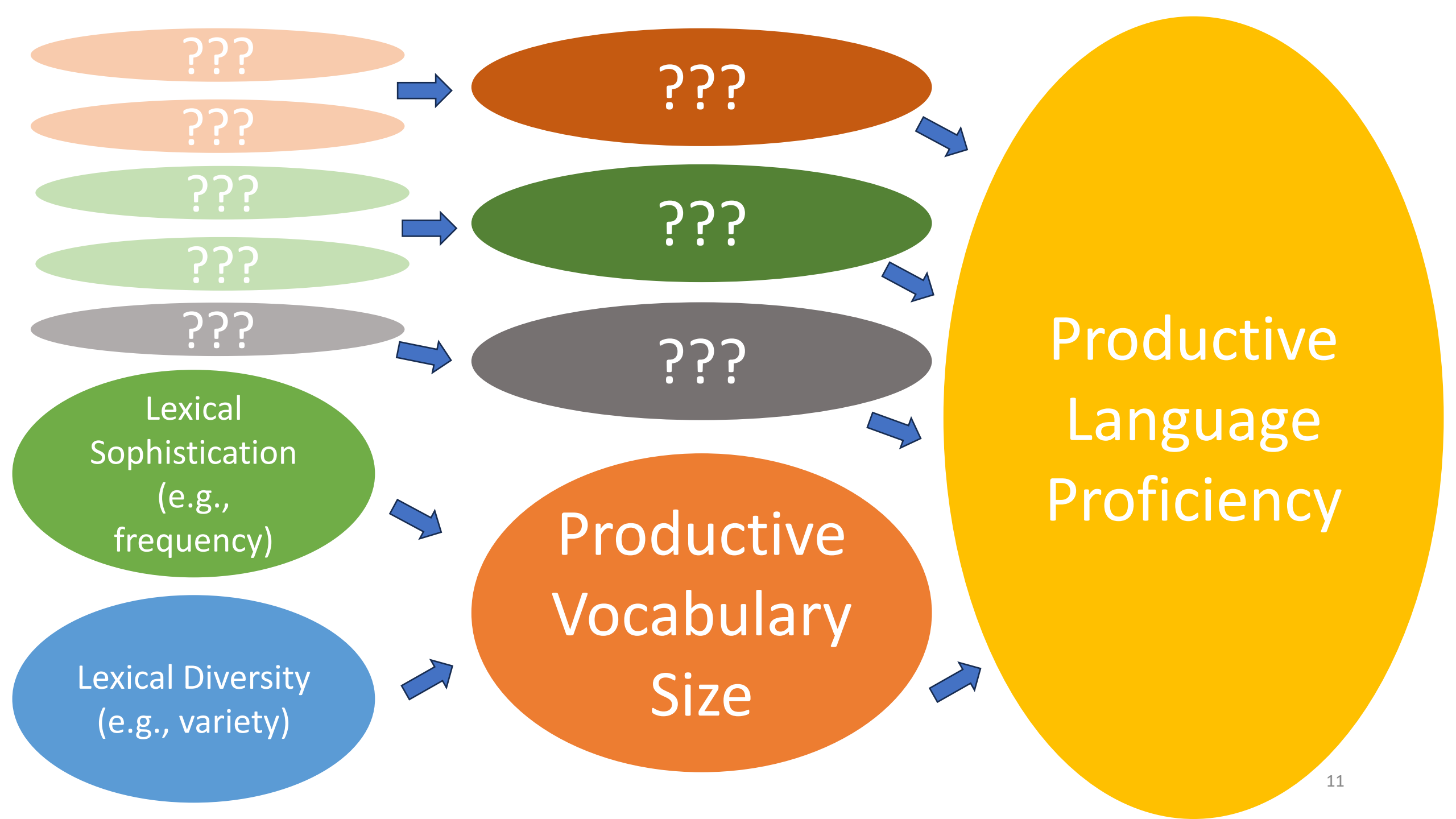
- How fast can Kris run?
- m/s? f/s? kph? mph?
- 100m?, 5k?, 50k?
- Fast enough to outrun a hungry velociraptor? (practical goal)
- Faster than other people in a group being hunted by a velociraptor? (norm-referenced practical goal)
- Possibly related, but construct irrelevant information:
  - Swimming speed
  - Height
  - Color of shoes



What are some constructs related to productive language proficiency?


???
???
???
???
???
???
???
???
???
???
Productive Language Proficiency

What are some constructs related to productive lexical proficiency?


???
???
???
???
???
???
???
???
Lexical Sophistication (e.g., frequency)
Lexical Diversity (e.g., variety)
Productive Vocabulary Size
Productive Language Proficiency

The Construct of Lexical Diversity

- Conceptualized since at least the 1930s (Carroll, 1938; Johnson, 1939; Yule, 1944; Zipf, 1937)
- Early on, it was conceptualized to measure the vocabulary size of a writer under the term *lexical richness* (see Yule, 1944):
  - More diverse vocabulary use = more words in mental lexicon (larger vocabulary)
- Later used in child development studies (e.g., Hunt, 1960) and in L2 development studies (e.g., Engber, 1995; Montanari et al., 2024)
- Usually refers to *lexical variety* though the construct has been argued to be multidimensional (see Jarvis, 2013, 2017a,b)

What is a “lexical item”?

*They **reanalyzed** the definition of a lexical item.*

- Word
  - definition: raw form of the item (usually in lower-case)
  - *reanalyzed*
- Lemma
  - definition: uninflected form of the item
  - *reanalyze (past tense inflection removed)*
- Family
  - definition: form of item with inflection and derivational affixes removed
  - *analyze*

What does each of these tell us about productive lexical knowledge?

What is a “lexical item”?

The smoker puffed smoke while smoking the smokes.

- How many different words?
- How many different lemmas?
- How many different word families?

What is a “lexical item”?

The smoker puffed smoke while smoking the smokes.

- How many different words?
  - [the, smoker, puffed, smoke, while, smoking, smokes] (7)
- How many different lemmas?
  - [the, smoker, puff, smoke, while] (5)
  - [the, smoker, puff, smoke_noun, while, smoke_verb] (6)
  - [the, smoker, puff, smoke_noun_s1, while, smoke_verb, smoke_noun_s2] (7)
- How many different word families?
  - [the, smoke, puff, while] (4)

Evaluating evidence for the reliability and validity of lexical diversity indices in L2 oral task responses

Kyle, Sung, Eguchi, & Zenker (2024, SSLA)

doi:10.1017/S0272263123000402

https://github.com/LCR-ADS-Lab/

UNIVERSITY OF OREGON

Overview

- We evaluate the reliability and validity of several indices of lexical diversity:
  - text-length stability (reliability)
  - relationship with human proficiency judgements (validity)
  - stability across task types (reliability)
- In a large corpus of L2 oral proficiency interviews (OPIs)

Reliability: Text length stability

- Indices of lexical diversity are often used as measures of lexical proficiency and/or lexical development in studies of SLA (e.g., Bulté & Roothooft, 2020; Lambelet, 2021; Tracy-Ventura et al., 2021; Vidal & Jarvis, 2020)
- As learners become more proficient, we expect that:
  - the size of their productive vocabulary will grow
  - they will use a wider variety of lexical items to complete a language task
  - they will also produce longer texts (for a variety of reasons)
- A well-known issue is that many indices of lexical diversity are intrinsically related to text length (e.g., Hess et al., 1986; McCarthy & Jarvis, 2010; Zenker & Kyle, 2021)

As language users become more proficient…

Larger productive vocabulary → More lexically diverse texts → Higher lexical diversity scores

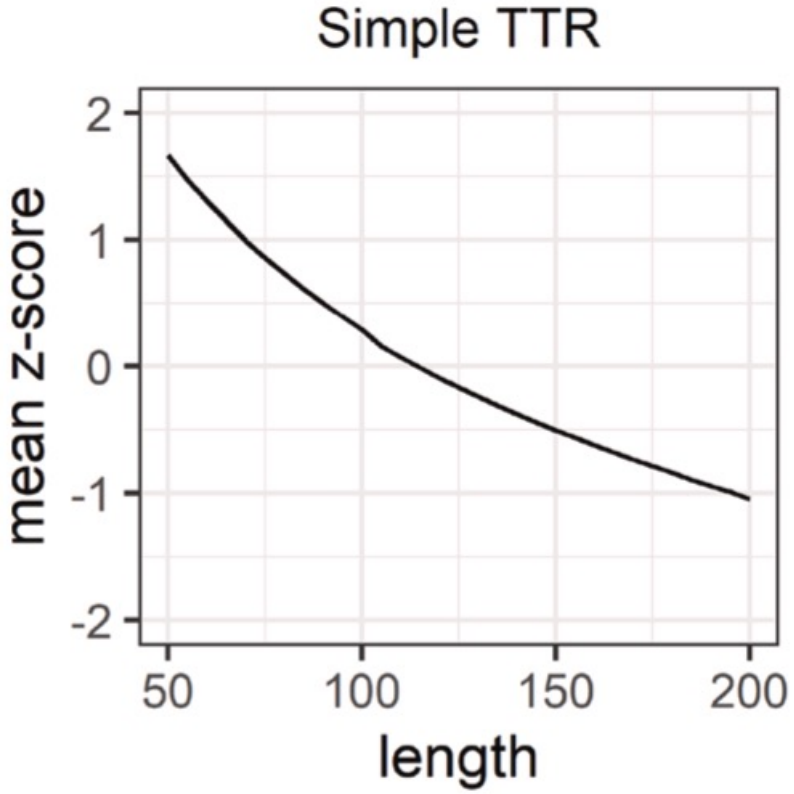
---

Increased fluency (broadly construed)
Increased register awareness
etc.

→ Longer texts → Higher lexical diversity scores?

Type-token ratio (TTR)

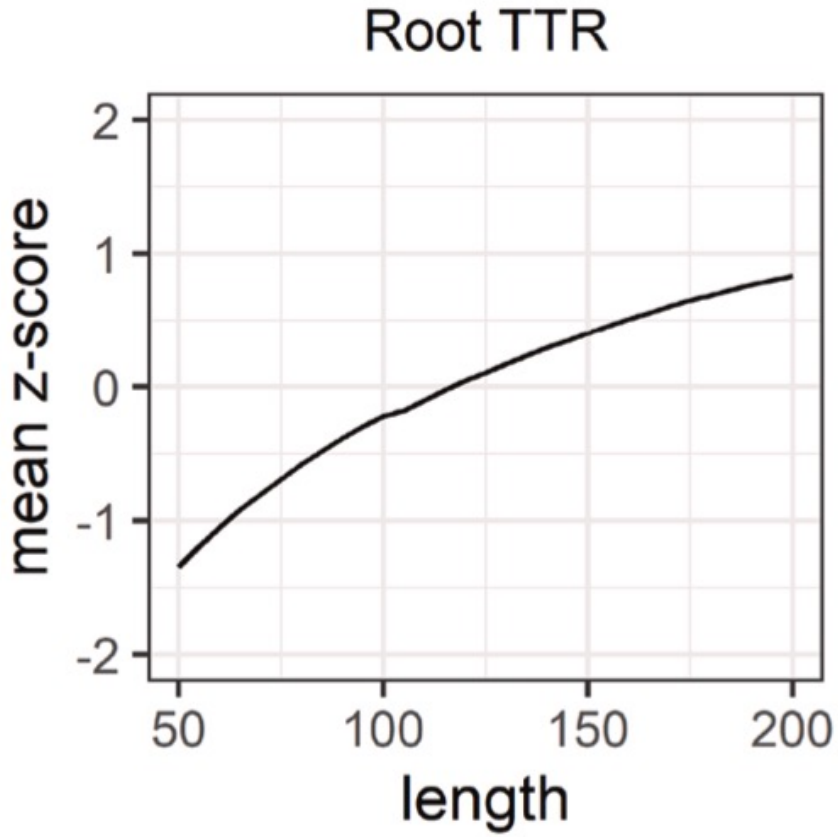
- The type-token ratio (TTR) is negatively (and intrinsically) correlated with text length
- *number of different words / total number of words*
- as texts get longer, TTR values decrease, suggesting that lexical variety decreases (e.g., In'nami & Koizumi, 2012; McCarthy & Jarvis, 2010; Zenker & Kyle, 2021)



(Zenker & Kyle, 2021)

Root TTR (Guiraud's Index)

- Root TTR (Guiraud's Index) overcorrects this effect, and is positively (and intrinsically) correlated with text length
- *number of different words / √(total number of words)*
- as text get longer, Root TTR values increase, conflating lexical diversity and text length (e.g., In'nami & Koizumi, 2012; McCarthy & Jarvis, 2010; Zenker & Kyle, 2021)



(Zenker & Kyle, 2021)

Is an index measuring LD or text length?

- It isn't unexpected for lexical diversity and text length to be correlated as both are indicative of proficiency gains
- However, we don't want lexical diversity indices that vary ***intrinsically*** due to text length
- One method of determining whether indices vary due to text length: Parallel Sampling (Hess et al., 1986)

Parallel sampling (Hess et al., 1986)

1. Clip text to first *n* tokens (e.g., 400)
2. Subdivide text into equal segments of 50 tokens (8 segments), 55 tokens (7 segments), and so on all the way up to the final segment of 400 tokens
3. Calculate lexical diversity values for each resulting segment
4. Average values from equal-length segments
5. Analyze relationship between lexical diversity values and segment lengths

Some promising indices of lexical diversity

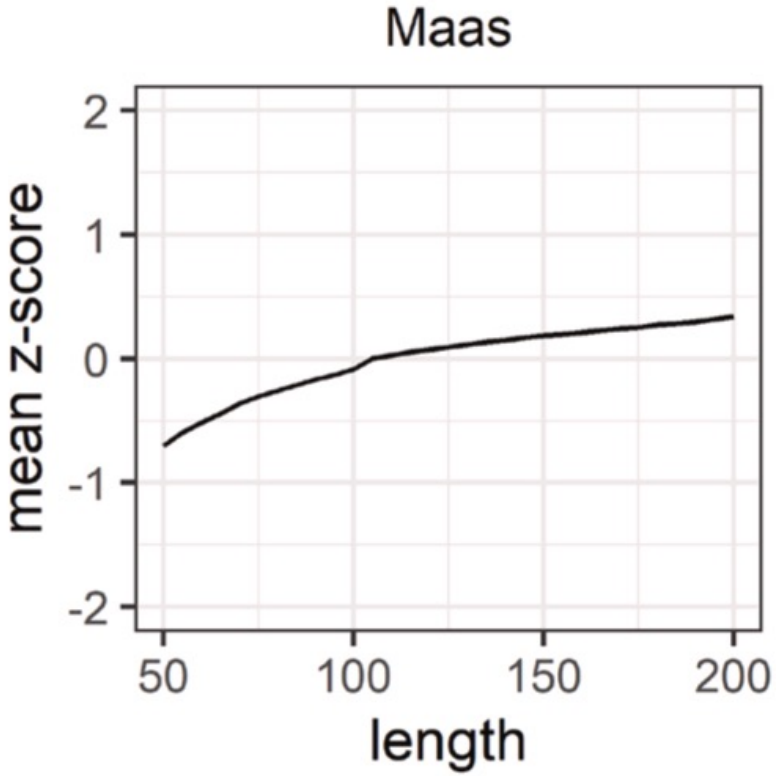
- Maas's index (Maas, 1971)
- D (Malvern & Richards, 1997; McCarthy & Jarvis, 2007)
- MATTR (Covington and McFall, 2010)
- MTLD (McCarthy & Jarvis, 2010)

Maas's index (Maas, 1972)

- Attempts to fit the TTR value to a logarithmic curve

$$Maas = \frac{Log(ntokens) - Log(ntypes)}{Log(ntokens)^2}$$

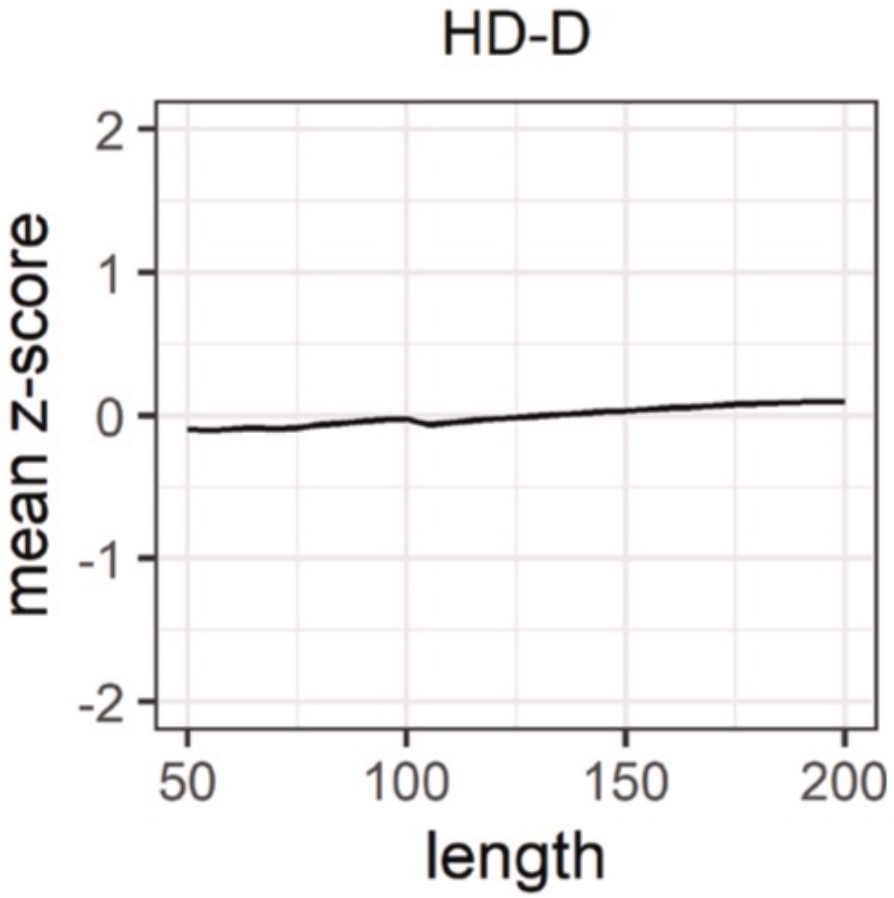
- A number of studies have demonstrated that Maas scores are affected by text length, though these effects tend to be lower than other transformations of TTR (Koizumi & In'nami, 2012; McCarthy & Jarvis 2007; 2010; Zenker & Kyle, 2021)



(Zenker & Kyle, 2021)

D

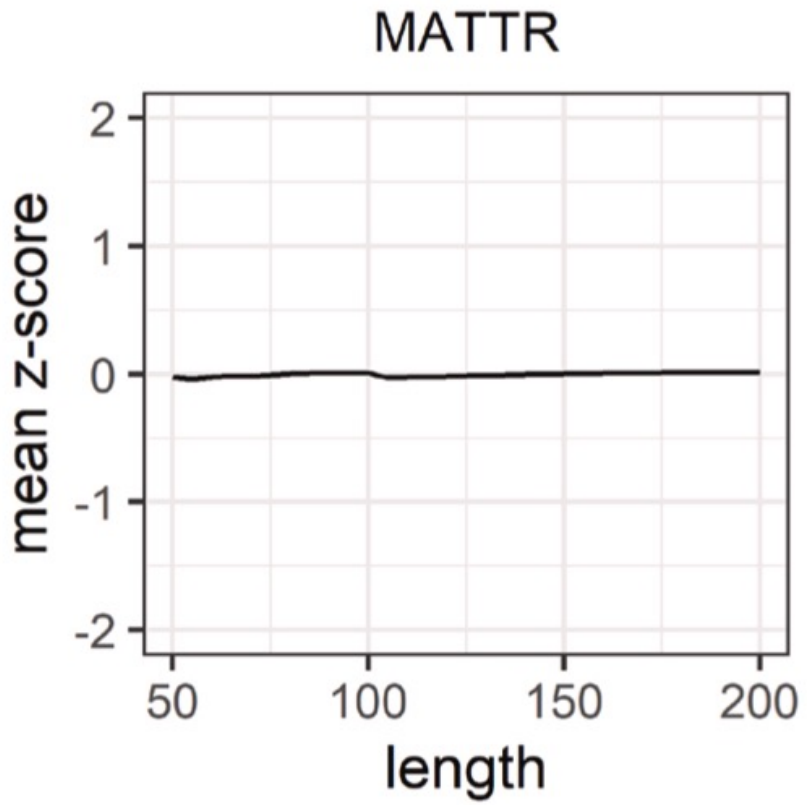
- vocd-D (Malvern & Richards, 1997)
  - Uses a bootstrapped approach that repeatedly fits the rate of decline in TTR values (i.e., the TTR curve) in random samples of varying lengths from a text
- HD-D (McCarthy & Jarvis, 2010)
  - For each word type in a text, uses hypergeometric distribtution to calculate probability of encountering one of its tokens in a random sample of 42 tokens
  - Probabilities added together to produce final HD-D value
  - McCarthy and Jarvis (2010) make a strong argument that HD-D and vocd-D measure the same construct, but HD-D does so in a more precise (and straightforward) manner
- A number of studies have shown that D is relatively stable across writing samples (e.g., McCarthy & Jarvis, 2010; Zenker & Kyle, 2021), but there is preliminary evidence that it may be less stable in spoken samples (e.g., Koizumi & In'nami, 2012)



(Zenker & Kyle, 2021)

Moving-average TTR (MATTR)

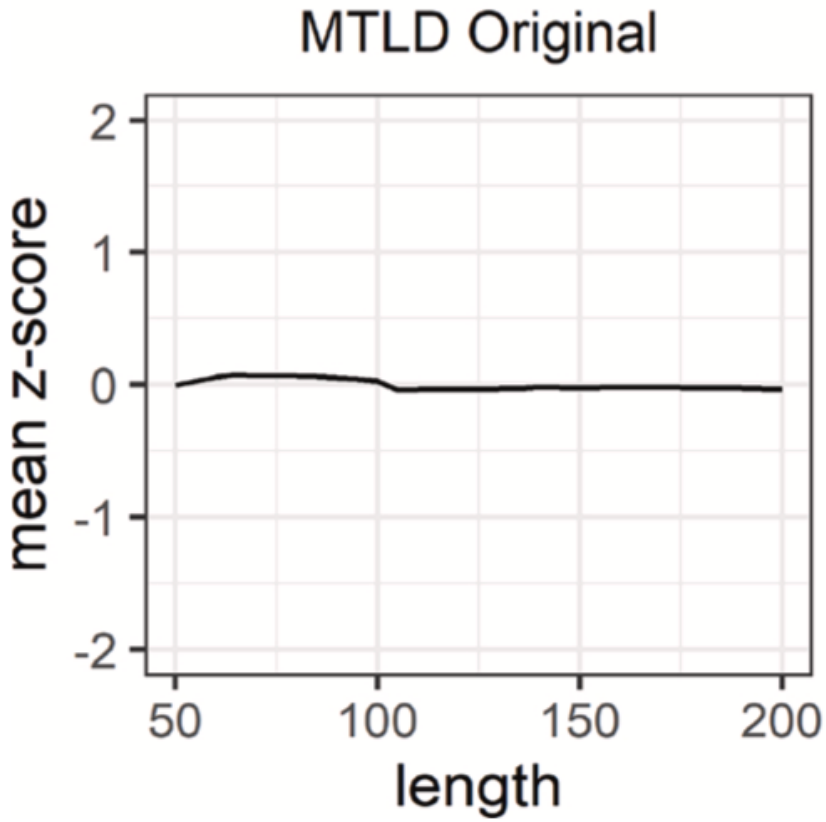
- Uses a moving-window approach that averages the TTR values for all progressive windows of a particular size
- For example (window size = 50): Mean of TTR values for words 1–50, 2–51, 3–52, etc.
- Fifty-word windows are common, but smaller windows can be used (e.g., Fergadiotis et al., 2015)



(Zenker & Kyle, 2021)

Measure of textual lexical diversity (MTLD)

- Average number of words in a text needed for TTR values to drop below a pre-determined value (e.g., .720)
- Calculated forwards and backwards in a text
- Cut-off value of .720 almost exclusively used, but based on the point of TTR stabilization in L1 texts (McCarthy & Jarvis, 2010)



(Zenker & Kyle, 2021)

Validity: Relationship with human judgements

- Relationship between LD score and direct human judgements of LD
  - Written L2 narrative retellings (e.g., Jarvis, 2017; Jarvis & Hashimoto, 2021)
  - L1 and L2 argumentative essays (Kyle, Crossley, & Jarvis, 2021)
  - Correlations as high as *r* = .600 for text-length resistant indices
- Relationship between LD score and analytic or holistic productive proficiency scores (e.g., Bulté & Roothooft, 2020; Engber, 1995; Jarvis, 2002; Treffers-Daller et al., 2018; Zenker & Kyle, 2021)
  - Correlations as high as *r* = .600 for text-length resistant indices

This study…

- Although a number of studies have looked at:
  - text-length stability
  - validity
  - across-task stability
- Most studies have used written data
- Very few have analyzed L2 spoken data
  - Koizumi & In'nami (2012; *n* = 38)
  - Koizumi (2012; *n* = 20)

Research questions

1. What is the relationship between lexical diversity indices and text length in oral proficiency interviews?
2. To what degree are text-length stable indices of lexical diversity predictive of oral proficiency interview scores?
3. To what degree are text-length stable indices of lexical diversity stable across oral proficiency interview sub-tasks?

Method: Learner corpus

- NICT JLE Corpus (*n* = 1,281)
- Japanese L1, English L2 Standardized Speaking Test (adapted from ACTFL OPI)
- Included 3 stages/tasks
  - single-picture description task
  - role play task
  - sequential picture storytelling task

Method: Linguistic analysis

- Each text was cleaned by removing
  - Pauses, disfluencies, Japanese utterances
  - Spelling errors (during transcription)
- Lemmatized and pre-processed using the Python package *pylats* (Kyle, 2022)
- Lexical diversity indices calculated by the Python package *taaled* (Kyle, Sung, & Eguchi, 2022)
- Included optimized versions of MATTR (11) and MTLD (.92)

TAALED
Search docs...
QUICK START GUIDES
Calculate Revised LD Indices
Calculate Classic (but Flawed) LD Indices
Batch Processing and Writing Results to a File
RESOURCES
Related Studies
TAALED main

README.md

Tool for the Automatic Analysis of Lexical Diversity (TAALED)

TAALED is a Python package for calculating lexical diversity (LD) indices. The package is designed for the researchers, students, and teachers in (applied) linguistics needing to calculate LD indices that are stable across different text lengths (i.e., revised LD indices) as well as classic LD indices. The package was developed by Kristopher Kyle. Many thanks to Scott Jarvis, who provided valuable insights about the calculation of MTLD and HD-D. This documentation page is contributed by Hakyung Sung and Masaki Eguchi in the LCR-ADS lab at the University of Oregon.

Quick Start Guides

How to Install TAALED

To install TAALED, you can use `pip (a package installer for Python)`:

```bash
pip install taaled
```

How to Install Related Packages

While not strictly necessary, this tutorial will presume that you have also installed a few helpful packages for text preprocessing and visualization. These are optional but recommended.

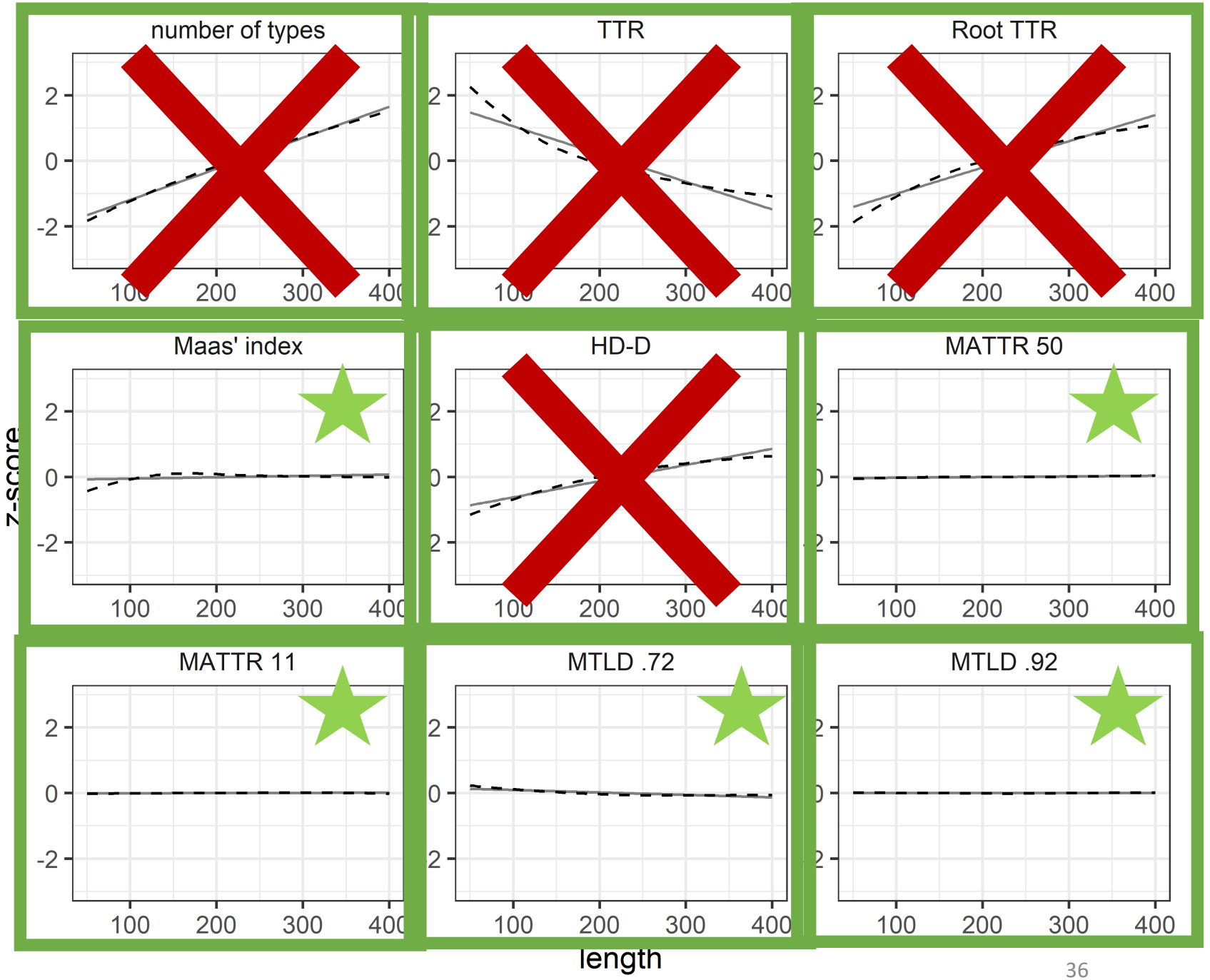
TAALED takes a list of strings as input and returns various indices of LD (and diagnostic information). In the rest of the tutorial, we will use pylats for preprocessing of texts (e.g., tokenization, lemmatization, word disambiguation, checking for misspelled words, etc.). Currently pylats only supports advanced features for English (models for other languages are forthcoming). Pylats was tested using spacy version 3.2 and by default uses the "en_core_web_sm" model. To install spacy and a language model, see the spacy installation instructions.

However, TAALED can work with any language, as long as texts are tokenized (and appropriately preprocessed). See tools such as

Method: Statistical analysis

- RQ1: Correlation between text length segments (via parallel analysis) and lexical diversity index score
- RQ2: Correlation between OPI proficiency score and lexical diversity score
- RQ3: Linear mixed-effects models with post-hocs for lexical diversity scores across the three tasks

Results: Text length stability



Results: Relationship with OPI score

Correlations between text-length stable indices and OPI Score							
	Score	number of tokens	number of types	Maas' Index	MATTR 11	MATTR 50	MTLD .72
number of tokens	0.831						
number of types	0.828	0.959					
Maas' Index	0.146	0.243	0.049				
MATTR 11	0.504	0.466	0.555	-0.225			
MATTR 50	0.258	0.181	0.320	-0.695	0.666		
MTLD .72	0.296	0.234	0.353	-0.582	0.726	0.862	
MTLD .92	0.430	0.405	0.499	-0.308	0.814	0.696	0.716

Discussion

- RQ1:
  - TTR, Root TTR, and D are NOT stable across texts of different lengths in L2 spoken contexts
  - Maas, MATTR (11 and 50), and MTLD (.72 and .92) are stable across texts of different lengths
- RQ2:
  - MATTR 11 (*r* = .504) and MTLD .92 (*r* = .403) demonstrated the strongest relationships with OPI score

Conclusion

- When we measure linguistic constructs, we should ensure that they are both reliable and arguably valid
- When measuring lexical diversity, we should NOT use:
  - TTR, Root TTR, Maas
  - vocd-D, HD-D
- We SHOULD use:
  - MATTR 11 or MTLD .92
  - MATTR 50 and MTLD .72 are also reasonable (but have less validity support)

Limitations and future directions

- We investigated one production mode and context (OPIs)
- We investigated a single L1 group (Japanese L1, English L2)

LCR-ADS-Lab

github.com/LCR-ADS-Lab

Update

Gmail | Mail - Kris Kyle -... | Google Play Books | Maps | Calendar | ScholarOne Manu...

Search or jump to... / Pull requests Issues Marketplace Explore

LCR-ADS-Lab

Follow

Overview | Repositories 10 | Projects | Packages | Teams | People 3 | Settings

Popular repositories

SL2E-Dependency-Treebank Public

☆ 3 ⑂ 1

LCR-ADS-Home Public

Landing page for information about the Learner Corpus Research and Applied Data Science Lab at the University of Oregon

☆ 1

pylats Public

Text pre-processing for downstream linguistic analyses

Python ☆ 1

TAALED Public

Tool for the automatic assessment of lexical diversity

Python ☆ 1 ⑂ 1

l2-nlp-training-spacy Public

Python

Repositories

Find a repository... | Type | Language | Sort | New

ld-project Private

Python ☆ 0 ⑂ 0 ⊙ 0 ⑂ 0 Updated on Aug 5

TAALED Public

View as: **Public**

You are viewing the README and pinned repositories as a public user.

You can create a README file or pin repositories visible to anyone.

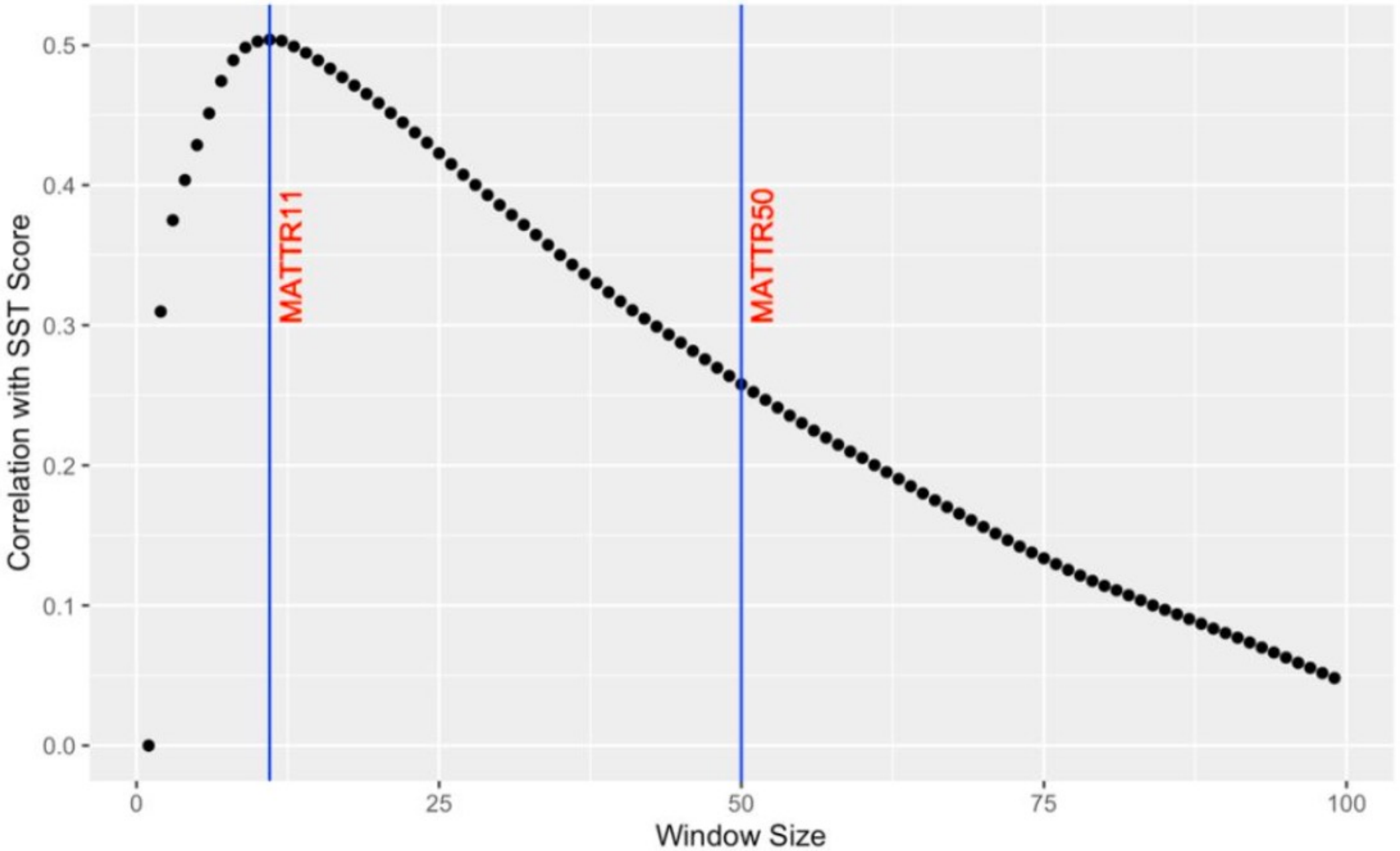
Get started with tasks that most successful organizations complete.

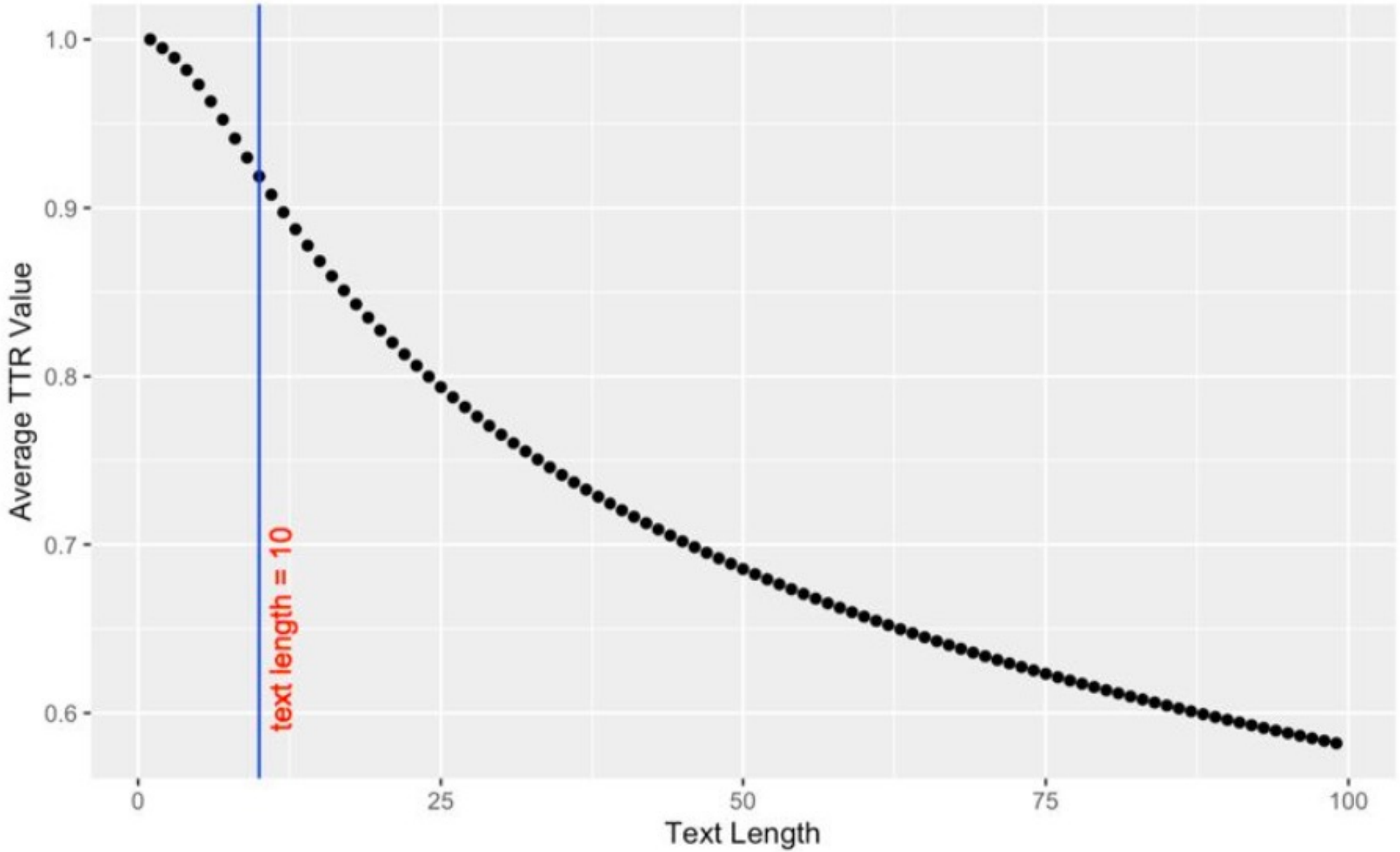
People

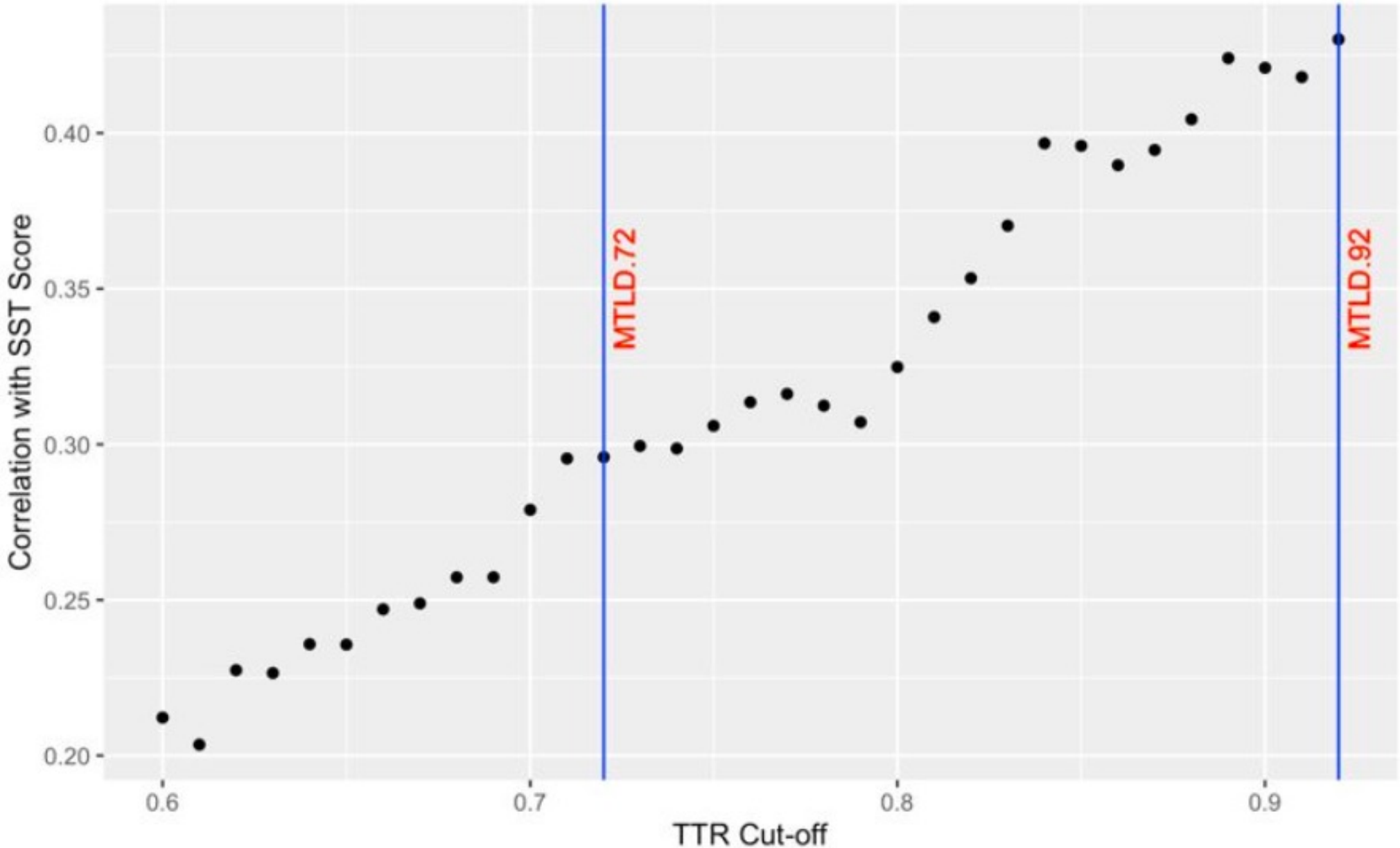
0 seats left — Buy more

Top languages

Python HTML Jupyter Notebook


Correlation with SST Score
0.5
0.4
0.3
0.2
0.1
0.0
MATTR11
MATTR50
0
25
50
75
100
Window Size


1.0
0.9
0.8
0.7
0.6
Average TTR Value
text length = 10
0
25
50
75
100
Text Length

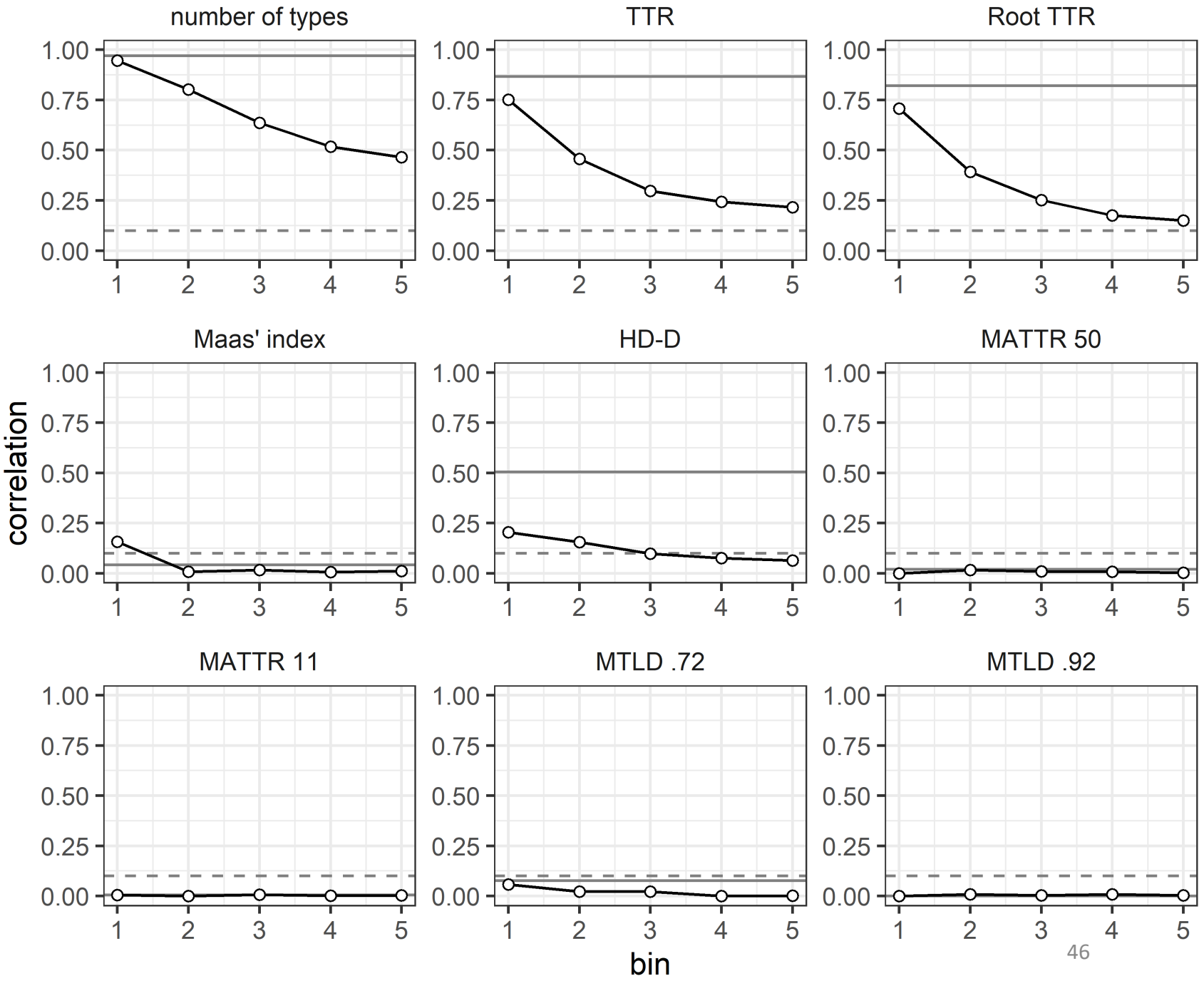

Correlation with SST Score
0.40
0.35
0.30
0.25
0.20
MTLD.72
MTLD.92
0.6
0.7
0.8
0.9
TTR Cut-off

Reliability across tasks

- In some studies, lexical diversity scores may be compared across tasks
  - Longitudinal studies (e.g., Tracy-Ventura et al., 2016)
  - Some cross-sectional studies (Lu, 2012; Verspoor et al., 2012)
- Relatively few studies in this area (all written)
  - Alexopoulou (2017) found differences in MTLD scores across EFCAMDAT tasks
  - Zenker and Kyle (2021) found differences across the two prompts in the ICNALE corpus for a range of LD indices
  - Yoon (2017) found negligible differences in D values across writing prompts

Results: Text length stability

Bin	Word Range
1	50-115
2	120-185
3	190-255
4	260-325
5	330-400



Results: Differences across tasks

Summary of differences across stages					
Index	$R^2_{Marginal}$	$R^2_{Conditional}$	d (Stage 2-3)	d (Stage 2-4)	d (Stage 3-4)
Maas' Index	.011	.226	-0.255	-0.136	0.119
MATTR 11	.003	.294	-0.113	0.013	0.127
MATTR 50	.010	.288	0.053	0.230	0.177
MTLD .72	< .001	.254	0.047	0.029	-0.018
MTLD .92	.002	.237	-0.069	0.041	0.110

Discussion

- RQ3:
  - MATTR 11 and MTLD .92 were stable across the three OPI tasks investigated

Limitation

- The tasks included in OPIs are reasonably similar (so across-task stability may be overstated)