Conceptualizing and measuring lexical complexity as a multidimensional construct

Dr. Kristopher Kyle
Linguistics Department, University of Oregon
Director, Learner Corpus Research and Applied Data Science Lab

www.kristopherkyle.com
www.linguisticanalysistools.org
Colleagues involved in related projects

Scott Crossley (GSU)
Cindy Berger (DuoLingo)
Masaki Eguchi (UO)
Minkyung Kim (NUCB)
Katia Monteiro (GSU)

Danielle McNamara (ASU)

And MANY Others!
Outline of talk

• What is lexical complexity?
• How has it been traditionally measured?
• Multivariate approaches to the measurement of lexical sophistication
• Overview of a range of indices related to lexical sophistication
• Example Study 1: Modeling TOEFL iBT writing scores
• Example Study 2: Modeling Oral Proficiency Interview scores
• Concluding thoughts
Productive lexical proficiency

Productive lexical proficiency is an essential aspect of communicating in a particular language (e.g., Nation, 2001)

Scholars have tended to refer to productive lexical proficiency using the terms “lexical complexity” or “lexical richness” (Yule, 1944; Ure, 1971; Laufer & Nation, 1995)

Lexical complexity/richness includes three broad constructs:

• lexical density
• lexical diversity
• lexical sophistication
Lexical Complexity/Richness

Lexical diversity
(see e.g., Jarvis, 2013, Koizumi & In’nami; McCarthy & Jarvis, 2010; Zenker & Kyle, 2021)

Lexical Density
(see Johansson, 2008; Lu, 2012)

Lexical Sophistication
(see Crossley, Salsbury, McNamara, & Jarvis, 2011a, b; Eguchi & Kyle, 2020; Kyle & Crossley, 2015; Meara, 1996; Nation, 1990; Read, 2000)
Lexical sophistication

Lexical sophistication has traditionally been referred to as “the proportion of unusual or advanced words in the learner’s text” (Read, 2000, pp. 203)

Hypothesis: More proficient writers will use a higher proportion of low frequency words (and therefore a lower proportion of high frequency words)

Lexical sophistication has often been measured using tools such as:

- Range (Heatley & Nation, 1994)
- AntWordProfiler (Anthony, 2014)
- TAALES (Kyle et al., 2018)
- VocabProfile (Cobb, 2019)
Word frequency is clearly an important indicator of proficient written lexical use...

For example, small to medium effect sizes reported in:

• Guo, Crossley, & McNamara (2013)
• Kyle & Crossley (2015, Study 1)
• Kyle, Crossley, & Berger (2018)
• Laufer & Nation (1995)
• etc.
But, word frequency isn’t everything

Theoretical perspectives and empirical research on language development support the importance of frequency (e.g., Ellis, 2002; Laufer & Nation, 1995; Kyle & Crossley, 2015)

However, theoretical perspectives outline other factors as well (e.g., salience)

Additionally, empirical research in applied linguistics and related fields (e.g., Balota et al., 2007; Brysbaert et al., 2013) has indicated that other factors affect difficulty at the word level

Research in corpus linguistics (e.g., Bestgen & Granger, 2014; Nation, 2001; Sinclair, 1991) has indicated that productive lexical proficiency includes how (and which) words fit together
Also, frequency measures are not consistent across all language use domains

• **Written Tasks**
  • More proficient writers tend to use words that are less frequent:
    • Guo, Crossley, & McNamara (2013)
    • Kyle & Crossley (2015; Study 1)
    • Kyle, Crossley, & Berger (2018)
    • Laufer & Nation (1995)

• **Spoken Tasks**
  • More proficient speakers tend to use words that are more frequent:
    • Kyle & Crossley (2015; Study 2)
    • Berger et al. (2019)
    • Eguchi & Kyle (2020)
A revised definition of lexical sophistication

In recent work (e.g., Crossley et al., 2011a,b; Kyle & Crossley, 2015; Kyle et al., 2018), colleagues and I have adopted a (slightly) revised definition of lexical sophistication, which refers to the:

- the relative difficulty of learning and/or using a word in a particular language use domain
- raters’/interlocutors’ perception of sophisticated word use in a particular language use domain (which seems to be related to the concept above)

This allows for word sophistication to be measured using a variety of features in addition to corpus frequency (and to vary based on domain) This also allows for lexicogrammatical features to be considered
In other words...

Factors that affect ease/difficulty of learning/use

Factors that affect perceptions of proficiency/sophistication

(We can see longitudinal trends; Crossley et al, 2011; Crossley & Kyle, 2016, etc.)

We can see cross-sectional, rating-based trends (Crossley et al.; Kyle et al., ad nauseum)

lexical sophistication
Beyond word frequency

Which one of these is more “sophisticated”?  
- unemployment
- intonation

What about these?  
- million
- home

And these?  
- intonation
- eggshell

unemployment: 3,137 instances (36.16 per million words)  
intonation: 42 instances (.47 per million words)

million: 80,701 instances (921.97 per mill)  
contextual diversity score of 3
home: 81,104 instances (926.58 per mill)  
contextual diversity score of 97

intonation: 42 instances (.47 per million words)  
concreteness of 2.46
eggshell: 42 instances (.47 per million words)  
concreteness of 4.77
A sample of lexical sophistication indices

Data from:

• Kyle & Eguchi (2021a)
  • In house scripts based on TAALES 3.0 (under development)
  • TOEFL Public Use Dataset (480 argumentative essays)
  • COCA news + magazine as reference corpus

• Kyle & Eguchi (2021b, under review)
  • In house scripts based on TAALES 3.0 (under development)
  • NICT JLE (1,281 transcribed oral proficiency interviews)
  • COCA spoken as reference corpus
Word Features: Content Word Frequency
$r = -0.193$
Word features: Contextual distinctiveness

Contextual distinctiveness refers to the diversity of lexical or semantic contexts in which a word occurs. Words that occur in more contexts are more easily learned, and therefore are considered to be less sophisticated.

More contexts

food (324)
work (196)
nice (115)
sweat (36)
want (33)

More sophisticated

tan (15)
ancient (13)
scissors (7)
satin (5)
inlet (2)
Some research findings with contextual distinctiveness

A negative relationship was found between USF norms for contextual distinctiveness and:

• spoken lexical proficiency scores (Berger, Crossley, & Kyle, 2017)

This is a rich area for future research

• (e.g., Eguchi & Kyle, 2020)
Concreteness refers to the degree to which a word refers to a perceptible entity (a nice paraphrase is *abstractness*).

Words with high concreteness values tend to be more salient, and therefore easier to learn.

Accordingly, words with lower concreteness values are considered more sophisticated.*
Some research findings with concreteness

Negative relationships have been observed between concreteness scores and:

- written lexical proficiency scores (Kyle, et al., 2018)
- time spent studying an L2 (Crossley, Kyle, & Salsbury, 2016)
- Oral proficiency interview scores (Eguchi & Kyle, 2020; Kyle & Eguchi, 2021)

Concreteness scores

Productive proficiency

Negative relationship
Beyond single words

Knowing sophisticated words is an important aspect of productive lexical proficiency

However, knowing how words fit together is also essential (e.g., Nation, 2001; Sinclair, 1991)

Reading essays produced by students who overuse their electronic dictionaries
Lexicogrammatically features: N-grams

N-grams are contiguous word sequences of n words
Most often two-word (bigram) or three-word (trigram) sequences are investigated
More lexically proficient users tend to use more strongly associated n-grams

More strongly associated:
- vouch for (.89)
- tend to (.87)
- reliance on (.83)
- thinly slice (.79)
- adversely affect (.72)

More sophisticated:
- hazard be (0.0)
- debate to (-.01)
- degree be (-.02)
- crisis a (-.02)
- sugar the (-.04)
$r = 0.165$
Some research findings with n-grams

Positive relationships have been found between indices of n-gram frequency and strength of association and:

• Written lexical proficiency scores
  • Kyle & Crossley (2015)
• Argumentative writing quality scores
  • Garner, Crossley, & Kyle (2020)
  • Garner, Crossley, & Kyle (2019)
  • Granger & Bestgen (2014)
• Narrative writing quality scores
  • Bestgen & Granger (2014)
• Monologic speaking proficiency
  • Kyle & Crossley (2015)
• Oral Proficiency Interviews
  • Eguchi & Kyle (2020)
Lexicogrammatical features: Dependency relations

Dependency relations capture direct grammatical relationships between two words:

- adjective – noun
- modifier – verb
- verb – direct object
- etc.

Dependency relations are likely clearer indicators of appropriate use than n-gram measures

More strongly associated

- persuasively + argue (.48)
- enviously + watch (.42)
- soundly + sleep (.38)

More sophisticated

- politically + go (-.02)
- evenly + go (-.02)
- efficiently + have (-.02)
The scatter plot shows a positive correlation between writing proficiency score and the variable `n_amod_MI`, with a correlation coefficient $r = 0.254$. The data points are scattered around a line that indicates a weak positive relationship.
Some research findings with dependency relations

Positive relationships have been found between indices of dependency relation strength of association and:

• Argumentative writing quality
  • Kyle & Eguchi (in press)
  • Paquot (2017, 2019)

This is a rich area for further research

• e.g., Kyle & Eguchi, (under review)
Putting the pieces together...

Small to moderate correlations exist between
  • various individual indices of lexical sophistication
  • Spoken and written productive lexical proficiency

Powerful, multivariate predictor models can be created by using these indices (and more):
  • 41% of the variance in spoken lexical proficiency scores (Berger et al., 2017)
  • 58% of the variance in OPI speaking scores (Kyle & Eguchi, under review)
  • 51% of the variance in TOEFL monologic speaking scores (Kyle & Crossley, 2015)
  • 58% of the variance in written lexical proficiency scores (Kyle et al., 2018)

Many more studies are warranted!
Study 1 (TOEFL iBT Writing)

- A combined model using four indices accounted for 22% of the variance ($R^2 = .222$, $r = .471$) in writing proficiency scores
  - contextual diversity (USF)
  - noun-adjective SOA
  - verb-adverb SOA
  - verb-direct object SOA
$r = 0.471$
Study 2 (OPI Scores)

- A combined model with six indices accounted for 49% of the variance in OPI scores ($R^2 = .486, r = .694$).
  - concreteness
  - content word frequency
  - contextual diversity (USF)
  - bigram SOA
  - verb-direct object SOA
  - verb-subject SOA
This is only represents an introduction to what is available

• Word level
  • lexical diversity
  • lexical decision times
  • word neighborhoods
  • semantic categories
  • etc.

• Lexicogrammatical level
  • verb argument construction use
  • other phrasal categories
  • cohesion
  • etc.
Beyond assessing productive lexical proficiency

Taking a multivariate approach to assessing productive lexical proficiency has proven to be extremely fruitful.

Accordingly, multivariate approaches have begun to be taken in other areas of lexical proficiency such as:

- Choosing items for vocabulary tests (Peters & Webb, 2018)
- Reading difficulty (Crossley, Greenfield, and McNamara, 2008)

However, these have been small steps, and there are many areas that need further work:

- creation of target vocabulary lists for curriculum development
- assessing L2 reading difficulty
- etc.
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 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 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 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 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 phrases. Starting with version 2.2, TAALES also provides comprehensive index diagnostics. (TAALES 2.2 now available) Click here to learn more.

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.
References:
Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904–911.


