**Title:** Assessing Writing with the Tool for the Automatic Analysis of Lexical Sophistication (TAALES)

Author 1: Scott A. Crossley
Department of Applied Linguistics/ESL
Georgia State University
25 Park Place, Suite 1500
Atlanta, GA 30303, USA
Email: sacrossley@gmail.com

Author 2: Kristopher Kyle
Department of Second Language Studies
University of Hawai’i Manoa
Moore Hall 570
1890 East-West Road
Honolulu, Hawai□i 96822
Email: kristopherkyle1@gmail.com

Synopsis

| | |
|---|---|
| Tool | Tool for the Automatic Analysis of Lexical Sophistication (TAALES) |
| Tool Purpose | To calculate a wide range of classic and newly developed indices of lexical sophistication (e.g., frequency, range, academic register, concreteness and familiarity, psycholinguistic word properties, and semantic network norms). |
| Key Premise | Lexical sophistication features are strong predictors of writing quality in both first language (L1) and second language (L2) contexts. |
| Research Connections | Inspired by Coh-Metrix but developed to improve on the limitations of the Coh-Metric tool. |
| Limitations | Does not examine accuracy of lexical use or words not available in selected databases. |
| Future Developments | Continue to add lexical databases as they become available. |

In this review, we provide an overview of the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015; Kyle, Crossley, & Berger, 2017) and discuss its applications for automatically assessing features of written text. TAALES is a is context and learner independent natural language processing (NLP) tool that provides counts for 100s of lexical features. Development of TAALES began in the summer of 2013 as a result of an independent study in NLP taught at Georgia State University. At that time, initial efforts were made to improve upon the lexical features reported by the NLP tool Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004), which the first author had helped develop and had extensively tested in earlier research (see the Connections Section for additional details).

The current version of TAALES (TAALES 2.8) is freely available[1] and works on a number of operating systems such as Linux, Mac, and Windows. The tool is also user friendly,

[1] TAALES 2.8 is freely available at www.kristopherkyle.com/taales.html under a Creative Commons Attribution-NonCommercial-ShareAlike International license. All included databases are free for non-commercial, research purposes at the time of writing but may fall under a use license other than that of TAALES. Researchers should check each database source (available in the supplementary material document entitled "TAALES_2.8_Index_Guide.xlsx") to determine whether their project falls within the guidelines and/or license for each database.

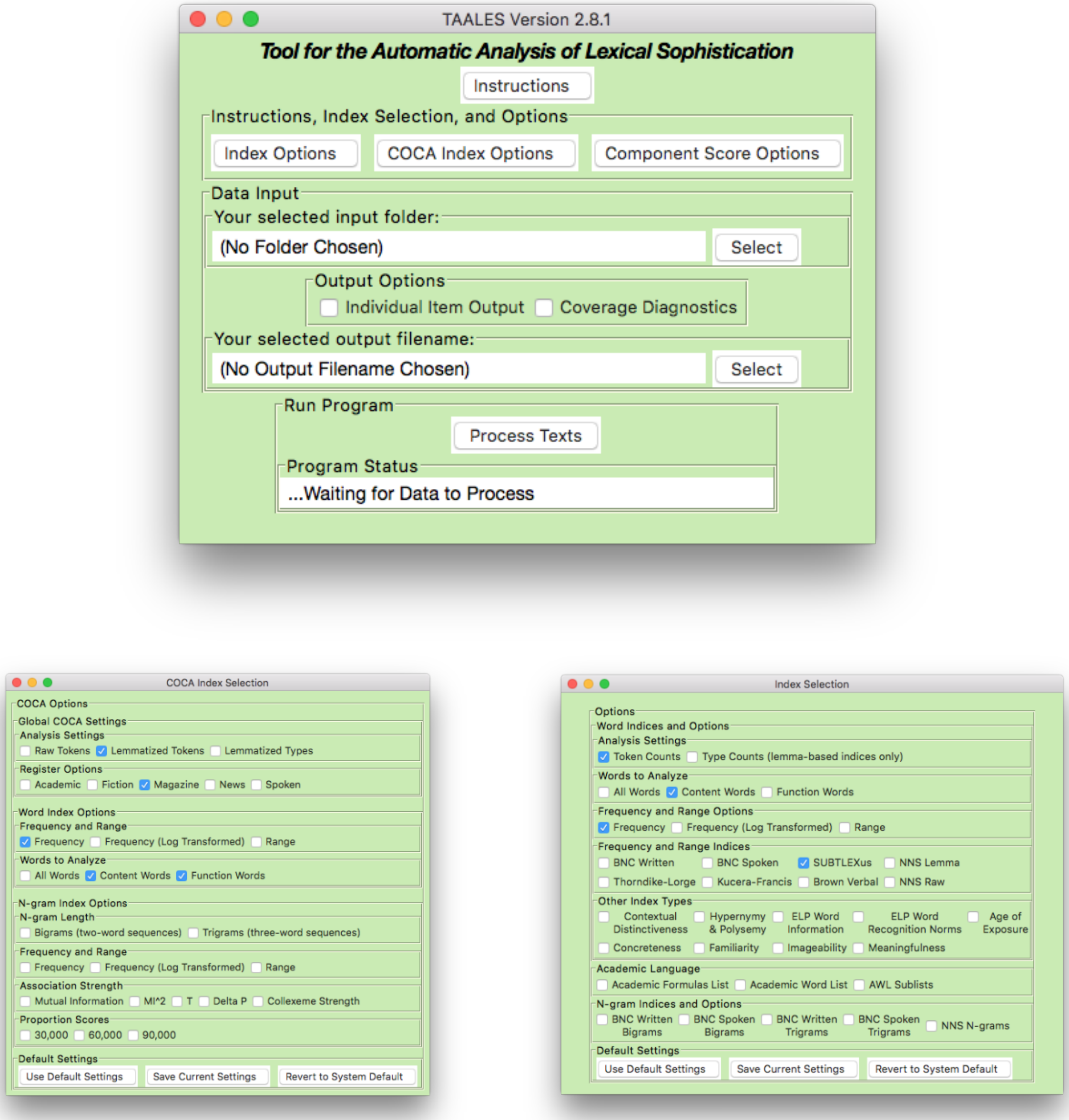
includes both a user manual and an index description guide, and has been widely used in a number of research studies in disciplines ranging from writing assessment (Bestgen, 2017; Kim & Crossley, 2018; Kyle & Crossley, 2016), speech analysis (Hsieh & Wang, 2017; Kyle & Crossley, 2015), creativity and humor (Ravi & Ravi, 2016; Skalicky, Crossley, McNamara, & Muldner, 2016), and text readability (Crossley, Skalicky, Dascalu, McNamara, & Kyle, 2017).

The purpose of the tool is to calculate a wide range of classic and newly developed indices of lexical sophistication. For instance, TAALES calculates indices related to lexical properties at both the word and phrase level (see Table 1 for an overview). These features are accessed using a simple and intuitive graphical user interface and no programming knowledge on the part of the user is required (see Figure 1). Because TAALES is stored on a user's hard drive allowing secure data processing without the need for an Internet connection. Another strength of TAALES is that it provides supplementary word-level output in addition to text-level output. This allows end users to see precisely how each text-level score was calculated.

Table 1
*Overview of TAALES 2.8.1 indices*

| Index type | Indices | Corpora/databases represented | Example indices |
|---|---|---|---|
| Word Frequency | 206 | 11 | Average lemma frequency of content words-academic subcorpus of COCA |
| Word Range | 178 | 8 | Average lemma range of content words-magazine subcorpus of COCA |
| Psycholinguistic Norms | 14 | 2 | Average concreteness score for content words |
| Age of Acquisition/Exposure | 14 | 2 | Average age of exposure score for all words |
| Academic Language | 26 | 2 | Normed academic word list counts |
| Contextual Distinctiveness | 33 | 5 | Average number of elicited word for content words |
| Word Recognition Norms | 24 | 1 | Average word recognition score for content words |

|  |  |  |  |
|---|---|---|---|
| Semantic Network | 14 | 1 | Average polysemy score for nouns and verbs |
| Ngram Frequency | 248 | 6 | Average bigram frequency – newspaper subcorpus of COCA |
| Ngram Range | 76 | 5 | Average trigram range – fiction subcorpus of COCA |
| Ngram Association Strength | 225 | 5 | Average bigram collexeme strength score – spoken subcorpus of COCA |
| Word Neighbor Information | 42 | 1 | Average number of phonological neighbors for content words |
| Other | 3 | 1 | Average character bigram frequency – HAL corpus |
| **Total number of indices:** | 1103 |  |  |



TAALES Version 2.8.1

Tool for the Automatic Analysis of Lexical Sophistication

Instructions, Index Selection, and Options

Data Input

Your selected input folder: (No Folder Chosen)

Output Options

□ Individual Item Output □ Coverage Diagnostics

Your selected output filename: (No Output Filename Chosen)

Run Program

Program Status

...Waiting for Data to Process

COCA Index Selection

COCA Options

Global COCA Settings

Analysis Settings

□ Raw Tokens ☑ Lemmatized Tokens □ Lemmatized Types

Register Options

□ Academic □ Fiction ☑ Magazine □ News □ Spoken

Word Index Options

Frequency and Range

☑ Frequency □ Frequency (Log Transformed) □ Range

Words to Analyze

□ All Words ☑ Content Words ☑ Function Words

N-gram Index Options

N-gram Length

□ Bigrams (two-word sequences) □ Trigrams (three-word sequences)

Frequency and Range

□ Frequency □ Frequency (Log Transformed) □ Range

Association Strength

□ Mutual Information □ MI^2 □ T □ Delta P □ Collexeme Strength

Proportion Scores

□ 30,000 □ 60,000 □ 90,000

Default Settings

Index Selection

Options

Word Indices and Options

Analysis Settings

☑ Token Counts □ Type Counts (lemma-based indices only)

Words to Analyze

□ All Words ☑ Content Words □ Function Words

Frequency and Range Options

☑ Frequency □ Frequency (Log Transformed) □ Range

Frequency and Range Indices

□ BNC Written □ BNC Spoken ☑ SUBTLEXus □ NNS Lemma

□ Thorndike-Lorge □ Kucera-Francis □ Brown Verbal □ NNS Raw

Other Index Types

□ Contextual Distinctiveness □ Hypernymy & Polysemy □ ELP Word Information □ ELP Word Recognition Norms □ Age of Exposure

□ Concreteness □ Familiarity □ Imageability □ Meaningfulness

Academic Language

□ Academic Formulas List □ Academic Word List □ AWL Sublists

N-gram Indices and Options

□ BNC Written Bigrams □ BNC Spoken Bigrams □ BNC Written Trigrams □ BNC Spoken Trigrams □ NNS N-grams

Default Settings

Figure 1. TAALES 2.8.1 graphical user interface

## Learning Objectives and Related Research

Lexical sophistication is generally defined as the production of advanced and difficult words (Laufer & Nation, 1995; Read, 2000). Lexical sophistication partners with, but is separate from lexical diversity, which is a measure of the range of unique words used (Jarvis, 2013). The prototypical measure of lexical sophistication is word frequency (i.e., how infrequent words are;

Laufer & Nation, 1995), but more recent studies have broadened the number of features that inform lexical sophistication. These studies suggest that sophisticated words include those that are less contextually diverse (McDonald & Shillcock, 2001), less concrete, imageable, and familiar (Crossley & Skalicky, in press; Saito, Webb, Trofimovich, & Isaacs, 2016), less specific (Fellbaum, 1998) and those that have fewer orthographical and phonological and neighbors and elicit slower response times in behavioral tasks (Balota et al., 2007).

Lexical sophistication features are strong predictors of writing quality in both first language (L1) and second language (L2) contexts. Lexical sophistication is a subcomponent of language knowledge, which is an important component of writing proficiency (Bachman & Palmer, 1996) in that more proficient writers have greater vocabulary skills that allow them to express ideas more succinctly and clearly (Schoonen, Gelderen, Stoel, Hulstijn, & Glopper, 2011). A number of studies have shown strong links between lexical sophistication features and human judgments of writing quality. The most common finding is that more proficient writers use lower frequency words than less proficient writers (Crossley, Allen, Snow, & McNamara, 2016; S. Crossley & Cai, 2012.; Guo, Crossley, & McNamara, 2013; Kyle & Crossley, 2016; Laufer & Nation, 1995). Beyond simple word frequency, a number of other lexical feature are strong predictors of writing quality including word properties such as age of acquisition scores, concreteness, familiarity, meaningfulness, and imageability (Crossley & McNamara, 2011; Crossley et al., 2016; Guo et al., 2013; Kyle & Crossley, 2016), word range scores (Kyle & Crossley, 2016), and word polysemy and hypernymy (Guo et al., 2013; Kyle & Crossley, 2016; Reynolds, 1995). Importantly, most studies show that lexical and phrasal features are stronger predictors of writing quality scores than other linguistic features such as syntactic complexity or cohesion. For instance, Crossley, Kyle, and McNamara (2015) reported that n-gram features and lexical features were the two strongest predictors of essay quality scores after text length in L1 writing while Crossley et al., (2016) found that the two strongest predictors of L1 essay quality were frequency of spoken bigrams and word concreteness. In terms of L2 writing, Kim & Crossley 2018) reported that the lexical decision reaction times explained the greatest amount of variance in the quality L2 writing samples.

The possibilities afforded by TAALES and by other, similar NLP tools is the ability to analyze writing samples for 100s of lexical features automatically. Since the majority of these lexical features overlap strongly with constructs of writing and show strong evidence of validity (i.e., they predict writing quality or grade level), the features can be used to better understand student writing and the construct of writing proficiency. The insights provided by TAALES help support the importance of lexical features in constructing quality writing samples and provides specific information about the importance of individual lexical features.

In terms of student writing, the features reported by TAALES have been used to inform both automatic essay scoring systems and automatic writing evaluation systems (Crossley, Allen, & McNamara, 2016), providing contributions to automatic text analyses. Such systems provide students with both summative feedback (i.e., overall scores) and formative feedback (i.e., specific feedback on how to improve an essay). The information in these systems can help instructors better understand what features correspond to writing quality at a theoretical level as well as at the student level.

## Connections

TAALES, like most NLP tools, is built upon its predecessors to provide discernable improvements. Specifically, TAALES was inspired by the Coh-Metrix tool (Graesser et al., 2004), which was a tool developed in the early 2000s to measure text cohesion. In addition to text cohesion, Coh-Metrix also reported on a number of lexical features including word frequency, lexical properties, and polysemy and hypernymy scores. TAALES builds on Coh-Metrix in a number of ways. First, its focus is specifically on lexical features and, as such, it expands well beyond the lexical features reported in Coh-Metrix. For instance, it reports on n-gram features, word association scores, academic lists, psycholinguistic features, contextual diversity measures, and range scores. Where TAALES overlaps with Coh-Metrix, we have made efforts to improve the measures. For instance, Coh-Metrix reports on a number of frequency variables, but these variables are based on relatively small and/or old frequency dictionaries. We improved on these features by including larger and more robust frequency dictionaries derived from the British National Corpus (Burnard, 2000) and the Corpus of Contemporary American English (Davies, 2008) or reported by SUBTLEXus (Brysbaert & New, 2009). We also updated word property lists reported for concreteness and age of acquisition (among others).

A primary motivation for developing TAALES was that previous NLP tools were either impractical for large datasets or were not made available to the community. For instance, Coh-Metrix is a freely available online, but its use is limited because the online tool does not allow for batch processing (i.e., each text had to be individually uploaded to the system). This constraint effectively limits the scale of analysis. Second, the on-line version of Coh-Metrix is a pared down version of an internal version available to a core number of Coh-Metrix researchers.

### Limitations and Future Steps

As an automatic text analysis tool, TAALES is limited in scope and in depth. In terms of scope, TAALES focuses specifically on lexical sophistication and does not report on a number of other language features that are important predictors of writing quality including text cohesion, discourse structures, and syntactic complexity. However, TAALES is part of a suite of tools developed by Crossley and Kyle that reported on a variety of larger linguistic and language constructs including syntactic complexity (Kyle, 2016; Kyle & Crossley, 2017), text cohesion (Crossley, Kyle, & McNamara, 2016), and sentiment (Crossley, Kyle, & McNamara, 2017).

Another limitation of TAALES is depth of analysis. While TAALES reports on hundreds of lexical features, it simply reports on incidence counts of these features within a text. It does not distinguish the context in which words are used and whether or not they are used appropriately (e.g., it does not identify discourse structures). Thus, TAALES does not distinguish words that are used in a thesis from those used supporting arguments nor does it distinguish between nouns used as subjects, indirect objects, direct objects, or objects of prepositions.

In addition, if a student produces an infrequent word (e.g., *assent*), TAALES will not distinguish if it was used accurately or inaccurately (perhaps confused with the word *ascent*). This brings up another limitation of TAALES; it will only calculate lexical features for words within its database. So, if a word is misspelled and that misspelled word is not found within TAALES' databases, it will ignored. If a word is misspelled and the misspelled word is in the TAALES database (e.g., *dual* as *duel*), the misspelled word will be calculated.

Another limitation of TAALES is that it is designed to be used by researchers and not teachers or administrators (although the indices reported by TAALES have been used in educational technologies aimed at teachers and administrators). The output of TAALES is a large spreadsheet that contains 100s of numbers for each text. These numbers are generally uninterpretable without inferential statistics that compare, for instance, group differences (e.g., differences between 9th and 11th grade writing samples) or attempt to predict a single variable (e.g., human scores of essay quality).

Another limitation is the breadth of indices report by TAALES. The current version reports over 300 indices and many of these indices are extremely similar (i.e., there are over 150 of frequency variables). Thus, researchers need to be well informed about the assumptions underlying many statistical analyses including normal distributions, overfitting, and multicollinearity. Recent attempts have been made to distill the number of variable in TAALES to a more manageable number (Kim, Crossley, & Kyle, 2018) using statistical analysis that convert related TAALES indices into linearly uncorrelated, aggregated variables. For instance, Kim et al., (2018) found that TAALES variables could be combined into twelve core lexical components. These are now available in the newest version of TAALES.

TAALES is under constant development and version 2.8 is currently available. As new lexical resources become available, they will be added. For instance, a new list of academic words from spoken discourse was recently released (Dang, Coxhead, & Webb, 2017) and will soon be added to TAALES. New indices related to normed frequencies and ranges specific to L2 writing and speaking will also be added to TAALES along with eye-tracking norms for both L1 and L2 readers. Beyond simply expanding the breadth of features reported by TAALES, we plan on adding additional features that will allow users to develop graphic representations of texts and automatically compare texts using machine learning techniques. Additional developments will include text cleaners to remove various problematic formats (i.e., xml formatting) and automatic spell checkers to better represent intended meaning. We also look forward to the tool being integrated into a number of educational technologies that will be used to teach reading and writing skills to in-need student populations.

# References

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests* (1 edition). Oxford ; New York: Oxford University Press.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459.

Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, 65–78.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Burnard, L. (2000). *Reference guide for the British National Corpus (world edition)*. Oxford University Computing Services Oxford.

Crossley, S. A., Allen, L. K., & McNamara, D. S. (2016). The writing pal: A writing strategy tutor. *Adaptive Educational Technologies for Literacy Instruction*. https://doi.org/10.4324/9781315647500

Crossley, S. A., Kyle, K., & McNamara, D. S. (2015). To aggregate or not? Linguistic features in automatic essay scoring and feedback systems. *Journal of Writing Assessment*, *8*(1).

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*, 1227–1237. https://doi.org/10.3758/s13428-015-0651-7

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods*, (49), 803–821.

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life Long Learning*, *21*(2–3), 170–191.

Crossley, S. A., & Skalicky, S. (in press). Examining lexical development in second language learners: An approximate replication of Salsbury, Crossley & McNamara (2011). *Language Teaching*.

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017). Predicting Text Comprehension, Processing, and Familiarity in Adult Readers: New Approaches to Readability Formulas. *Discourse Processes*, *54*(5–6), 340–359. https://doi.org/10.1080/0163853X.2017.1296264

Crossley, S., Allen, L. K., Snow, E. L., & McNamara, D. S. (2016). Incorporating learning characteristics into automatic essay scoring models: What individual differences and linguistic features tell us about writing quality. *JEDM | Journal of Educational Data Mining*, *8*(2), 1–19.

Crossley, S., & Cai, Z. (n.d.). Syntagmatic, Paradigmatic, and Automatic N-gram Approaches to Assessing Essay Quality, 6.

Dang T. N. Y., Coxhead, A., & Webb, S. (2017). The Academic Spoken Word List. *Language Learning*, *67*(4), 959–997. https://doi.org/10.1111/lang.12253

Davies, M. (2008). *The corpus of contemporary American English*. BYE, Brigham Young University.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3758/BF03195564

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*(3), 218–238. https://doi.org/10.1016/j.asw.2013.05.002

Hsieh, C.-N., & Wang, Y. (2017). Speaking proficiency of young language students: A discourse-analytic study. *Language Testing*, 0265532217734240. https://doi.org/10.1177/0265532217734240

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, *63*(s1), 87–106.

Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. *Assessing Writing*, *37*, 39–56. https://doi.org/10.1016/j.asw.2018.03.002

Kim, M., Crossley, S. A., & Kyle K. (2018). Lexical Sophistication as a Multidimensional Phenomenon: Relations to Second Language Lexical Proficiency, Development, and Writing Quality. *The Modern Language Journal*, *102*(1), 120–141. https://doi.org/10.1111/modl.12447

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication*. Georgia State University. Retrieved from http://scholarworks.gsu.edu/alesl_diss/35

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12–24.

Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, *34*(4), 513–535.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757–786.

Kyle, K., Crossley, S., & Berger, C. (2017). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior Research Methods*. https://doi.org/10.3758/s13428-017-0924-4

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322.

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the Word Frequency Effect: The Neglected Role of Distributional Information in Lexical Processing. *Language and Speech*, *44*(3), 295–322. https://doi.org/10.1177/00238309010440030101

Ravi, K., & Ravi, V. (2016). A novel automatic satire and irony detection using ensembled feature selection and data mining. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2016.12.018

Read, J. (2000). *Assessing vocabulary*. Cambridge, MA: Cambridge University Press.

Reynolds, D. W. (1995). Repetition in Nonnative Speaker Writing: More than Quantity. *Studies in Second Language Acquisition*, *17*(2), 185–209. https://doi.org/10.1017/S0272263100014157

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical profiles of comprehensible second language speech: The role of appropriateness, fluency, variation, sophistication, abstractness, and sense relations. *Studies in Second Language Acquisition*, *38*(4), 677–701.

Schoonen, R., Gelderen, A. van, Stoel, R. D., Hulstijn, J., & Glopper, K. de. (2011). Modeling the Development of L1 and EFL Writing Proficiency of Secondary School Students. *Language Learning*, *61*(1), 31–79. https://doi.org/10.1111/j.1467-9922.2010.00590.x

Skalicky, S., Crossley, S. A., McNamara, D. S., & Muldner, K. (2016, July). *Predicting creativity in task-based problem solving using linguistic features*. Poster presented at the 26th Annual Meeting of the Society of Text and Discourse, Kassel, German.

Dr. Scott Crossley is a Professor of Applied Linguistics at Georgia State University. Professor Crossley’s primary research focus is on natural language processing and the application of computational tools and machine learning algorithms in language learning, educational success, writing, and text comprehensibility.

Dr. Kristopher Kyle is an Assistant Professor of Second Language Studies at the University of Hawaii, Manoa. His research interests include second language acquisition, second language writing, corpus linguistics, computational linguistics, and second language assessment.