Assessing writing with the tool for the automatic analysis of lexical sophistication (TAALES)

Scott A. Crossley⁎, Kristopher Kyle

⁎ Corresponding author.

E-mail addresses: sacrossley@gmail.com (S.A. Crossley), kristopherkyle1@gmail.com (K. Kyle).

Available online 20 June 2018

1075-2935/ © 2018 Elsevier Inc. All rights reserved.

ARTICLE INFO

Keywords:
Lexical sophistication
Natural language processing
Automatic writing assessment

ABSTRACT

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.

1. Introduction

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 context and learner independent natural language processing (NLP) tool that provides counts for 100 s 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 available1 and works on a number of operating systems such as Linux, Mac, and Windows. The tool is also user friendly, 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, Kyle, & McNamara, 2017; Crossley, Skalicky, Dascalu, McNamara, & Kyle, 2017).

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.
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.

Table 1
Overview of TAALES 2.8.1 indices.

<table>
<thead>
<tr>
<th>Index type</th>
<th>Indices</th>
<th>Corpora/databases represented</th>
<th>Example indices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Frequency</td>
<td>206</td>
<td>11</td>
<td>Average lemma frequency of content words-academic subcorpus of COCA</td>
</tr>
<tr>
<td>Word Range</td>
<td>178</td>
<td>8</td>
<td>Average lemma range of content words-magazine subcorpus of COCA</td>
</tr>
<tr>
<td>Psycholinguistic Norms</td>
<td>14</td>
<td>2</td>
<td>Average concreteness score for content words</td>
</tr>
<tr>
<td>Age of Acquisition/Exposure</td>
<td>14</td>
<td>2</td>
<td>Average age of exposure score for all words</td>
</tr>
<tr>
<td>Academic Language</td>
<td>26</td>
<td>2</td>
<td>Normed academic word list counts</td>
</tr>
<tr>
<td>Contextual Distinctiveness</td>
<td>33</td>
<td>5</td>
<td>Average number of elicited word for content words</td>
</tr>
<tr>
<td>Word Recognition Norms</td>
<td>24</td>
<td>1</td>
<td>Average word recognition score for content words</td>
</tr>
<tr>
<td>Semantic Network</td>
<td>14</td>
<td>1</td>
<td>Average polysemy score for nouns and verbs</td>
</tr>
<tr>
<td>Ngram Frequency</td>
<td>248</td>
<td>6</td>
<td>Average bigram frequency – newspaper subcorpus of COCA</td>
</tr>
<tr>
<td>Ngram Range</td>
<td>76</td>
<td>5</td>
<td>Average trigram range – fiction subcorpus of COCA</td>
</tr>
<tr>
<td>Ngram Association Strength</td>
<td>225</td>
<td>5</td>
<td>Average bigram collexeme strength score – spoken subcorpus of COCA</td>
</tr>
<tr>
<td>Word Neighbor Information</td>
<td>42</td>
<td>1</td>
<td>Average number of phonological neighbors for content words</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td>1</td>
<td>Average character bigram frequency – HAL corpus</td>
</tr>
<tr>
<td><strong>Total number of indices:</strong></td>
<td><strong>1103</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 Fig. 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.

2. Learning objectives and related research

Lexical sophistication is generally defined as the production of advanced and difficult words (Laufner & 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; Laufner & 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, 2018; Saito, Webb, Trofimovich, & Isaacs, 2016), less specific (Fellbaum, 1998) and those that have fewer orthographical and phonological 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, van Stoel, Hulstijn, & de 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 & Cai, 2018; Crossley, Allen, & McNamara, 2016; Crossley, Kyle, & McNamara, 2016; Crossley, Allen, Snow, & McNamara, 2016; Guo, Crossley, & McNamara, 2013; Kyle & Crossley, 2016; Laufner & 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, Allen, McNamara, 2012; Crossley, Kyle et al., 2016; Crossley, Allen, Snow et al., 2016; Guo et al., 2013; Kyle & Crossley, 2016), word range scores (Kyle & Crossley, 2016), and word polysemy and hyponymy (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 n-gram features of lexical features were the two strongest predictors of essay quality scores after text length in L1 writing while Crossley, Allen, McNamara (2016), Crossley, Kyle et al. (2016), Crossley, Allen, Snow 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 100 s 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; Crossley, Kyle et al., 2016; Crossley, Allen, Snow et al., 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.

3. 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 2000’s 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 hyponymy 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.
4. Limitations and future steps

As an automatic text analysis tool, TAALES is limited in scope and in depth. In terms of scope, TAAALES 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, TAAALES 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 & Crossley, 2017; Kyle, 2016), text cohesion (Crossley, Allen, McNamara, 2016; Crossley, Kyle et al., 2016; Crossley, Allen, Snow et al., 2016), and sentiment (Crossley, Kyle et al., 2017; Crossley, Skalkicky et al., 2017).

Another limitation of TAALES is depth of analysis. While TAAALES 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, TAAALES 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), TAAALES will not distinguish if it was used accurately or inaccurately (perhaps confused with the word ascent). This brings up another limitation of TAAALES; 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 TAAALES’ 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 TAAALES is that it is designed to be used by researchers and not teachers or administrators (although the indices reported by TAAALES have been used in educational technologies aimed at teachers and administrators). The output of TAAALES is a large spreadsheet that contains 100 s 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 TAAALES. 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 TAAALES. New indices related to normed frequencies and ranges specific to L2 writing and speaking will also be added to TAAALES along with eye-tracking norms for both L1 and L2 readers. Beyond simply expanding the breadth of features reported by TAAALES, 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


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.