Tool for the Automatic Analysis of Lexical Sophistication (TAALES)

User Manual for TAALES 2.0 (updated 11-2-2016)

Kristopher Kyle and Scott Crossley
Georgia State University

This document is intended to assist users of TAALES. It includes a brief explanation of how to use the tool. Additional information about TAALES is included in the supplementary Index Description Spreadsheet (available at www.kristopherkyle.com).

Please use the following citation when referencing TAALES in your work:

Kyle, K. & Crossley, S.A. (in press). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly.* doi: 10.1002/tesq.194

Getting Started

First, download the version of TAALES that is appropriate for your operating system. Second, download and install the version of the Java Development Kit (JDK) that is appropriate for your operating system. Note that this is NOT the same version of Java that is normally installed on one's computer. JDK is used when calculating Hypernymy and Polysemy indices (which are tagged by a part of speech tagger). It is possible to run TAALES 2.0 without installing JDK, but if the user selects the Hypernymy and Polysemy indices without first installing JDK, TAALES will crash.

Indices and Options

TAALES 2.0 calculates 484 indices of lexical sophistication in a number of categories. All indices in TAALES 2.0 are normed. Please see the supplementary Index Description Spreadsheet for more information. Also, see the article referenced above for more information.

The user may choose to include indices from any (or all) of a number of categories.

Note that the hypernymy and polysemy indices are pre-processed using a part of speech tagger. These indices may be less accurate when applied to transcribed spoken texts.

Included index categories are (please reference the Index Description Spreadsheet for detailed information about each index and index type):

Academic Language
Academic Formulas List (AFL)
Academic Word List (AWL)
Academic Word List (AWL) Sublists

COCA Indices
COCA Word Frequency and Range for five registers
COCA Bigram Frequency, Range, and Association Strength for five registers
COCA Trigram Frequency, Range, and Association Strength for five registers

Frequency and Range Indices (from sources other than COCA)
BNC Word Frequencies
BNC Ngram Frequencies
MRC Frequencies (includes range for some lists)
SUBTLEXus Frequencies (includes range)

Other Index Types
Age of Exposure (AOE)
Contextual Distinctiveness
ELP Word Information
ELP Response Time Norms
Hypernymy and Polysemy
Psycholinguistic Word Information Norms

Input

All input files must be text files (.txt) that do not include any type of markup (e.g., XML, HTML, etc.). Files must be located in a single folder. TAALES will process all .txt files in the chosen input folder.

Saving Your Output

TAALES provides output in the form of a comma-separated (.csv) file that can be opened with any spreadsheet software. The default output file name is “results.csv”, though we would recommend changing this file name each time you run TAALES to ensure that the file is not overwritten.

Diagnostics

Index Coverage Diagnostics

TAALES also provides index coverage diagnostics, which (where appropriate) indicates the percent of words in a target text that are represented in each database. This data allows the user to make informed choices when selecting between similar indices (e.g., Kucera-Francis written frequencies and COCA written frequencies). Index coverage data is saved in the same folder as the normal output file, and includes the user-generated output filename with “_index_coverage appended (e.g., “results_index_coverage.csv”). Note that LSA and AOE indices will indicate lower than expected text coverage. This is due to the fact that LSA and AOE calculations undergo dimension reduction, which results in the preclusion of high-frequency words.

Individual Item Diagnostics

TAALES also provides the ability to obtain the index score for each item (word, bigram, trigram) in a text, and to determine which items are included in the associated database. Any item not included in the associated database is listed as “n/a”.

To obtain individual item diagnostics, click the “Include Individual Item Output?” box, which is located under the “Select Output Filename” button. During file processing, TAALES will create a file folder with the same name as your output filename. TAALES will then create a tab-delimited text file for each input file (with the same name). Within this file, indices are listed in alphabetical order. To find a particular index, one can either scroll down the file OR search for the particular index name. Because these files can become unwieldy, it is suggested that individual item analysis is conducted only with a small number of target indices.

Currently AFL, EAT, and USF indices are the only index types not supported.