Using Learner Corpora to Study SLA

SLRF 2019 Pre-Conference Workshop

September 19th, 2019

Kristopher Kyle

Department of Linguistics

Colleagues involved in TAALES, TAALED, and TAASSC projects

Scott Crossley (GSU)

Cindy Berger (DuoLingo)

Masaki Eguchi (UO)

Scott Jarvis (UU)

Minkyung Kim (GSU)

Katia Monteiro (GSU)

Danielle McNamara (ASU)

Fred Zenker (UHM)

And MANY Others!

Workshop outline

- Brief discussion of research in SLA and LCR
- Introduction to LCR
- Introduction to complexity analysis
- Work with some data!

SLA and LCR

Second language acquisition (SLA) and learner corpus research (LCR):

Related in that both fields have been interested in topics such as linguistic development

Distinct

- SLA grew out of educational psychology (among other fields)
- SLA has had a focus on hypothesis testing through controlled studies (with typically small n-sizes)
- LCR grew out of corpus linguistics
- LCR has tended to use large data sets with less control
- LCR has often focused on describing language use by larger populations

SLA and LCR

Second language acquisition (SLA) and learner corpus research (LCR):

Are **converging**

- SLA researchers are beginning to use larger, pre-collected datasets to help address wider research questions
- LCR research is becoming more informed by SLA research
- Learner corpus compilers are now controlling for a variety of learner variables, enabling questions in SLA to be addressed more clearly

Edited volumes for further reading:

Second Language Acquisition and Learner Corpora (Paquot & LeBruyn, 2019)

Routledge Handbook of SLA & Corpora Tracy-Ventura & Paquot (Spring 2020)

Learner Corpus Research

Learner corpora consist of generally large collections of learner productions. Learner corpora can be categorized based on

- L2 (a wide variety, but still predominately **English**)
- L1 (a wide variety)
- Mode (spoken, ***written***, online chat)
- Natural vs. Structured (free conversation, ***prompt-based essays***, monologues, etc.)
- longitudinal vs. ***cross-sectional***
- Available metadata (age of onset, proficiency scores, etc.)

The CECL website has an excellent, fairly up to date list of available learner corpora!

Learner Corpus Research

An edited volume entitled *The Cambridge Handbook of Learner Corpus Research* (Granger, Gilquin, & Meunier, 2015) is an excellent resource for getting acquainted with areas of interest and critical issues in the field.

Analyzing Linguistic Complexity In Learner Corpora

Learner corpora are unwieldy without automated analysis techniques.

Learner corpus research is mostly confined by these techniques

But, many advancements have been made in the last 10 years (yay, Silicon Valley!)

- Part of speech tagging
- Syntactic annotation
- Sentiment analysis
- Voice to text
- etc.

Reminder of workshop website

Data, links, and index descriptions are available on the workshop website!

Lexical Richness and Syntactic Complexity

Lexical sophistication (TAALES, AntWordProfiler)

- frequency
- concreteness
- etc.

Lexical diversity (TAALED, CLAN)

- type-token ratio
- length-independent indices (voc-D/HD-D, MTLD, MATTR)

Syntactic complexity (TAASSC, L2SCA, CLAN)

- Clausal subordination
- Phrasal elaboration
- Frequency, contingency, and salience of verb argument constructions (VACs)

Analysis steps

1. Research questions (*e.g., what is the relationship between receptive vocabulary size and lexical use?)*
2. Choose/collect appropriate corpus (e.g., *ICNALE corpus*)
3. Refine research questions as needed
4. Use analysis tools to annotate data/calculate indices (*e.g, TAALES*)
5. Check accuracy, etc. Visualize relationships (*e.g., using R, Excel, etc.*)
6. Conduct statistical analyses (*e.g., using R, Excel, etc.*)
7. Conduct fine-grained follow up analyses (*e.g., using fine-grained output in TAALES*)

8. Interpret results

Workshop activity: Receptive vocabulary and productive linguistic complexity

1. Choose a complexity index of interest (see workshop website)
2. Visualize the relationship between VST score and that index
3. Calculate a correlation between VST score and that index
4. Choose candidates for a fine-grained (text-level) analysis
5. Conduct fine-grain analysis
6. Interpret the results of your analysis based on the correlations and your fine-grained analysis

Workshop activity: Receptive vocabulary and productive linguistic complexity

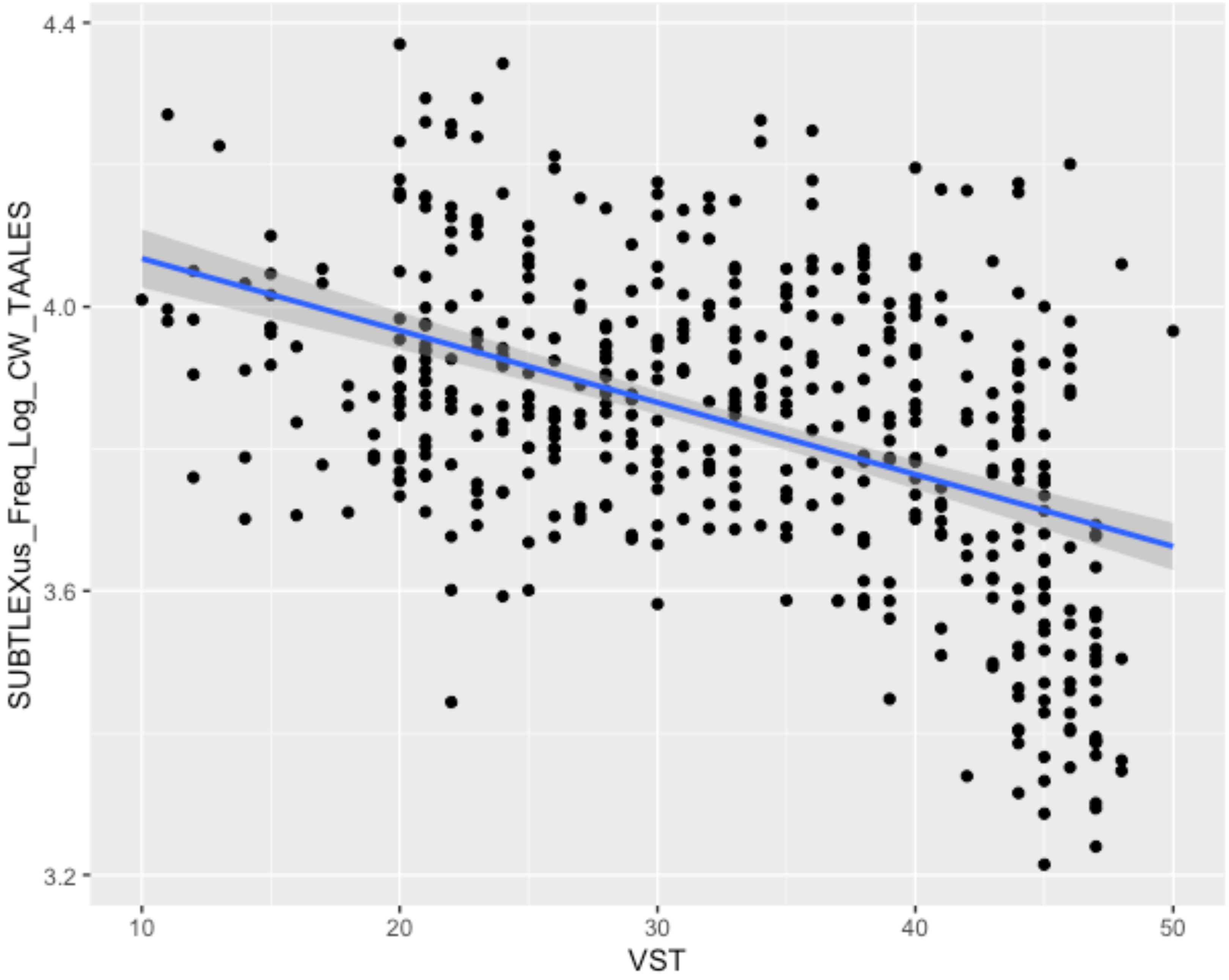
1. Choose a complexity index of interest (***SUBTLEXus Frequency***)
2. Visualize the relationship between VST score and that index (***Use R, Excel, etc.***)
3. Calculate a correlation between VST score and that index (***Use R, Excel, etc.***)
4. Choose candidates for a fine-grained (text-level) analysis
5. Conduct fine-grained analysis
6. Interpret the results of your analysis based on the correlations and your fine-grained analysis

1. Choose a complexity index of interest

- We will first look at the index ***SUBTLEXus_Freq_Log_CW***
- This index represents the mean frequency score for content words based on the SUBTLEXus corpus (a corpus of movie and sitcom subtitles)
- Content words include lexical verbs, nouns, adjectives, and most adverbs
- Frequency scores are logarithmically transformed to help account for the Zipfian nature of frequency distributions

2. Visualize the relationship between VST score and that index

- Note that R scripts are available on the workshop website
- I will demonstrate how to create a scatterplot in Excel (though it is much easier in R!)


4.4
4.0
3.6
3.2
SUBTLEXus_Freq_Log_CW_TAALES
10
20
30
40
50
VST

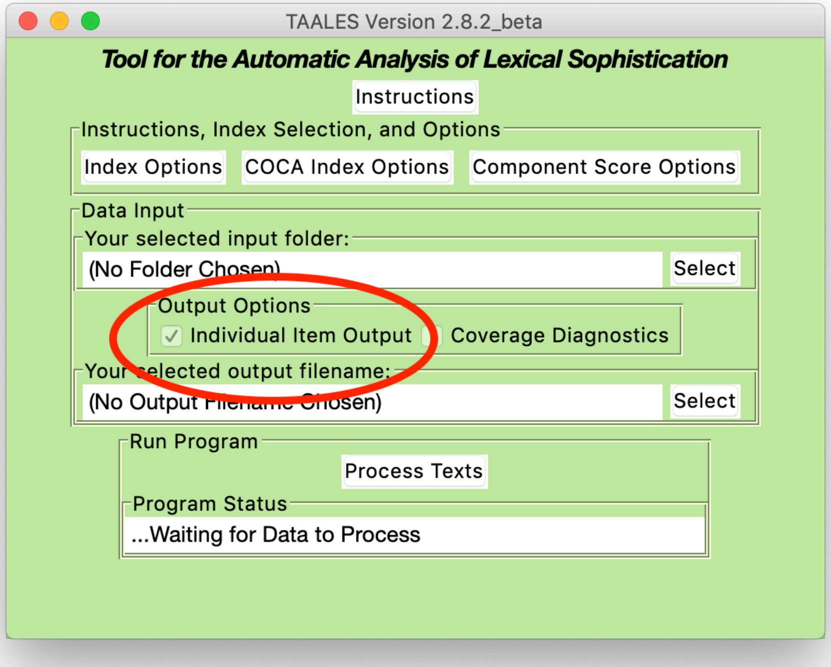
3. Calculate a correlation between VST score and that index (*Use R, Excel, etc.*)

- Note that R scripts are available on the workshop website
- I will demonstrate how to calculate a correlation in Excel (though it is easier in R!)

4. Choose candidates for a fine-grained (text-level) analysis

- Lets do some sorting in Excel…

5. Conduct fine-grained analysis



TAALES Version 2.8.2_beta

Tool for the Automatic Analysis of Lexical Sophistication

Instructions

Instructions, Index Selection, and Options

Index Options | COCA Index Options | Component Score Options

Data Input

Your selected input folder:

(No Folder Chosen) Select

Output Options

☑ Individual Item Output □ Coverage Diagnostics

Your selected output filename:

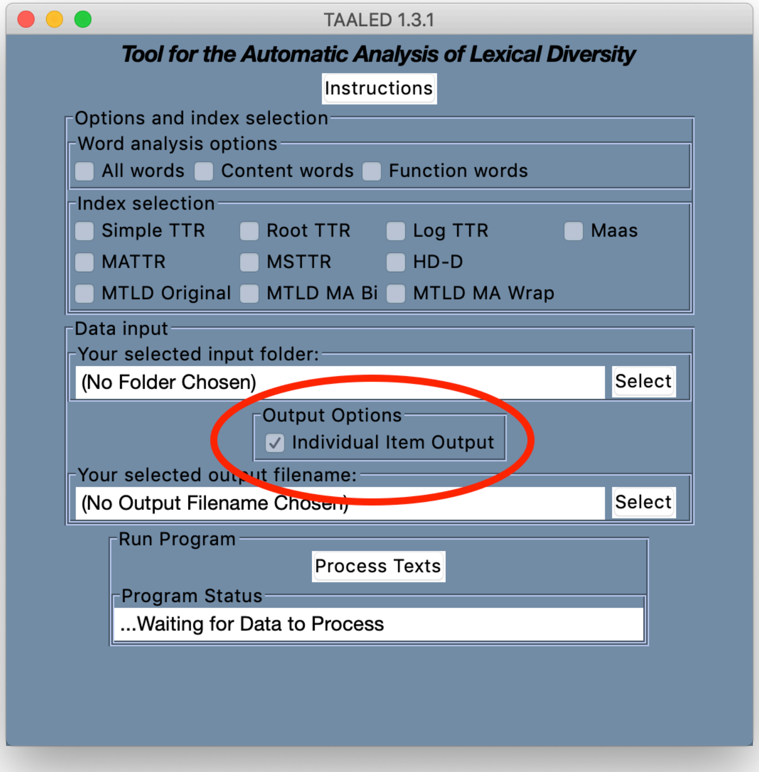
(No Output Filename Chosen) Select

Run Program

Process Texts

Program Status

...Waiting for Data to Process



TAALED 1.3.1

Tool for the Automatic Analysis of Lexical Diversity

Instructions

Options and index selection

Word analysis options

□ All words □ Content words □ Function words

Index selection

□ Simple TTR □ Root TTR □ Log TTR □ Maas

□ MATTR □ MSTTR □ HD-D

□ MTLD Original □ MTLD MA Bi □ MTLD MA Wrap

Data input

Your selected input folder:

(No Folder Chosen) Select

Output Options

☑ Individual Item Output

Your selected output filename:

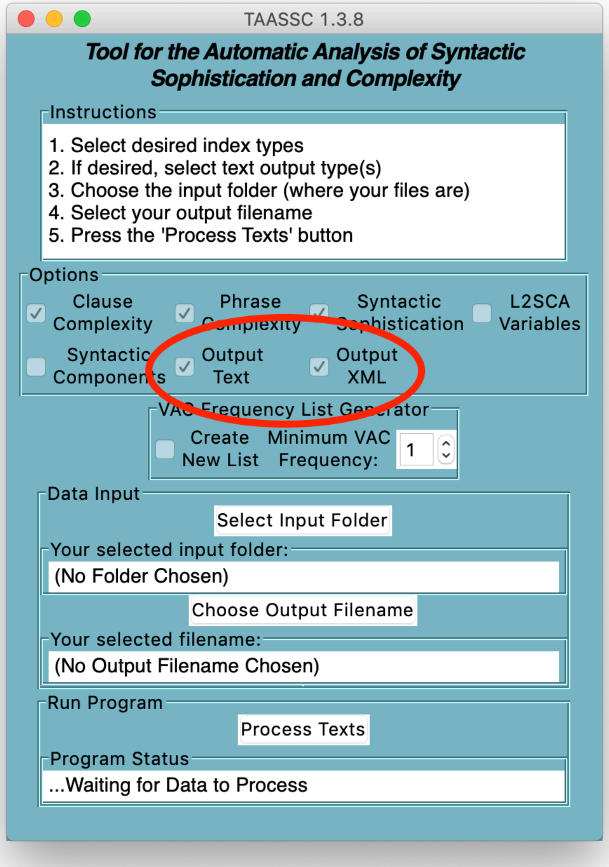
(No Output Filename Chosen) Select

Run Program

Process Texts

Program Status

...Waiting for Data to Process



TAASSC 1.3.8

Tool for the Automatic Analysis of Syntactic Sophistication and Complexity

Instructions

1. Select desired index types
2. If desired, select text output type(s)
3. Choose the input folder (where your files are)
4. Select your output filename
5. Press the 'Process Texts' button

Options

☑ Clause Complexity ☑ Phrase Complexity ☑ Syntactic Sophistication □ L2SCA Variables

□ Syntactic Components ☑ Output Text ☑ Output XML

VAC Frequency List Generator

□ Create New List Minimum VAC Frequency: 1

Data Input

Select Input Folder

Your selected input folder:

(No Folder Chosen)

Choose Output Filename

Your selected filename:

(No Output Filename Chosen)

Run Program

Process Texts

Program Status

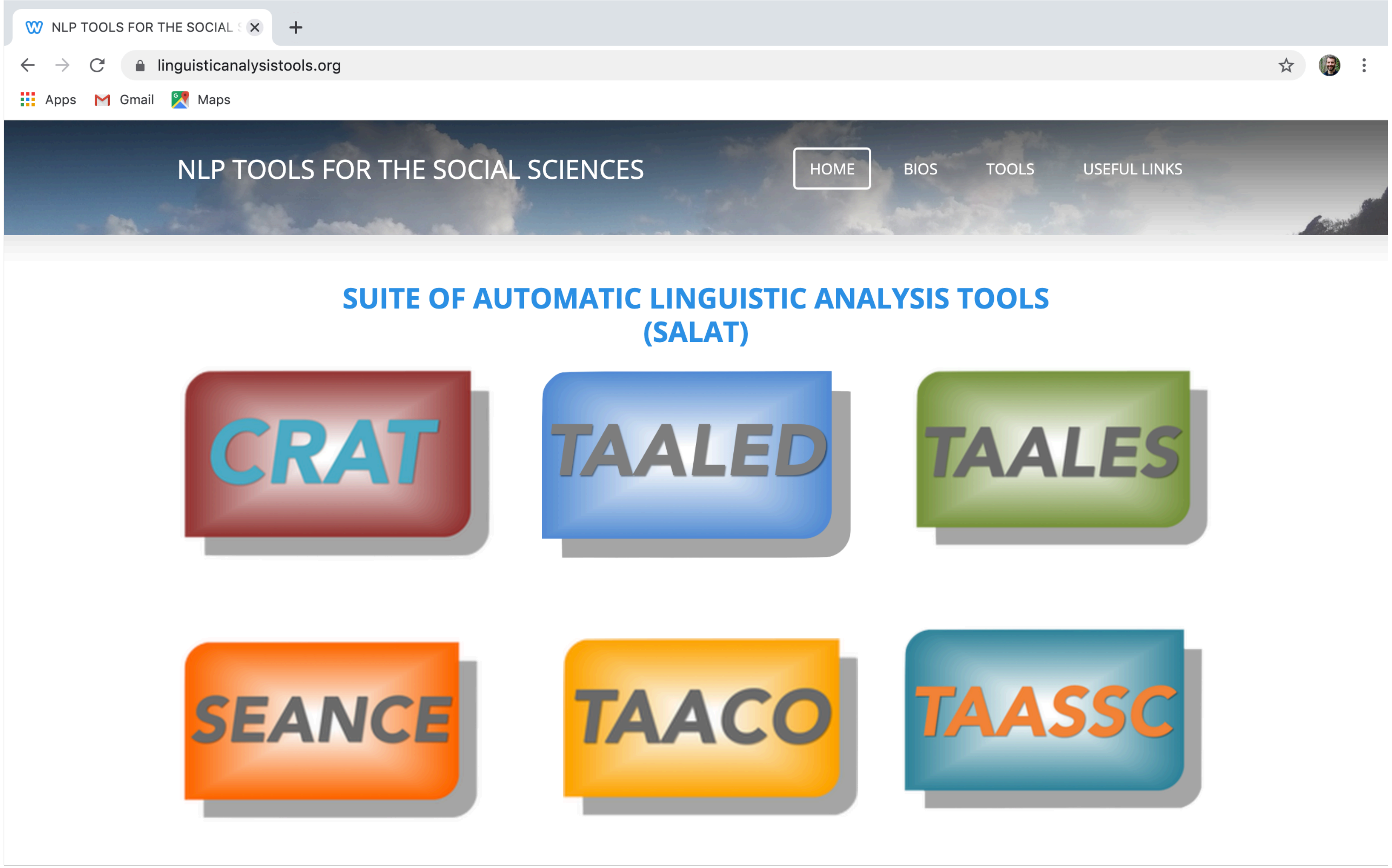
...Waiting for Data to Process

6. Interpret the results of your analysis based on the correlations and your fine-grained analysis

- What general relationships did we find?
- What are some low-frequency content words that are used by individuals with larger receptive vocabularies, but not those with smaller receptive vocabularies?

Now, choose another index, and repeat!

Thanks!



SUITE OF AUTOMATIC LINGUISTIC ANALYSIS TOOLS (SALAT)