Assessment as Ethical and Diagnostic Practice

by Fredrik deBoer

Academic Assessment Manager
Brooklyn College
City University of New York
Learning Goals

Theory
• For attendees to understand why we might want to use text analysis on student artifacts gathered in the assessment process

Application
• For attendees to learn the names of several programs that can be used in text analysis
Let’s Look at a Student Artifact

• Please take a few moments to read a very brief piece of student writing that might be used in a freshman writing seminar.

• Rate that response, to the best of your ability, from 1 (lowest rating) to 6 (highest rating).
Going Beyond Good and Bad

• Due to the practical consequences of our work, assessment literature can fixate on simplistic success/failure binary

• There are ample opportunities to explore assessment data for deeper, more generative information without expending significant resources
Assessment Creates Data

• When we assess, we are frequently gathering a large amount of data
• Because this data gathering tends to be resource-intensive, we should strive to make it as useful as possible
• Faculty often define themselves as researchers and may be more inclined to help if assessment is defined as research
• Publication is a possibility
Uniting the Three Major Tasks of the Academic Life
Sort Out IRB First

- Often, this type of investigation will prove to be IRB exempt
- Your IRB will have to grant that exemption
- “Know Before You Go”
An End User, Not a Developer

- I use these corpus/computational linguistics software as someone who wants to learn about students and their data, not to understand fundamental properties of language or to develop algorithms.
- I know how to use these and understand their overall mechanisms, not their specific algorithms.
The Best Way to Learn is to Play

• When getting familiar with these programs, there is no substitute for just trying things out and seeing what happens
• I promise you won’t break your computer
Tool #1: Concordancers

- “A concordance is an alphabetical list of the principal words used in a book or body of work, listing every instance of each word with its immediate context.” – Wikipedia
- Concordancers generate concordances quickly and efficiently.
ANTCONC

• A free, stable, full-featured concordance
• Pros: Great tutorial videos, solid program
• Cons: Concordancers are limited in what they can do
Example: Validating Directed Self Placement

Local Assessment: Using Genre Analysis to Validate Directed Self-Placement

Grounded in the principle that writing assessment should be locally developed and controlled, this article describes a study that contextualizes and validates the decisions that students make in the modified Directed Self-Placement (DSP) process used at the University of Michigan. The authors present results of a detailed text analysis of students' DSP essays, showing key differences between the writing of students who self-selected into a mainstream first-year writing course and that of students who self-selected into a preparatory course. Using both rhetorical move analysis and corpus-based text analysis, the examination provides information that can, in addition to validating student decisions, equip students with a rhetorically reflexive awareness of genre and offer an alternative to externally imposed writing assessment.
University of Michigan researchers used a concordancer to look for differences between two groups of students sorted into different outcomes by their directed self placement system.
Tool #2: Lexical Diversity Measures

- “A concordance is an alphabetical list of the principal words used in a book or body of work, listing every instance of each word with its immediate context.” – Wikipedia
- Concordancers generate concordances quickly and efficiently.
CLAN (Computerized Language Analysis) by the CHILDES Project

- A full-featured analysis software suite developed originally for analyzing the speech of children
- Pros: Polish software with a comprehensive manual and tutorial videos
- Cons: Uses a command line interface
Example: Ensuring Fairness in Writing Assessment of Second Language Writers
Does Limited Vocabulary Disproportionately Impact Second Language Writers?

• A persistent question in higher education, particularly for diverse institutions like Brooklyn College, is whether different populations of students are receiving equal educational treatment.

• Assessment data can help answer such questions.
Vocabulary Diversity and Essay Ratings

• The relationship between the range of vocabulary used in standardized essays and the ratings those essays receive is a good example.

• Lexical diversity = the displayed range of diversity in vocabulary in a given language sample, defined according to the relationship between types of words and total number of words.

• Several competing metrics and methods of evaluation exist.
An Analysis of 50 Real Student Essays

- 50 real standardized essays – 25 by Chinese learners of English, 25 by native English writers – produced as part of internal research within Purdue’s 106i (Introductory Composition for international students) program
- Comparing ratings to lexical diversity to find relationships between displayed vocabulary and quality
vocd, a Popular LD Metric

- Simple indices of lexical diversity, such as Number of Different Words (NDW) and Type-to-Token Ratio (TTR), have been deprecated as unstable over varying text lengths.
- A popular algorithmic approach is vocd, developed by David Malvern and Brian Richards and implemented in many popular computational linguistics application.
Effective Assessment of Lexical Diversity Remains Controversial

Evaluating the comparability of two measures of lexical diversity

Fredrik deBoer*

*Purdue University, United States
A Restricted Range

• Essays are scored on a 10-60 point scale by two raters, discrepant scores adjudicated by a third rater, with final ratings the average of two ratings

• Owing to the fact that the test takers were English-language college students, the lack of any 10s, the combined average rating of 41.56, and combined median rating of 40 are to be expected
Descriptive Statistics

Chinese L2 Writers
• $n = 25$
• Mean Essay Rating: 35.4
• Average vocab: 59.71

English L1 Writers
• $n = 25$
• Mean Essay Rating: 47.5
• Average vocab: 68.74
# Chinese Writers: A Moderate, Significant Correlation

\[ r = 0.498 \]

\[ 0.01 < p < 0.05 \]

## Correlations

<table>
<thead>
<tr>
<th></th>
<th>Types</th>
<th>Tokens</th>
<th>TTR</th>
<th>Deptimum</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spearman's rho</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tokens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.907</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>25</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>-.373</td>
<td>-.539</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deptimum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.800</td>
<td>.446</td>
<td>.401</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rating</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.950</td>
<td>.949</td>
<td>.452</td>
<td>.490</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).**  
*. Correlation is significant at the 0.05 level (2-tailed).
English Writers: No Similar Correlation

\[ \rho = .357 \]
\[ p > .05 \]
Combined: Moderate and Significant

\[ \rho = .468 \]

\[ p < .01 \]

<table>
<thead>
<tr>
<th></th>
<th>Types</th>
<th>Tokens</th>
<th>TTR</th>
<th>Deoptimum</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Spearman's rho</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Types</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Tokens</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation Coefficient</td>
<td>.952**</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td></td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>50</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TTR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation Coefficient</td>
<td>-.291</td>
<td>-.510**</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td></td>
<td>0.041</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>50</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Deoptimum</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation Coefficient</td>
<td>.058</td>
<td>.483**</td>
<td>.353</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td></td>
<td>0.000</td>
<td>.012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Rating</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation Coefficient</td>
<td>.039**</td>
<td>.352**</td>
<td>-.345</td>
<td>.468**</td>
<td>1.000</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td></td>
<td>0.000</td>
<td>.014</td>
<td>.001</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td></td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).
Combined Data Set Scatterplot
Indicative Example:
Low LD, Low Score

vocd = 41.382
score = 20/60

From file `<file path>`
every people need to know news in the world
some people like to read newspaper but some people like to watch television or listen radio
some people like to read the newspaper because they can know each different kinds of news
they can spent all day read different kind of newspaper
the reader and newspaper that can deep the expression of news
some people like to watch television or listen radio that can gains news before they read newspaper
the people think the news that have time limite
if you can gain news before others people i know
the former can gain some kinds of benefit
for example one day the war happen
you can gain news from television or radio so you can buy some kinds of material
this method may be let you gain a large or money
if you just read newspaper your news is gained lower others people
you may be have a large of demangement
i think that we need to watch television and listen radio at the same time we need to read the newspaper
such as we may be not lost the sources of news

<table>
<thead>
<tr>
<th>tokens</th>
<th>samples</th>
<th>ttr</th>
<th>st.dev</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>100</td>
<td>0.7706</td>
<td>0.065</td>
<td>45.291</td>
</tr>
<tr>
<td>36</td>
<td>100</td>
<td>0.7483</td>
<td>0.058</td>
<td>40.053</td>
</tr>
<tr>
<td>37</td>
<td>100</td>
<td>0.7449</td>
<td>0.063</td>
<td>40.231</td>
</tr>
<tr>
<td>38</td>
<td>100</td>
<td>0.7405</td>
<td>0.062</td>
<td>40.155</td>
</tr>
<tr>
<td>39</td>
<td>100</td>
<td>0.7528</td>
<td>0.058</td>
<td>44.710</td>
</tr>
<tr>
<td>40</td>
<td>100</td>
<td>0.7502</td>
<td>0.060</td>
<td>45.075</td>
</tr>
<tr>
<td>41</td>
<td>100</td>
<td>0.7332</td>
<td>0.061</td>
<td>41.298</td>
</tr>
<tr>
<td>42</td>
<td>100</td>
<td>0.7460</td>
<td>0.055</td>
<td>45.997</td>
</tr>
<tr>
<td>43</td>
<td>100</td>
<td>0.7237</td>
<td>0.056</td>
<td>40.760</td>
</tr>
<tr>
<td>44</td>
<td>100</td>
<td>0.7130</td>
<td>0.059</td>
<td>39.958</td>
</tr>
<tr>
<td>45</td>
<td>100</td>
<td>0.7118</td>
<td>0.056</td>
<td>39.550</td>
</tr>
<tr>
<td>46</td>
<td>100</td>
<td>0.7172</td>
<td>0.052</td>
<td>41.827</td>
</tr>
<tr>
<td>47</td>
<td>100</td>
<td>0.7123</td>
<td>0.051</td>
<td>41.454</td>
</tr>
<tr>
<td>48</td>
<td>100</td>
<td>0.6954</td>
<td>0.046</td>
<td>38.166</td>
</tr>
<tr>
<td>49</td>
<td>100</td>
<td>0.6996</td>
<td>0.054</td>
<td>39.866</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>0.6912</td>
<td>0.057</td>
<td>38.879</td>
</tr>
</tbody>
</table>

D: average = 41.382; std dev. = 2.446
D_optimum <41.24; min least sq val = 0.001
Indicative Example:
High LD, High Score

vocd = 97.912
score = 60/60
Outlier Example:
High LD, Below Average Score

vocd = 92.582
score = 35/60
Outlier Example: Low LD, Above Average Score

vocd = 38.59
score = 50/60
OK, what do we do with it?

- If students in a program (or certain subpopulations) demonstrate an inability to utilize vocabulary, and this contributes to difficulty on standardized writing assessments, techniques for effectively introducing vocabulary into the writing classroom might be discussed.
- This would of course be implemented voluntarily, through brown bags, in-services, emailing of relevant literature, etc.
Tool #3: Contrastive Corpus Analysis

- Generates indicative lexical features of human-tagged large data sets, then tests whether students' texts are more or less like one group or the other
- Identifies statistically improbable n-grams
- Differentials: lexical features (words and phrases) that are common to one tested corpus but not common to the other (hence contrastive) after commonalities have been algorithmically eliminated
The Gramulator

- Comprehensive freeware text analysis tool
- Developed by Philip McCarthy, the University of Memphis and the Institute for Intelligent Systems
Example: Ensuring Efficient Use of Scarce Institutional Resources
Do Natively Bilingual Indian Students Resemble First or Second Language Students in Their Writing?

• Many of Purdue’s students from India report a first language other than English (Hindi, Gujarati, Tamil, Bengali, etc.) but identify English as a native second language

• Administrators have questioned whether placing such students in our designated second language sections of freshman composition

• Contrastive corpus analysis and Genre Purity Analysis could be used to determine the “nativeness” of these writers
Contrasting L1 and L2

• A corpus of 167 essays written by L1 writers was contrasted with a corpus of 102 L2 writers using the Gramulator.

• Differentials were generated, and sufficient statistically improbable n-grams were discovered in order to define L1 and L2 genre features.

• L2 genre had more distinct n-gram features than L1.
Failed Attempt to Distinguish Natively Bilingual Indian Student Writing

• Using the Genre Purity Assessment Tool (GPAT) in the Gramulator, texts can be analyzed to predict their fit, based on differentials, into previously analyzed corpora

• 27 texts by natively bilingual Indian students were analyzed

• Unfortunately, these texts did not produce significant differentials to be accurately placed into either genre

• Limited sample size is a likely contributing factor
The Possibilities Are Endless

• We have data. We have expertise. We only need to ask interesting questions and look for answers.
Thanks!
Questions?

Fredrik deBoer
Academic Assessment Manager & WAC Coordinator
Brooklyn College
City University of New York
fredrik.deboer@brooklyn.cuny.edu