Meaningful levels of analysis in (corpus) linguistics

Jesse Egbert
Northern Arizona University
Introduction

• Corpus linguistics = computers for linguistics
• Computational linguistics = linguistics for computers
Online language

“[...] our predictions may be more prone to failure in the era of Big Data. As there is an exponential increase in the amount of available information, there is likewise an exponential increase in the number of hypotheses to investigate. [...] there isn’t any more truth in the world than there was before the internet or the printing press. Most of the data is just noise, as most of the universe is filled with empty space.”

Nate Silver
The text

• Definition: a written or spoken unit of discourse that is:
  • Naturally occurring
  • Recognizably self-contained
  • Functional

• The text is the ideal unit of observation for corpus linguistic research.
  1. Fundamental unit of discourse
  2. Important social construct
  3. Situational and linguistic integrity

  Egbert, forthcoming; Biber & Conrad (2009)
Language sample (corpus)

Texts

Linguistic characteristics
Levels of analysis

• Levels of analysis within texts (i.e. leaves)
  • Discourse
  • Syntax
  • Lexico-grammar
  • Phraseology
  • Lexis
  • Morphology
  • Phonology
What meaningful levels of analysis exist between the text and the corpus?
Levels of analysis

• Levels of analysis *containing* texts (i.e. species)
  • Defined by *user*
    • Geographic region
    • Socioeconomic status
    • Gender
    • Age
    • Race
  • Defined by *use*
    • Register
Register

• Definition: Varieties of language defined by their situation of use (Biber & Conrad, 2009)

• Functional link between situation and language (Egbert & Biber, 2017)

• Valid social construct (Egbert, Biber & Davies, 2015)

• Strong(est?) predictor of linguistic variation (Biber, 2012)
Register—functionally interpretable

• Functional link between situation and language

Adapted from Biber & Conrad (2009)
Register and probability

“Register variation can in fact be defined as systematic variation in probabilities”

Halliday (1991)

• Language varies across registers at every linguistic level

• Probabilities based on “general” language are inaccurate
CORE: Corpus of Online Registers of English

• Large corpus of English web documents
  • ~50,000 documents
  • ~50 million words

• Random sample from the searchable web

• Situational characteristics coded by non-experts
  • 8 register categories
    • At least 3-way agreement: 69.2%
  • 33 sub-register categories
    • At least 3-way agreement: 51.4%

Biber & Egbert (in press)
Register—strong predictor of variation

- 1\textsuperscript{st} person pronouns across registers
Register—strong predictor of variation

• 1st person pronouns across sub-registers
## Blog registers

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Personal Blog</th>
<th>Travel Blog</th>
<th>Religious Blog</th>
<th>Opinion Blog</th>
<th>Info. Blog</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purpose</td>
<td>Narrative</td>
<td>Narrative/Description</td>
<td>Opinion</td>
<td>Opinion</td>
<td>Description</td>
</tr>
<tr>
<td>Subject</td>
<td>Author’s life</td>
<td>Travel</td>
<td>Religion</td>
<td>Author’s stance</td>
<td>Topic to be explained</td>
</tr>
<tr>
<td>Audience</td>
<td>Friends/Family/Followers</td>
<td>Travelers</td>
<td>Religious adherents</td>
<td>Various</td>
<td>Students/Non-experts</td>
</tr>
</tbody>
</table>
Register—strong predictor of variation

• 1st person pronouns across sub-registers
Register and probability

“Register variation can in fact be defined as systematic variation in probabilities”  
Halliday (1991)

• Language varies across registers at every linguistic level

• Probabilities based on “general” language are inaccurate

• Baseline probabilities should be conditioned on register:

\[ P(FREQ_{TEXT} | FREQ_{REGISTER}) \]
Determining the ideal level of analysis

• Which level accounts for the most variance?
  • Mode (spoken v. written)
  • Register (8 levels)
  • Sub-register (33 levels)

• Six linguistic variables
  • Contractions
  • 1\textsuperscript{st} person pronouns
  • 2\textsuperscript{nd} person pronouns
  • 3\textsuperscript{rd} person pronouns
  • Nouns
  • Attributive adjectives

• Coefficient of determination ($R^2$)
Determining the ideal level of analysis

![Graph showing R² values for different levels of analysis: Mode, Register, Sub-register, with different types of linguistic features: contractions, 2nd person, Att. Adjectives, Nouns, 1st person, 3rd person.](image)
Analyzing multiple levels of analysis

• Multi-Dimensional analysis (Biber, 1988)
• Cluster analysis (Biber & Egbert, in press)
• Factorial designs (Egbert, 2014)
• Hierarchical mixed effects models (Gries, 2015)
• Machine learning (Argamon, Koppel & Pennebaker, 2007)
Take away messages

• The text is the ideal unit of observation
• (Online) language is noisy; register can provide signal
• Accuracy improves when linguistic probabilities are conditioned on register
• Statistical methods can:
  • help identify the ideal level of analysis
  • simultaneously account for multiple levels of analysis
• Keep the “linguistics” in computational linguistics!
References


Thank you

Jesse Egbert

Jesse.Egbert@nau.edu
http://oak.ucc.nau.edu/jae89

This research was funded in part by the National Science Foundation (Grant No. 1147581)