ENGLISH/LINGUISTICS
4886/6886: Text and Corpus Analysis

Dr. John Hale
Keiko Bridwell
Katie Ireland Kuiper

DIGI Colloquium April 17, 2020
What is Text and Corpus Analysis?

- **Corpus**: (pl. corpora) a computerized collection of writing or speech transcripts
- ENGL/LING {4,6}886: use corpora to answer **literary** and **linguistic** questions
- This course builds upon **text analysis** activities here in the Digilab
- No matter your background, this course is for you!
Who Should Take This Class?

- Padawan learners seeking to master an elegant digital humanities toolbox
- Experimental philosophers who wonder what people do with words
- Researchers who want Data to support their claims about English, Spanish, German...
- Academic tourists who want something fun to do on Mondays, Wednesdays and Fridays
Corpus Linguistics in Action

Which gendered nouns are used more often in English girl(s), boy(s) ? Baker (2010)

For many other examples, check out Dr. Hale’s previous DIGI Colloquium here
Do you Reckon? DASS as a Searchable Corpus Tool

Keiko Bridwell & Katie Ireland Kuiper
Variation in Speech

The way that people speak is affected by a wide range of social factors, including:

● Where they live
● Who they are (age, gender, ethnicity, occupation, etc.)
● Who they are talking to
● What the conversation is about
● The setting the conversation takes place in
Project Goal

What patterns can we see in how Southern people use the word “reckon”?

I reckon that he’ll be back by Sunday.
He’ll be back by Sunday, I reckon.
Our Corpus: DASS

The Digital Archive of Southern Speech

- Part of the Linguistic Atlas Project
- 64 interviews recorded from 1968-1983

For more information about DASS, see:

- “Mapping Phonetic Variation in the American South” by Leah Dudley
- Linguistic Atlas Project website
<table>
<thead>
<tr>
<th>Speaker 025</th>
<th>I reckon, but you know used to people made their own soap.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interviewer</td>
<td>Mmhmm.</td>
</tr>
<tr>
<td>Speaker:</td>
<td>Make it out of lye. Ashes and lye. Make their own soap. Oh it'd sure would clean stuff. [...] People them days you know they had to - well, we'd get a old woman to wash for us only after our children was born.</td>
</tr>
<tr>
<td>Interviewer</td>
<td>Right.</td>
</tr>
<tr>
<td>Speaker:</td>
<td>Way back yonder they'd get great big [...] blocks you know. We'd call 'em battling blocks. Then lay the clothes on that and take a big paddle and beat 'em out. Then later on we got the washboards.</td>
</tr>
<tr>
<td>Interviewer</td>
<td>Uh-huh.</td>
</tr>
<tr>
<td>Speaker:</td>
<td>Then later on we got the washing machine. Now then we've got the washer and the dryer.</td>
</tr>
</tbody>
</table>
Reckon Results

- Resulting DASS corpus: ~3.5 million words
- 526 tokens of “reckon”
- **Reckon-to-think ratio**: how often speakers used “reckon” vs. its more frequent synonym, “think”
- We looked at this ratio across location, gender, and age
Reckon Results

Reckon-to-Think Ratio by Gender

Reckon-to-Think Ratio by Age Group
Conclusions

- **Regional variation**: “reckon” is a word associated with Southern American English
- **Change over time**: older speakers used “reckon” frequently, while younger speakers barely used it at all
- **Social variation**: men adopted this change later than women
T&C: Course Topics & Plans

- Corpus Linguistics
- Spoken Language
- Syntax
- Data Visualization
- Corpus Methods
- British National Corpus
- Social Media
- Forensic Linguistics
- Data
- COCA
- News
- CQP
- Health-related data
- Treebanks
- Analysis
- SLA
Frequently Asked Questions

- Do I need programming experience?
  - No programming required!

- Is it OK if I'm not a Linguistics major?
  - It is OK! We welcome students from across the humanities and sciences.

- What if I am only interested in a specific subfield of Linguistics?
  - Annotated corpora allow you to study your favorite subfield: phonology, morphology, syntax... Your final project can focus on just that area.
Q. Does the course count toward my degree?

A. Quite likely. It counts as...

- A 4000-level course as required for the ENGL ugrad major
- One of the two required courses for the LING ugrad major (second list)
- It also counts for the Humanities Computing Area of Emphasis within ENGL.
- For the Digital Humanities certificate for undergraduates at UGA!

For grad students, this counts towards the basic 9 courses for ENGL and as a Research Skill!
Go ahead and sign up!

CRN 42695

Join the corpus revolution!
Questions?

- Feel free to reach out with any questions or concerns at:
  - Dr. John Hale [jthale@uga.edu](mailto:jthale@uga.edu)
  - Katie Kuiper (TA) [katherine.kuiper25@uga.edu](mailto:katherine.kuiper25@uga.edu)
  - Keiko Bridwell [keiko.bridwell@uga.edu](mailto:keiko.bridwell@uga.edu)
The DASS Project is supported by: NSF BCS #1625680 to co-PIs Kretzschmar and Renwick, the UGA Graduate School, and the American Dialect Society.


