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)

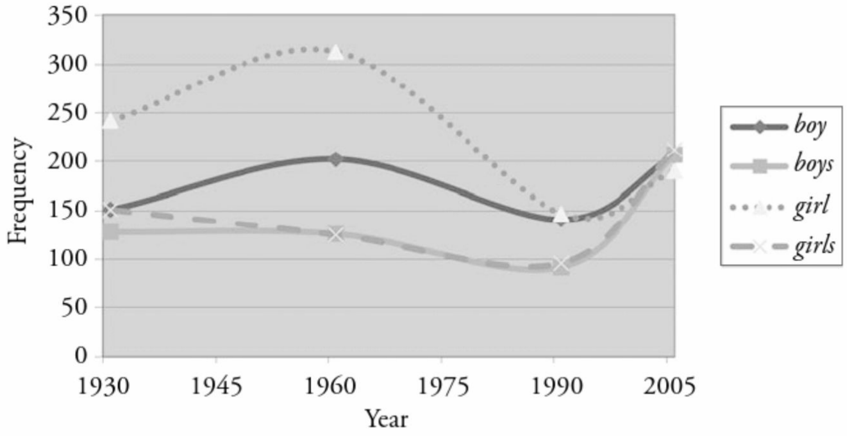


Figure 3.2 Frequencies of *girl(s)* and *boy(s)* over time

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***.

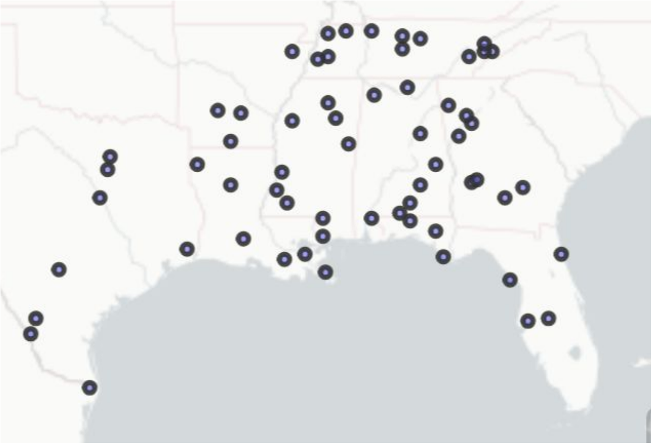
Our Corpus: DASS

The Digital Archive of Southern Speech

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

For more information about DASS, see:

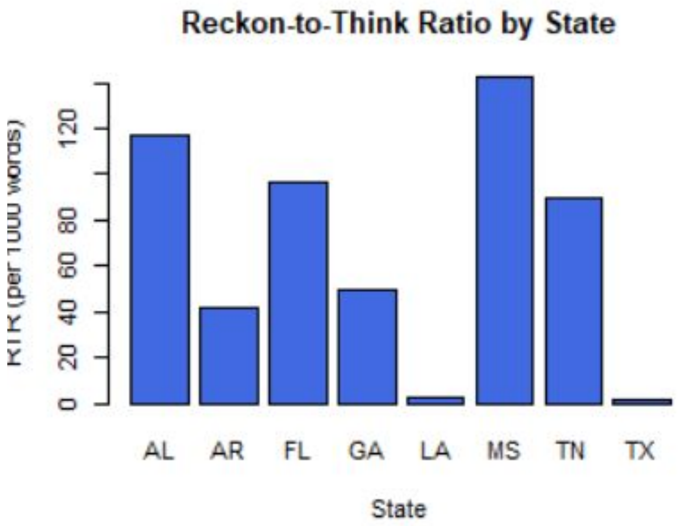
- "Mapping Phonetic Variation in the American South" by Leah Dudley
- Linguistic Atlas Project website



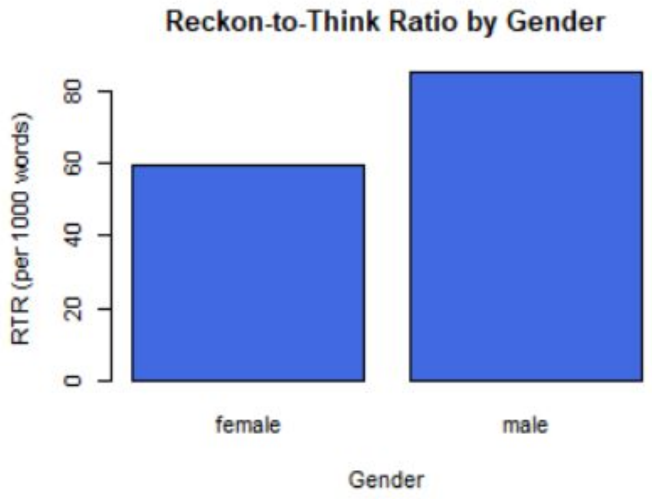
Speaker 025:	I reckon, but you know used to people made their own soap.
Interviewer:	Mmhmm.
Speaker:	Make it out of lye. Ashes and lye. Make their own soap. Oh it'd sure would clean stuff. [...] People them days you know they had to - well, we'd get a old woman to wash for us only after our children was born.
Interviewer:	Right.
Speaker:	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.
Interviewer:	Uh-huh.
Speaker:	Then later on we got the washing machine. Now then we've got the washer and the dryer.

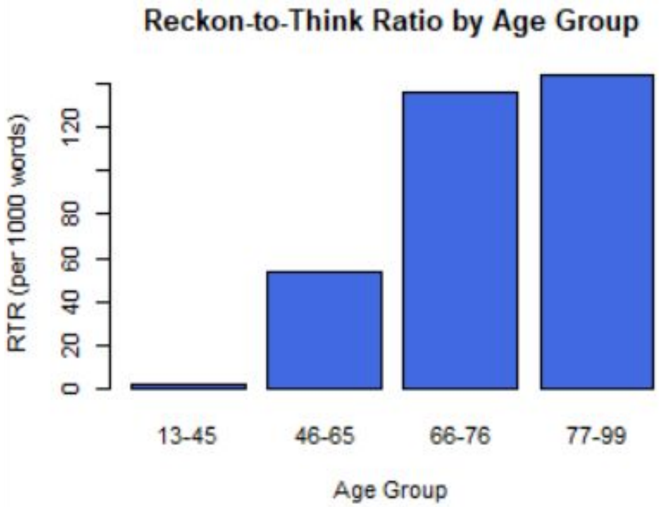
Reckon Results

- 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

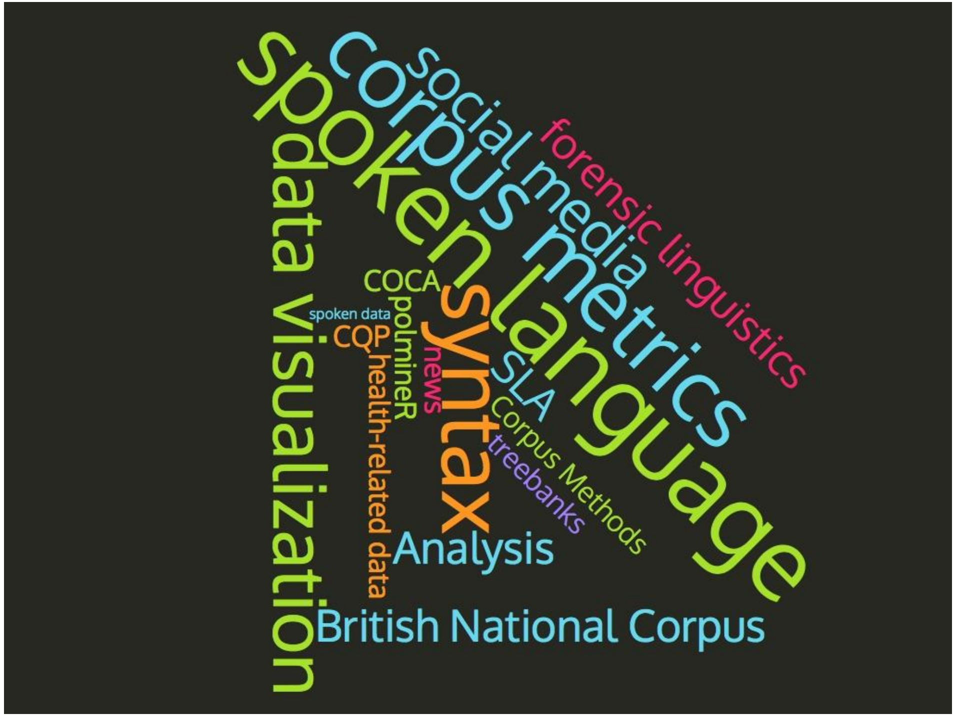




Conclusions

- **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



Frequently Asked Questions

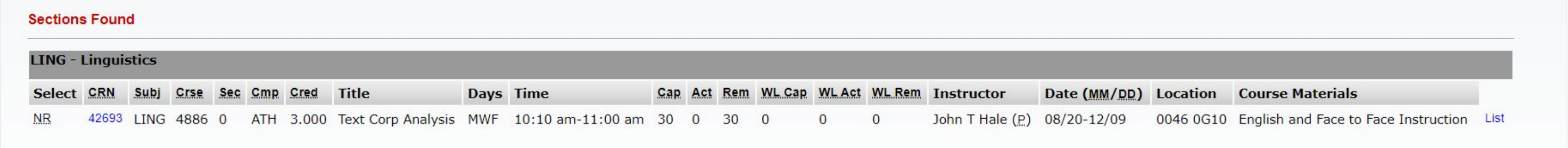
- Do I need programming experience?
- Is it OK if I’m not a Linguistics major?
- What if I am only interested in a specific subfield of Linguistics?

- No programming required!
- It is OK! We welcome students from across the humanities and sciences.
- Annotated corpora allow you to study your favorite subfield: phonology, morphology, syntax... Your final project can focus on just that area.

Frequently Asked Questions

- 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!



Sections Found

LING - Linguistics

Select	CRN	Subj	Crse	Sec	Cmp	Cred	Title	Days	Time	Cap	Act	Rem	WL Cap	WL Act	WL Rem	Instructor	Date (MM/DD)	Location	Course Materials	
NR	42693	LING	4886	0	ATH	3.000	Text Corp Analysis	MWF	10:10 am-11:00 am	30	0	30	0	0	0	John T Hale (P)	08/20-12/09	0046 0G10	English and Face to Face Instruction	List

CRN 42695

Join the corpus revolution!

Questions?

- Feel free to reach out with any questions or concerns at:
  - Dr. John Hale jthale@uga.edu
  - Katie Kuiper (TA) katherine.kuiper25@uga.edu
  - Keiko Bridwell keiko.bridwell@uga.edu

Works Cited

- *The DASS Project is supported by: NSF BCS #1625680 to co-PIs Kretzschmar and Renwick, the UGA Graduate School, and the American Dialect Society.*
- Baker, Paul. 2010. Sociolinguistics and Corpus Linguistics. Edinburgh University Press.
- Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press.
- Evert, Stefan and the CWB Development Team. 2017. The IMS Open Corpus Workbench (CWB) Corpus Encoding Tutorial.
- Pederson, L., McDaniel, S. L., and Adams, C. M. (Eds.) 1986. *Linguistic Atlas of the Gulf States*, University of Georgia Press, Athens, Georgia, Vols. 1–7.
- Stanley, Joseph A., Margaret E. L. Renwick, William A. Kretzschmar Jr., Rachel M. Olsen, & Michael Olsen. 2017. “The Gazetteer of Southern Vowels.” The American Dialect Society Annual Meeting. Salt Lake City, UT.