CS 293/EDUC 473

Discovery & Exploration in Educational Text Data

Parsing, Lexical Analyses

Stanford GRADUATE SCHOOL OF EDUCATION

Stanford | NLP

Quiz time!

Reminders

- Final chance to sign up for your own discussant slot!
  - ~4 students still haven't signed up; we'll assign folks on Friday
- A0 due **next Monday** at 5pm
  - How is everyone doing on teacher recruitment?
- A1 due the following Monday (Jan 27)
- Dan Meyer from the Amplify/Desmos will join the Thursday class!

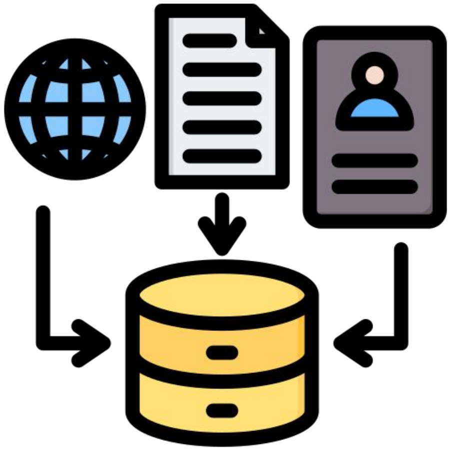
Today's class

- Lecture by Dora on data exploration and textbook study
- Discussion!

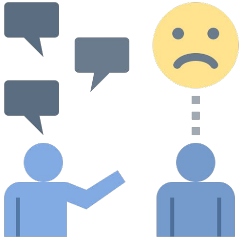
Data Exploration

What counts as data?

- Qualitative data
  - Conversations & interviews with stakeholders (teachers, students, peers, researchers)
  - Results and insights presented by related work
  - Surveys
- Quantitative data
  - Text data from the target domain, e.g. classroom transcripts, text books, lesson plans
  - Metadata associated with the text, e.g. demographic information, learning outcome data, satisfaction ratings



Things you learn from qualitative information gathering

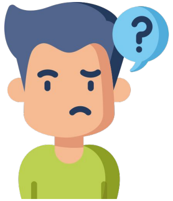


"We already think we know that"

"That's too naive"

"that doesn't reflect social reality"

"Text analysis is unlikely to answer that question"



"Two major camps in the field would give different answers to that question"

"We tried to look at that back in the 1960s but we didn't have the technology"

"That sounds like something teachers would love"

"That's a really fundamental question"

How we do things with words (Nguyen et al., 2020)

Challenges in obtaining & using educational text data

- Privacy
- Noisiness
  - Transcription errors
  - Data often requires lots and lots of cleaning
- Sparsity
  - Especially for self-reported data
- Bridging Expertise (Pedagogical and Technical)
- Operationalizing Ambiguous Concepts

For your review

Building a solution requires **having the right data & understanding that data**

- ➔ What are the ethical or legal considerations for using this data?
- ➔ Do you have access to your desired sample (size)?
- ➔ What is the distribution of the data? Is it representative of your target population?
- ➔ How clean is the data?
- ➔ How do variables in the data relate to each other?
- ➔ Do you have the required outcomes for answering the research question / estimating impact?
- ➔ What hypotheses or assumptions can be made based on initial exploration?

For your review

Identifying a research question

Think about these questions as you’ll need to address them in the Project Rationale.

➔ Who is waiting for the solution / answer to your question? What would this solution / knowing the answer would change, both in your field of study and in the wider world?

➔ Are these questions answerable with text?

➔ Why is computational text analysis necessary or valuable for this solution / answering this question?

➔ Do you have access to the data that will support these research questions?

➔ What are the ethical implications of your research / solution? Who will be affected by decisions made based on your solution / results?

How we do things with words (Nguyen et al., 2020)

For your review

Guiding questions for conceptualization

➔ What are the core concepts you are addressing? And are you being true to their core meaning?

➔ What are competing definitions? Which is best suited to the task and why?

➔ Does the systematized concept you’ve selected reflect an adequate understanding of the background concept?

➔ How do domain experts approach the topic? Does your research connect to this wider context? Have you considered relevant methods and theories in other domains?

➔ Is it possible to speak of “ground truth” for the concept(s) in question?

How we do things with words (Nguyen et al., 2020)

For your review

Guiding questions for data exploration

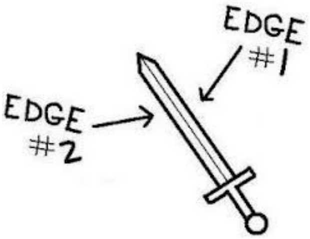
- ➔ Are sources representative? Are they disproportionately of one form? Are all relevant time windows covered? Does the data represent all relevant groups, including those often marginalized?
- ➔ When metadata is available: Are there errors, inconsistencies, biases, or missing information? Is this quality of metadata consistent across the dataset, or are some parts better or worse?
- ➔ When labels are available: How were the labels created? Do the labels actually mean what you are using them to represent?
- ➔ If you are filtering, subsampling, or selecting from the original data, is the remaining subset representative? Can you describe how selective removal alters the data and the interpretation of the data? Are you losing anything that might be valuable at a later stage?
- ➔ Who created the data, and do they have agency over its use? Should this data be used for research? How does respect for document creators affect how you conduct and share your research?

How we do things with words (Nguyen et al., 2020)

For your review

Think about dual use

- What if the tool was used in any type of high stakes decision making?
- Could the tool be used to surveil teachers?
- Could it be used to punish students or teachers?
- Could the tool be used to harm vulnerable populations, aggravate biases and inequities?

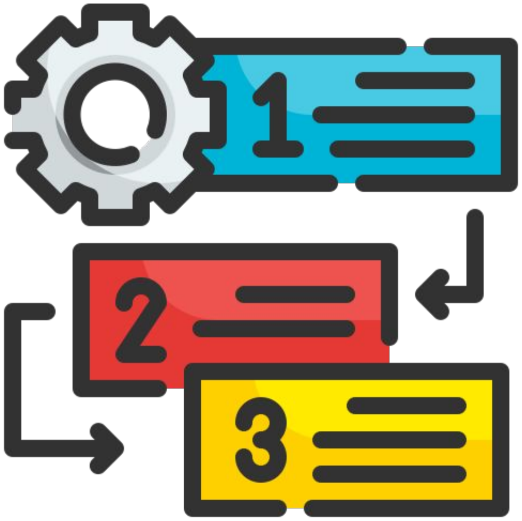


How we do things with words (Nguyen et al., 2020)

Best practices for data exploration

#1 **Prioritize.**

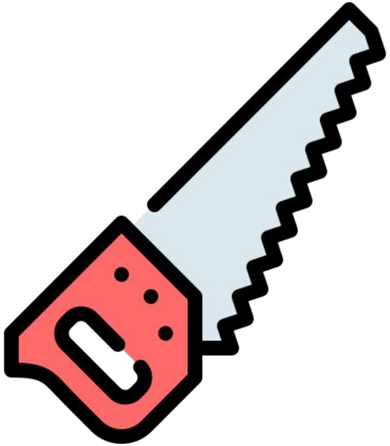
- It’s easy to get lost in the weeds — pause to **take a step back** and keep your core question / goal in mind
- Each separate measure requires a lot of resources to develop and validate (e.g. each pedagogical practice).
- Focus on constructs that seem to be the **highest leverage** and can serve as proxies for other important constructs (e.g. via lit review)



***Example**: teachers’ uptake of student ideas*

#2 Don’t be afraid of doing a LOT of manual work.

- Automating tasks is sometimes more work and less precise than doing the clean-up manually
- When you need data for validation, it’s usually best to clean it manually to avoid circularity
- Close reading and qualitative coding is often the best way to understand the contents of your data and do error analyses. Don’t be afraid to do that a lot even if you’re not an expert.



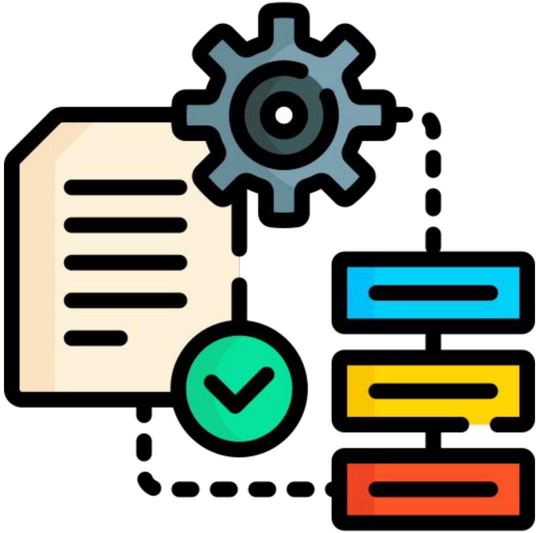
***Example**: identifying nouns referring to people in textbooks*

#3 **Create visualizations.**

- Visualizations are oftentimes best way to understand the distribution of your data
- Graphs and plots are usually much better at explaining patterns than regressions
- Visualizations can help a lot with debugging, too

#4 **Don’t be hand-wavy about pre-processing.**

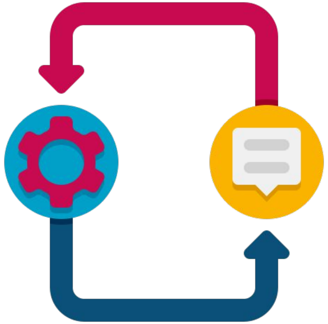
- Preprocessing decisions (e.g. removing stopwords) can make a big difference (see A1)
- Explore how pre-processing decisions affect your data
- Often it’s helpful to report the core analyses with different pre-processing decisions in your results (e.g. in supplement)



***Example**: stemming in topic models*

#5 **Reinforce the feedback loop.**

- Use insights from your explorations to finetune your goals / research questions
- Seek feedback from educators after you've looked at the data
  - Best is to show them some data and ask what they observe



Talk to your teachers buddy!

#6 **Start with the most interpretable methods first.**

- It’s hard to “debug” and interpret methods with multiple layers of abstraction
- **Words** often explain a lot of the variance — lexical analysis is often the best starting point

***Example**: log odds ratios; cosine similarities of word embeddings*

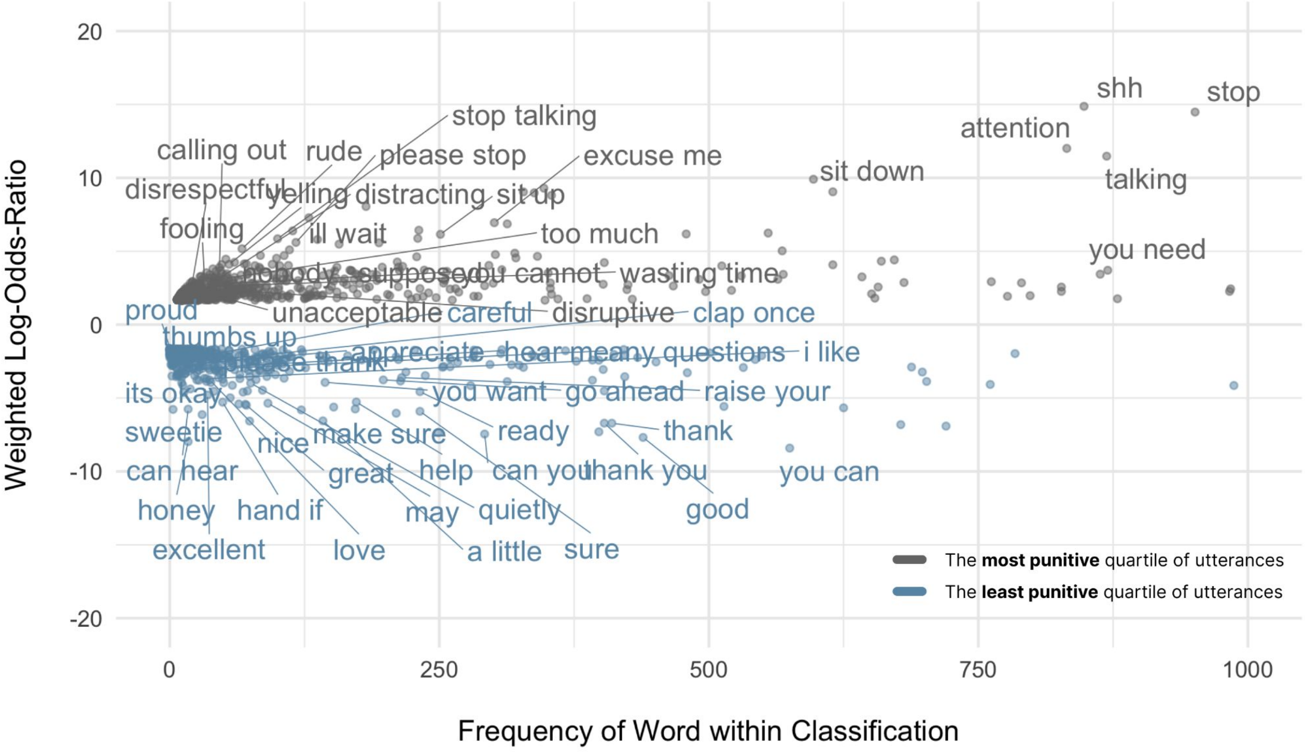
Example: Z-scored log odds ratios

It answers **how word usage is different along a particular dimension?**

1. **Sample two different groups from the data** (e.g. teacher utterances from transcripts with high ratings for instruction quality vs ones with low ratings) + create a third group (**prior**) that includes your entire dataset (e.g. all teacher utterances in data)
2. **Obtain counts** for words / phrases in the data (you can use it to count any other feature, too, e.g. lexical categories);
    a. see Demszky et al. (2019) for the use of the log odds method on different features (LIWC, topics)
3. **Compute the z-scored log odds ratios** for each word / phrase. Positive values indicate association with group 1 and negative values indicate association with group 1. Magnitude represents number of standard deviations (discard those with < 1).

My favorite data exploration method!!! Will share the script on Ed!

Z-scored log odds ratios (Section 4.3 in Monroe et al., 2017)



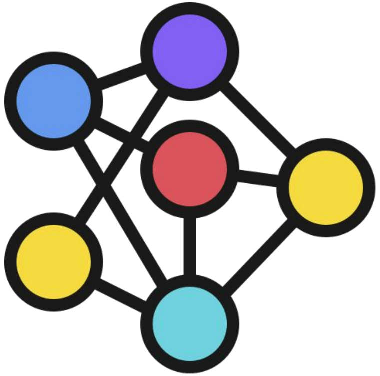
(Tan & Demszky, 2023)

#7 **Be scientific about debugging.**

- If you (don’t) observe something, that can be due to many things
  - Noise in the data
  - Imprecise definitions of your construct
  - Imprecise measures
    - e.g. lexicons or ML classifiers are not perfect
- Be systematic to explore all the possible causes of both positive and negative results

#8 **Triangulate several data sources as much as possible**

- E.g. self-reported data + student outcome data + language features
- Doing so can help
  - provide a more nuanced understanding of relationships in your data
  - validate measures
  - corroborate hypotheses
  - debug issues



Tools for data exploration

- **General Python packages (ChatGPT may be your best friend :))**
  - pandas (data wrangling)
  - statsmodels (regressions, although it's better to do them in R/Stata)
  - scipy (e.g. for t-tests, correlations)
  - seaborn (for visualization)
- **Lexical analysis:**
  - log odds method to identify words/phrases that distinguish two groups (see homework)
  - lexicons (e.g. NRC valence arousal dominance lexicon, concreteness lexicon)
- **Unsupervised methods (next class)**

Relevant resources

- Dirk Hovy’s Github repository
- Introduction to Cultural Analytics by Melanie Walsh
- DLATK command line text analysis tool
- Text analysis tutorial on scikit-learn
- Computational text analysis course by Adam Poliak
- NLP + CSS tutorials by Katie Keith and Ian Stewart
- StatQuest Youtube Channel
- Computational and Inferential Thinking

Additional data sources

Text corpora:

- Open Syllabus Project (old Github version)
- Teacher-student Chatroom Corpus
- Coursera Forum Dataset
- CIMA
- TalkMoves Dataset
- DRYAD dataset
- PERSUADE dataset of student writing
- KhanAcademy Math Tutoring Dataset
- DLP Catalog of Math-Teaching Related Datasets

Pedagogical resources:

- Middle school math misconceptions
- Achieve The Core
- TLE dataset of recordings
- MQI video library

Browse the following conference proceedings: BEA, LAK, EDM for more datasets

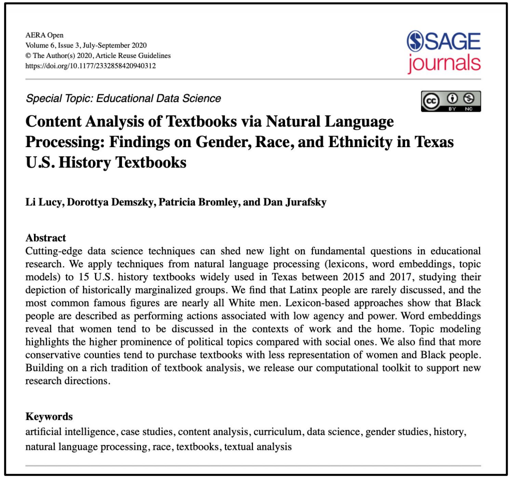
Case Study

Lucy Li

Tricia Bromley

Dan Jurafsky

Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas US History Textbooks



AERA Open
Volume 6, Issue 3, July-September 2020
© The Author(s) 2020, Article Reuse Guidelines
https://doi.org/10.1177/2332858420940312

Special Topic: Educational Data Science

Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas U.S. History Textbooks

Li Lucy, Dorottya Demszky, Patricia Bromley, and Dan Jurafsky

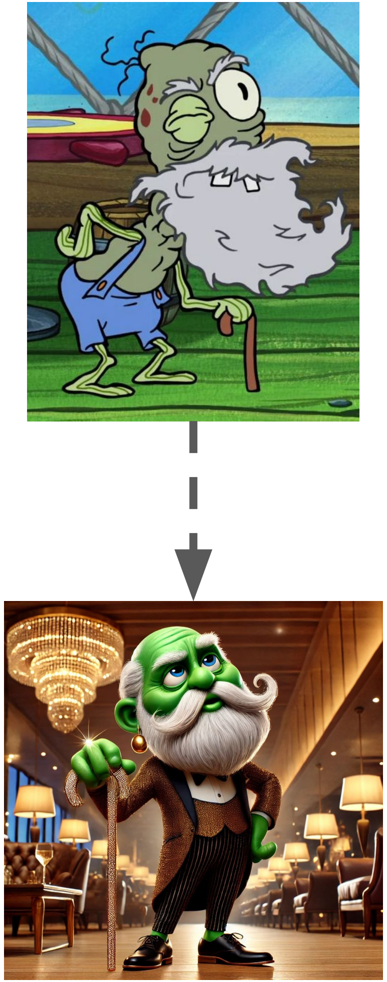
Abstract
Cutting-edge data science techniques can shed new light on fundamental questions in educational research. We apply techniques from natural language processing (lexicons, word embeddings, topic models) to 15 U.S. history textbooks widely used in Texas between 2015 and 2017, studying their depiction of historically marginalized groups. We find that Latinx people are rarely discussed, and the most common famous figures are nearly all White men. Lexicon-based approaches show that Black people are described as performing actions associated with low agency and power. Word embeddings reveal that women tend to be discussed in the contexts of work and the home. Topic modeling highlights the higher prominence of political topics compared with social ones. We also find that more conservative counties tend to purchase textbooks with less representation of women and Black people. Building on a rich tradition of textbook analysis, we release our computational toolkit to support new research directions.

Keywords
artificial intelligence, case studies, content analysis, curriculum, data science, gender studies, history, natural language processing, race, textbooks, textual analysis

Why are we reading this old paper???

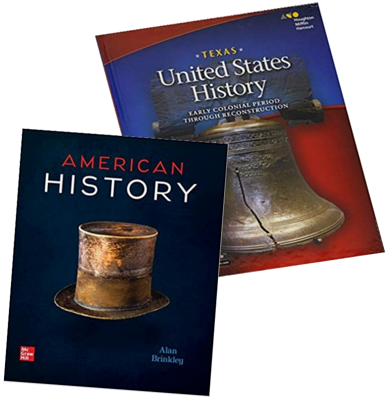
2020 is pre-ChatGPT… ancient times.

- Broad survey of lexical methods
- Illustrates how to apply NLP to **small** data w/ statistical rigor
- Great starting point for discussing the relevance of word-based analyses!



Motivation

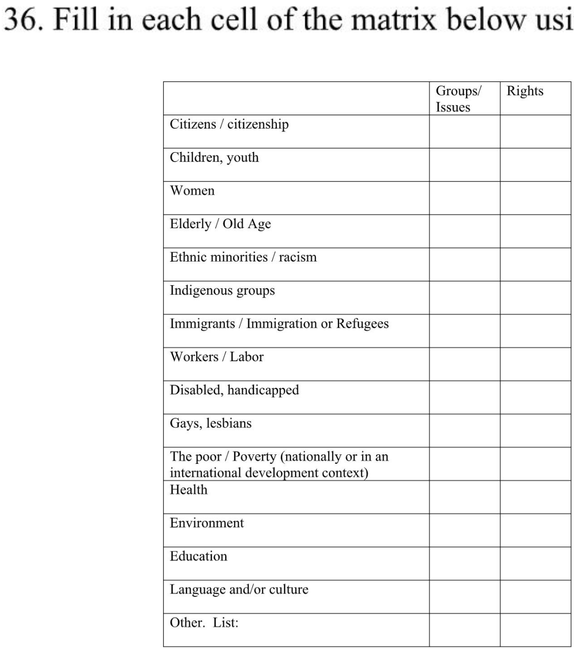
Textbooks are the most widely used instructional tool around the world


Houghton Mifflin Harcourt
TEXAS
United States History
EARLY COLONIAL PERIOD THROUGH RECONSTRUCTION
AMERICAN HISTORY
Mc Graw Hill
Alan Brinkley

Social & cultural values

Traditional Methods

Coding protocols, e.g:



36. Fill in each cell of the matrix below using the following codes:

	Groups/ Issues	Rights
Citizens / citizenship		
Children, youth		
Women		
Elderly / Old Age		
Ethnic minorities / racism		
Indigenous groups		
Immigrants / Immigration or Refugees		
Workers / Labor		
Disabled, handicapped		
Gays, lesbians		
The poor / Poverty (nationally or in an international development context)		
Health		
Environment		
Education		
Language and/or culture		
Other. List:		

1 = mentioned
0 = not mentioned

0 = no mention
1 = one or two sentences
2 = at least a paragraph
3 = at least one subheading
4 = at least one chapter heading
5 = over half the chapters

From Meyer, Bromley, & Ramirez (2010)

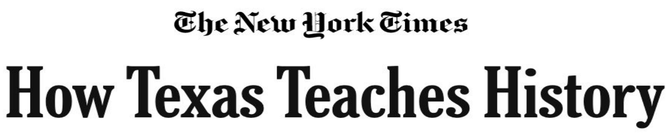
Texas

- 5.4M K-12 students (2017), 2nd largest in US
- Major textbook market
- Large influence on textbooks in U.S.

TEA®

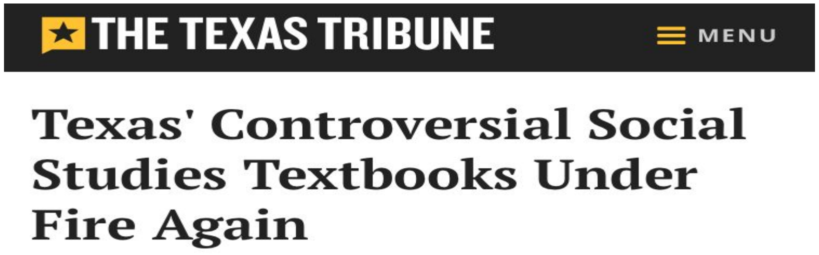
Texas Education Agency

Texas



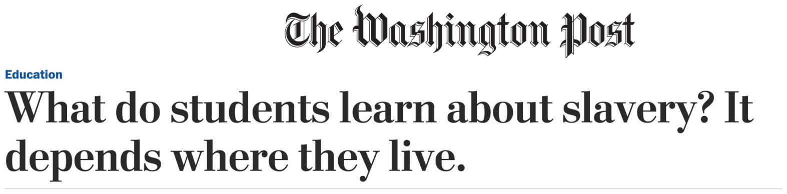
The New York Times

How Texas Teaches History



THE TEXAS TRIBUNE

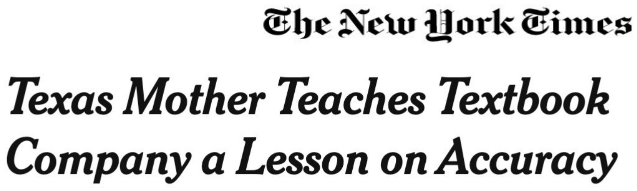
Texas' Controversial Social Studies Textbooks Under Fire Again



The Washington Post

Education

What do students learn about slavery? It depends where they live.



The New York Times

Texas Mother Teaches Textbook Company a Lesson on Accuracy

Research Questions

RQ1 How much are different groups of people **mentioned**?

Research Questions

RQ1 How much are different groups of people **mentioned**?

RQ2 How are different groups and individuals **described**?

Research Questions

RQ1 How much are different groups of people **mentioned**?

RQ2 How are different groups and individuals **described**?

RQ3 Which **topics** are prominent and how do they relate to groups of people?

American History Textbook Data (2015-17)

American History Textbook Data (2015-17)

Messy purchase data from Texas districts

8th gr give me liberty	T.ISD
pearson US hsitory coloniz...	B.ISD
Pearson us history texas ed	B.ISD

American History Textbook Data (2015-17)

Messy purchase data from Texas districts

8th gr give me liberty	T.ISD
pearson US hsitory coloniz...	B.ISD
Pearson us history texas ed	B.ISD

manual clean-up & disambiguation

book	district	count
Am. Hist.	T.ISD	30
Give me lib.	B.ISD	100

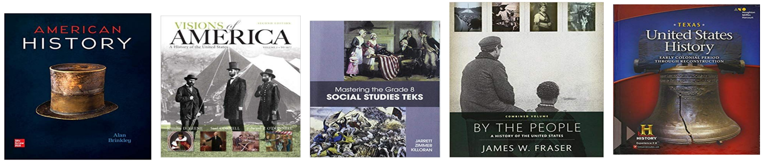
American History Textbook Data (2015-17)

Messy purchase data from Texas districts

8th gr give me liberty	T.ISD
pearson US hsitory coloniz...	B.ISD
Pearson us history texas ed	B.ISD

book	district	count
Am. Hist.	T.ISD	30
Give me lib.	B.ISD	100

keep **15** most widely purchased textbooks

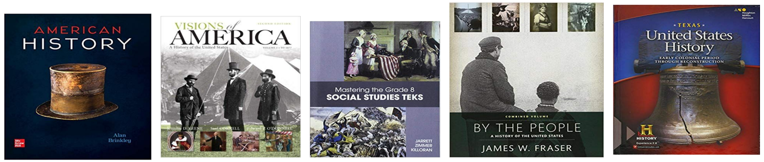

AMERICAN HISTORY
Alan Brinkley
VISIONS of AMERICA
Mastering the Grade 8 SOCIAL STUDIES TEKS
BY THE PEOPLE
A HISTORY OF THE UNITED STATES
JAMES W. FRASER
TEXAS
United States History
EARLY COLONIAL PERIOD THROUGH RECONSTRUCTION

American History Textbook Data (2015-17)

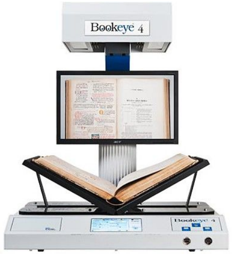
Messy purchase data from Texas districts

8th gr give me liberty	T.ISD
pearson US hsitory coloniz...	B.ISD
Pearson us history texas ed	B.ISD

book	district	count
Am. Hist.	T.ISD	30
Give me lib.	B.ISD	100



scan

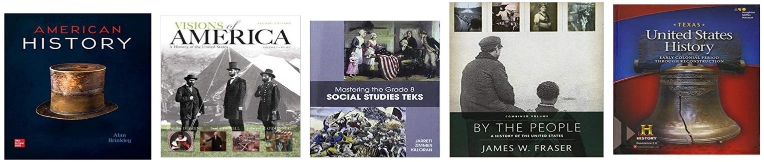


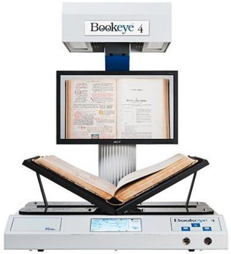
American History Textbook Data (2015-17)

Messy purchase data from Texas districts

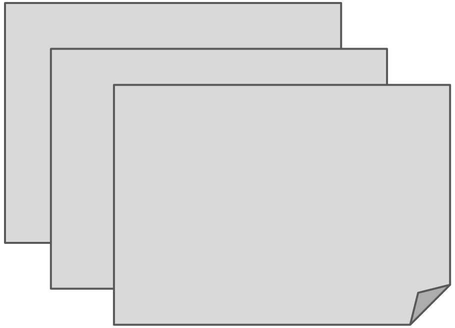
8th gr give me liberty	T.ISD
pearson US hsitory coloniz...	B.ISD
Pearson us history texas ed	B.ISD

book	district	count
Am. Hist.	T.ISD	30
Give me lib.	B.ISD	100





OCR w/ ABBYY® FineReader®

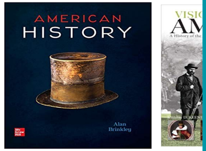


American History Textbook Data (2015-17)

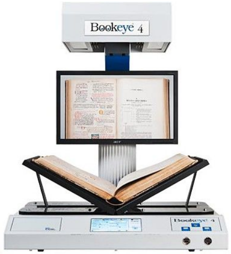
Messy purchase data from Texas districts

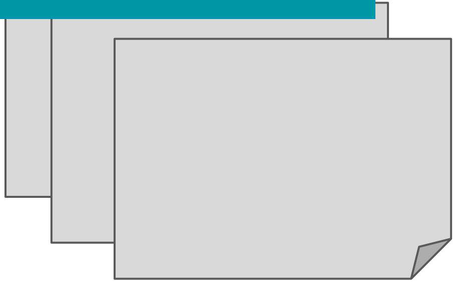
8th gr give me liberty	T.ISD
pearson US hsitory coloniz...	B.ISD
Pearson us his	

book	district	count
Am. Hist.	T.ISD	30
Give me lib	B.ISD	100



2 MONTHS OF WORK FOR TWO PEOPLE





Demographic Data

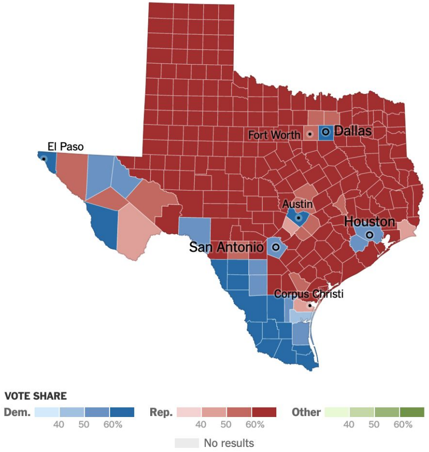
- district-level student demographic data
  - the National Center for Education Statistics (NCES), for AY 2016-17

NCES

source: New York Times

Demographic Data

- district-level student demographic data
  - the National Center for Education Statistics (NCES), for AY 2016-17
- county-level political leaning
  - two party vote shares in 2016 elections



source: New York Times

Example

Progress toward feminist goals was limited in the antebellum years, but individual women did manage to break the social barriers to advancement. Elizabeth Blackwell, born in England, gained acceptance and fame as a physician. Her sister-in-law Antoinette Brown Blackwell became the first ordained woman minister in the United States; and another sister-in-law, Lucy Stone, took the revolutionary step of retaining her maiden name after marriage. Stone became a successful and influential lecturer on women's rights. (Brinkley, 2015: p. 330)

RQ1

How Much Are Different Groups of People **Mentioned**?

Progress toward feminist goals was limited in the antebellum years, but individual women did manage to break the social barriers to advancement. Elizabeth Blackwell, born in England, gained acceptance and fame as a physician. Her sister-in-law Antoinette Brown Blackwell became the first ordained woman minister in the United States; and another sister-in-law, Lucy Stone, took the revolutionary step of retaining her maiden name after marriage. Stone became a successful and influential lecturer on women’s rights. (Brinkley, 2015: p. 330)

RQ1

How Much Are Different Groups of People **Mentioned**?

Progress toward feminist goals was limited in the antebellum years, but individual women did manage to break the social barriers to advancement. Elizabeth Blackwell, born in England, gained acceptance and fame as a physician. Her sister-in-law Antoinette Brown Blackwell became the first ordained woman minister in the United States; and another sister-in-law, Lucy Stone, took the revolutionary step of retaining her maiden name after marriage. Stone became a successful and influential lecturer on women's rights. (Brinkley, 2015: p. 330)

Coreference Resolution

RQ1

How Much Are Different Groups of People **Mentioned**?

Progress toward feminist goals was limited in the antebellum years, but individual women did manage to break the social barriers to advancement. Elizabeth Blackwell, born in England, gained acceptance and fame as a physician. Her sister-in-law Antoinette Brown Blackwell became the first ordained woman minister in the United States; and another sister-in-law, Lucy Stone, took the revolutionary step of retaining her maiden name after marriage. Stone became a successful and influential lecturer on women’s rights. (Brinkley, 2015: p. 330)

Identifying people-related common nouns
(WordNet, 95% accuracy)

RQ1

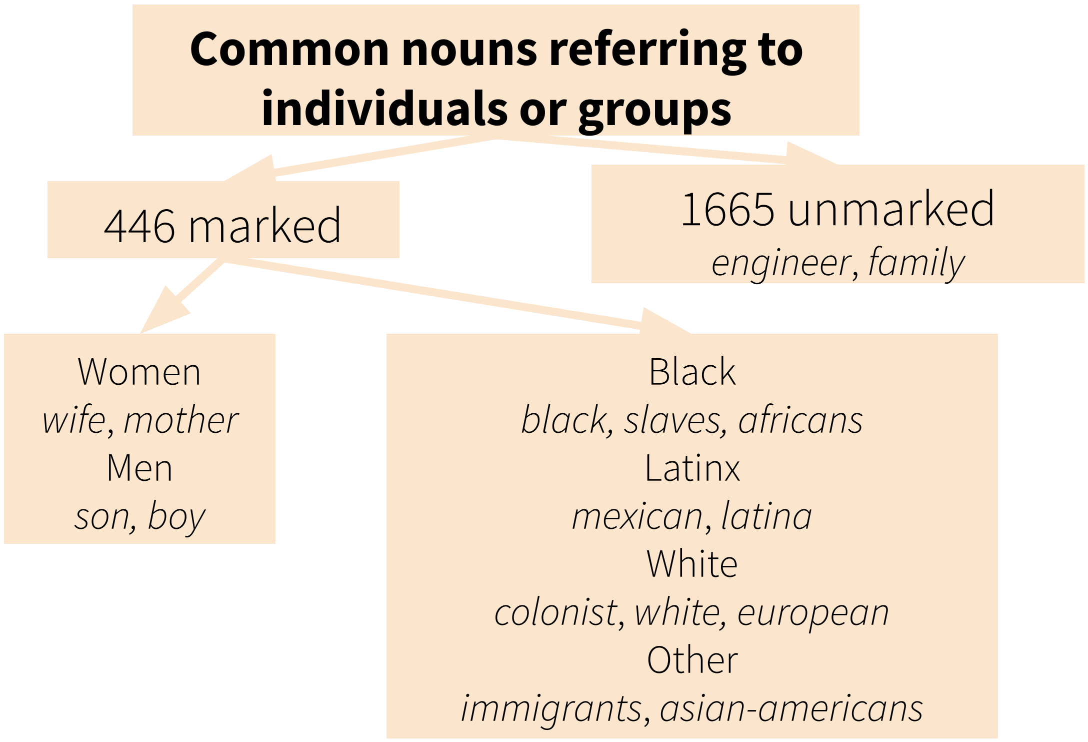
How Much Are Different Groups of People **Mentioned**?

Progress toward feminist goals was limited in the antebellum years, but individual women did manage to break the social barriers to advancement. Elizabeth Blackwell, born in England, gained acceptance and fame as a physician. Her sister-in-law Antoinette Brown Blackwell became the first ordained woman minister in the United States; and another sister-in-law, Lucy Stone, took the revolutionary step of retaining her maiden name after marriage. Stone became a successful and influential lecturer on women’s rights. (Brinkley, 2015: p. 330)

Named Entity Recognition

RQ1

Race/Ethnicity & Gender



Common nouns referring to individuals or groups

446 marked

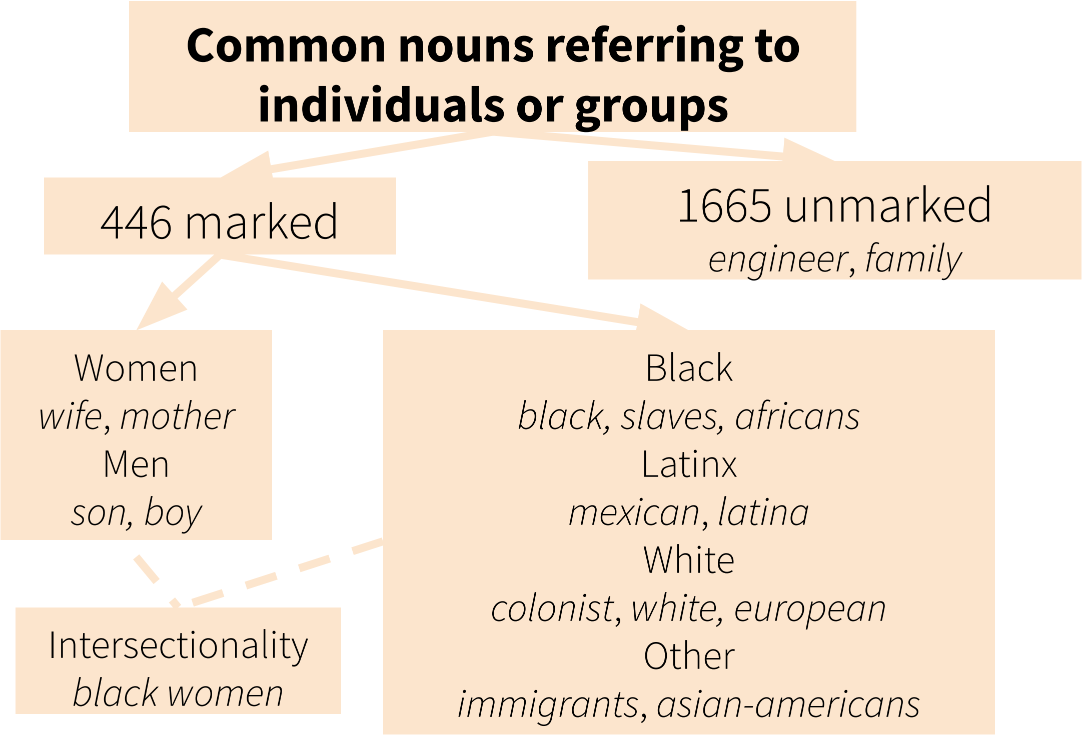
1665 unmarked
engineer, family

Women
wife, mother
Men
son, boy

Black
black, slaves, africans
Latinx
mexican, latina
White
colonist, white, european
Other
immigrants, asian-americans

RQ1

Race/Ethnicity & Gender



Common nouns referring to individuals or groups

446 marked

1665 unmarked
engineer, family

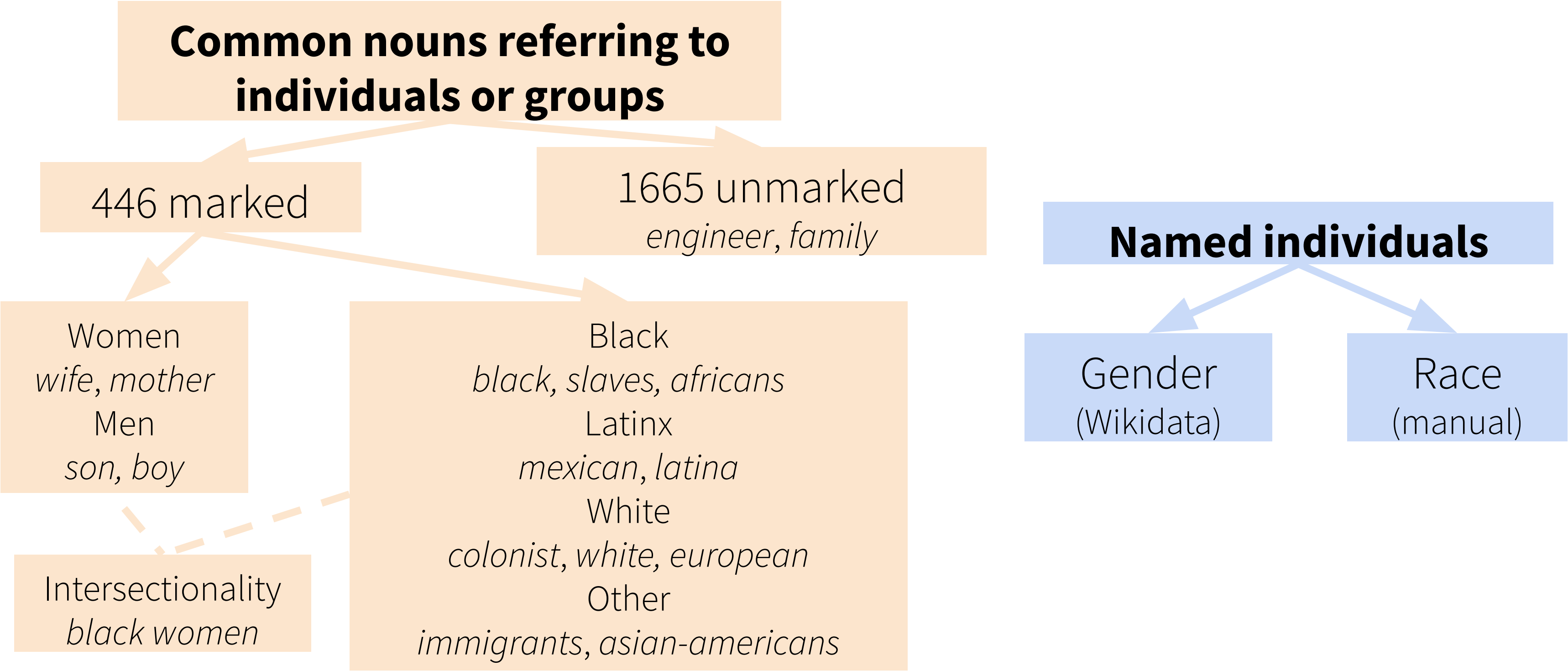
Women
wife, mother
Men
son, boy

Black
black, slaves, africans
Latinx
mexican, latina
White
colonist, white, european
Other
immigrants, asian-americans

Intersectionality
black women

RQ1

Race/Ethnicity & Gender



Common nouns referring to individuals or groups

446 marked

1665 unmarked
engineer, family

Women
wife, mother
Men
son, boy

Black
black, slaves, africans
Latinx
mexican, latina
White
colonist, white, european
Other
immigrants, asian-americans

Intersectionality
black women

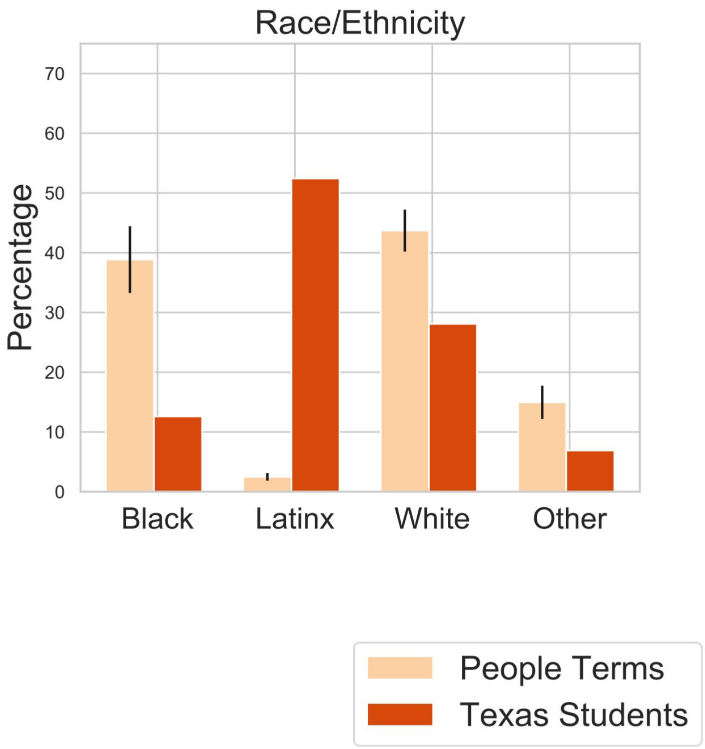
Named individuals

Gender
(Wikidata)

Race
(manual)

RQ1

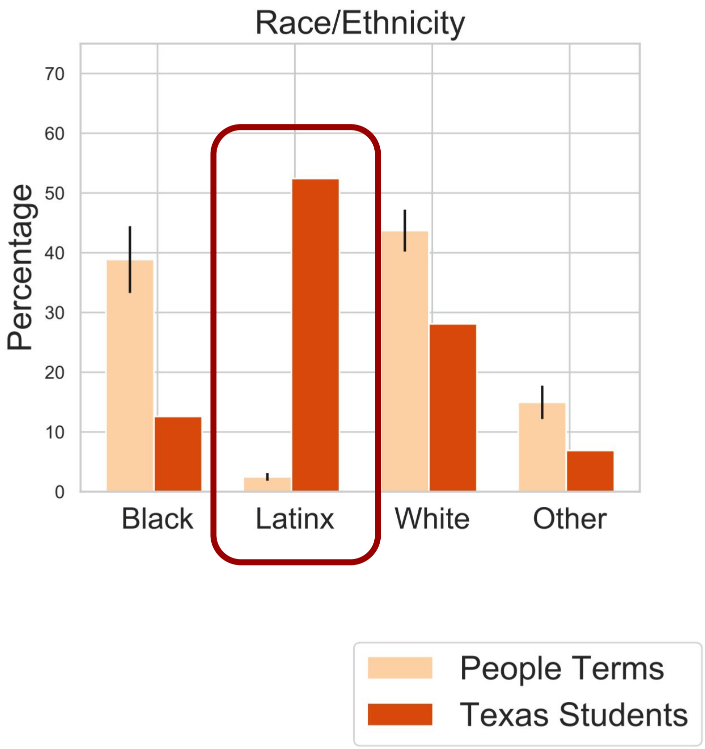
Comparing Student Demographics w/ Representation in Text



Common nouns referring to individuals or groups

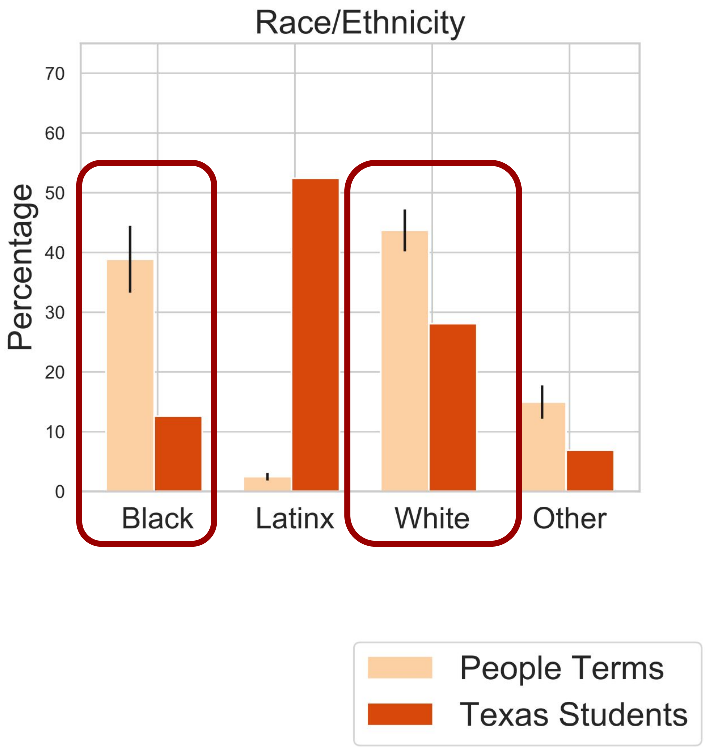
RQ1

Hispanic / Latinx Students are Disproportionately Underrepresented



RQ1

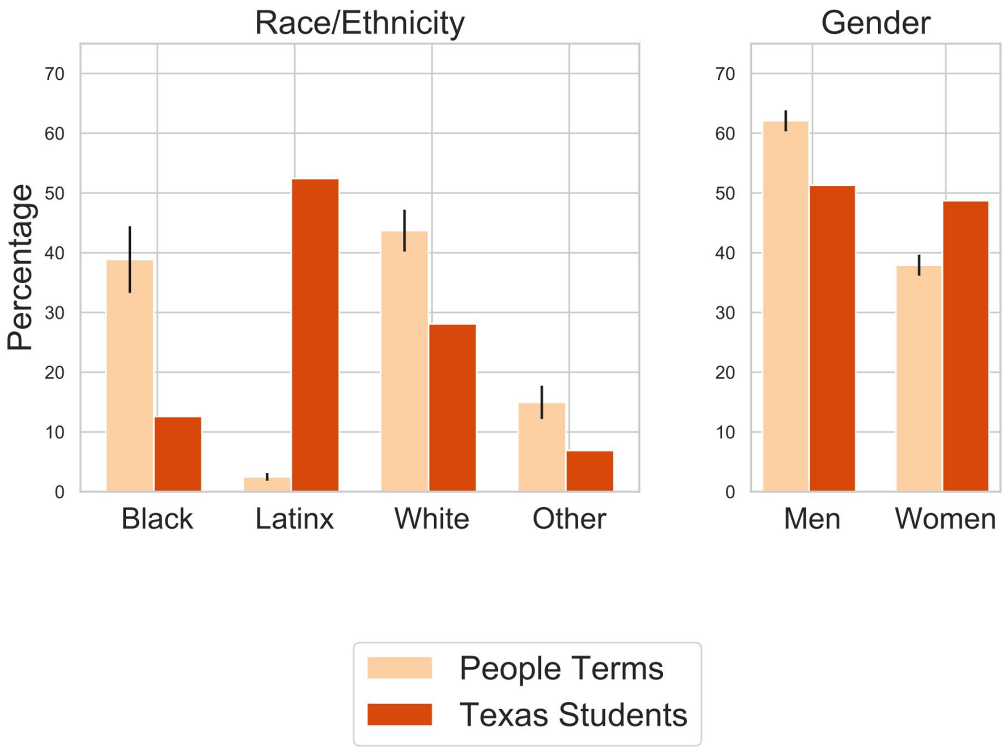
African Americans and White People are Mentioned Disproportionately More



white people are mentioned even more often than the plot shows (since this ethnicity is often unmarked)

RQ1

Men Are Mentioned Disproportionately More Often Than Women

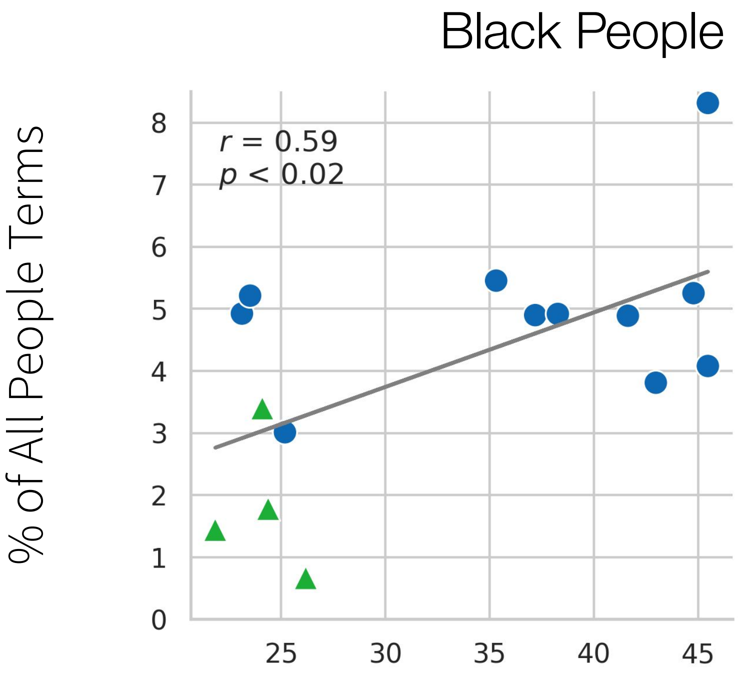


RQ1

Top 50 Named People

RQ1

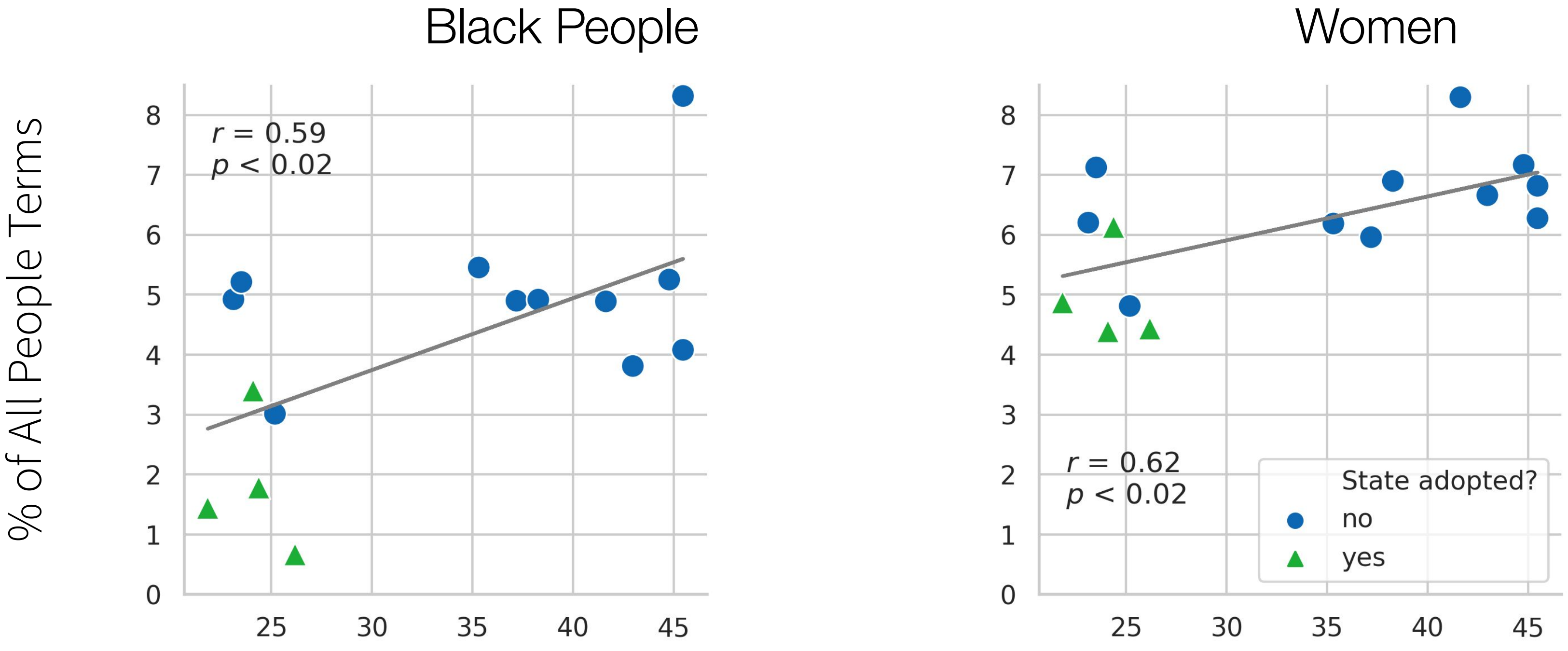
Books in More Democratic Counties Mention Black People and Women More



Median % of Democrats Across Counties Where Textbook is Bought

RQ1

Books in More Democratic Counties Mention Black People and Women More



Median % of Democrats Across Counties
Where Textbook is Bought

RQ2

How Are Different Groups and Individuals **Described**?

Progress toward feminist goals was limited in the antebellum years, but individual women did manage to break the social barriers to advancement. Elizabeth Blackwell, born in England, gained acceptance and fame as a physician. Her sister-in-law Antoinette Brown Blackwell became the first ordained woman minister in the United States; and another sister-in-law, Lucy Stone, took the revolutionary step of retaining her maiden name after marriage. Stone became a successful and influential lecturer on women's rights. (Brinkley, 2015: p. 330)

RQ2

How Are Different Groups and Individuals **Described**?

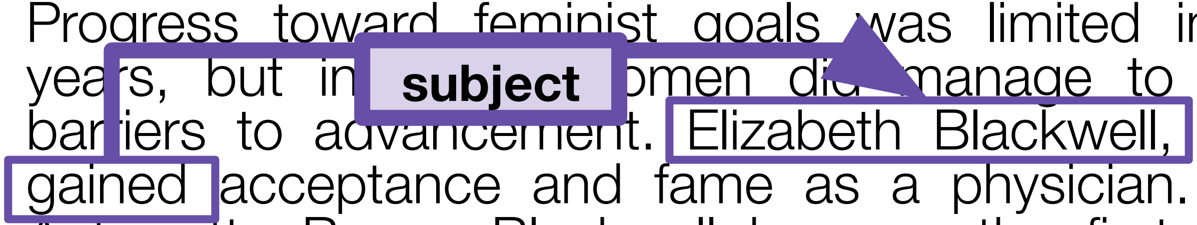
Progress toward feminist goals was limited in the antebellum years, but individual women did manage to break the social barriers to advancement. Elizabeth Blackwell, born in England, gained acceptance and fame as a physician. Her sister-in-law Antoinette Brown Blackwell became the first ordained woman minister in the United States; and another sister-in-law, Lucy Stone, took the revolutionary step of retaining her maiden name after marriage. Stone became a successful and influential lecturer on women's rights. (Brinkley, 2015: p. 330)

Dependency Parsing

RQ2

How Are Different Groups and Individuals **Described**?

Progress toward feminist goals was limited in the antebellum years, but individual women did manage to break the social barriers to advancement. Elizabeth Blackwell, born in England, gained acceptance and fame as a physician. Her sister-in-law Antoinette Brown Blackwell became the first ordained woman minister in the United States; and another sister-in-law, Lucy Stone, took the revolutionary step of retaining her maiden name after marriage. Stone became a successful and influential lecturer on women's rights. (Brinkley, 2015: p. 330)



Dependency Parsing

RQ2

How Are Different Groups and Individuals **Described**?

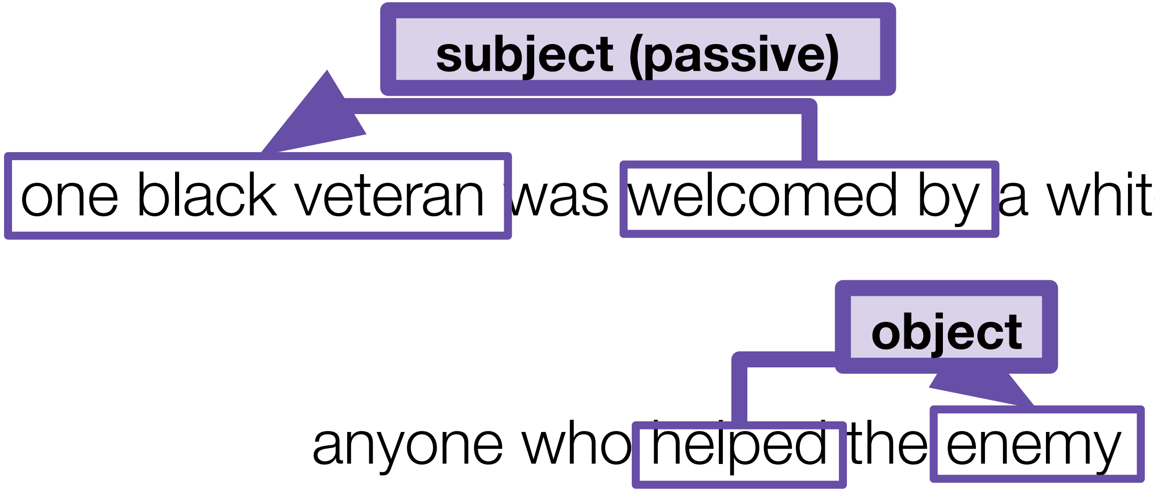
adj modifier

Progress toward feminist goals was limited in the antebellum years, but individual women did manage to break the social barriers to advancement. Elizabeth Blackwell, born in England, gained acceptance and fame as a physician. Her sister-in-law Antoinette Brown Blackwell became the first ordained woman minister in the United States; and another sister-in-law, Lucy Stone, took the revolutionary step of retaining her maiden name after marriage. Stone became a successful and influential lecturer on women’s rights. (Brinkley, 2015: p. 330)

Dependency Parsing

RQ2

How Are Different Groups and Individuals **Described**?



one black veteran was welcomed by a white neighbor

anyone who helped the enemy

Dependency Parsing

RQ2

Lexicons

adjectives

NRC Valence, Arousal, Dominance lexicons (Mohammad, 2018)

amazing (↑ valence) *asleep* (↓ arousal) *competitive* (↑ dominance)

RQ2

Lexicons

adjectives

NRC Valence, Arousal, Dominance lexicons (Mohammad, 2018)

amazing (↑ valence) *asleep* (↓ arousal) *competitive* (↑ dominance)

verbs

Connotation frames (Rashkin et al, 2016; Sap et al., 2017)

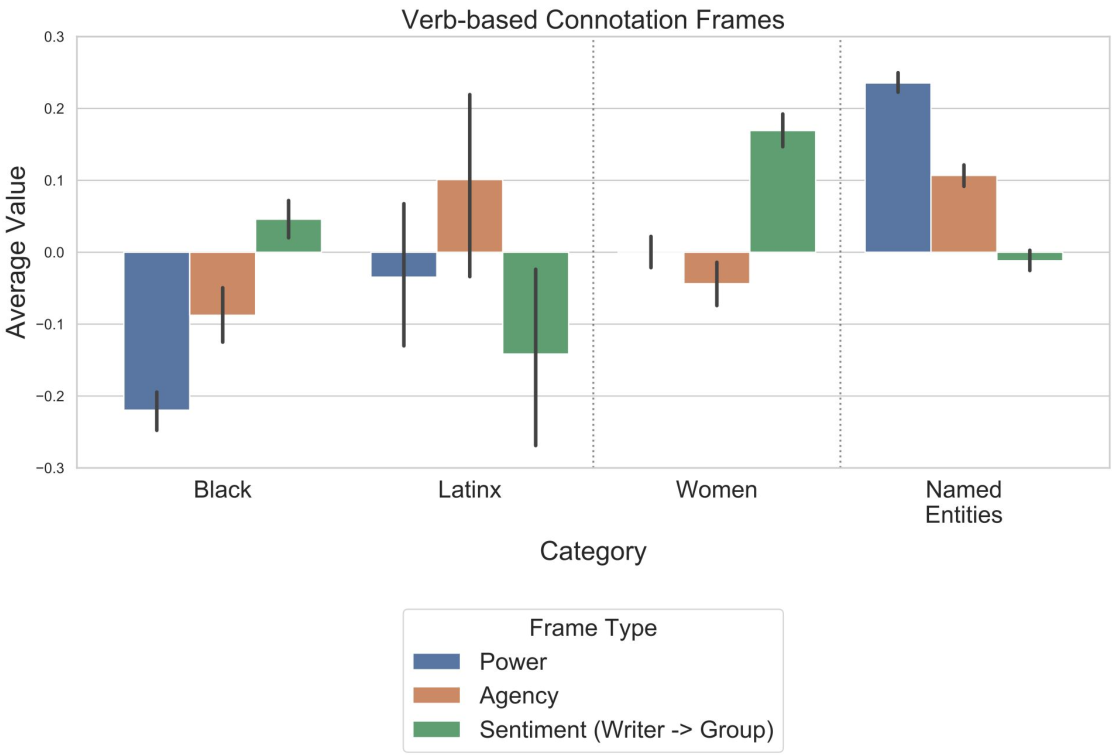
X (-1 agency) *obeys*

X (-1 power) *applauds Y* (+1 power)

X (↑ sentiment) *suffered*

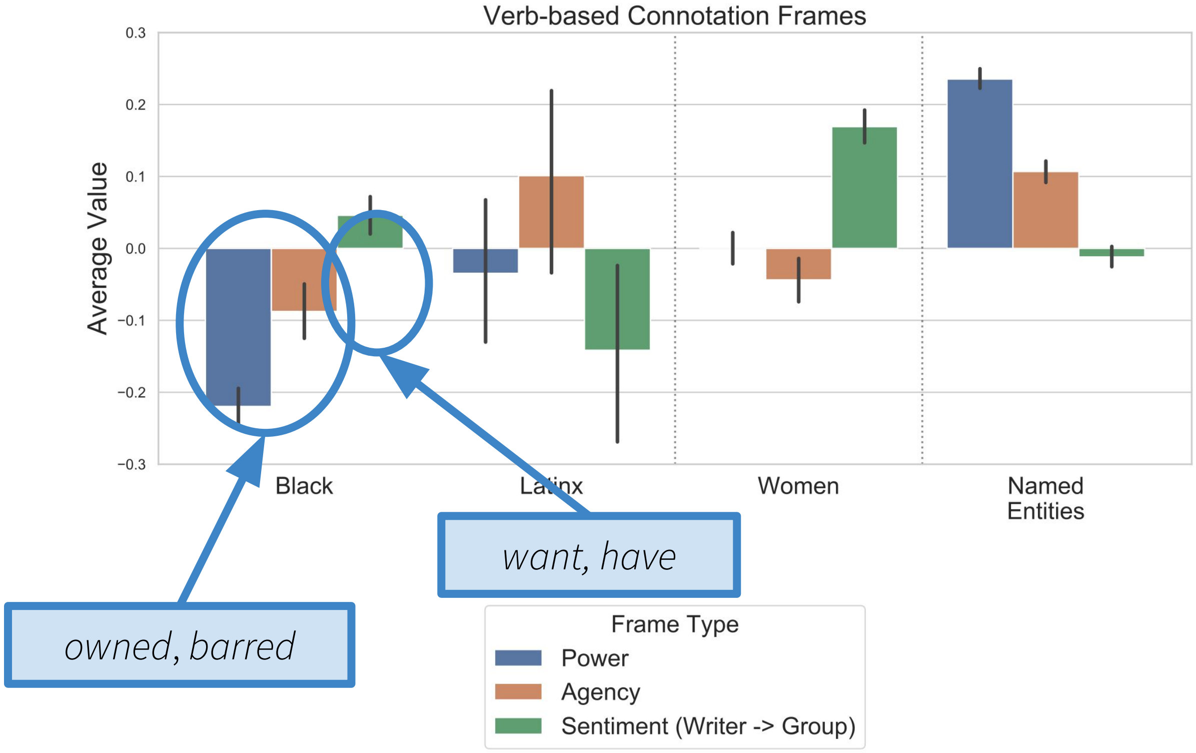
RQ2

Power & Agency



RQ2

Power & Agency

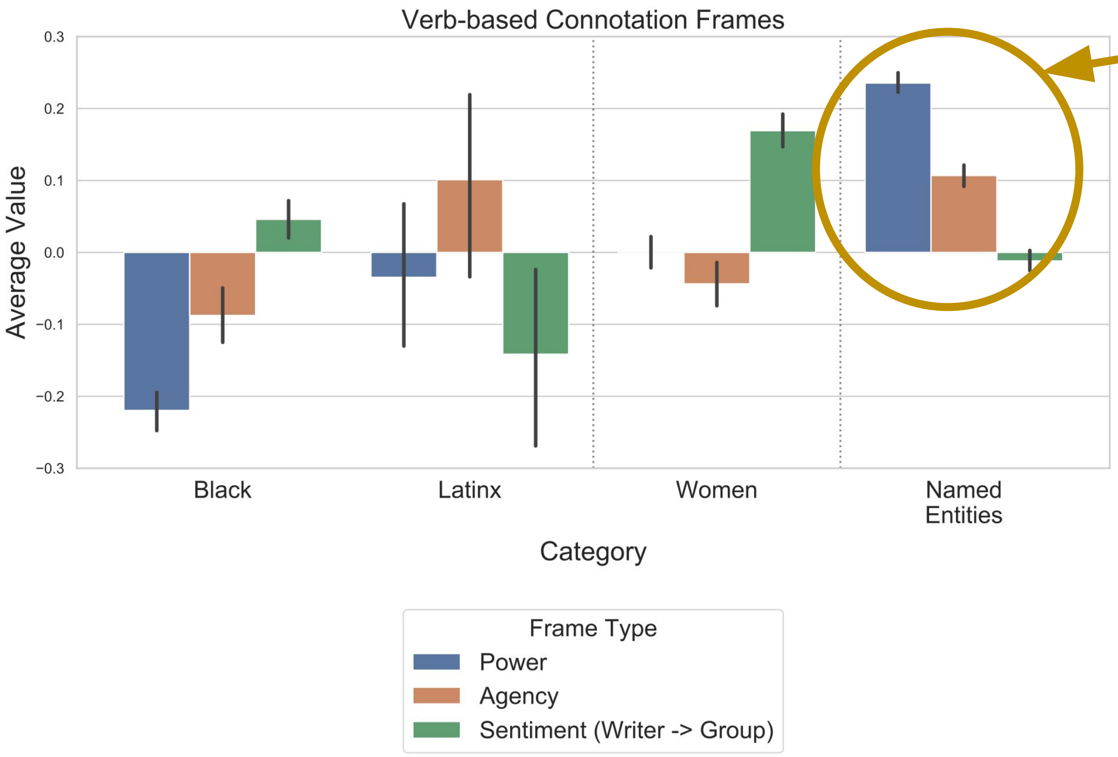


owned, barred

want, have

RQ2

Power & Agency



veto, initiate

RQ2

Other Lexicon Findings

African Americans (↓ adjective dominance)

Ex: *slave*, *inferior*

Famous people (↑ adjective arousal)

Ex: *worried*, *victorious*, *furious*

Women (↑ verb sentiment)

Ex: women *marry* or *help*

RQ2

GloVe Embeddings w/ Bootstrapping

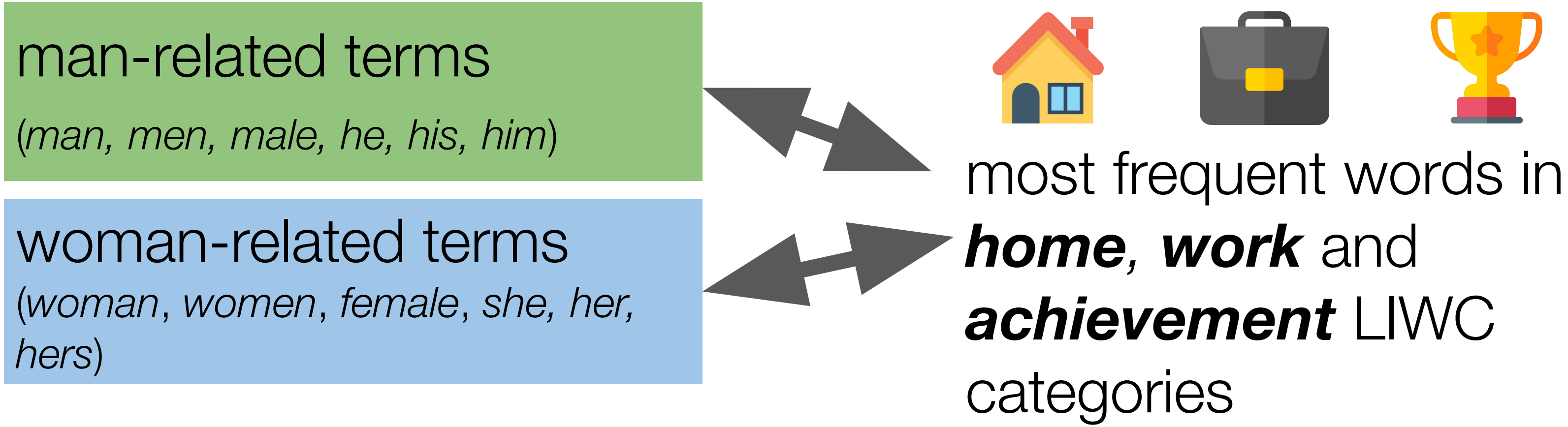
- unigrams & bigrams (skip stopwords)
- GloVe training w/ bootstrapping (Antoniak & Mimno, 2018)
- mean cosine similarity across 50 runs, between:

Bootstrapping helps mitigate data sparsity!
Create samples of the data (e.g. sample 50 times with replacement), train model on each and aggregate results.

RQ2

GloVe Embeddings w/ Bootstrapping

- unigrams & bigrams (skip stopwords)
- GloVe training w/ bootstrapping (Antoniak & Mimno, 2018)
- mean cosine similarity across 50 runs, between:



man-related terms

(*man, men, male, he, his, him*)

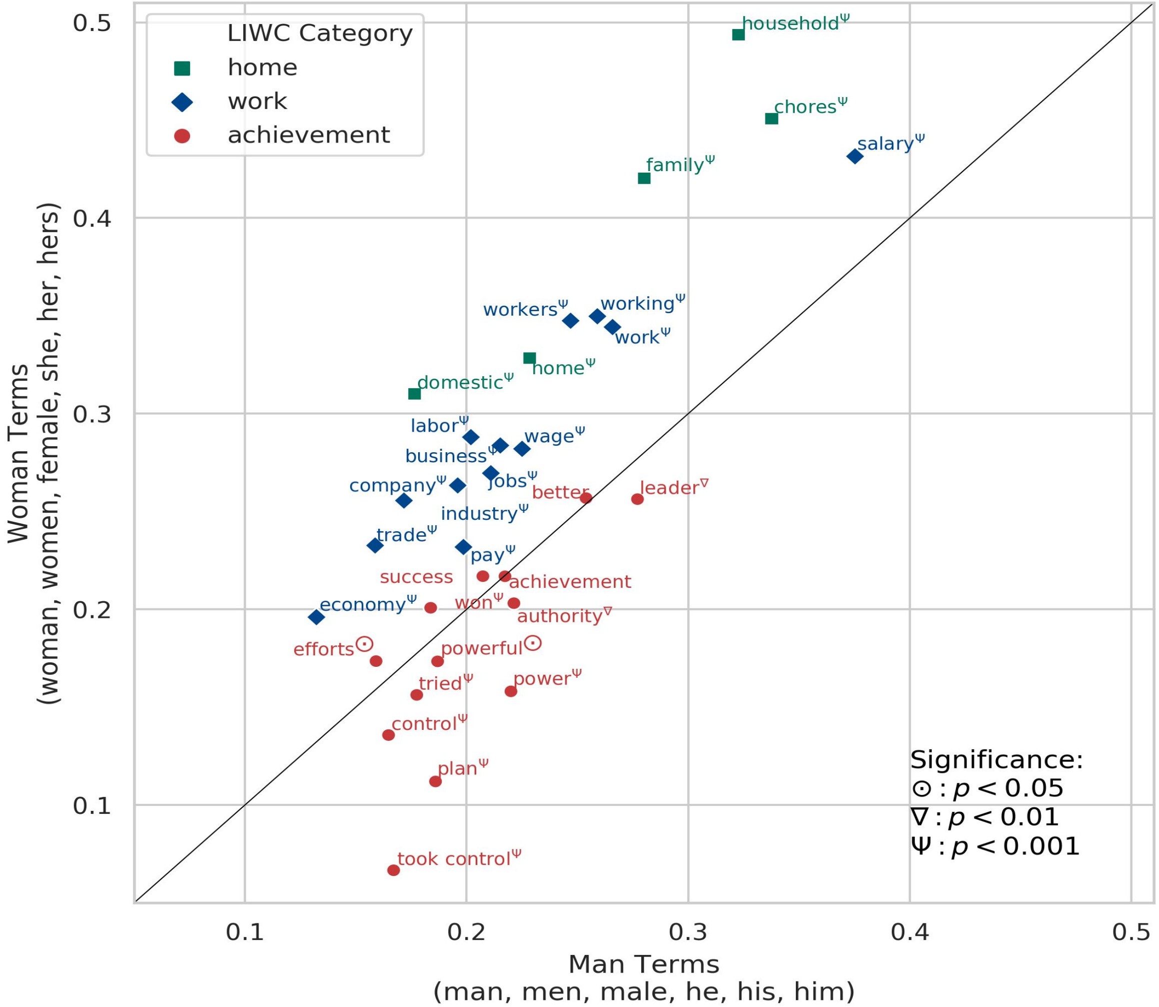
woman-related terms

(*woman*, *women*, *female*, *she*, *her*, *hers*)

most frequent words in ***home***, ***work*** and ***achievement*** LIWC categories

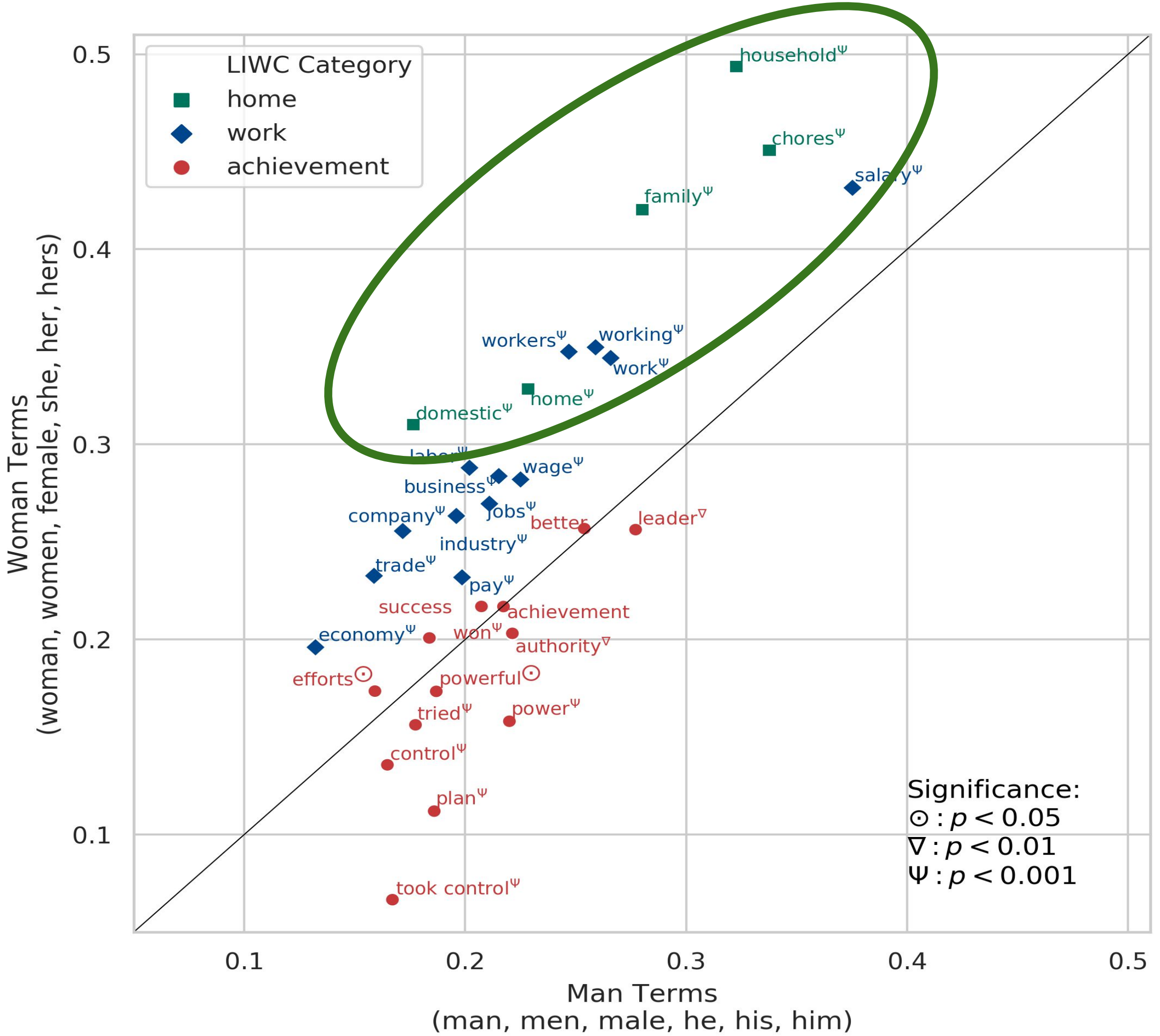
(Pennebaker et al., 2015)

RQ2


LIWC Category
🟩 home
◆ work
● achievement
household$^{\Psi}$
chores$^{\Psi}$
salary$^{\Psi}$
family$^{\Psi}$
workers$^{\Psi}$
working$^{\Psi}$
work$^{\Psi}$
home$^{\Psi}$
domestic$^{\Psi}$
labor$^{\Psi}$
wage$^{\Psi}$
business$^{\Psi}$
jobs$^{\Psi}$
company$^{\Psi}$
better
leader$^{\nabla}$
industry$^{\Psi}$
trade$^{\Psi}$
pay$^{\Psi}$
success
achievement
economy$^{\Psi}$
won$^{\Psi}$
authority$^{\nabla}$
efforts$^{\odot}$
powerful$^{\odot}$
tried$^{\Psi}$
power$^{\Psi}$
control$^{\Psi}$
plan$^{\Psi}$
took control$^{\Psi}$
Significance:
$\odot : p < 0.05$
$\nabla : p < 0.01$
$\Psi : p < 0.001$
Woman Terms
(woman, women, female, she, her, hers)
0.1
0.2
0.3
0.4
0.5
Man Terms
(man, men, male, he, his, him)
0.1
0.2
0.3
0.4
0.5

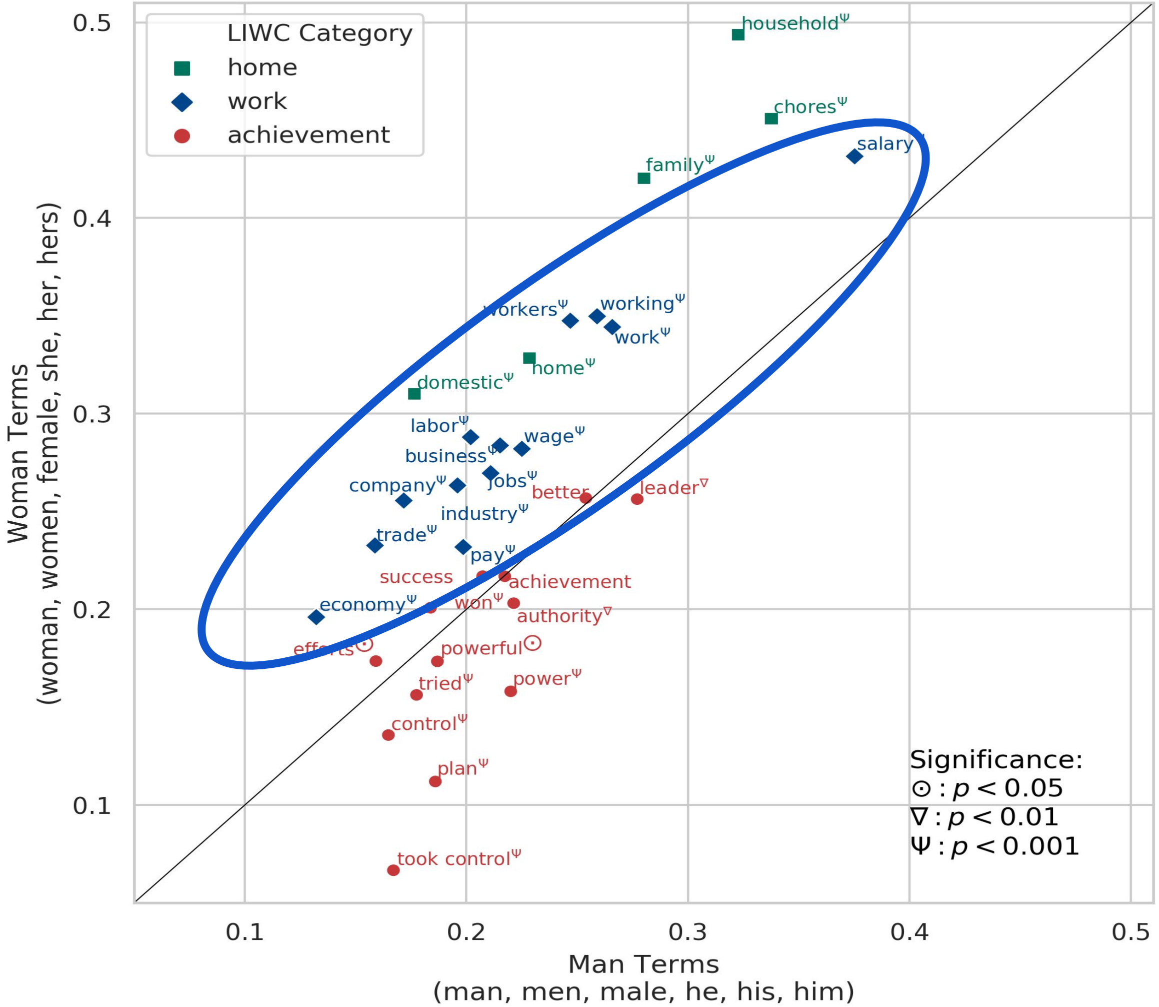
RQ2

Home related terms are more closely related to women, with very high significance 🏠



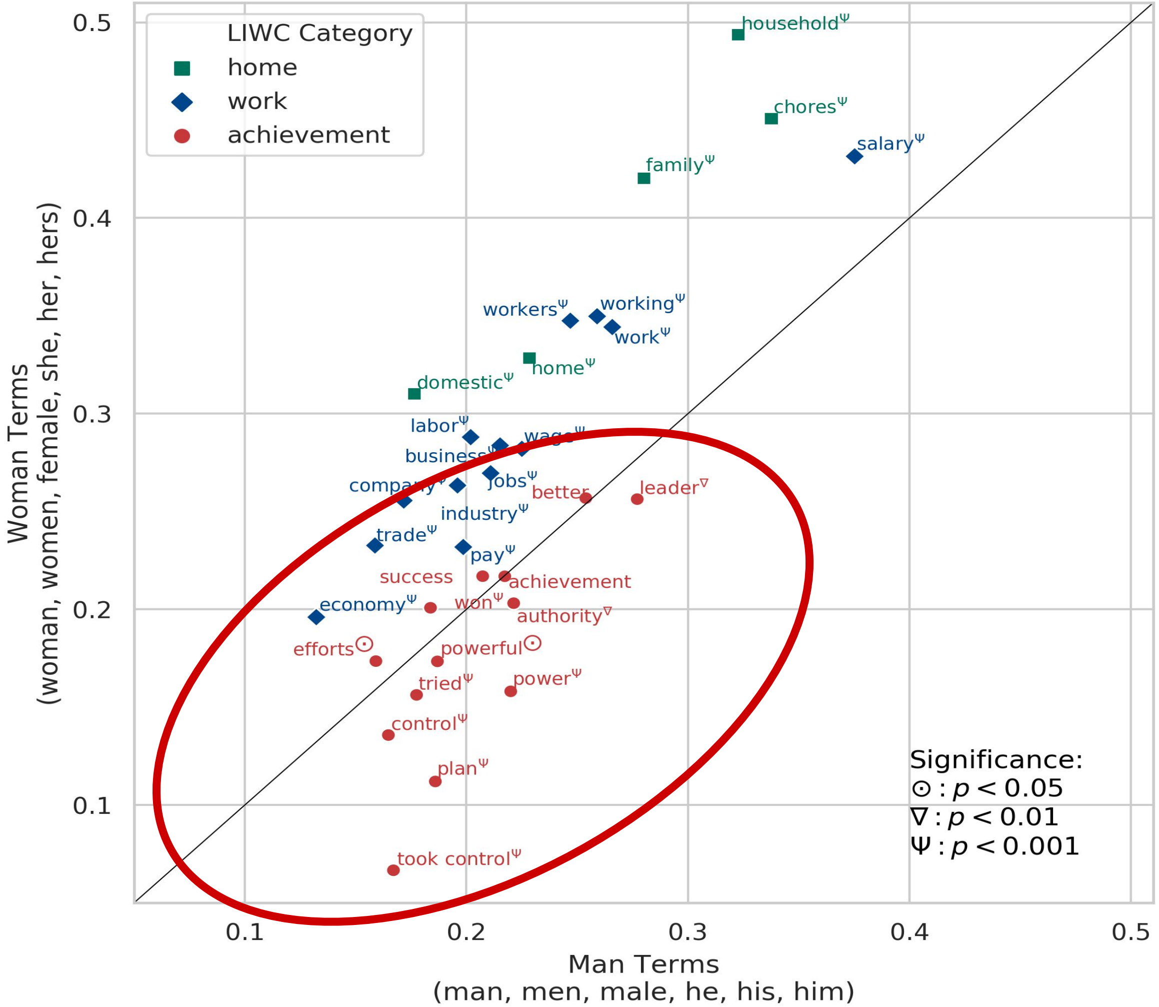
RQ2

Work related terms are more closely related to women, with very high significance



RQ2

Most **achievement** related terms are more closely related to men but not all



RQ3

Which **Topics** Are Prominent and How Do They Relate to Groups of People?

Progress toward feminist goals was limited in the antebellum years, but individual women did manage to break the social barriers to advancement. Elizabeth Blackwell, born in England, gained acceptance and fame as a physician. Her sister-in-law Antoinette Brown Blackwell became the first ordained woman minister in the United States; and another sister-in-law, Lucy Stone, took the revolutionary step of retaining her maiden name after marriage. Stone became a successful and influential lecturer on women's rights. (Brinkley, 2015: p. 330)

Topics

reform (reform,progress, social,...)

women's rights (women,right, movement,...)

religion (church,religion, christian,...)

marriage (women,men, young,...)

RQ3

Topic Modeling

- LDA (Blei & Jordan, 2003)
  - 50 topics, induced at the sentence-level
  - run together on all books

RQ3

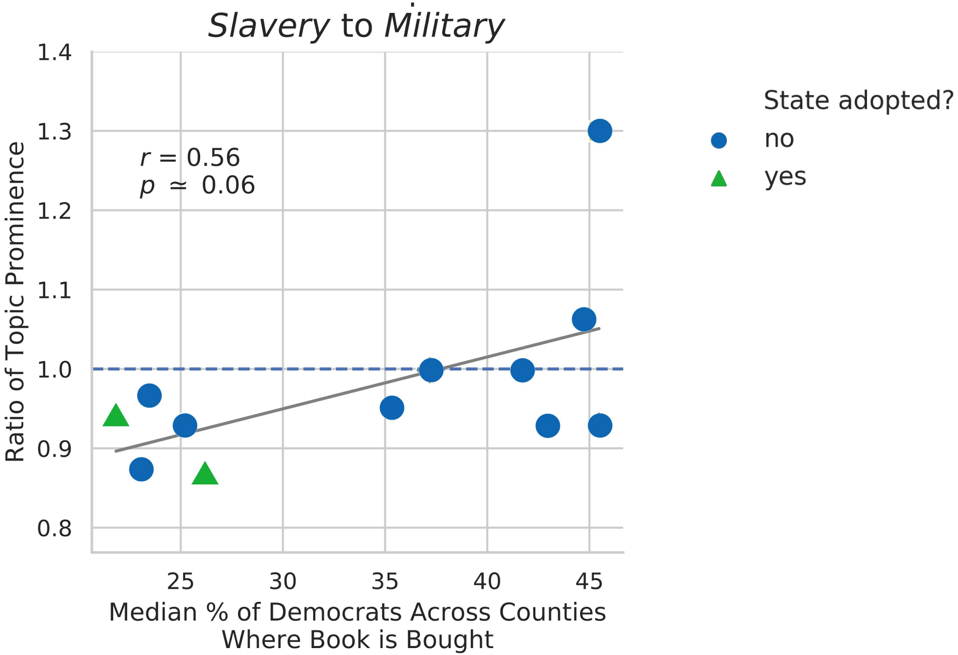
Comparing Topic Prominence Across Books

Using **ratio of relative frequencies** of topics helps control for noise

$$\frac{\text{Mean freq. of topic(s) related to } \textbf{X} \text{ in Book A}}{\text{Mean freq. of topic(s) related to } \textbf{Y} \text{ in Book A}}$$

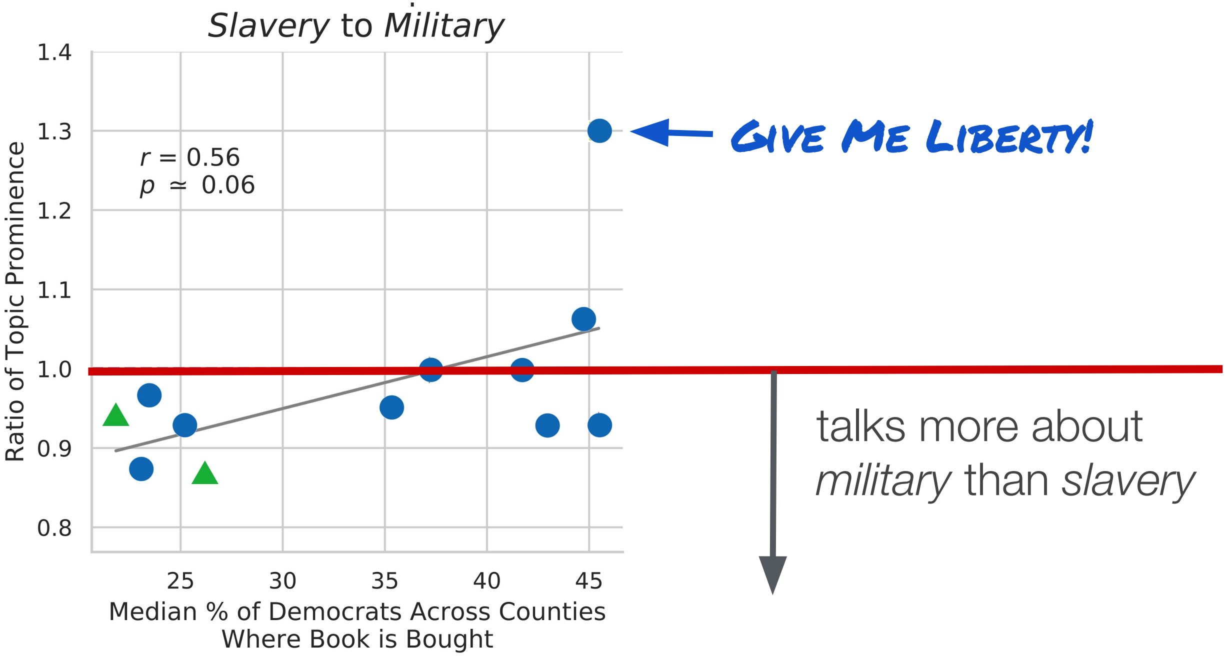
RQ3

Comparing Groups of Topics



RQ3

Besides One Outlier, The Ratio of These Topics is Similar Across Books

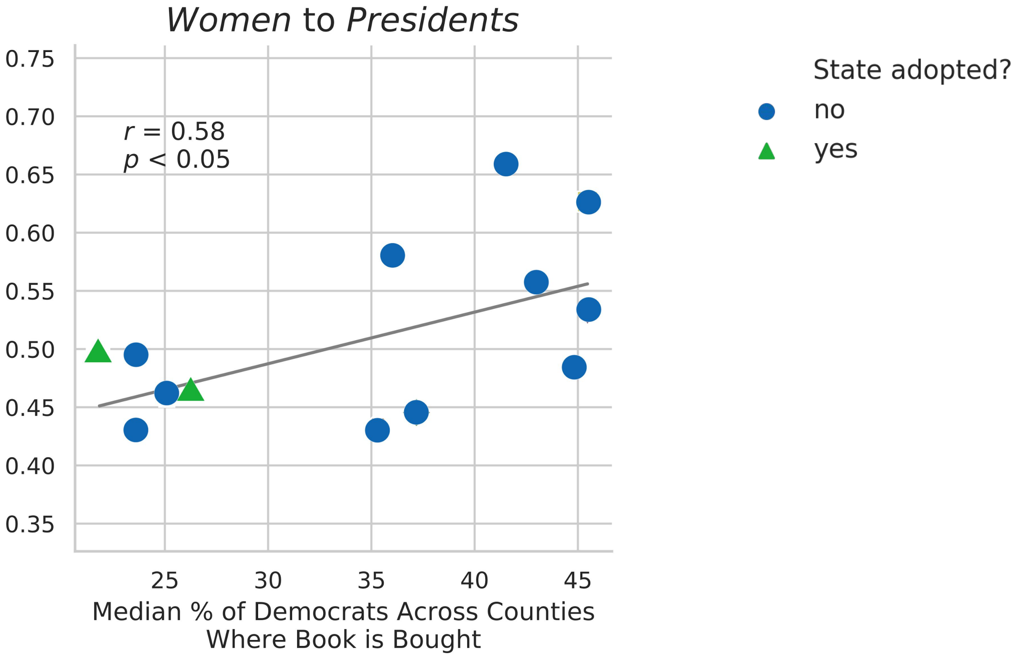


Give Me Liberty!

talks more about *military* than *slavery*

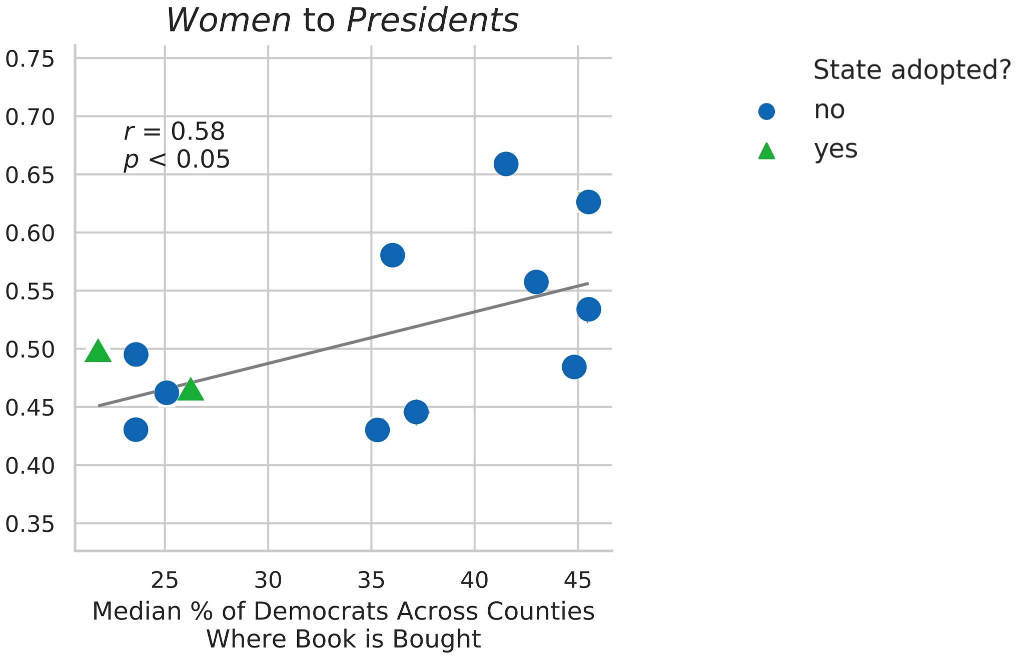
RQ3

Comparing Groups of Topics



RQ3

All Books Talk More about *Presidents* than about *Women*, but the Ratio is Closer to 1 in Books in Democratic Counties



Summary

Methods

Results

RQ1

How much are different groups of people **mentioned**?

Methods

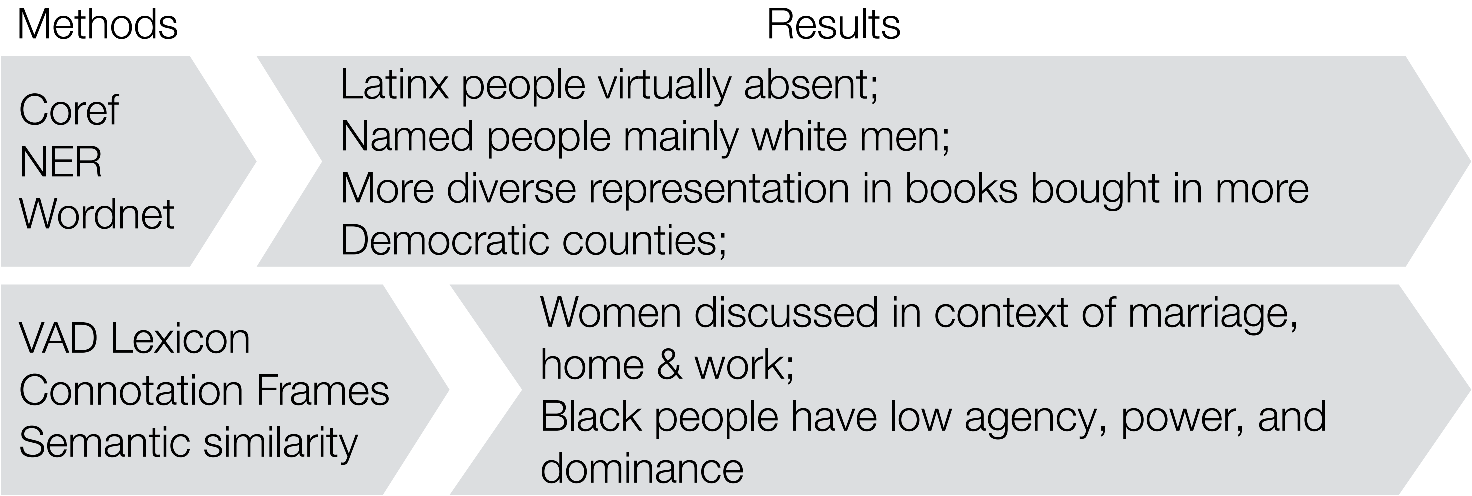
Coref
NER
Wordnet

Results

Latinx people virtually absent;
Named people mainly white men;
More diverse representation in books bought in more Democratic counties;

RQ2

How are different groups and individuals **described**?



Methods

Results

Coref
NER
Wordnet

Latinx people virtually absent;
Named people mainly white men;
More diverse representation in books bought in more Democratic counties;

VAD Lexicon
Connotation Frames
Semantic similarity

Women discussed in context of marriage, home & work;
Black people have low agency, power, and dominance

RQ3

Which **topics** are prominent and how do they relate to groups of people?

Methods	Results
Coref NER Wordnet	Latinx people virtually absent; Named people mainly white men; More diverse representation in books bought in more Democratic counties;
VAD Lexicon Connotation Frames Semantic similarity	Women discussed in context of marriage, home & work; Black people have low agency, power, and dominance
Topic modeling	Social history topics tend be more prominent in books bought in more Democratic counties, but similarities across books are greater than their differences

What questions do you have?

Discussion time!

Ben & Mahathi