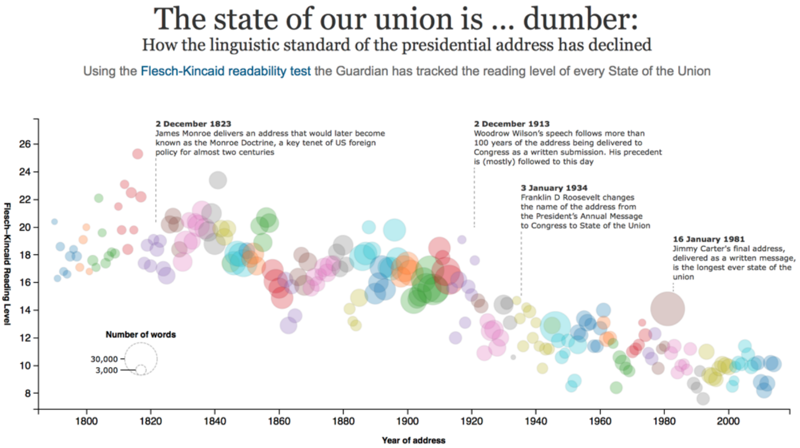
Measuring and Explaining Political Sophistication Through Textual Complexity

Kenneth Benoit Kevin Munger Arthur Spirling

SSRC Anxieties of Democracy Conference
Princeton October 28-29

Political sophistication in the public mind



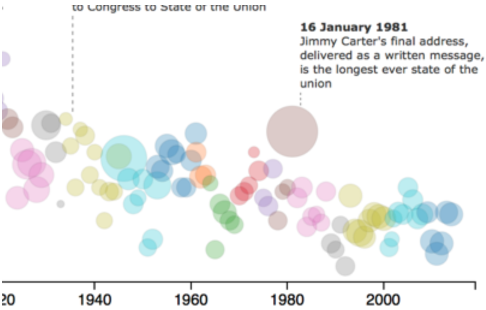
Source: The Guardian, February 2013

Does this make sense?

Does this make sense?

- Since 1913, SOTU given aloud

Post-1913

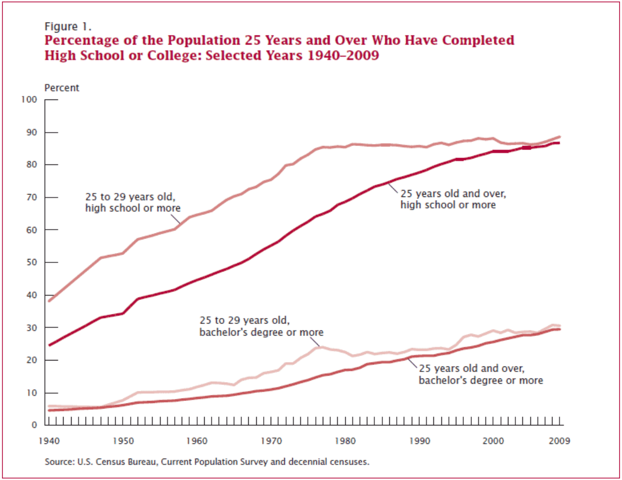


to Congress to State of the Union
16 January 1981
Jimmy Carter's final address, delivered as a written message, is the longest ever state of the union
20
1940
1960
1980
2000

Does this make sense?

- Since 1913, SOTU given aloud
- Audiences becoming *more* sophisticated, better-educated

Education



Camille L. Ryan and Julie Siebens - U.S. Census Bureau

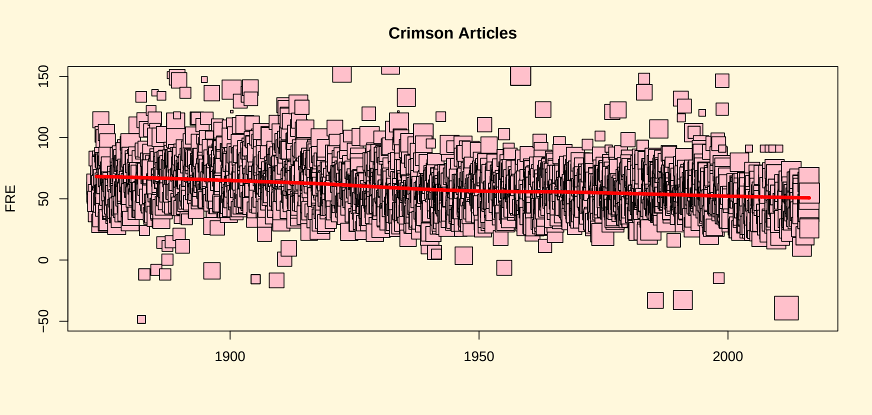
Does this make sense?

- Since 1913, SOTU given aloud
- Audiences becoming *more* sophisticated, better-educated
- Political issues more numerous and complicated

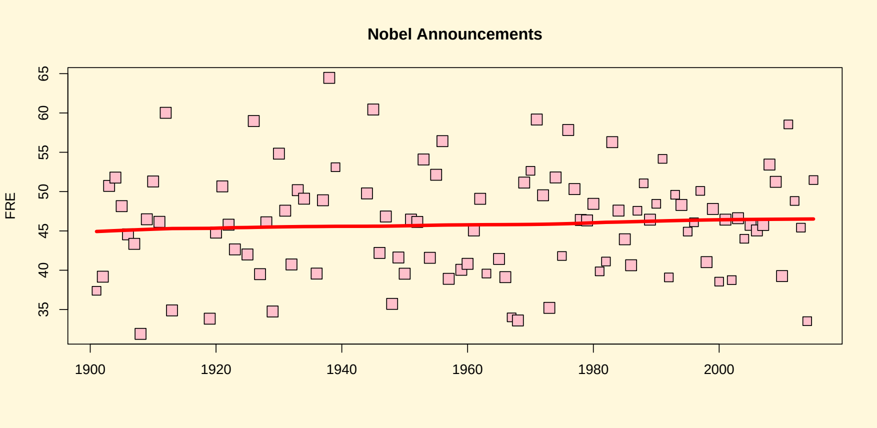
Does this make sense?

- Since 1913, SOTU given aloud
- Audiences becoming *more* sophisticated, better-educated
- Political issues more numerous and complicated
- Is this a general trend?

Is this a general trend?



Is this a general trend?



Does this make sense?

- Since 1913, SOTU given aloud
- Audiences becoming *more* sophisticated, better-educated
- Political issues more numerous and complicated
- Is this a general trend?
- **What exactly are we measuring?**

Existing Measures

Name of Method	Author	Year	Citations
Flesch Reading Ease	Flesch	1948/49	3,793
SMOG	McLaughlin	1969	1,402
Dale-Chall	Dale and Chall	1948	1,389
Gunning Fog Index	Gunning	1952	1,232
Flesch-Kincaid Level	Kincaid et al	1975	1,093
Fry Graph	Fry	1968	1,007
Spache Formula	Spache	1953	355
Coleman-Liau	Coleman and Liau	1975	261

Commonly used ‘reading ease’ measures in order of citation via Google scholar at the time of writing.

Exploring a Measure: Flesch Reading Ease

- Flesch Reading Ease (FRE) Score
  - Developed to measure average grade level of students based on ability to answer multiple-choice questions after reading a text

Exploring a Measure: Flesch Reading Ease

- Flesch Reading Ease (FRE) Score
  - Developed to measure average grade level of students based on ability to answer multiple-choice questions after reading a text
  - In 1948

Exploring a Measure: Flesch Reading Ease

- Flesch Reading Ease (FRE) Score
  - Developed to measure average grade level of students based on ability to answer multiple-choice questions after reading a text
  - In 1948

$$206.835 - 1.015\left(\frac{\#\text{ of words}}{\#\text{ of sentences}}\right) - 84.6\left(\frac{\#\text{ of syllables}}{\#\text{ of words}}\right)$$

Breaking the FRE Score

- Consider this sentence
    - Indeed, the shoemaker was frightened.

Breaking the FRE Score

- Consider this sentence
  - Indeed, the shoemaker was frightened.
  - FRE $= 16.23$

Breaking the FRE Score

- Consider this sentence
  - Indeed, the shoemaker was frightened.
  - FRE $= 16.23$
  - Forsooth, the cordwainer was afeared.

Breaking the FRE Score

- Consider this sentence
  - Indeed, the shoemaker was frightened.
  - FRE $= 16.23$
  - Forsooth, the cordwainer was afeared.
  - FRE $= 16.23$

Breaking the FRE Score

- Consider this sentence
  - Indeed, the shoemaker was frightened.
  - FRE $=$ 16.23
  - Forsooth, the cordwainer was afeared.
  - FRE $=$ 16.23
- No measure of the difficulty of the words (or any other grammatical challenges)

Breaking the FRE Score

- Consider this sentence
    - Indeed, the shoemaker was frightened.
    - FRE $= 16.23$
    - Forsooth, the cordwainer was afeared.
    - FRE $= 16.23$
- No measure of the difficulty of the words (or any other grammatical challenges)
- Is this really the quantity we're interested in?

Political Communication and Textual Complexity

- Citizen comprehension of political speech

Political Communication and Textual Complexity

- Citizen comprehension of political speech
- Changes over time, differences between speakers:

Political Communication and Textual Complexity

- Citizen comprehension of political speech
- Changes over time, differences between speakers:
  - We need to collect judicial and (especially) legislative speech over time

Political Communication and Textual Complexity

- Citizen comprehension of political speech
- Changes over time, differences between speakers:
  - We need to collect judicial and (especially) legislative speech over time
- Problems with existing measures of textual complexity

Political Communication and Textual Complexity

- Citizen comprehension of political speech
- Changes over time, differences between speakers:
  - We need to collect judicial and (especially) legislative speech over time
- Problems with existing measures of textual complexity
- Preview of our solution:

Political Communication and Textual Complexity

- Citizen comprehension of political speech
- Changes over time, differences between speakers:
  - We need to collect judicial and (especially) legislative speech over time
- Problems with existing measures of textual complexity
- Preview of our solution:
  - Crowdsource comparisons of relevant political text

Political Communication and Textual Complexity

- Citizen comprehension of political speech
- Changes over time, differences between speakers:
  - We need to collect judicial and (especially) legislative speech over time
- Problems with existing measures of textual complexity
- Preview of our solution:
  - Crowdsource comparisons of relevant political text
  - Scale those texts and learn what features best predict easiness

Political Communication and Textual Complexity

- Citizen comprehension of political speech
- Changes over time, differences between speakers:
  - We need to collect judicial and (especially) legislative speech over time
- Problems with existing measures of textual complexity
- Preview of our solution:
  - Crowdsource comparisons of relevant political text
  - Scale those texts and learn what features best predict easiness
  - Fit a model that can be applied to other texts

The "Out-of-Domain" prediction problem

We want to measure how well adult citizens are able to understand political texts. Previous measures were:

The “Out-of-Domain” prediction problem

We want to measure how well adult citizens are able to understand political texts. Previous measures were:

- designed for educational research, not political texts;

The “Out-of-Domain” prediction problem

We want to measure how well adult citizens are able to understand political texts. Previous measures were:

- designed for educational research, not political texts;
- validated on schoolchildren, not adults; and

The "Out-of-Domain" prediction problem

We want to measure how well adult citizens are able to understand political texts. Previous measures were:

- designed for educational research, not political texts;
- validated on schoolchildren, not adults; and
- mostly designed in the 1940s and 50s, which is a long time ago.

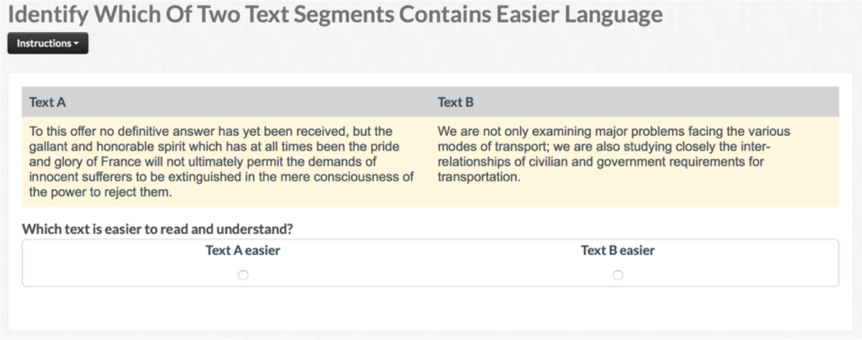
The "Out-of-Domain" prediction problem

We want to measure how well adult citizens are able to understand political texts. Previous measures were:

- designed for educational research, not political texts;
- validated on schoolchildren, not adults; and
- mostly designed in the 1940s and 50s, which is a long time ago.

These problems are straightforward to fix.

A modern solution: crowdsourcing binary comparisons



Crowdflower specifics

1. Created pairwise comparisons between 2,000 randomly sampled snippets from the SOTU corpus, with coarse matching on snippet length and FRE score

Crowdflower specifics

1. Created pairwise comparisons between 2,000 randomly sampled snippets from the SOTU corpus, with coarse matching on snippet length and FRE score
2. Coded these comparisons three separate times, so 6,000 total data points

Problems for inference

Extant measures have undesirable statistical properties.

Problems for inference

Extant measures have undesirable statistical properties.

1. No way of evaluating "model fit" for measures applied to a new context

Problems for inference

Extant measures have undesirable statistical properties.

1. No way of evaluating "model fit" for measures applied to a new context

    - No way of comparing different measures in a given context

Problems for inference

Extant measures have undesirable statistical properties.

1. No way of evaluating "model fit" for measures applied to a new context
   - No way of comparing different measures in a given context
2. No natural interpretation of fine-grained differences in document scores

Problems for inference

Extant measures have undesirable statistical properties.

1. No way of evaluating "model fit" for measures applied to a new context
   - No way of comparing different measures in a given context
2. No natural interpretation of fine-grained differences in document scores
   - eg FRE of 75 vs 70 vs 80

Problems for inference

Extant measures have undesirable statistical properties.

1. No way of evaluating "model fit" for measures applied to a new context
    - No way of comparing different measures in a given context
2. No natural interpretation of fine-grained differences in document scores
    - eg FRE of 75 vs 70 vs 80

We can model this!

Our approach: Bradley-Terry Regression

1. Consider determining which of two texts, i and j, is "easier"

Our approach: Bradley-Terry Regression

1. Consider determining which of two texts, i and j, is "easier"
2. Defining $\lambda_i = \log \alpha_i$, the regression model can be rewritten:

$$\text{logit}[\Pr(i \text{ easier than } j)] = \lambda_i - \lambda_j$$

Our approach: Bradley-Terry Regression

1. Consider determining which of two texts, i and j, is "easier"
2. Defining $\lambda_i = \log \alpha_i$, the regression model can be rewritten:

$$\text{logit}[\Pr(i \text{ easier than } j)] = \lambda_i - \lambda_j$$

3. Using only the labels from crowdsourcing, we fit an unstructured Bradley Terry model to scale the snippets and generate a rank ordering and λ score for each

Variable selection

- We begin with all constituent variables of the traditional models, add in some new ones

Variable selection

- We begin with all constituent variables of the traditional models, add in some new ones
- 29 possible variables

Variable selection

- We begin with all constituent variables of the traditional models, add in some new ones
- 29 possible variables
- Use a machine learning technique called random forests to select the variables that best fit the snippets scaled through unstructured Bradley-Terry regression

Word Rarity: Google nGram

- A collection of word counts in the Google books corpus

Word Rarity: Google nGram

- A collection of word counts in the Google books corpus
- Word frequency by year, smoothing by decade

Word Rarity: Google nGram

- A collection of word counts in the Google books corpus
- Word frequency by year, smoothing by decade
- Word frequency in the 2000s–the closest decade to the present–to measure the presence of words that are rare from the perspective of our coders

Structured Bradley-Terry Model

- We have our covariates

Structured Bradley-Terry Model

- We have our covariates
- We can model λ_i as a function of the covariates r that we selected using a structured Bradley-Terry model:

$$\lambda_i = \sum_{r=1}^{p} \beta_r x_{ir}$$

Results

	Simple Model
Characters per sentence	-0.01^{*}
	(0.00)
Proportion of 3-syllable words	-1.31^{*}
	(0.28)
Proportion of adpositions	-1.11^{*}
such as *to*, *with*, *from*, *under*	(0.46)
Mean word frequency (/'the')	-1.68^{*}
	(0.35)
Percent Correctly Predicted	0.662

Standard errors in parentheses. $*$ *indicates significance at* $p < 0.05$

Evaluating traditional measures

We can check the predictive ability of extant measures on our ranked snippets

	% Correct
FRE	0.602
Dale-Chall	0.603
FOG	0.638
SMOG	0.574
Spache	0.635
Coleman-Liau	0.552

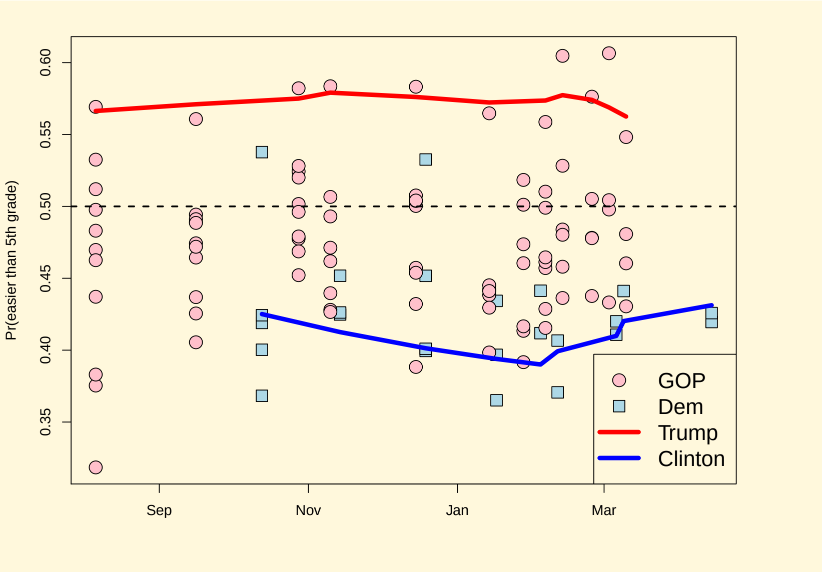
Structured Bradley-Terry Model

- We have our covariates
- We can model λ_i as a function of the covariates r that we selected using a structured Bradley-Terry model:

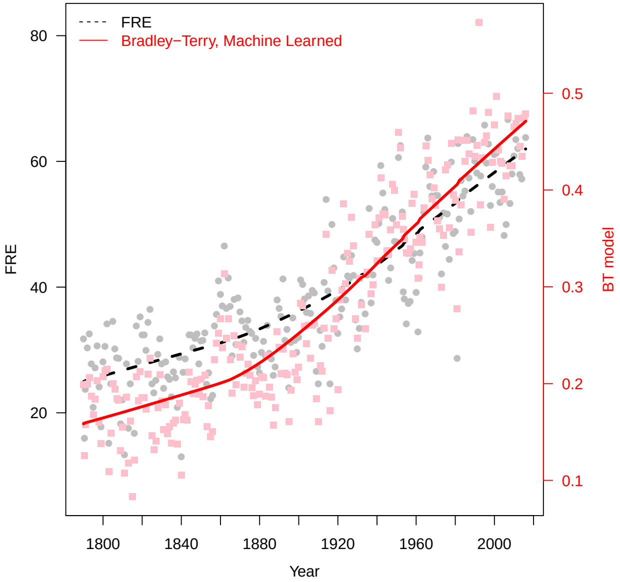
$$\lambda_i = \sum_{r=1}^{p} \beta_r x_{ir}$$

- We have estimated the relevant $\hat{\beta}_r$'s and can then "plug in" covariates to evaluate other texts

Speeches in 2016 Campaign Debates



SOTU Re-evaluated



Conclusions

- Political discourse may well be getting dumber

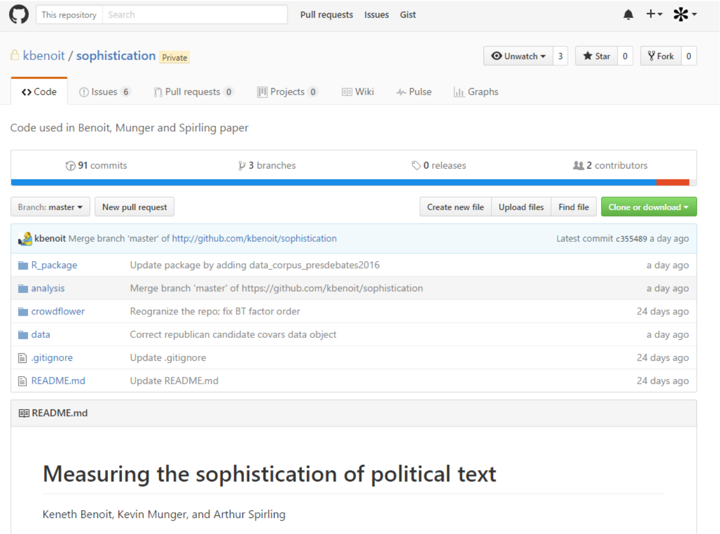
Conclusions

- Political discourse may well be getting dumber but now we're aware of the difficulty in saying so

Conclusions

- Political discourse may well be getting dumber but now we're aware of the difficulty in saying so
- General lesson is not to draw strong conclusions from measures applied out of domain

R package


This repository
Search
Pull requests
Issues
Gist
kbenoit / sophistication
Private
Unwatch 3
Star 0
Fork 0
Code
Issues 6
Pull requests 0
Projects 0
Wiki
Pulse
Graphs
Code used in Benoit, Munger and Spirling paper
91 commits
3 branches
0 releases
2 contributors
Branch: master
New pull request
Create new file
Upload files
Find file
Clone or download
kbenoit Merge branch 'master' of http://github.com/kbenoit/sophistication
Latest commit c355489 a day ago
R_package
Update package by adding data_corpus_presdebates2016
a day ago
analysis
Merge branch 'master' of https://github.com/kbenoit/sophistication
a day ago
crowdflower
Reogranize the repo; fix BT factor order
24 days ago
data
Correct republican candidate covars data object
a day ago
.gitignore
Update .gitignore
24 days ago
README.md
Update README.md
24 days ago
README.md
Measuring the sophistication of political text
Keneth Benoit, Kevin Munger, and Arthur Spirling