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

The state of our union is ... dumber: How the linguistic standard of the presidential address has declined

Using the Flesch-Kincaid readability test the Guardian has tracked the reading level of every State of the Union
Does this make sense?
Since 1913, SOTU given aloud
16 January 1981
Jimmy Carter’s final address, delivered as a written message, is the longest ever state of the union.
Since 1913, SOTU given aloud

Audiences becoming *more* sophisticated, better-educated
Figure 1. *Percentage of the Population 25 Years and Over Who Have Completed High School or College: Selected Years 1940–2009*


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
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?
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

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

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

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

Indeed, the shoemaker was frightened.
Breaking the FRE Score

Consider this sentence

Indeed, the shoemaker was frightened.

FRE = 16.23
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

No measure of the difficulty of the words (or any other grammatical challenges) is this really the quantity we're interested in?
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)
Consider this sentence

- Indeed, the shoemaker was frightened.
  \[ \text{FRE} = 16.23 \]

- Forsooth, the cordwainer was afeared.
  \[ \text{FRE} = 16.23 \]

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
Citizen comprehension of political speech

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

Changes over time, differences between speakers:

We need to collect judicial and (especially) legislative speech over time
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:
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:
  
  Crowdsourcing comparisons of relevant political text
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
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:

- Crowdsourced 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.
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**

[The text is a comparison task where participants are asked to identify which of two text segments contains easier language. The left segment reads: **Text A**

To this offer no definitive answer has yet been received, but the gallant and honorable spirit which has at all times been the pride and glory of France will not ultimately permit the demands of innocent sufferers to be extinguished in the mere consciousness of the power to reject them.

The right segment reads: **Text B**

We are not only examining major problems facing the various modes of transport; we are also studying closely the inter-relationships of civilian and government requirements for transportation.

Participants are asked to choose which text is easier to read and understand.]
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!
Consider determining which of two texts, \( i \) and \( j \), is “easier”
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:

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

3. Using only the labels from crowdsourcing, we fit an unstructured Bradley Terry model to scale the snippets and generate a rank ordering and $\lambda$ 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
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
We have our covariates

We can model $\lambda_i$ 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

<table>
<thead>
<tr>
<th>Character</th>
<th>Simple Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characters per sentence</td>
<td>−0.01*</td>
</tr>
<tr>
<td>Proportion of 3-syllable words</td>
<td>−1.31*</td>
</tr>
<tr>
<td>Proportion of adpositions</td>
<td>−1.11*</td>
</tr>
<tr>
<td>such as <em>to</em>, <em>with</em>, <em>from</em>, <em>under</em></td>
<td>(0.46)</td>
</tr>
<tr>
<td>Mean word frequency ('the')</td>
<td>−1.68*</td>
</tr>
<tr>
<td>Per cent Correctly Predicted</td>
<td>0.662</td>
</tr>
</tbody>
</table>

*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

<table>
<thead>
<tr>
<th></th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRE</td>
<td>0.602</td>
</tr>
<tr>
<td>Dale-Chall</td>
<td>0.603</td>
</tr>
<tr>
<td>FOG</td>
<td>0.638</td>
</tr>
<tr>
<td>SMOG</td>
<td>0.574</td>
</tr>
<tr>
<td>Spache</td>
<td>0.635</td>
</tr>
<tr>
<td>Coleman-Liau</td>
<td>0.552</td>
</tr>
</tbody>
</table>
Structured Bradley-Terry Model

- We have our covariates

- We can model $\lambda_i$ 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

Pr(easier than 5th grade)

GOP
Dem
Trump
Clinton
SOTU Re-evaluated

FRE

Bradley–Terry, Machine Learned

Year

1800 1840 1880 1920 1960 2000

FRE

0.1 0.2 0.3 0.4 0.5

BT model

1800 1840 1880 1920 1960 2000

Year
Conclusions

- Political discourse may well be getting dumber
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

Code used in Benoit, Munger and Spirling paper

- **91 commits**
- **3 branches**
- **0 releases**
- **2 contributors**

Branch: master — New pull request

- **kbnoit** Merge branch 'master' of http://github.com/kbnoit/sophistication
- **R_package** Update package by adding data_corpus_presdebates2016
- **analysis** Merge branch 'master' of https://github.com/kbnoit/sophistication
- **crowdfower** Reorganize the repo: fix BT factor order
- **data** Correct republican candidate covars data object
- **.gitignore** Update .gitignore
- **README.md** Update README.md

---

**Measuring the sophistication of political text**

Keneth Benoit, Kevin Munger, and Arthur Spirling