Measuring and Explaining Political Sophistication Through Textual Complexity

Kenneth Benoit  Kevin Munger  Arthur Spirling

New Directions in Analyzing Text as Data Conference
Northeastern University, October 14-15, 2016
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

Source: The Guardian, February 2013
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

Problems with existing measures of textual complexity
Political Communication and Textual Complexity

- Citizen comprehension of political speech
- Changes over time, differences between speakers
- Problems with existing measures of textual complexity
- Preview of our solution:
Citizen comprehension of political speech
Changes over time, differences between speakers
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
- 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
Political Communication and Textual Complexity

- Citizen comprehension of political speech
- Changes over time, differences between speakers
- 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
Other measures of reading ease

<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

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

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

Ostensibly bounded between 0 and 100.
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

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

- Ostensibly bounded between 0 and 100
- Updated by Kincaid et al. 1975 as a linear rescaling to US grade school level
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

No measure of the difficulty of the words (or any other grammatical challenges) is this really the quantity we're interested in?
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.
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?
The “Out-of-Domain” prediction problem

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

Identify Which Of Two Text Segments Contains Easier Language

Overview

Below, you will see two short passages of text.

Your task is to read two short passages of text, and to judge which you think would be easier for a native English speaker to read and understand. An easier text is one that takes a reader less time to comprehend fully, requires less re-reading, and can be more easily understood by someone with a lower level of education and language ability.

If you think text A is easier, click the button next to A. If you think text B is easier, click the button next to B. In every case, you must make a decision: there is no 'equal' or 'don't know' option. You cannot move to the next question until you give a response.

The tasks contain some "gold" questions that our internal panel judged to be very clear, but ultimately we want to leave this question open, so the threshold for getting the test questions correct is set at just 60%. Watch out for screener questions (see last Example 4 below). These re designed to ensure that you read all of the text carefully!

Example 1

<table>
<thead>
<tr>
<th>Text A</th>
<th>Text B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Special relativity (or the special theory of relativity) is a theory in physics that was developed and explained by Albert Einstein in 1905 because of some weaknesses that had been discovered in older physics. For example, older physics could not explain the fact that the speed of light never changes, even when the observer (the person looking at it) is moving toward the light source, or when the light source is moving toward the observer. Einstein figured out that this was because older physical theories only considered one group of observers and assumed that their view point, or reference frame, was the &quot;right&quot; one.</td>
<td>In physics, special relativity (SR, also known as the special theory of relativity or STR) is the generally accepted physical theory regarding the relationship between space and time. It is based on two postulates: (1) that the laws of physics are invariant (i.e. identical) in all inertial systems (non-accelerating frames of reference); and (2) that the speed of light in a vacuum is the same for all observers, regardless of the motion of the light source. It was originally proposed in 1905 by Albert Einstein in the paper &quot;On the Electrodynamics of Moving Bodies&quot;.</td>
</tr>
</tbody>
</table>

Answer: Text A is easier.
Crowdflower specifics

1. Formed all possible 1- and 2- sentence snippets from the SOTU corpus
Crowdflower specifics

1. Formed all possible 1- and 2- sentence snippets from the SOTU corpus

2. Discarded those with extreme FRE scores, and those containing large numbers
Crowdflower specifics

1. Formed all possible 1- and 2- sentence snippets from the SOTU corpus

2. Discarded those with extreme FRE scores, and those containing large numbers

3. Created 10,000 pairwise comparisons between 2,000 randomly sampled snippets, with coarse matching on snippet length and FRE score
Crowdflower specifics

1. Formed all possible 1- and 2- sentence snippets from the SOTU corpus

2. Discarded those with extreme FRE scores, and those containing large numbers

3. Created 10,000 pairwise comparisons between 2,000 randomly sampled snippets, with coarse matching on snippet length and FRE score
   - sufficient connectivity that we could scale all of them
Crowdflower specifics

1. Formed all possible 1- and 2- sentence snippets from the SOTU corpus

2. Discarded those with extreme FRE scores, and those containing large numbers

3. Created 10,000 pairwise comparisons between 2,000 randomly sampled snippets, with coarse matching on snippet length and FRE score

   - sufficient connectivity that we could scale all of them

4. Coded 2,000 of these comparisons three separate times, so 6,000 total data points
A modern solution: crowdsourcing binary comparisons

Identify Which Of Two Text Segments Contains Easier Language

**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.

**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.

Which text is easier to read and understand?

- Text A easier
- Text B easier
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

3. No standard errors or other measure of uncertainty
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

3. No standard errors or other measure of uncertainty

We can model this!
Consider determining which of two texts, $i$ and $j$, is “easier”
Consider determining which of two texts, $i$ and $j$, is “easier”

If the ‘easiness’ of $i$ is $\alpha_i$, and the ‘easiness’ of $j$ is $\alpha_j$, then the odds that snippet $i$ is deemed easier than $j$ may be written as $\frac{\alpha_i}{\alpha_j}$.
Our approach: Bradley-Terry Regression

1. Consider determining which of two texts, $i$ and $j$, is “easier”

2. If the ‘easiness’ of $i$ is $\alpha_i$, and the ‘easiness’ of $j$ is $\alpha_j$, then the odds that snippet $i$ is deemed easier than $j$ may be written as $\frac{\alpha_i}{\alpha_j}$

3. 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. If the ‘easiness’ of $i$ is $\alpha_i$, and the ‘easiness’ of $j$ is $\alpha_j$, then the odds that snippet $i$ is deemed easier than $j$ may be written as $\frac{\alpha_i}{\alpha_j}$

3. Defining $\lambda_i = \log \alpha_i$, the regression model can be rewritten:

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

4. 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

- There are millions of potential variables to consider
There are millions of potential variables to consider.

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

- There are millions of potential variables to consider
- We begin with all constituent variables of the traditional models, add in some new ones
- 29 possible variables
There are millions of potential variables to consider.

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

29 possible variables.

Use a variant of random forests to select the variables that best fit the snippets scaled through unstructured Bradley-Terry regression.

VSURF package developed by Genuer, Poggi and Tuleau-Malot.
Variable selection

- There are millions of potential variables to consider.
- We begin with all constituent variables of the traditional models, add in some new ones.
- 29 possible variables.
- Use a variant of random forests to select the variables that best fit the snippets scaled through unstructured Bradley-Terry regression.
  - VSURF package developed by Genuer, Poggi and Tuleau-Malot.
Google nGram

- A collection of word counts in the Google books corpus
A collection of word counts in the Google books corpus

1790 onward, filter out tokens that appeared fewer than five times or did not match a dictionary of 133,000 English words/word forms
Google nGram

- A collection of word counts in the Google books corpus

- 1790 onward, filter out tokens that appeared fewer than five times or did not match a dictionary of 133,000 English words/word forms \( \sim 615,362,456,717 \) token counts from 85,623 word types

- Normalize each word frequency to its frequency relative to the word "the" in that 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

- When "plugging in" values of covariates to evaluate older texts, we will use the word frequency from the decade in which they originate
A collection of word counts in the Google books corpus

1790 onward, filter out tokens that appeared fewer than five times or did not match a dictionary of 133,000 English words/word forms \( \sim \) 615,362,456,717 token counts from 85,623 word types

Normalize each word frequency to its frequency relative to the word “the” in that year, smoothing by decade
Google nGram

- A collection of word counts in the Google books corpus
- 1790 onward, filter out tokens that appeared fewer than five times or did not match a dictionary of 133,000 English words/word forms \( \sim 615,362,456,717 \) token counts from 85,623 word types
- Normalize each word frequency to its frequency relative to the word “the” in that 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
Google nGram

- A collection of word counts in the Google books corpus
- 1790 onward, filter out tokens that appeared fewer than five times or did not match a dictionary of 133,000 English words/word forms $\sim 615,362,456,717$ token counts from 85,623 word types
- Normalize each word frequency to its frequency relative to the word “the” in that 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
- When “plugging in” values of covariates to evaluate older texts, we will use the word frequency from the decade in which they originate
Structured Bradley-Terry Model

We have our covariates

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

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 thus estimate the relevant \( \hat{\beta_r} \)'s and can then "plug in" covariates to evaluate other texts.
We have our covariates

We still model this comparison:

\[
\text{logit}[\Pr(i \text{ easier than } j)] = \lambda_i - \lambda_j
\]
Structured Bradley-Terry Model

- We have our covariates

- We still model this comparison:

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

- 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 our covariates

We still model this comparison:

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

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 thus estimate the relevant \( \hat{\beta}_r \)'s and can then “plug in” covariates to evaluate other texts
<table>
<thead>
<tr>
<th></th>
<th>All variables</th>
<th>Simple Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characters per sentence</td>
<td>$-0.01^*$</td>
<td>$-0.01^*$</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>Proportion of 3-syllable</td>
<td>$-1.04^*$</td>
<td>$-1.31^*$</td>
</tr>
<tr>
<td>words</td>
<td>(0.34)</td>
<td>(0.28)</td>
</tr>
<tr>
<td>Proportion of words from</td>
<td>-0.41</td>
<td></td>
</tr>
<tr>
<td>Dale-Chall</td>
<td>(0.28)</td>
<td></td>
</tr>
<tr>
<td>Proportion of adpositions</td>
<td>$-0.99^*$</td>
<td>$-1.11^*$</td>
</tr>
<tr>
<td>such as <em>to, with, from, under</em></td>
<td>(0.48)</td>
<td>(0.46)</td>
</tr>
<tr>
<td>Mean word frequency ('the')</td>
<td>$-1.74^*$</td>
<td>$-1.68^*$</td>
</tr>
<tr>
<td></td>
<td>(0.35)</td>
<td>(0.35)</td>
</tr>
<tr>
<td>Proportion of conjunctions</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.71)</td>
<td></td>
</tr>
<tr>
<td>PCP</td>
<td>0.663</td>
<td>0.662</td>
</tr>
<tr>
<td>AIC</td>
<td>7419.90</td>
<td>7419.09</td>
</tr>
</tbody>
</table>

Standard errors in parentheses. $^*$ indicates significance at $p < 0.05$. 
Speeches in 2016 Campaign Debates

![Graph showing speeches in 2016 campaign debates]

- **GOP**
- **Dem**
- **Trump**
- **Clinton**

Pr(easier than 5th grade)
Evaluating traditional measures

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

<table>
<thead>
<tr>
<th>Measure</th>
<th>AIC</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRE</td>
<td>7,893</td>
<td>0.602</td>
</tr>
<tr>
<td>Dale-Chall</td>
<td>7,895</td>
<td>0.603</td>
</tr>
<tr>
<td>FOG</td>
<td>7,619</td>
<td>0.638</td>
</tr>
<tr>
<td>SMOG</td>
<td>7,726</td>
<td>0.574</td>
</tr>
<tr>
<td>Spache</td>
<td>7,665</td>
<td>0.635</td>
</tr>
<tr>
<td>Coleman-Liau</td>
<td>8,219</td>
<td>0.552</td>
</tr>
</tbody>
</table>
# Results—Refit FRE Model

<table>
<thead>
<tr>
<th></th>
<th>FRE, refit</th>
<th>Sentence only</th>
<th>Syllables only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean syllables/word</td>
<td>$-1.34^*$</td>
<td>$-0.71^*$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.12)</td>
<td>(0.11)</td>
<td></td>
</tr>
<tr>
<td>Mean words/sentence</td>
<td>$-0.07^*$</td>
<td>$-0.06^*$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td></td>
</tr>
<tr>
<td>PCP</td>
<td>0.66</td>
<td>0.64</td>
<td>0.53</td>
</tr>
<tr>
<td>AIC</td>
<td>7494.81</td>
<td>7625.82</td>
<td>8275.97</td>
</tr>
</tbody>
</table>

Standard errors in parentheses. * indicates significance at $p < 0.05$
Sophistication is a normatively important component of political speech.
Moving forward

- Sophistication is a normatively important component of political speech
- We have demonstrated the insufficiency of existing measures, and developed a framework for creating a better one
Sophistication is a normatively important component of political speech.

We have demonstrated the insufficiency of existing measures, and developed a framework for creating a better one.

However, the predictive accuracy of our best model is underwhelming; improvements include:

- Calculating word rarity for different parts of speech
- Performing comparisons between longer snippets of text
- Incorporating more syntactic information
Sophistication is a normatively important component of political speech

We have demonstrated the insufficiency of existing measures, and developed a framework for creating a better one

However, the predictive accuracy of our best model is underwhelming; improvements include:

- Calculating word rarity for different parts of speech
Sophistication is a normatively important component of political speech.

We have demonstrated the insufficiency of existing measures, and developed a framework for creating a better one.

However, the predictive accuracy of our best model is underwhelming; improvements include:

- Calculating word rarity for different parts of speech
- Performing comparisons between longer snippets of text
Moving forward

- Sophistication is a normatively important component of political speech

- We have demonstrated the insufficiency of existing measures, and developed a framework for creating a better one

- However, the predictive accuracy of our best model is underwhelming; improvements include:
  - Calculating word rarity for different parts of speech
  - Performing comparisons between longer snippets of text
  - Incorporate more syntactic information
Measuring the sophistication of political text

Kenneth Benoit, Kevin Munger, and Arthur Spirling