Lexical Diversity

Lexical Diversity

Recall that the elementary components of a text are called tokens.

Lexical Diversity

Recall that the elementary components of a text are called tokens. These are generally words, but they may also include numbers, sums of money, etc.

Lexical Diversity

Recall that the elementary components of a text are called tokens. These are generally words, but they may also include numbers, sums of money, etc.

The types in a document are the set of unique tokens.

Lexical Diversity

Recall that the elementary components of a text are called tokens. These are generally words, but they may also include numbers, sums of money, etc.

The types in a document are the set of unique tokens.

thus we typically have many more tokens than types,

Lexical Diversity

Recall that the elementary components of a text are called tokens. These are generally words, but they may also include numbers, sums of money, etc.

The types in a document are the set of unique tokens.

thus we typically have many more tokens than types, because authors repeat tokens.

Lexical Diversity

Recall that the elementary components of a text are called tokens. These are generally words, but they may also include numbers, sums of money, etc.

The types in a document are the set of unique tokens.

thus we typically have many more tokens than types, because authors repeat tokens.

TTR we can use the type-to-token ratio as a measure of lexical diversity. This is:

Lexical Diversity

Recall that the elementary components of a text are called tokens. These are generally words, but they may also include numbers, sums of money, etc.

The types in a document are the set of unique tokens.

thus we typically have many more tokens than types, because authors repeat tokens.

TTR we can use the type-to-token ratio as a measure of lexical diversity. This is:

$$\boxed{TTR = \frac{\text{total types}}{\text{total tokens}}}$$

Lexical Diversity

Recall that the elementary components of a text are called tokens. These are generally words, but they may also include numbers, sums of money, etc.

The types in a document are the set of unique tokens.

thus we typically have many more tokens than types, because authors repeat tokens.

TTR we can use the type-to-token ratio as a measure of lexical diversity. This is:

$$\boxed{TTR = \frac{\text{total types}}{\text{total tokens}}}$$

e.g. authors with limited vocabularies will have a low lexical diversity.

Tabloid vs Broadsheet

Tabloid vs Broadsheet

RCH NEW YORK POST

NEWS

Iraqi troops retake key government complex in Ramadi

By Associated Press December 28, 2015 | 6:34am | Updated

Members of Iraq's elite counter-terrorism service secure a neighborhood in the city of Ramadi.

Photo: Getty Images

MORE ON:
ISIS

BAGHDAD — Iraqi military forces on Monday retook a strategic government complex in the city of

Tabloid vs Broadsheet

RCH NEW YORK POST

NEWS

Iraqi troops retake key government complex in Ramadi

By Associated Press December 28, 2015 | 6:34am | Updated

Members of Iraq's elite counter-terrorism service secure a neighborhood in the city of Ramadi.

Photo: Getty Images

MORE ON: ISIS

BAGHDAD — Iraqi military forces on Monday retook a strategic government complex in the city of

$$TTR = \frac{250}{491} = 0.51$$

Tabloid vs Broadsheet

NEW YORK POST

NEWS

Iraqi troops retake key government complex in Ramadi

By Associated Press December 28, 2015 | 6:34am | Updated

Members of Iraq's elite counter-terrorism service secure a neighborhood in the city of Ramadi.

Photo: Getty Images

MORE ON: ISIS

BAGHDAD — Iraqi military forces on Monday retook a strategic government complex in the city of

The New York Times

Obama's 'Boots on the Ground': U.S. Special Forces Are Sent to Tackle Global Threats

Japan and South Korea Settle Dispute Over Wartime 'Comfort Women'

ITINERARIES T.S.A. Moves Closer to Rejecting Some State Driver's Licenses for...

MIDDLE EAST

Iraqi Forces Retake Center of Ramadi From ISIS

By FALIH HASSAN and SEWELL CHAN DEC. 28, 2015

Iraqi soldiers at the Anbar police headquarters in Ramadi, Iraq, on Monday, after seizing a government complex from the Islamic State. Ahmad Al-Rubaye/Agence France-Presse — Getty Images

Email

Share

BAGHDAD — Iraqi forces said on Monday they had seized a strategic government complex in the western city of Ramadi from the Islamic State after a fierce weeklong battle, putting them on the verge of a crucial victory following a brutal seven-month occupation of the city by the extremist group

$$TTR = \frac{250}{491} = 0.51$$

Tabloid vs Broadsheet

NEW YORK POST

NEWS

Iraqi troops retake key government complex in Ramadi

By Associated Press December 28, 2015 | 6:34am | Updated

Members of Iraq's elite counter-terrorism service secure a neighborhood in the city of Ramadi.

Photo: Getty Images

MORE ON: ISIS

BAGHDAD — Iraqi military forces on Monday retook a strategic government complex in the city of

$$TTR = \frac{250}{491} = 0.51$$

The New York Times

Obama's 'Boots on the Ground': U.S. Special Forces Are Sent to Tackle Global Threats

Japan and South Korea Settle Dispute Over Wartime 'Comfort Women'

ITINERARIES T.S.A. Moves Closer to Rejecting Some State Driver's Licenses for...

MIDDLE EAST

Iraqi Forces Retake Center of Ramadi From ISIS

By FALIH HASSAN and SEWELL CHAN DEC. 28, 2015

Iraqi soldiers at the Anbar police headquarters in Ramadi, Iraq, on Monday, after seizing a government complex from the Islamic State. Ahmad Al-Rubaye/Agence France-Presse — Getty Images

Email

Share

BAGHDAD — Iraqi forces said on Monday they had seized a strategic government complex in the western city of Ramadi from the Islamic State after a fierce weeklong battle, putting them on the verge of a crucial victory following a brutal seven-month occupation of the city by the extremist group

$$TTR = \frac{428}{978} = 0.43$$

Hmm. . .

Hmm. . .

Unexpected, and mostly product of different text lengths:

Hmm...

Unexpected, and mostly product of different text lengths: shorter texts tend to have fewer repetitions (of e.g. common words).

Hmm...

Unexpected, and mostly product of different text lengths: shorter texts tend to have fewer repetitions (of e.g. common words).

but also case that longer documents cover more topics which presumably adds to richness (?)

Hmm...

Unexpected, and mostly product of different text lengths: shorter texts tend to have fewer repetitions (of e.g. common words).

but also case that longer documents cover more topics which presumably adds to richness (?)

so make denominator non-linear:

Measurement of Linguistic Complexity

Measurement of Linguistic Complexity

- over a hundred years of literature on measurement of 'readability':

Measurement of Linguistic Complexity

- over a hundred years of literature on measurement of 'readability': general issue was assigning school texts to pupils of different ages and abilities.

Measurement of Linguistic Complexity

- over a hundred years of literature on measurement of 'readability': general issue was assigning school texts to pupils of different ages and abilities.
- Flesch (1948) suggests *Flesch Reading Ease* statistic

Measurement of Linguistic Complexity

- over a hundred years of literature on measurement of 'readability': general issue was assigning school texts to pupils of different ages and abilities.
- Flesch (1948) suggests *Flesch Reading Ease* statistic

FRE

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

Measurement of Linguistic Complexity

- over a hundred years of literature on measurement of 'readability': general issue was assigning school texts to pupils of different ages and abilities.
- Flesch (1948) suggests *Flesch Reading Ease* statistic

FRE

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

based on $\hat{\beta}$s from linear model where $y =$ average grade level of school children who could correctly answer at least 75% of mc qs on texts.

Measurement of Linguistic Complexity

- over a hundred years of literature on measurement of 'readability': general issue was assigning school texts to pupils of different ages and abilities.
- Flesch (1948) suggests *Flesch Reading Ease* statistic

FRE

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

based on $\hat{\beta}$s from linear model where $y =$ average grade level of school children who could correctly answer at least 75% of mc qs on texts. Scaled s.t. a document with score of 100 could be understood by fourth grader (9yo).

Measurement of Linguistic Complexity

- over a hundred years of literature on measurement of 'readability': general issue was assigning school texts to pupils of different ages and abilities.
- Flesch (1948) suggests *Flesch Reading Ease* statistic

FRE

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

based on $\hat{\beta}$s from linear model where y = average grade level of school children who could correctly answer at least 75% of mc qs on texts. Scaled s.t. a document with score of 100 could be understood by fourth grader (9yo).

- Kincaid et al later translate to US School grade level that would be (on average) required to comprehend text.

Readability Guidelines

Readability Guidelines

in practice,

Readability Guidelines

in practice, estimated FRE can be outside $[0, 100]$.

Readability Guidelines

in practice, estimated FRE can be outside $[0, 100]$.

However. . .

Readability Guidelines

in practice, estimated FRE can be outside $[0, 100]$.

However. . .

Score	Education	Description	Clve % US popn
0–30	college graduates	very difficult	28
31–50		difficult	72
51–60		fairly difficult	85
61–70	9th grade	standard	–
71–80		fairly easy	–
81–90		easy	–
91–100	4th grade	very easy	–

Examples

Examples

Score	Text
-800	Molly Bloom's (3.6K) Soliloquy, *Ulysses*

Examples

Score	Text
-800	Molly Bloom's (3.6K) Soliloquy, *Ulysses*
33	mean political science article; judicial opinion

Examples

Score	Text
-800	Molly Bloom's (3.6K) Soliloquy, *Ulysses*
33	mean political science article; judicial opinion
37	Spirling
45	life insurance requirement (FL)

Examples

Score	Text
-800	Molly Bloom's (3.6K) Soliloquy, *Ulysses*
33	mean political science article; judicial opinion
37	Spirling
45	life insurance requirement (FL)
48	*New York times*
65	*Reader's Digest*
67	*Al Qaeda* press release

Examples

Score	Text
-800	Molly Bloom's (3.6K) Soliloquy, *Ulysses*
33	mean political science article; judicial opinion
37	Spirling
45	life insurance requirement (FL)
48	*New York times*
65	*Reader's Digest*
67	*Al Qaeda* press release
77	Dickens' works
80	children's books: e.g. *Wind in the Willows*

Examples

Score	Text
-800	Molly Bloom's (3.6K) Soliloquy, *Ulysses*
33	mean political science article; judicial opinion
37	Spirling
45	life insurance requirement (FL)
48	*New York times*
65	*Reader's Digest*
67	*Al Qaeda* press release
77	Dickens' works
80	children's books: e.g. *Wind in the Willows*
90	death row inmate last statements (TX)
100	this entry right here.

Notes

Notes

Flesch scoring only uses syllable information:

Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "`Indeed, the shoemaker was frightened`" would score similarly to

Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "`Indeed, the shoemaker was frightened`" would score similarly to "`Forsooth, the cordwainer was afeared`"

Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "`Indeed, the shoemaker was frightened`" would score similarly to "`Forsooth, the cordwainer was afeared`"

Widely used because it 'works',

Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "`Indeed, the shoemaker was frightened`" would score similarly to "`Forsooth, the cordwainer was afeared`"

Widely used because it 'works', not because it is justified from first principles

Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "`Indeed, the shoemaker was frightened`" would score similarly to "`Forsooth, the cordwainer was afeared`"

Widely used because it 'works', not because it is justified from first principles

One of many such indices:

Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "`Indeed, the shoemaker was frightened`" would score similarly to "`Forsooth, the cordwainer was afeared`"

Widely used because it 'works', not because it is justified from first principles

One of many such indices: Gunning-Fog,

Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "`Indeed, the shoemaker was frightened`" would score similarly to "`Forsooth, the cordwainer was afeared`"

Widely used because it 'works', not because it is justified from first principles

One of many such indices: Gunning-Fog, Dale-Chall,

Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "`Indeed, the shoemaker was frightened`" would score similarly to "`Forsooth, the cordwainer was afeared`"

Widely used because it 'works', not because it is justified from first principles

One of many such indices: Gunning-Fog, Dale-Chall, Automated Readability Index,

Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "`Indeed, the shoemaker was frightened`" would score similarly to "`Forsooth, the cordwainer was afeared`"

Widely used because it 'works', not because it is justified from first principles

One of many such indices: Gunning-Fog, Dale-Chall, Automated Readability Index, SMOG.

Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "`Indeed, the shoemaker was frightened`" would score similarly to "`Forsooth, the cordwainer was afeared`"

Widely used because it 'works', not because it is justified from first principles

One of many such indices: Gunning-Fog, Dale-Chall, Automated Readability Index, SMOG. Typically highly correlated (at text level).

Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "`Indeed, the shoemaker was frightened`" would score similarly to "`Forsooth, the cordwainer was afeared`"

Widely used because it 'works', not because it is justified from first principles

One of many such indices: Gunning-Fog, Dale-Chall, Automated Readability Index, SMOG. Typically highly correlated (at text level).

Surprisingly little effort to describe statistical behavior of estimator:

Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "`Indeed, the shoemaker was frightened`" would score similarly to "`Forsooth, the cordwainer was afeared`"

Widely used because it 'works', not because it is justified from first principles

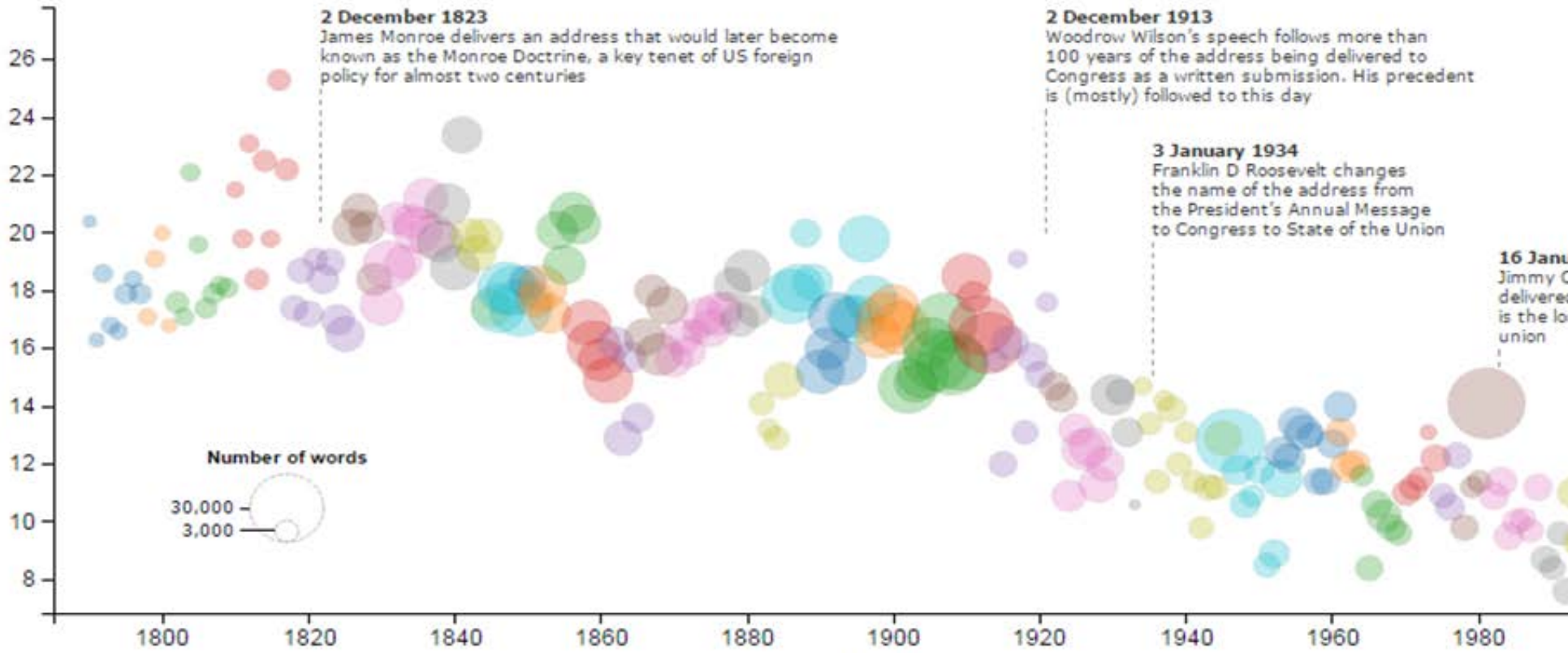
One of many such indices: Gunning-Fog, Dale-Chall, Automated Readability Index, SMOG. Typically highly correlated (at text level).

Surprisingly little effort to describe statistical behavior of estimator: sampling distribution etc.

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



Leaders and their incentives

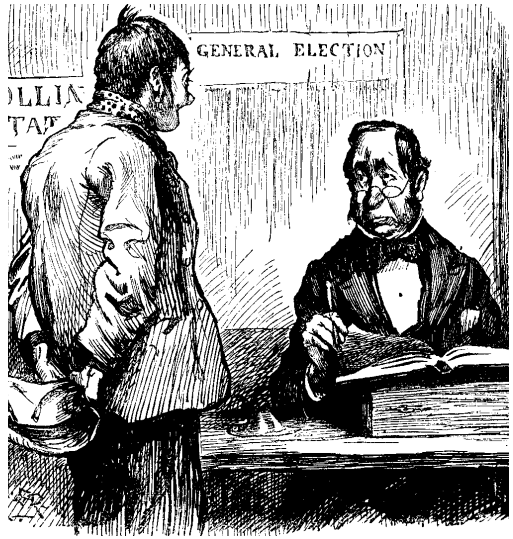
Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

new voters tended to be poorer and less literate



Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

new voters tended to be poorer and less literate

↓ local, clientalistic appeals via bribery...

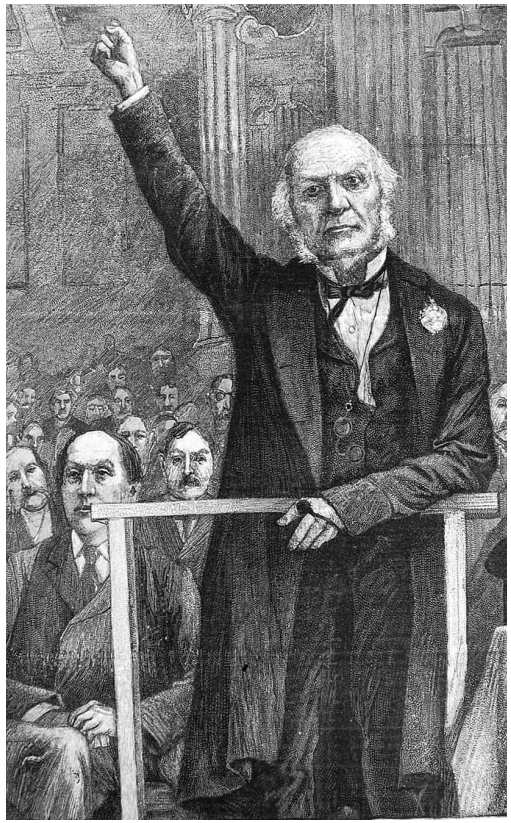
Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

new voters tended to be poorer and less literate

↓ local, clientalistic appeals via bribery...

↑ 'party orientated electorate', with national policies and national leaders



Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

new voters tended to be poorer and less literate

↓ local, clientalistic appeals via bribery...

↑ 'party orientated electorate', with national policies and national leaders

Q how did these leaders respond to new voters?

Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

new voters tended to be poorer and less literate

↓ local, clientalistic appeals via bribery...

↑ 'party orientated electorate', with national policies and national leaders

Q how did these leaders respond to new voters?

A by changing nature of their speech:

Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

new voters tended to be poorer and less literate

↓ local, clientalistic appeals via bribery...

↑ 'party orientated electorate', with national policies and national leaders

Q how did these leaders respond to new voters?

A by changing nature of their speech: simpler,

Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

new voters tended to be poorer and less literate

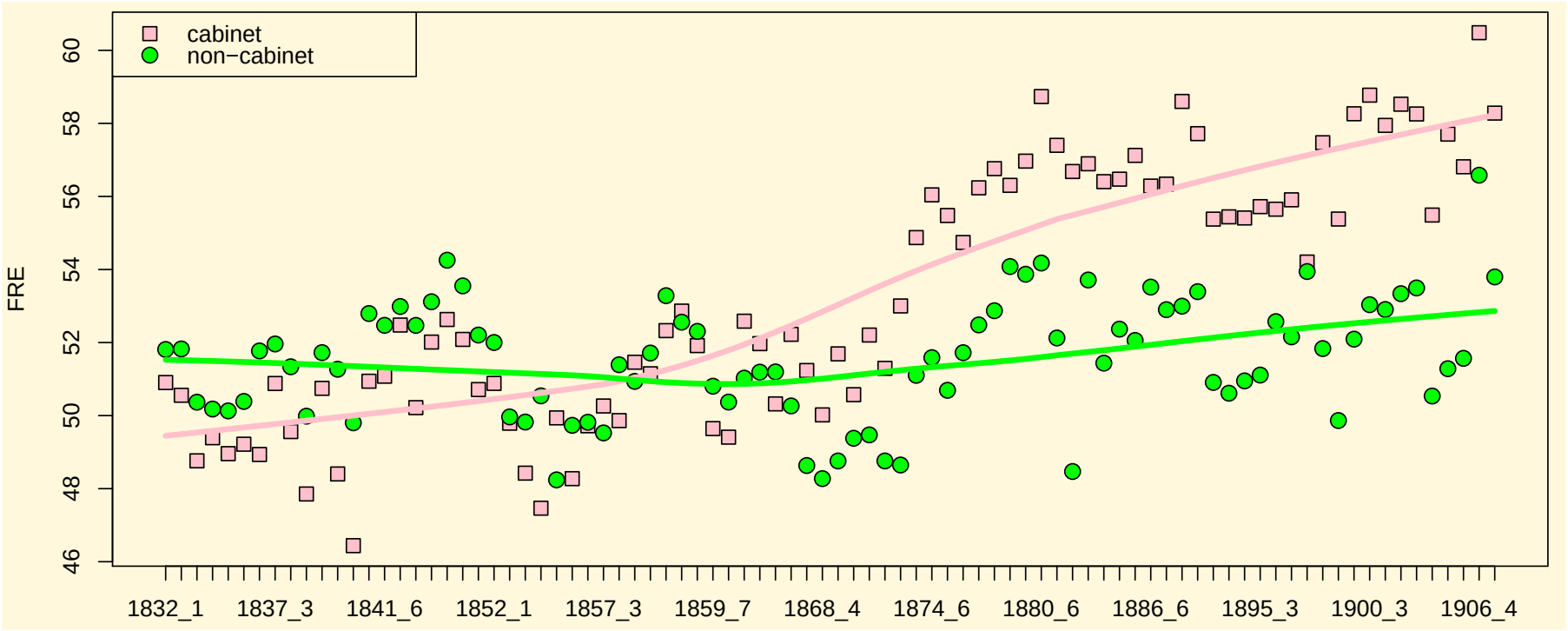
↓ local, clientalistic appeals via bribery...

↑ 'party orientated electorate', with national policies and national leaders

Q how did these leaders respond to new voters?

A by changing nature of their speech: simpler, less complex expressions in parliament

Flesch overtime plot



Dale-Chall, 1948

Dale-Chall, 1948

yields grade level of text sample.

Dale-Chall, 1948

yields grade level of text sample.

DC

$$0.1579 \times (\text{PDW}) + 0.0496 \times \left(\frac{\text{total words}}{\text{total sentences}} \right)$$

Dale-Chall, 1948

yields grade level of text sample.

DC

$$0.1579 \times (\text{PDW}) + 0.0496 \times \left(\frac{\text{total words}}{\text{total sentences}} \right)$$

where PDW is percentage of difficult words,

Dale-Chall, 1948

yields grade level of text sample.

DC

$$0.1579 \times (\text{PDW}) + 0.0496 \times \left(\frac{\text{total words}}{\text{total sentences}} \right)$$

where PDW is percentage of difficult words,

and a 'difficult' word is one that does not appear on Dale & Chall's list of 763

Dale-Chall, 1948

yields grade level of text sample.

DC

$$0.1579 \times (\text{PDW}) + 0.0496 \times \left(\frac{\text{total words}}{\text{total sentences}} \right)$$

where PDW is percentage of difficult words,

and a 'difficult' word is one that does not appear on Dale & Chall's list of 763 (later updated to 3000)

Dale-Chall, 1948

yields grade level of text sample.

DC

$$0.1579 \times (\text{PDW}) + 0.0496 \times \left(\frac{\text{total words}}{\text{total sentences}} \right)$$

where PDW is percentage of difficult words,

and a 'difficult' word is one that does not appear on Dale & Chall's list of 763 (later updated to 3000) 'familiar' words.

Dale-Chall, 1948

yields grade level of text sample.

DC

$$0.1579 \times (\text{PDW}) + 0.0496 \times \left(\frac{\text{total words}}{\text{total sentences}} \right)$$

where PDW is percentage of difficult words,

and a 'difficult' word is one that does not appear on Dale & Chall's list of 763 (later updated to 3000) 'familiar' words.

e.g. `about`, `back`, `call`, etc.

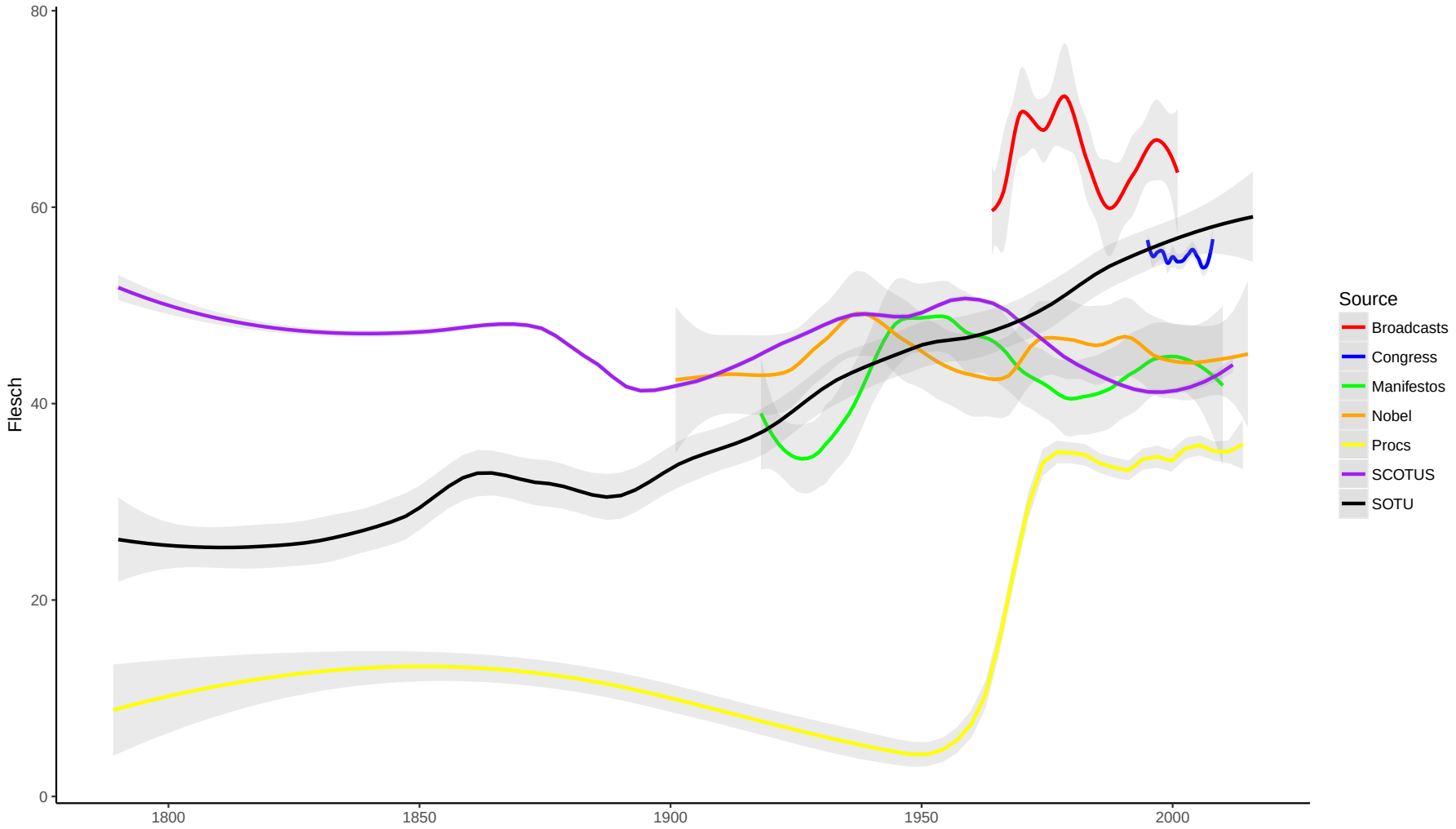
The Great Sentence Length Shift (Benoit, Munger & Spirling)

The Great Sentence Length Shift (Benoit, Munger & Spirling)

SOTU is simpler: is public discourse becoming 'dumbed down'? Probably not.

The Great Sentence Length Shift (Benoit, Munger & Spirling)

SOTU is simpler: is public discourse becoming 'dumbed down'? Probably not.

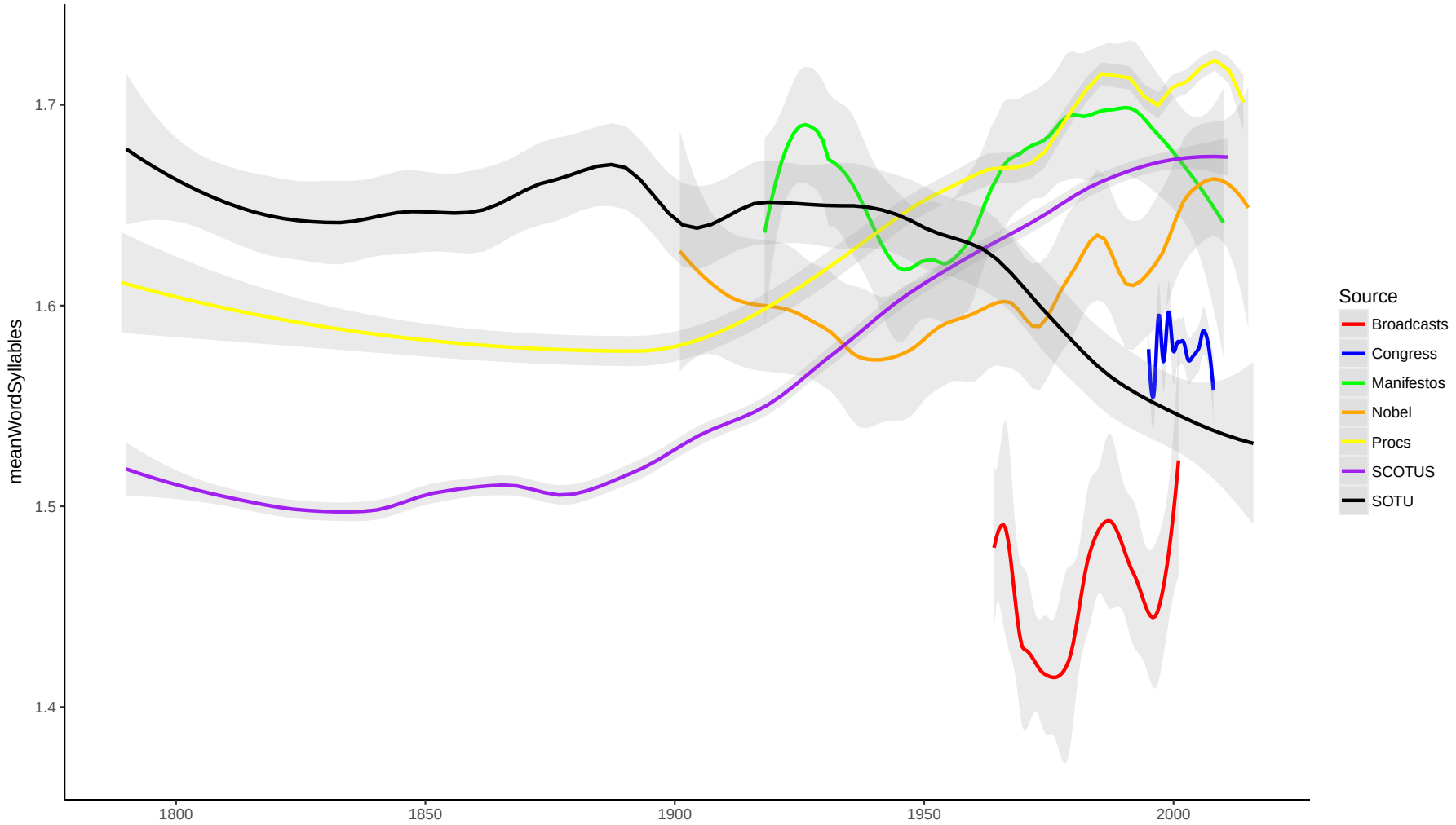


The Great Sentence Length Shift (Benoit, Munger & Spirling)

SOTU is simpler: is public discourse becoming 'dumbed down'? Probably not. What's driving these patterns?

The Great Sentence Length Shift (Benoit, Munger & Spirling)

SOTU is simpler: is public discourse becoming 'dumbed down'? Probably not. What's driving these patterns? Syllables?

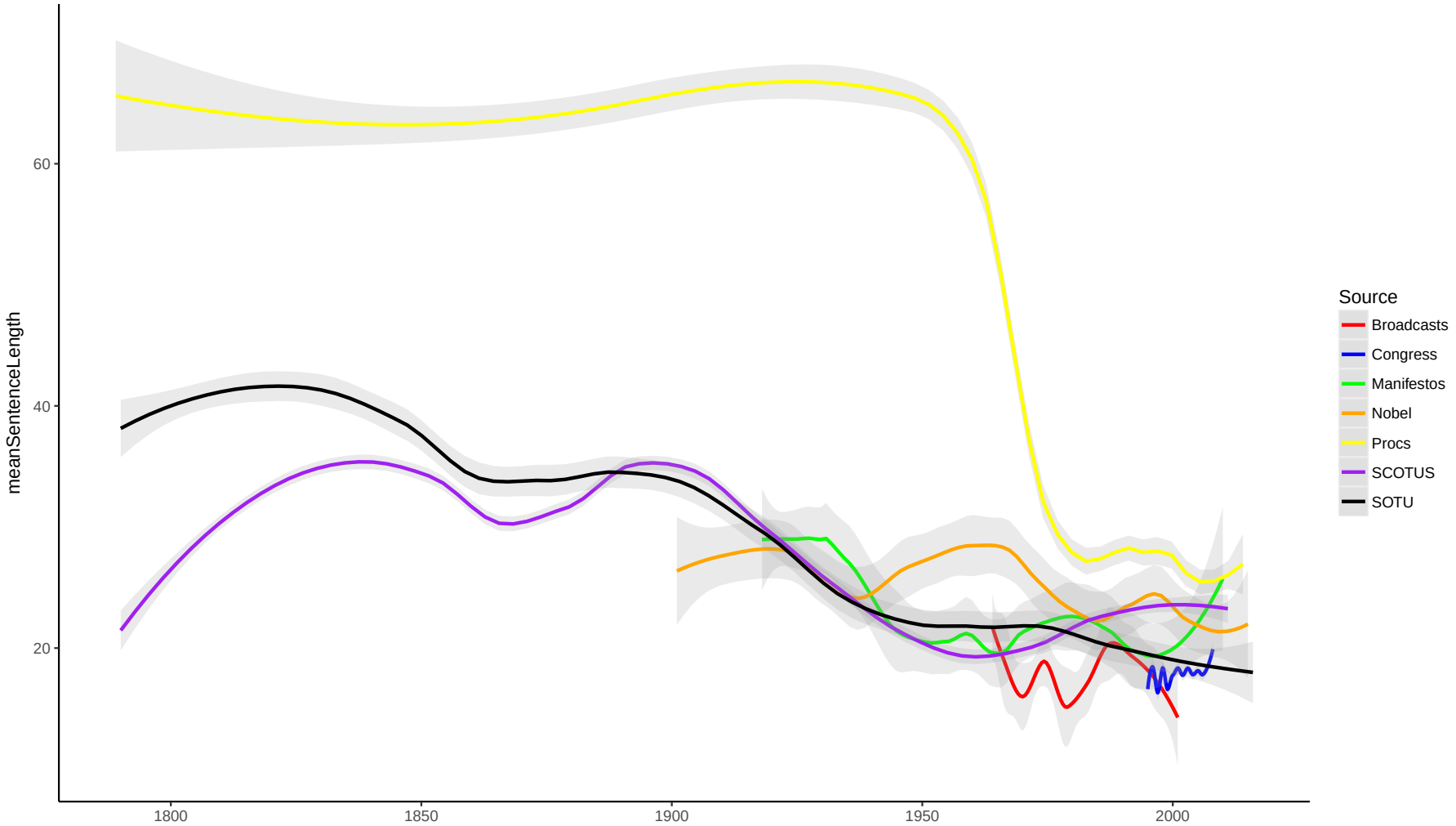


The Great Sentence Length Shift (Benoit, Munger & Spirling)

SOTU is simpler: is public discourse becoming 'dumbed down'? Probably not. What's driving these patterns? Syllables? Sentence length?

The Great Sentence Length Shift (Benoit, Munger & Spirling)

SOTU is simpler: is public discourse becoming 'dumbed down'? Probably not. What's driving these patterns? Syllables? Sentence length?



Can we do better?

Can we do better?

i.e. have I, personally, written a paper about this?

Can we do better?

i.e. have I, personally, written a paper about this?

YES!

Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS crowdsource thousands of snippet comparisons: ask raters which is more difficult,

Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS crowdsource thousands of snippet comparisons: ask raters which is more difficult, make that a function of covariates.

Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS crowdsource thousands of snippet comparisons: ask raters which is more difficult, make that a function of covariates.

Everything wrapped into well-known statistical model (Bradley-Terry) for pairwise comparisons: can make probabilistic statements about difficultly.

Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS crowdsource thousands of snippet comparisons: ask raters which is more difficult, make that a function of covariates.

Everything wrapped into well-known statistical model (Bradley-Terry) for pairwise comparisons: can make probabilistic statements about difficultly.

Model performance not hugely better than FRE,

Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS crowdsource thousands of snippet comparisons: ask raters which is more difficult, make that a function of covariates.

Everything wrapped into well-known statistical model (Bradley-Terry) for pairwise comparisons: can make probabilistic statements about difficultly.

Model performance not hugely better than FRE, but note we are at least able to make the comparison!

Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS crowdsource thousands of snippet comparisons: ask raters which is more difficult, make that a function of covariates.

Everything wrapped into well-known statistical model (Bradley-Terry) for pairwise comparisons: can make probabilistic statements about difficultly.

Model performance not hugely better than FRE, but note we are at least able to make the comparison!

Can ask: what is Pr(Eisenhower easier than Bush)? (is ~ 0.43)

Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS crowdsource thousands of snippet comparisons: ask raters which is more difficult, make that a function of covariates.

Everything wrapped into well-known statistical model (Bradley-Terry) for pairwise comparisons: can make probabilistic statements about difficultly.

Model performance not hugely better than FRE, but note we are at least able to make the comparison!

Can ask: what is Pr(Eisenhower easier than Bush)? (is ~ 0.43)

Key innovation:

Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS crowdsource thousands of snippet comparisons: ask raters which is more difficult, make that a function of covariates.

Everything wrapped into well-known statistical model (Bradley-Terry) for pairwise comparisons: can make probabilistic statements about difficultly.

Model performance not hugely better than FRE, but note we are at least able to make the comparison!

Can ask: what is Pr(Eisenhower easier than Bush)? (is ~ 0.43)

Key innovation: rarity is from google books corpus,

Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS crowdsource thousands of snippet comparisons: ask raters which is more difficult, make that a function of covariates.

Everything wrapped into well-known statistical model (Bradley-Terry) for pairwise comparisons: can make probabilistic statements about difficultly.

Model performance not hugely better than FRE, but note we are at least able to make the comparison!

Can ask: what is Pr(Eisenhower easier than Bush)? (is ~ 0.43)

Key innovation: rarity is from google books corpus, and fitted to local decade and domain (adults) that you care about.

Paper and Software

Paper and Software

Download This Paper | Open PDF in Browser | Share:

Measuring and Explaining Political Sophistication Through Textual Complexity

42 Pages • Posted: 1 Nov 2017

Kenneth Benoit
London School of Economics & Political Science (LSE); Trinity College Dublin

Kevin Munger
New York University (NYU)

Arthur Spirling
New York University

Date Written: October 30, 2017

Abstract

The sophistication of political communication has been measured using "readability" scores developed from other contexts, but their application out of domain is problematic. We systematically review the shortcomings of previous measures, before developing a new approach, with software, better suited to the task. We use the crowd to perform thousands of pairwise comparisons of text snippets and incorporate these results into a statistical model of comprehension. We include previously excluded features such as parts of speech, and a

Paper and Software

Download This Paper | Open PDF in Browser | Share:

Measuring and Explaining Political Sophistication Through Textual Complexity

42 Pages • Posted: 1 Nov 2017

Kenneth Benoit
London School of Economics & Political Science (LSE); Trinity College Dublin

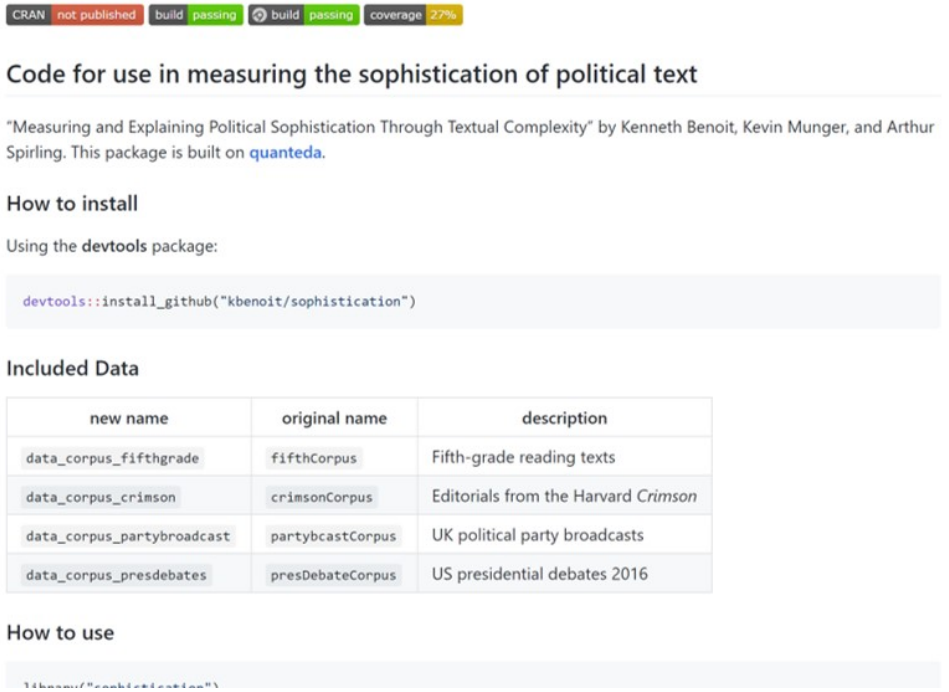
Kevin Munger
New York University (NYU)

Arthur Spirling
New York University

Date Written: October 30, 2017

Abstract

The sophistication of political communication has been measured using "readability" scores developed from other contexts, but their application out of domain is problematic. We systematically review the shortcomings of previous measures, before developing a new approach, with software, better suited to the task. We use the crowd to perform thousands of pairwise comparisons of text snippets and incorporate these results into a statistical model of comprehension. We include previously excluded features such as parts of speech, and a



CRAN not published | build passing | build passing | coverage 27%

Code for use in measuring the sophistication of political text

"Measuring and Explaining Political Sophistication Through Textual Complexity" by Kenneth Benoit, Kevin Munger, and Arthur Spirling. This package is built on quanteda.

How to install

Using the **devtools** package:

```
devtools::install_github("kbenoit/sophistication")
```

Included Data

new name	original name	description
data_corpus_fifthgrade	fifthCorpus	Fifth-grade reading texts
data_corpus_crimson	crimsonCorpus	Editorials from the Harvard *Crimson*
data_corpus_partybroadcast	partybcastCorpus	UK political party broadcasts
data_corpus_presdebates	presDebateCorpus	US presidential debates 2016

How to use

Paper and Software

Download This Paper | Open PDF in Browser | Share:

Measuring and Explaining Political Sophistication Through Textual Complexity

42 Pages • Posted: 1 Nov 2017

Kenneth Benoit
London School of Economics & Political Science (LSE); Trinity College Dublin

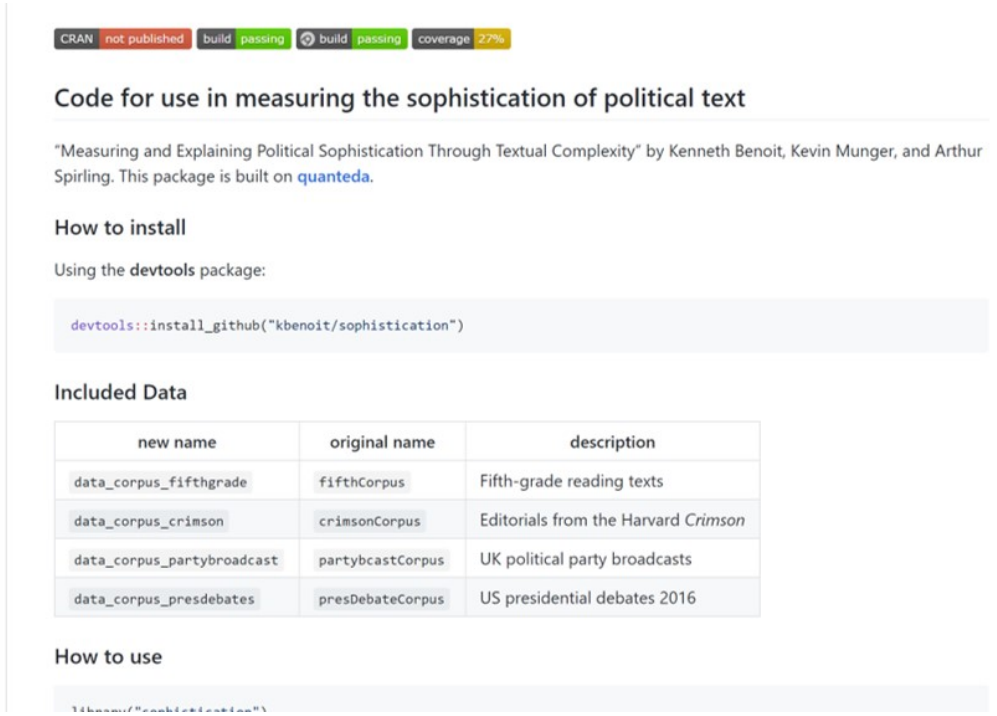
Kevin Munger
New York University (NYU)

Arthur Spirling
New York University

Date Written: October 30, 2017

Abstract

The sophistication of political communication has been measured using "readability" scores developed from other contexts, but their application out of domain is problematic. We systematically review the shortcomings of previous measures, before developing a new approach, with software, better suited to the task. We use the crowd to perform thousands of pairwise comparisons of text snippets and incorporate these results into a statistical model of comprehension. We include previously excluded features such as parts of speech, and a



CRAN not published | build passing | build passing | coverage 27%

Code for use in measuring the sophistication of political text

"Measuring and Explaining Political Sophistication Through Textual Complexity" by Kenneth Benoit, Kevin Munger, and Arthur Spirling. This package is built on quanteda.

How to install

Using the **devtools** package:

```
devtools::install_github("kbenoit/sophistication")
```

Included Data

new name	original name	description
data_corpus_fifthgrade	fifthCorpus	Fifth-grade reading texts
data_corpus_crimson	crimsonCorpus	Editorials from the Harvard *Crimson*
data_corpus_partybroadcast	partybcastCorpus	UK political party broadcasts
data_corpus_presdebates	presDebateCorpus	US presidential debates 2016

How to use

`github.com/kbenoit/sophistication`