EC999: Describing Text

Thiemo Fetzer

University of Chicago & University of Warwick

April 6, 2017

Descriptive Statistics for Text data

Before performing analysis, you want to get to know your data - this may inform you as to what are the necessary steps for dimensionality reduction. Some simple stats may be...

Word (relative) frequency

Theme (relative) frequency

Length in characters, words, lines, sentences, paragraphs, pages, sections, chapters, etc.

Vocabulary diversity (At its simplest) involves measuring a type-to-token ratio (TTR) where unique words are types and the total words are tokens.

Readability Use a combination of syllables and sentence length to indicate "readability" in terms of complexity

Formality Measures relationship of different parts of speech.

Vocabulary diversity

(At its simplest) involves measuring a type-to-token ratio (TTR) where unique words are types and the total words are tokens.

We have already talked about this in the section on Text normalization (pre-processing.)

Type-Token Ratio in Congressional speaches

```
dat
```

```
##        Text Types Tokens Sentences   speaker_name speaker_party
## text1 text1  4658  34151      1370     Mike Pence             R
## text2 text2 12509 440340     18343 Bernie Sanders             I
## text3 text3 11849 350175     18239      Rand Paul             R
## text4 text4  8212 182977      8843 Lindsey Graham             R
## text5 text5 10788 270801     12671    Marco Rubio             R
## text6 text6  5003  41051      1613       Jim Webb             D
## text7 text7 12862 304637     14101       Ted Cruz             R
```

⇒ this highlights that there is a negative correlation between the TTR and the total corpus length as measured by the number of sentences. We have seen this previously as *Heap's Law*.

Alternative Lexical Diversity Measures

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

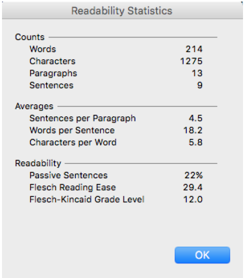
Guiraud $\frac{\text{total types}}{\sqrt{\text{total tokens}}}$

D iversity: Randomly sample a fixed number of tokens and count number of types.

MTLD the mean length of sequential word strings in a text that maintain a given TTR value (McCarthy and Jarvis, 2010) ??? fixes the TTR at 0.72 and counts the length of the text required to achieve it

Complexity and Readability

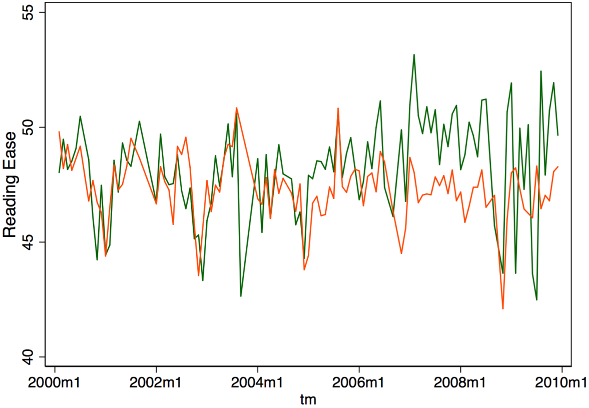
- Use of language is endogenous, and electoral incentives may affect the *communication strategies* chosen by elected officials.
- Readability scores us a combination of syllables and sentence length to indicate “complexity“ of text
- Common in educational research, but could also be used to describe textual complexity and increasingly some political science applications.
- No natural scale, so most are calibrated in terms of some interpretable metric



Readability Statistics

Counts	
Words	214
Characters	1275
Paragraphs	13
Sentences	9
Averages	
Sentences per Paragraph	4.5
Words per Sentence	18.2
Characters per Word	5.8
Readability	
Passive Sentences	22%
Flesch Reading Ease	29.4
Flesch-Kincaid Grade Level	12.0

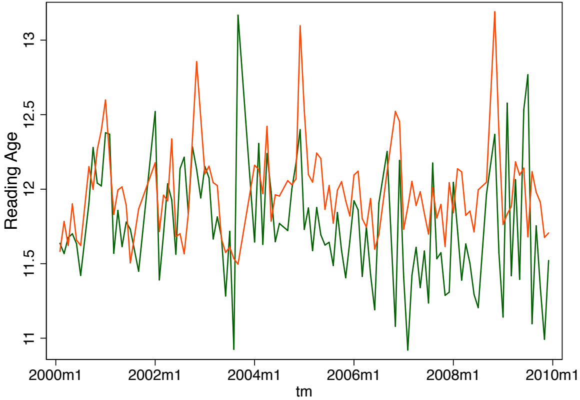
Reading Ease in Congress By Party



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

⇒ corpus data obtained via the Capitolwords API.

Reading Age in Congress By Party



$$\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$$

⇒ corpus data obtained via the Capitolwords API.

Gunning fog index

- Measures the readability in terms of the years of formal education required for a person to easily understand the text on first reading
- Usually taken on a sample of around 100 words, not omitting any sentences or words
- Computed as

$$0.4[(\frac{\text{total words}}{\text{total sentences}})] + 100\frac{\text{complex words}}{\text{total words}}$$

- Complex words are defined as those having three or more syllables, not including proper nouns (for example, Ljubljana), familiar jargon or compound words, or counting common suffixes such as -es, -ed, or -ing as a syllable.
- in *R* all readability features are embedded in the `quanteda` function `readability()`.

Example Readability computation

```
class(CORPUS.COMBINED)

## [1] "corpus" "list"

# can compute various readability indices on a corpus index in quanteda package
TEMP <- readability(CORPUS.COMBINED, measure = "Flesch.Kincaid")
TEMP

## text1 text2 text3 text4 text5 text6 text7
## 11.50 10.57  8.32  9.02  9.32 12.21 10.03

# can add this as piece of meta information
CORPUS.COMBINED[["readability"]] <- TEMP

summary(CORPUS.COMBINED)

## Corpus consisting of 7 documents.
##
##   Text Types Tokens Sentences   speaker_name speaker_party readability
##  text1  4658  34151      1370     Mike Pence             R       11.50
##  text2 12509 440340     18343 Bernie Sanders             I       10.57
##  text3 11849 350175     18239      Rand Paul             R        8.32
##  text4  8212 182977      8843 Lindsey Graham             R        9.02
##  text5 10788 270801     12671    Marco Rubio             R        9.32
##  text6  5003  41051      1613       Jim Webb             D       12.21
##  text7 12862 304637     14101       Ted Cruz             R       10.03
##
## Source:  /Users/thiemo/Dropbox/Teaching/Quantitative Text Analysis/Week 2d/* on x86_64 by thiemo
## Created: Mon Nov 21 16:25:05 2016
## Notes:
```

Formality of Language

This is to inform you that your book has been rejected by our publishing company as it was not up to the required standard. In case you would like us to reconsider it, we would suggest that you go over it and make some necessary changes.

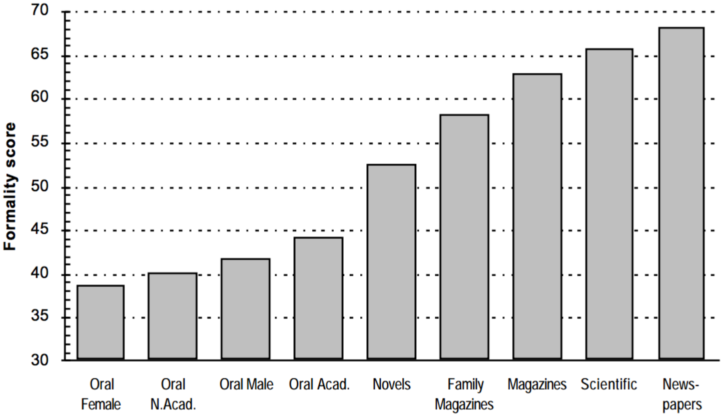
Formality of Language

You know that book I wrote? Well, the publishing company rejected it. They thought it was awful. But hey, I did the best I could, and I think it was great. I???m not gonna redo it the way they said I should.

Features of (In)formal language

- A formal style is characterized by detachment, accuracy, rigidity and heaviness
- Nouns, adjectives, articles and prepositions are more frequent in formal language
- an informal style is more flexible, direct, implicit, and involved, but less informative
- Pronouns, adverbs, verbs and interjections are more frequent in informal styles.

Formality Score



Heylighen, F., & Dewaele, J. (1999). Formality of Language : definition , measurement and behavioral determinants.

Formality Score

Language is considered more formal when it contains much of the information directly in the text, whereas, contextual language relies on shared experiences to more efficiently dialogue with others.

A candidate measure is the Heylighen & Dewaele's (1999) F-measure.

$$F = 50(\frac{nf - nc}{N} + 1)$$

Where:

- $f = \{$noun, adjective, preposition, article$\}$
- $c = \{$pronoun, verb, adverb, interjection$\}$
- $N = nf + nc$

This yields an F-measure between 0 and 100%, with completely contextualized language on the zero end and completely formal language on the 100 end.

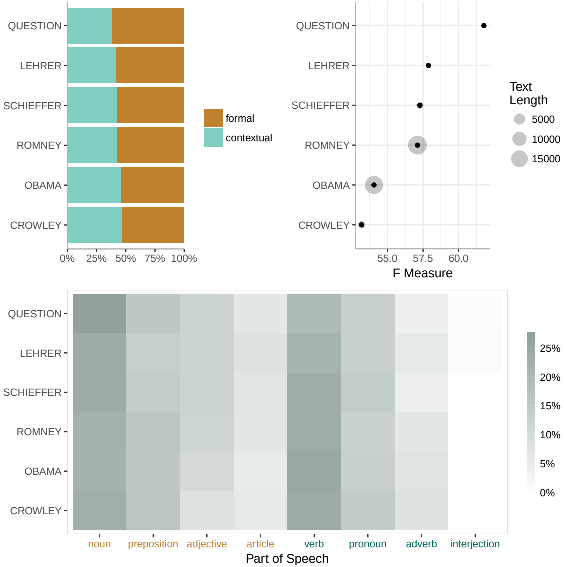
As is evident, this requires known *Parts of Speech*.

Computing Formality Scores in R

```r
# installing the formality package which is in developmental state
if (!require("pacman")) install.packages("pacman")
pacman::p_load_gh(c("trinker/formality"))
library(formality)
data(presidential_debates_2012)
debateformality <- formality(presidential_debates_2012$dialogue, presidential_debates_2012$person)
```

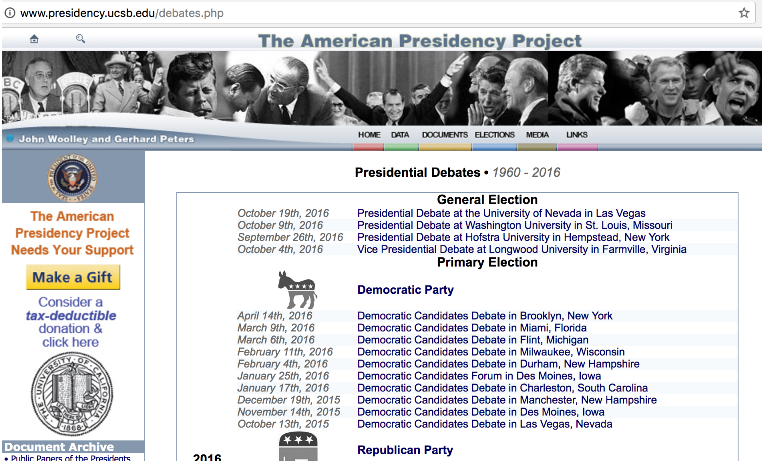
Some plotting capability

```
plot(debateformality)
```



Presidential Debates Online

Last course iteration, scraping and building the 2016 Presidential Debates corpus was one of the assignments.



www.presidency.ucsb.edu/debates.php

The American Presidency Project

John Woolley and Gerhard Peters

HOME DATA DOCUMENTS ELECTIONS MEDIA LINKS

The American Presidency Project Needs Your Support

Make a Gift

Consider a *tax-deductible* donation & click here

Document Archive

- Public Papers of the Presidents

Presidential Debates • *1960 - 2016*

General Election

October 19th, 2016	Presidential Debate at the University of Nevada in Las Vegas
October 9th, 2016	Presidential Debate at Washington University in St. Louis, Missouri
September 26th, 2016	Presidential Debate at Hofstra University in Hempstead, New York
October 4th, 2016	Vice Presidential Debate at Longwood University in Farmville, Virginia

Primary Election

Democratic Party

April 14th, 2016	Democratic Candidates Debate in Brooklyn, New York
March 9th, 2016	Democratic Candidates Debate in Miami, Florida
March 6th, 2016	Democratic Candidates Debate in Flint, Michigan
February 11th, 2016	Democratic Candidates Debate in Milwaukee, Wisconsin
February 4th, 2016	Democratic Candidates Debate in Durham, New Hampshire
January 25th, 2016	Democratic Candidates Forum in Des Moines, Iowa
January 17th, 2016	Democratic Candidates Debate in Charleston, South Carolina
December 19th, 2015	Democratic Candidates Debate in Manchester, New Hampshire
November 14th, 2015	Democratic Candidates Debate in Des Moines, Iowa
October 13th, 2015	Democratic Candidates Debate in Las Vegas, Nevada

2016

Republican Party

The 2016 Debates

Last course iteration, scraping and building the 2016 Presidential Debates corpus was one of the assignments.

```r
load("../../Data/PRESIDENTIAL-DEBATES.rdata")

debates_2016_final[, .N, by = debate][order(N, decreasing = TRUE)][1:10]
```

```
##                                                                                              debate
##  1:                                        Republican Candidates Debate in Simi Valley, California
##  2:                                              Republican Candidates Debate in Las Vegas, Nevada
##  3:                                      Republican Candidates Debate in Manchester, New Hampshire
##  4:                                                 Republican Candidates Debate in Houston, Texas
##  5:                                              Republican Candidates Debate in Detroit, Michigan
##  6:                                               Republican Candidates Debate in Des Moines, Iowa
##  7: Democratic Presidential Candidates Debate at Saint Anselm College in Manchester, New Hampshire
##  8:                         Vice Presidential Debate at Longwood University in Farmville, Virginia
##  9:         Democratic Presidential Candidates Debate at The Citadel in Charleston, South Carolina
## 10:                               Presidential Debate at Hofstra University in Hempstead, New York
##       N
##  1: 968
##  2: 756
##  3: 585
##  4: 535
##  5: 531
##  6: 516
##  7: 509
##  8: 500
##  9: 471
## 10: 470
```

Cleaning HTML fragments

Last course iteration, scraping and building the 2016 Presidential Debates corpus was one of the assignments.

```r
cleanfragment <- function(htmlString) {

    htmlString <- gsub("<.*?>", "", htmlString)
    htmlString <- gsub("\\[.*]", "", htmlString)
    htmlString <- gsub("&.*;", "", htmlString)

    return(htmlString)
}

debates_2016_final$fragment <- cleanfragment(debates_2016_final$fragment)
```

Cleaning HTML fragments

Last course iteration, scraping and building the 2016 Presidential Debates corpus was one of the assignments.

```
library(formality)

head(debates_2016_final[speaker %in% c("TRUMP", "CLINTON")]$fragment)

## [1] "Thank you very much, Chris. And thanks to UNLV for hosting us.You know, I think when we talk about the
## [2] "Well, first of all, it's great to be with you, and thank you, everybody. The Supreme Court"
## [3] "Well, first of all, I support the Second Amendment. I lived in Arkansas for 18 wonderful years. I repr
## [4] "Well, the D.C. vs. Heller decision was very stronglyand she was extremely angry about it. I watched. I
## [5] "Well, I was upset because, unfortunately, dozens of toddlers injure themselves, even kill people with
## [6] "Well, let me just tell you before we go any further. In Chicago, which has the toughest gun laws in th

FINAL <- debates_2016_final[pid == 119039][speaker %in% c("TRUMP", "CLINTON")]

formality2016 <- formality(FINAL$fragment, FINAL$speaker)
```

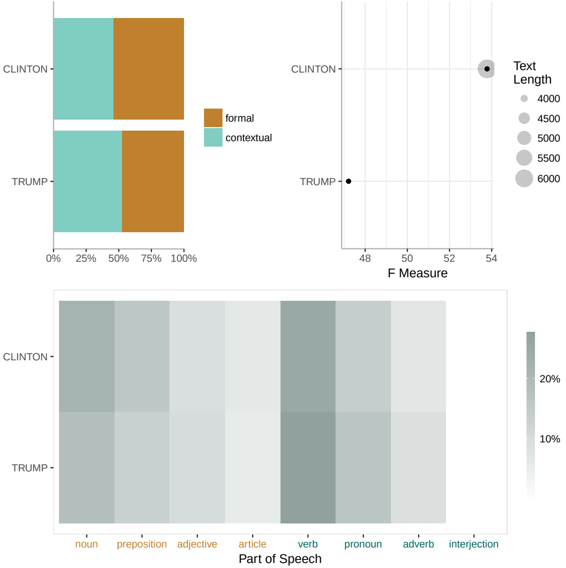
Guess who speaks more informally?

```
formality2016
```

```
##    speaker noun preposition adjective article verb pronoun adverb interjection formal
## 1: CLINTON 1376        1006       577     411 1568     876    444            8   3370
## 2:   TRUMP  714         516       385     222 1066     639    342            7   1837
##    contextual    n    F
## 1:       2896 6266 53.8
## 2:       2054 3891 47.2
```

Guess who speaks more informally?

```
plot(formality2016)
```


CLINTON
TRUMP
formal
contextual
CLINTON
TRUMP
F Measure
Text Length
CLINTON
TRUMP
noun
preposition
adjective
article
verb
pronoun
adverb
interjection
Part of Speech