EC999: Describing Text

Thiemo Fetzer

University of Chicago & University of Warwick

January 12, 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 $\frac{\text{total types}}{\text{total tokens}}$

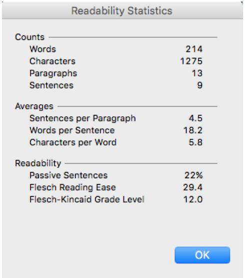
Guiraud $\frac{\text{total types}}{\sqrt{\text{total tokens}}}$

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



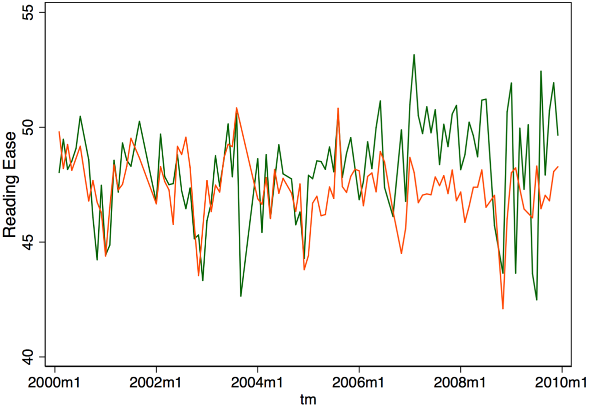
Readability Statistics

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

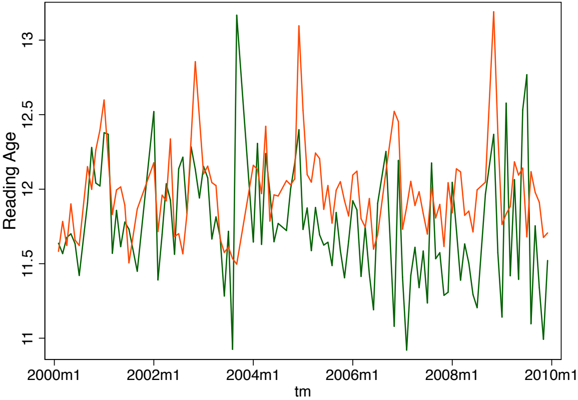
Reading Ease in Congress By Party



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

⇒ corpus data obtained via the Capitolwords API.

Reading Age in Congress By Part



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

⇒ 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 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$\}$
- $N = nf + nc$

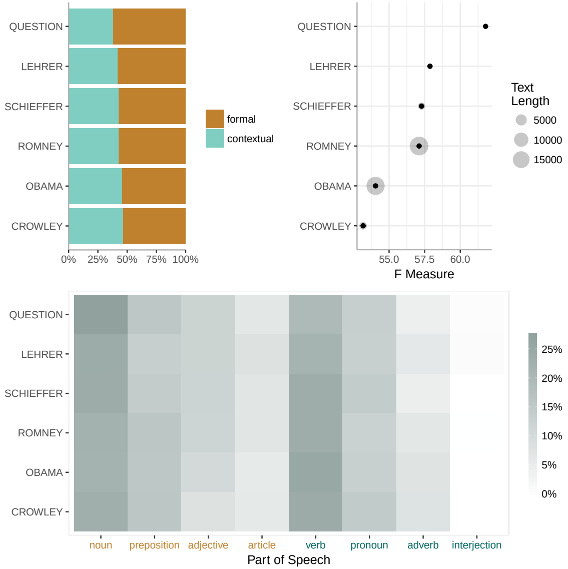
This yields an F-measure between 0 and 100%, with completely contextualized language on the zero end and completely formal language on the 100 end.
As is evident, this requires known *Parts of Speech*.

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)
```


QUESTION
LEHRER
SCHIEFFER
ROMNEY
OBAMA
CROWLEY
formal
contextual
F Measure
Text Length
noun
preposition
adjective
article
verb
pronoun
adverb
interjection
Part of Speech