Day 2: Descriptive statistical methods for textual analysis

Kenneth Benoit

Quantitative Analysis of Textual Data

September 30, 2014
Day 2 Outline

- Getting texts into quanteda
- Walk through Exercise 1
- Detecting collocations
- Exploring texts
- Describing textual data
- Quantifying lexical diversity
- Quantifying the complexity of texts
- Bootstrapping text
Getting texts into quanteda

- text format issue
  - text files
  - zipped text files
  - spreadsheets/CSV
  - (pdfs)
  - (Twitter feed)
- encoding issue
- metadata and document variable management
Identifying collocations

- Does a given word occur next to another given word with a higher relative frequency than other words?
- If so, then it is a candidate for a collocation
- We can detect these using measures of association, such as a likelihood ratio, to detect word pairs that occur with greater than chance frequency, compared to an independence model
- The key is to distinguish “true collocations” from uninteresting word pairs/triplets/etc, such as “of the”
- Implemented in quanteda as collocations
Table 5.1  Finding Collocations: Raw Frequency. $C(\cdot)$ is the frequency of something in the corpus.

(from Manning and Schütze, *FSNLP*, Ch 5)
Table 5.1  Finding Collocations: Raw Frequency. $C(\cdot)$ is the frequency of something in the corpus.

(from Manning and Schütze, *FSNLP*, Ch 5)
Detecting collocations: Constructing the association table

<table>
<thead>
<tr>
<th></th>
<th>Word 2</th>
<th>~ (Word 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word 1</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
</tr>
<tr>
<td>~ (Word 1)</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
</tr>
<tr>
<td></td>
<td>$n_{1.}$</td>
<td>$n_{2.}$</td>
</tr>
<tr>
<td></td>
<td>$n_{.1}$</td>
<td>$n_{.2}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$n$</td>
</tr>
</tbody>
</table>

where:

- $n_{ij}$ are observed counts
- $n_{i.}, n_{j}$ are row, column marginals
- $n$ is total token count
- $m_{ij} = \frac{n_{i.}n_{.j}}{n}$ is an expected count under the independence model
Method 1: Pearson’s chi-squared statistic

\[ X^2 = \sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \]

where \( X \sim \chi^2 \) with 1 d.f. [same as \( (I - 1)(J - 1) \)]
Method 2: Likelihood ratio test (Dunning)

\[
\begin{array}{ccc}
\text{Word 1} & \text{~ (Word 2)} \\
\hline
n_{11} & n_{12} & n_1. \\
\hline
n_{21} & n_{22} & n_2. \\
\hline
n_{.1} & n_{.2} & n \\
\end{array}
\]

\[
G^2 = 2 \sum_i \sum_j n_{ij} \ln \frac{n_{ij}}{m_{ij}}
\]

where \( G \sim \chi^2 \) with 1 d.f. [same as \((I - 1)(J - 1)\)]
Generalization to trigrams

\[ G^2 = 2 \sum_i \sum_j \sum_k n_{ijk} \ln \frac{n_{ijk}}{m_{ijk}} \]

where

\( G \sim \chi^2 \) with 1 d.f. [same as \((I - 1)(J - 1)(K - 1)\)]

\( m_{ijk} = \frac{n_{i..}n_{j..}n_{..k}}{n} \) is an expected count under the independence model

but the table of observed counts is slightly more complicated, as is the calculation of two words dependence but independence of the third – see Bautin and Hart for details
Other methods

- *t*-tests of frequencies (but assumes normality)
- mutual information, pointwise mutual information
- Pearson exact tests
- Many more: see Pecina (2005) for an exhaustive(ing) listing
Augmenting collocation detection with additional information

- Use parts of speech information

<table>
<thead>
<tr>
<th>Tag Pattern</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>A N</td>
<td>linear function</td>
</tr>
<tr>
<td>N N</td>
<td>regression coefficients</td>
</tr>
<tr>
<td>A A N</td>
<td>Gaussian random variable</td>
</tr>
<tr>
<td>A N N</td>
<td>cumulative distribution function</td>
</tr>
<tr>
<td>N A N</td>
<td>mean squared error</td>
</tr>
<tr>
<td>N N N</td>
<td>class probability function</td>
</tr>
<tr>
<td>N P N</td>
<td>degrees of freedom</td>
</tr>
</tbody>
</table>

Table 5.2 Part of speech tag patterns for collocation filtering. These patterns were used by Justeson and Katz to identify likely collocations among frequently occurring word sequences.

- other (machine prediction) tools
KWIC  *Key words in context* Refers to the most common format for concordance lines. A KWIC index is formed by sorting and aligning the words within an article title to allow each word (except the stop words) in titles to be searchable alphabetically in the index.

**lime (14)**

79[C.10] 4 /Which was builded of lime and sand;/Until they came to
247A.6 4 /That was well biggit with lime and stane.
303A.1 2 bower./Well built wi lime and stane./And Willie came
247A.9 2 /That was well biggit wi lime and stane./Nor has he stoln
305A.2 1 a castell biggit with lime and stane./O gin it stands not
305A.71 2 is my awin./I biggit it wi lime and stane;/The Tinnies and
79[C.10] 6 /Which was builded with lime and stone.
305A.30 1 a prittie castell of lime and stone./O gif it stands not
108.15 2 /Which was made both of lime and stone./Shee tooke him by
175A.33 2 castle then./Was made of lime and stone;/The vttermost
178[H.2] 2 near by./Well built with lime and stone;/There is a lady
178F.18 2 built with stone and lime!/But far mair pittie on Lady
178G.35 2 was biggit wi stane and lime!/But far mair pity o Lady
2D.16 1 big a cart o stane and lime./Gar Robin Redbreast trail it
Table 3
Example of Keyword in Context (KWIC) and associated word clusters display

Extracts from Keyword in Context (KWIC) list for the word ‘scan’
An MRI scan then indicated it had spread slightly
Fortunately, the MRI scan didn’t show any involvement of the lymph nodes
3 very worrying weeks later, a bone scan also showed up clear.
The bone scan is to check whether or not the cancer has spread to the bones.
The bone scan is done using a type of X-ray machine.
The results were terrific, CT scan and pelvic X-ray looked good
Your next step appears to be to await the result of the scan and I wish you well there.
I should go and have an MRI scan and a bone scan

Three-word clusters most frequently associated with keyword ‘scan’

<table>
<thead>
<tr>
<th>N</th>
<th>Cluster</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A bone scan</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>Bone scan and</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>An MRI scan</td>
<td>18</td>
</tr>
<tr>
<td>4</td>
<td>My bone scan</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>The MRI scan</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>The bone scan</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>MRI scan and</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>And MRI scan</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>Scan and MRI</td>
<td>9</td>
</tr>
</tbody>
</table>
Two cases:

1. I am really disappointed by what we have seen today. Instead of the Minister taking the radical
decisions, people on disability and even blind people... The Minister has some nerve quoting Ted
Kennedy, the champion of the poor and fairness in America; how much worse is it for the early school leaver and young unemployed person?

2. reminding everyone that I was the party that looked after child benefit. It would
be interesting how the Government would react.

I hear sports shops are doing a roaring trade in single golf clubs this Christmas. With a possible election next year, one never knows when a club might come in handy to
men who break their promises. The Minister should ask Tiger Woods about it.

I have read scores of articles by people who argue that child benefit payments are of little importance, including journalists and academics who argue it would make no difference if
the payment were restricted. Most of these articles were written by men, none of whom could state absolutely that he spoke for his wife or partner. I have yet to meet a mother
of young or teenage children who says casually that child benefit has no importance to her. Perhaps we do not live in the same world where this benefit is a trifle. Certainly, I do not represent
a constituency that places no value on the advantages of universal child benefit.

Almost every day I hear the voice of Marie Finucane on radio advertisements for the Simon Community, as I am sure everyone here does. She tells us that the current crisis has
brought community services to breaking point. I hear the same message from Professor John Monaghan of the Society of St. Vincent de Paul. Are these societies lying? Is the Simon
Community faking its message this Christmas? Is the Society of St. Vincent de Paul out of touch? Are they saying social welfare in Ireland is so generous that it can be cut? I have
Irish Budget Speeches KIWC in quanteda

```r
> data(iebudgets)
> iebudgets2010 <- subset(iebudgets, year==2010)
> kwic(iebudgets2010, "christmas", regex=TRUE)

preword  word                        postword
[2010_BUDGET_02_Richard_Bruton_FG.txt, 628] and to see out this Christmas in the hope of something
[2010_BUDGET_03_Joan_Burton_LAB.txt, 371] to suggest titles for a Christmas hit single. Fianna Fáil's hit
[2010_BUDGET_03_Joan_Burton_LAB.txt, 379] Fianna Fáil's hit single for Christmas will be, "I saw NAMA
[2010_BUDGET_03_Joan_Burton_LAB.txt, 922] women will say goodbye after Christmas because they must take the
[2010_BUDGET_03_Joan_Burton_LAB.txt, 1518] in single golf clubs this Christmas. With a possible election next
[2010_BUDGET_03_Joan_Burton_LAB.txt, 1726] Community faking its message this Christmas? bags. In previous years at Christmas
[2010_BUDGET_03_Joan_Burton_LAB.txt, 3159] €204 per week or the Christmas time people were laden down
[2010_BUDGET_04_Arthur_Morgan_SF.txt, 346] to social welfare payments this Christmas. The loss of the Christmas
[2010_BUDGET_04_Arthur_Morgan_SF.txt, 3244] Christmas. The loss of the Christmas bonus, a double payment which
[2010_BUDGET_04_Arthur_Morgan_SF.txt, 3272] streets on Santa presents and Christmas food. The Government's Scrooge measures
[2010_BUDGET_04_Arthur_Morgan_SF.txt, 5899] their jobs, who face this Christmas in debt, in poverty and
[2010_BUDGET_04_Arthur_Morgan_SF.txt, 5899] to implement the reduction before Christmas. I do not know whether
[2010_BUDGET_06_Enda_Kenny_FG.txt, 2629] from the change in the Christmas period. We suggested that the
[2010_BUDGET_07_Kieran_DODonnell_FG.txt, 1365] cut of €641, including the Christmas payment. A couple on invalidity
[2010_BUDGET_08_Eamon_Gilmore_LAB.txt, 550] are on social welfare, the Christmas payment is gone. Earnest lectures
[2010_BUDGET_08_Eamon_Gilmore_LAB.txt, 638] of emigration. Once again this Christmas, we will witness the scenes
[2010_BUDGET_08_Eamon_Gilmore_LAB.txt, 998] noted recently that over the Christmas recess work will be done
[2010_BUDGET_13_Ciaran_Cuffe_Green.txt, 911] will all be over by Christmas. If it is the last
[2010_BUDGET_14_Caoimhghin_OCaoalain_SF.txt, 148]```
Basic descriptive summaries of text

**Readability statistics** Use a combination of syllables and sentence length to indicate “readability” in terms of complexity

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

**Word (relative) frequency**

**Theme (relative) frequency**

**Length** in characters, words, lines, sentences, paragraphs, pages, sections, chapters, etc.
## Simple descriptive table about texts: Describe your data!

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Party</th>
<th>Tokens</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brian Cowen</td>
<td>FF</td>
<td>5,842</td>
<td>1,466</td>
</tr>
<tr>
<td>Brian Lenihan</td>
<td>FF</td>
<td>7,737</td>
<td>1,644</td>
</tr>
<tr>
<td>Ciaran Cuffe</td>
<td>Green</td>
<td>1,141</td>
<td>421</td>
</tr>
<tr>
<td>John Gormley (Edited)</td>
<td>Green</td>
<td>919</td>
<td>361</td>
</tr>
<tr>
<td>John Gormley (Full)</td>
<td>Green</td>
<td>2,998</td>
<td>868</td>
</tr>
<tr>
<td>Eamon Ryan</td>
<td>Green</td>
<td>1,513</td>
<td>481</td>
</tr>
<tr>
<td>Richard Bruton</td>
<td>FG</td>
<td>4,043</td>
<td>947</td>
</tr>
<tr>
<td>Enda Kenny</td>
<td>FG</td>
<td>3,863</td>
<td>1,055</td>
</tr>
<tr>
<td>Kieran O'Donnell</td>
<td>FG</td>
<td>2,054</td>
<td>609</td>
</tr>
<tr>
<td>Joan Burton</td>
<td>LAB</td>
<td>5,728</td>
<td>1,471</td>
</tr>
<tr>
<td>Eamon Gilmore</td>
<td>LAB</td>
<td>3,780</td>
<td>1,082</td>
</tr>
<tr>
<td>Michael Higgins</td>
<td>LAB</td>
<td>1,139</td>
<td>437</td>
</tr>
<tr>
<td>Ruairi Quinn</td>
<td>LAB</td>
<td>1,182</td>
<td>413</td>
</tr>
<tr>
<td>Arthur Morgan</td>
<td>SF</td>
<td>6,448</td>
<td>1,452</td>
</tr>
<tr>
<td>Caoimhghin O'Caolain</td>
<td>SF</td>
<td>3,629</td>
<td>1,035</td>
</tr>
<tr>
<td>All Texts</td>
<td></td>
<td>49,019</td>
<td>4,840</td>
</tr>
<tr>
<td><strong>Min</strong></td>
<td></td>
<td>919</td>
<td>361</td>
</tr>
<tr>
<td><strong>Max</strong></td>
<td></td>
<td>7,737</td>
<td>1,644</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td></td>
<td>3,704</td>
<td>991</td>
</tr>
<tr>
<td><strong>Hapaxes with Gormley Edited</strong></td>
<td></td>
<td>67</td>
<td></td>
</tr>
<tr>
<td><strong>Hapaxes with Gormley Full Speech</strong></td>
<td></td>
<td>69</td>
<td></td>
</tr>
</tbody>
</table>
Lexical Diversity

- Basic measure is the **TTR**: Type-to-Token ratio
- Problem: This is very sensitive to overall document length, as shorter texts may exhibit fewer word repetitions
- Special problem: length may relate to the introduction of additional subjects, which will also increase richness
Lexical Diversity: Alternatives to TTRs

TTR \( \frac{\text{total types}}{\text{total tokens}} \)

Guiraud \( \frac{\text{total types}}{\sqrt{\text{total tokens}}} \)

D (Malvern et al 2004) Randomly sample a fixed number of tokens and count those

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
Vocabulary diversity and corpus length

- In natural language text, the rate at which new types appear is very high at first, but diminishes with added tokens.

Fig. 1. Chart of vocabulary growth in the tragedies of Racine (chronological order, 500 token intervals).
Vocabulary Diversity Example

- Variations use automated segmentation – here approximately 500 words in a corpus of serialized, concatenated weekly addresses by de Gaulle (from Labbé et. al. 2004).
- While most were written, during the period of December 1965 these were more spontaneous press conferences.

Fig. 8. Evolution of vocabulary diversity in General de Gaulle’s broadcast speeches (June 1958–April 1969).
Complexity and Readability

- Use a combination of syllables and sentence length to indicate “readability” in terms of complexity
- Common in educational research, but could also be used to describe textual complexity
- Most use some sort of sample
- No natural scale, so most are calibrated in terms of some interpretable metric
- Not (yet) implemented in quanteda, but available from koRpus package
Flesch-Kincaid readability index

- **F-K** is a modification of the original Flesch Reading Ease Index:

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

**Interpretation:** 0-30: university level; 60-70: understandable by 13-15 year olds; and 90-100 easily understood by an 11-year old student.

- **Flesch-Kincaid** rescales to the US educational grade levels (1–12):

\[
0.39 \left( \frac{\text{total words}}{\text{total sentences}} \right) + 11.8 \left( \frac{\text{total syllables}}{\text{total words}} \right) - 15.59
\]
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
- Formula:

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

where 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
Sampling issues in existing measures

- Lexical diversity measures may take sample frames, or moving windows, and average across the windows.
- Readability may take a sample, or multiple samples, to compute readability measures.
- But rather than simulating the “sampling distribution” of a statistic, these are more designed to:
  - get a representative value for the text as a whole
  - normalize the length of the text relative to other texts
Bootstrapping text-based statistics
Simulation and bootstrapping

Used for:
- Gaining intuition about distributions and sampling
- Providing distributional information not distributions are not directly known, or cannot be assumed
- Acquiring uncertainty estimates

Both simulation and bootstrapping are numerical approximations of the quantities we are interested in. (Run the same code twice, and you get different answers)

Solution for replication: save the seed
Bootstrapping

- *Bootstrapping* refers to repeated resampling of data points with replacement
- Used to estimate the error variance (i.e. the **standard error**) of an estimate when the sampling distribution is unknown (or cannot be safely assumed)
- Robust in the absence of parametric assumptions
- Useful for some quantities for which there is no known sampling distribution, such as computing the standard error of a median
> ## illustrate bootstrap sampling
> set.seed(30092014)  # set the seed so that your results will match mine!
> # using sample to generate a permutation of the sequence 1:10
> sample(10)
[1]  4  2  1  9  8  5  7  3  6 10
> # bootstrap sample from the same sequence
> sample(10, replace=T)
[1]  8  6  6  2  5  8  4  8  4  9
> # bootstrap sample from the same sequence with probabilities that
> # favor the numbers 1-5
> prob1 <- c(rep(.15, 5), rep(.05, 5))
> prob1
[1] 0.15 0.15 0.15 0.15 0.15 0.05 0.05 0.05 0.05 0.05
> sample(10, replace=T, prob=prob1)
[1]  4  1  1  2  8  3  1  6  1  9
Bootstrapping the standard error of the median

Using a user-defined function:

```r
b.median <- function(data, n) {
  resamples <- lapply(1:n, function(i) sample(data, replace=T))
  sapply(resamples, median)
  std.err <- sqrt(var(r.median))
  list(std.err=std.err, resamples=resamples, medians=r.median)
}

summary(b.median(spending, 10))
summary(b.median(spending, 100))
summary(b.median(spending, 400))
median(spending)
```
Bootstrapping the standard error of the median

Using R’s `boot` library:

```r
library(boot)
samplemedian <- function(x, d) return(median(x[d]))
quantile(boot(spending, samplemedian, R=10)$t, c(.025, .5, .975))
quantile(boot(spending, samplemedian, R=100)$t, c(.025, .5, .975))
quantile(boot(spending, samplemedian, R=400)$t, c(.025, .5, .975))
```

**Note:** There is a good reference on using `boot()` from
Question: what is the “sampling distribution” of a text-based statistic? Examples:
- a term’s (relative) frequency
- lexical diversity
- complexity