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`

Example

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Table 5.1 Finding Collocations: Raw Frequency. $C(\cdot)$ is the frequency of something in the corpus.

(from Manning and Schütze, *FSNLP*, Ch 5)

Example

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

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

	Word 2	**~ (Word 2)**	
Word 1	n_{11}	n_{12}	$n_{1.}$
~ (Word 1)	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

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

	Word 2	**~ (Word 2)**	
Word 1	n_{11}	n_{12}	$n_{1.}$
~ (Word 1)	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

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

where $X \sim \chi^2$ with 1 d.f. [same as $(I-1)(J-1)$]

Method 2: Likelihood ratio test (Dunning)

	Word 2	~ (Word 2)	
Word 1	n_{11}	n_{12}	$n_{1.}$
~ (Word 1)	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

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

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

Tag Pattern	Example
A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*
N A N	*mean squared error*
N N N	*class probability function*
N P N	*degrees of freedom*

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

Exploring Texts: Key Words in Context

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	/W*hi*ch 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

Another KWIC Example (Seale et al (2006)

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

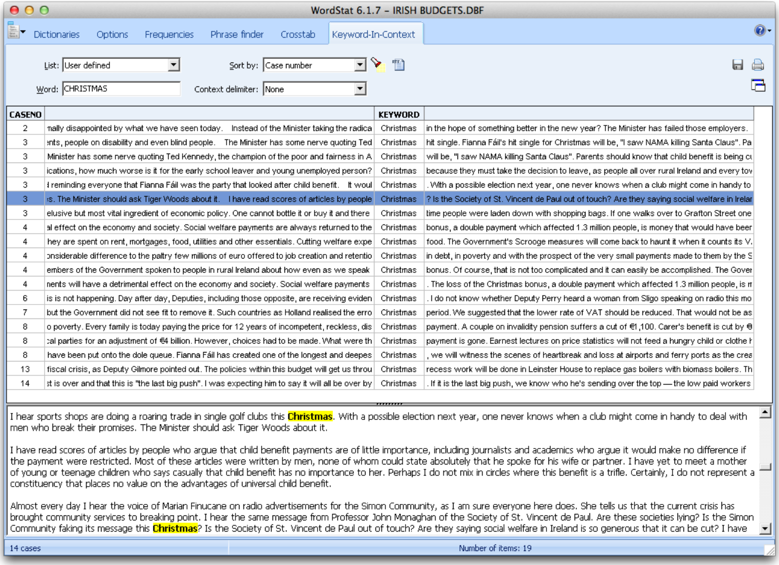
Your next step appears to be to await the result of the **scan** and I wish you well there.

I should go and have an MRI **scan** and a bone **scan**

Three-word clusters most frequently associated with keyword 'scan'

N	Cluster	Freq
1	A bone scan	28
2	Bone scan and	25
3	An MRI scan	18
4	My bone scan	15
5	The MRI scan	15
6	The bone scan	14
7	MRI scan and	12
8	And Mri scan	9
9	Scan and MRI	9

Another KWIC Example: Irish Budget Speeches



Irish Budget Speeches KIWC in quanteda

R Console

Help Search

```
> 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? Is the Society of St.
       [2010_BUDGET_03_Joan_Burton_LAB.txt, 3159]         bags. In previous years at Christmas  time people were laden down
       [2010_BUDGET_04_Arthur_Morgan_SF.txt, 346]              €204 per week or the Christmas  bonus. Of course, that is
      [2010_BUDGET_04_Arthur_Morgan_SF.txt, 3239]   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_06_Enda_Kenny_FG.txt, 2629] to implement the reduction before Christmas. I do not know whether
    [2010_BUDGET_07_Kieran_ODonnell_FG.txt, 1365]              from the change in the  Christmas period. We suggested that the
      [2010_BUDGET_08_Eamon_Gilmore_LAB.txt, 550]         cut of €641, including the Christmas  payment. A couple on invalidity
      [2010_BUDGET_08_Eamon_Gilmore_LAB.txt, 638]        are on social welfare, the Christmas  payment is gone. Earnest lectures
      [2010_BUDGET_08_Eamon_Gilmore_LAB.txt, 998]     of emigration. Once again this Christmas, we will witness the scenes
     [2010_BUDGET_13_Ciaran_Cuffe_Green.txt, 911]        noted recently that over the  Christmas recess work will be done
[2010_BUDGET_14_Caoimhghin_OCaolain_SF.txt, 148]              will all be over by Christmas. If it is the last
>
```

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!

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

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 introdution 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

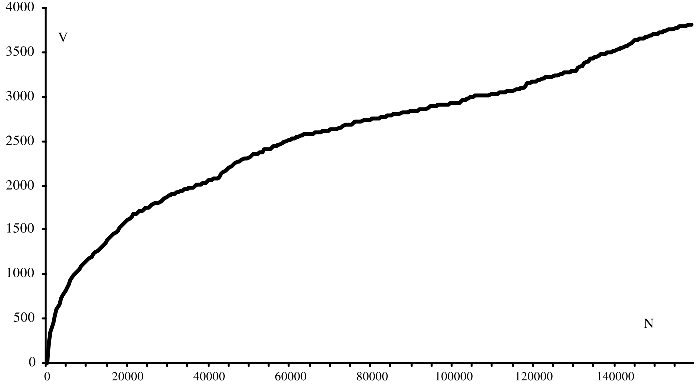


Fig. 1. Chart of vocabulary growth in the tragedies of Racine (chronological order, 500 token intervals).

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

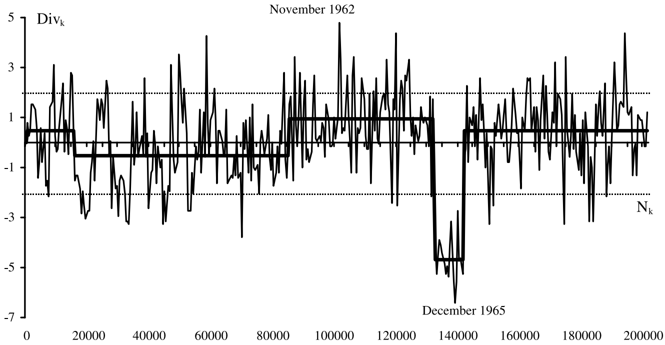


Fig. 8. Evolution of vocabulary diversity in General de Gaulle's broadcast speeches (June 1958–April 1969).

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

Bootstrapping illustrated

```
> ## illustrate bootstrap sampling
> set.seed(30092014)   # set the seed so that your results will match m
> # 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
> # boostrap 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:

```
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:

```
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 `http://www.mayin.org/ajayshah/KB/R/documents/boot.html`

Bootstrapping methods for textual data

- Question: what is the "sampling distribution" of a text-based statistic? Examples:
  - a term's (relative) frequency
  - lexical diversity
  - complexity