

# Day 3: Descriptive statistical methods for textual analysis

Kenneth Benoit

Essex Summer School 2014

July 23, 2014

## Day 3 Basic Outline

- ▶ Exploring texts
- ▶ Presenting overall text statistics
- ▶ Quantifying the complexity of texts
- ▶ Quantifying lexical diversity
- ▶ Summarizing and presenting results
- ▶ Graphical methods
- ▶ Demonstration with inaugural speeches
- ▶ Work through Exercise 2

## 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 /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

# 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'*

<i>N</i>	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

WordStat 6.1.7 - IRISH BUDGETS.DBF

Dictionary Options Frequencies Phrase finder Crosstab Keyword-In-Context

List: User defined Sort by: Case number

Word: CHRISTMAS Context delimiter: None

CASENO	TEXT	KEYWORD
2	nally disappointed by what we have seen today. Instead of the Minister taking the radical	Christmas
3	snts, people on disability and even blind people. The Minister has some nerve quoting Ted	Christmas
3	Minister has some nerve quoting Ted Kennedy, the champion of the poor and fairness in A	Christmas
3	ications, how much worse is it for the early school leaver and young unemployed person?	Christmas
3	r reminding everyone that Fianna Fáil was the party that looked after child benefit. It woul	Christmas
3	is. The Minister should ask Tiger Woods about it. I have read scores of articles by people	Christmas
3	elusive but most vital ingredient of economic policy. One cannot bottle it or buy it and there	Christmas
4	al effect on the economy and society. Social welfare payments are always returned to the	Christmas
4	hey are spent on rent, mortgages, food, utilities and other essentials. Cutting welfare expe	Christmas
4	nsiderable difference to the paltry few millions of euro offered to job creation and retento	Christmas
4	embers of the Government spoken to people in rural Ireland about how even as we speak	Christmas
4	nents will have a detrimental effect on the economy and society. Social welfare payments	Christmas
6	is is not happening. Day after day, Deputies, including those opposite, are receiving eviden	Christmas
7	but the Government did not see fit to remove it. Such countries as Holland realised the ero	Christmas
8	o poverty. Every family is today paying the price for 12 years of incompetent, reckless, dis	Christmas
8	cal parties for an adjustment of €6 billion. However, choices had to be made. What were th	Christmas
8	have been put onto the dole queue. Fianna Fáil has created one of the longest and deepest	Christmas
13	fiscal crisis, as Deputy Gilmore pointed out. The policies within this budget will get us throu	Christmas
14	it is over and that this is "the last big push". I was expecting him to say it will all be over by	Christmas

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 deal with 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 I do not mix in circles 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 Marian 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

14 cases Number of items: 19

# Irish Budget Speeches KIWC in quanteda

```
R Console
> 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_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
>
```

## Simple descriptive table about texts: Example

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 O'Donnell	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
<i>Min</i>		919	361
<i>Max</i>		7,737	1,644
<i>Median</i>		3,704	991
<i>Hapaxes with Gormley Edited</i>		67	
<i>Hapaxes with Gormley Full Speech</i>		69	

# 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.



## 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

# Scales and Indexes

- ▶ Involves characterizing the coded text units using additional quantification

- ▶ Examples

**Category frequencies** Coded category frequency measures, such as the proportion of times “economy” is mentioned in a speech, or the proportion of mentions of the environment

**Type/token measures** Frequency tabulations of token types and their frequencies

**Range/variance** Here we might be interested in the total number or the spread or variance of categories used in particular documents or by particular speakers

- ▶ May also involve scales or indexes constructed from summary information

## Summarizing: Scale Example

- ▶ A very simple example comes from the CMP, using PER110 “European Union: Positive Mentions” and PER108 “European Union: Negative Mentions”
- ▶ The overall pro- versus anti- EU-ness can be assessed as  $PER110 - PER108$ . Theoretical range is  $[-100, 100]$ .
- ▶ A more complicated example is the CMP’s famous “rile” index, which adds 26 categories of the “right” and subtracts from this the sum of 13 categories of the “left”.

# 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

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

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

**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 Example

- ▶ Variations use vocabulary diversity analysis (e.g. Labbé et. al. 2004)

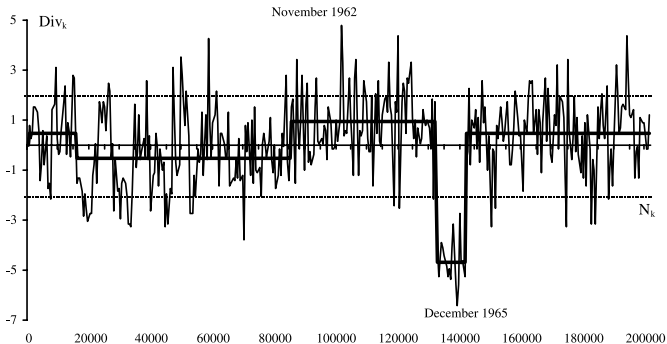


Fig. 8. Evolution of vocabulary diversity in General de Gaulle's broadcast speeches (June 1958–April 1969).

## Inference and Reporting

- ▶ This involves drawing conclusions from the research, and these conclusions will depend on the *validity* established by the research design
- ▶ Reporting means communicating the results in a clear and relevant fashion. (This can be challenging – see for instance the Schonhardt-Bailey article.)
- ▶ No iron-clad rules here – use your discretion as applied to a particular case



## Summarizing: Example

Democratic	Republican
iraq	consent
administration	ask
year	unanimous
health	bill
families	committee
program	senate
care	30
debt	2006
women	border
veterans	senator
help	vote
americans	law
country	hearing
children	authorized
new	further
education	states
funding	proceed
workers	order
programs	session
disaster	time

Top 20 Democratic and Republican words from the 2006 US Senate (source: Nicholas Beauchamp 2008)

## LIWC Example

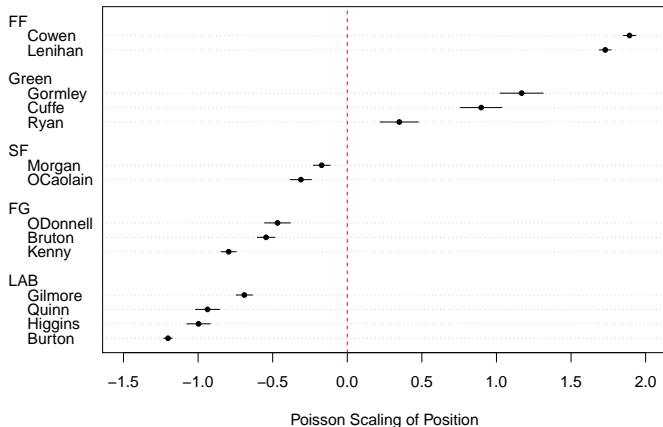
- From an application of the Linguistic Inquiry and Word Count dictionary to texts by Al Zawahiri and Bin Laden, benchmarked against a general corpus

	Bin Ladin (1988 to 2006) N = 28	Zawahiri (2003 to 2006) N = 15	Controls N = 17	p (two- tailed)
Word Count	2511.5	1996.4	4767.5	
Big words (greater than 6 letters)	21.2a	23.6b	21.1a	.05
Pronouns	9.15ab	9.83b	8.16a	.09
I (e.g. I, me, my)	0.61	0.90	0.83	
We (e.g. we, our, us)	1.94	1.79	1.95	
You (e.g. you, your, yours)	1.73	1.69	0.87	
He/she (e.g. he, hers, they)	1.42	1.42	1.37	
They (e.g., they, them)	2.17a	2.29a	1.43b	.03
Prepositions	14.8	14.7	15.0	
Articles (e.g. a, an, the)	9.07	8.53	9.19	
Exclusive Words (but, exclude)	2.72	2.62	3.17	
Affect	5.13a	5.12a	3.91b	.01
Positive emotion (happy, joy, love)	2.57a	2.83a	2.03b	.01
Negative emotion (awful, cry, hate)	2.52a	2.28ab	1.87b	.03
Anger words (hate, kill)	1.49a	1.32a	0.89b	.01
Cognitive Mechanisms	4.43	4.56	4.86	
Time (clock, hour)	2.40b	1.89a	2.69b	.01
Past tense verbs	2.21a	1.63a	2.94b	.01
Social Processes	11.4a	10.7ab	9.29b	.04
Humans (e.g. child, people, selves)	0.95ab	0.52a	1.12b	.05
Family (mother, father)	0.46ab	0.52a	0.25b	.08
Content				
Death (e.g. dead, killing, murder)	0.55	0.47	0.64	
Achievement	0.94	0.89	0.81	
Money (e.g. buy, economy, wealth)	0.34	0.38	0.58	
Religion (e.g. faith, Jew, sacred)	2.41	1.84	1.89	

Note. Numbers are mean percentages of total words per text file. Statistical tests are between Bin Ladin, Zawahiri, and Controls. Documents whose source indicates "Both" (n=3) or "Unknown" (n=2) were excluded due to their small sample sizes.

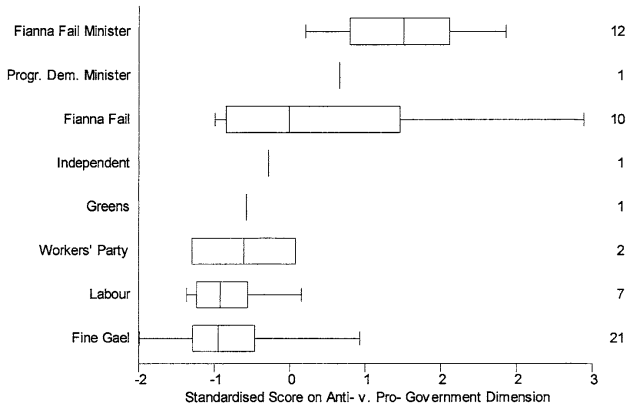
# Graphical Methods: Example

- ▶ From a uni-dimensional scaling model from a term-document matrix (Poisson scaling)



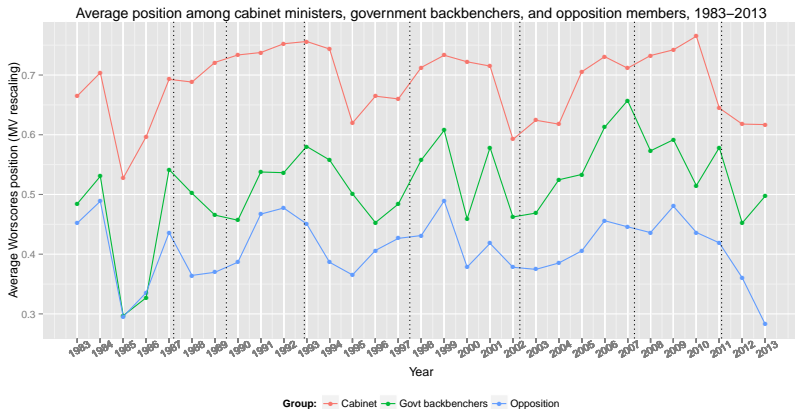
# No confidence debate speeches (Wordscores)

**FIGURE 3. Box Plot of Standardized Scores of Speakers in 1991 Confidence Debate on “Pro- versus Antigovernment” Dimension, by Category of Legislator**



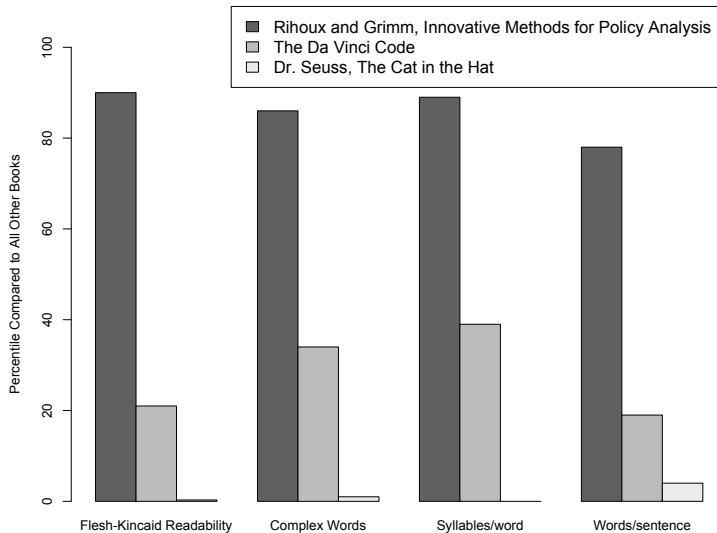
(from Benoit and Laver, *Irish Political Studies* 2002)

# Government v. Opposition in yearly budget debates



(from Herzog and Benoit EPSA 2013)

# Comparing Texts on the Basis of Quantitative Information



## Applied to political texts

Bush's second inaugural address:

freedom America

liberty nation American country world  
time free citizen hope history people day human right  
seen ideal work unite justice cause government move choice  
tyranny live act life accept defend duty generation great question honor  
states president fire character force power fellow enemy century witness excuse  
soul God division task define advance speak institution independence society serve

Obama's inaugural address:

nation America people  
work generation world common

time seek spirit day American peace crisis hard  
greater meet men remain job power moment women  
father endure government short hour life hope freedom carried  
journey forward force prosperity courage man question future friend  
service age history God oath understand ideal pass economy care  
promise children Earth stand demand purpose faith hand found interest

