

# Day 1: Introduction and Issues in Quantitative Text Analysis

Kenneth Benoit

Essex Summer School 2013

July 22, 2013

# Today's Basic Outline

- ▶ Motivation for this course
- ▶ Logistics
- ▶ Issues
- ▶ Examples
- ▶ Class exercise of working with texts

## Class schedule: Typical day

14:15–15:45 Lecture

15:55–16:35 Focus on Examples

16:45–17:45 In-class exercises (Lab)

# MOTIVATION

# Motivation

- ▶ Whom this class is for
- ▶ Learning objectives
- ▶ Prior knowledge
  - ▶ quantitative methods (intermediate statistics)
  - ▶ familiarity with some sort of quantitative analysis software, preferably R
  - ▶ ability and willingness to try to learn QDA Miner/Wordstat
  - ▶ ability to use a **text editor**
  - ▶ (optional) ability to process text files in a programming language such as Python or Perl

# What is Quantitative Text Analysis?

- ▶ A variant of **content analysis** that is expressly quantitative, not just in terms of representing textual content numerically but also in analyzing it, typically using computers
- ▶ “Mild” forms reduce text to quantitative information and analyze this information using quantitative techniques
- ▶ “Extreme” forms treat text units as data directly and analyze them using statistical methods
- ▶ Necessity spurred on by huge volumes of text available in the electronic information age
- ▶ (Particularly “text as data”) An emerging field with many new developments in a variety of disciplines

## What Quantitative Text Analysis is not

- ▶ Not discourse analysis, which is concerned with how texts as a whole represent (social) phenomena
- ▶ Not social constructivist examination of texts, which is concerned with the social constitution of reality
- ▶ Not rhetorical analysis, which focuses on how messages are delivered stylistically
- ▶ Not ethnographic, which are designed to construct narratives around texts or to discuss their “meaning” (what they *really* say as opposed to what they *actually* say)
- ▶ Any non-explicit procedure that cannot be approximately replicated

(more exactly on how to define **content analysis** later)

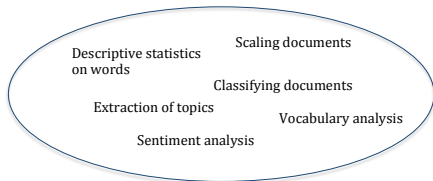
# Basic QTA Process: Texts → Feature matrix → Analysis

When I presented the supplementary budget to this House last April, I said we could work our way through this period of severe economic distress. Today, I can report that notwithstanding the difficulties of the past eight months, we are now on the road to economic recovery.

In this next phase of the Government's plan we must stabilise the deficit in a fair way, safeguard those worst hit by the recession, and stimulate crucial sectors of our economy to sustain and create jobs. The worst is over.

This Government has the moral authority and the well-grounded optimism rather than the cynicism of the Opposition. It has the imagination to create the new jobs in energy, agriculture, transport and construction that this green budget will create. It has the

docs	made	because	had	into	get	some	through	next	where	many	irish
t06_kenny_fg	12	11	5	4	8	4	3	4	5	7	10
t05_cowen_ff	9	4	8	5	5	5	14	13	4	9	8
t14_o'caolain_sf	3	3	3	4	7	3	7	2	3	5	6
t01_lenihan_ff	12	1	5	4	2	11	9	16	14	6	9
t11_gormley_green	0	0	0	3	0	2	0	3	1	1	2
t04_morgan_sf	11	8	7	15	8	19	6	5	3	6	6
t12_ryan_green	2	2	3	7	0	3	0	1	6	0	0
t10_quinn_lab	1	4	4	2	8	4	1	0	1	2	0
t07_odonne11_fg	5	4	2	1	5	0	1	1	0	3	0
t09_higgins_lab	2	2	5	4	0	1	0	0	2	0	0
t03_burton_lab	4	8	12	10	5	5	4	5	8	15	8
t13_cuffe_green	1	2	0	0	11	0	16	3	0	3	1
t08_gilmore_lab	4	8	7	4	3	6	4	5	1	2	11
t02_burton_fg	1	10	6	4	4	3	0	6	16	5	3





ISSUES

## Is there any difference between “qualitative” and “quantitative” text analysis?

- ▶ Ultimately all reading of texts is qualitative, even when we count elements of the text or convert them into numbers
- ▶ QTA may involve human judgment in the construction of the feature-document matrix
- ▶ But quantitative text analysis differs from more qualitative approaches in that it:
  - ▶ Involves large-scale analysis of many texts, rather than close readings of few texts
  - ▶ Requires no interpretation of texts in a non-positivist fashion
  - ▶ Does not explicitly concern itself with the social or cultural predispositions of the analysts (not critical or constructivist)
- ▶ Uses a variety of statistical techniques to extract information from the feature-document matrix

## Relationship to “content analysis”

- ▶ Classical content analysis receives a day (Day 3) but course is broader than classical content analysis
- ▶ Classical (quantitative) content analysis consists of applying explicit coding rules to classify content, then summarizing these numerically. Examples:
  - ▶ Frequency analysis of article types in an academic journal (this is content analysis at the unit of the *article*)
  - ▶ Determination of different forms of affect in sets of speeches, for instance positive or negative evaluations in free-form text responses on surveys, by applying a dictionary
  - ▶ Machine coding of texts using dictionaries and complicated rules sets (e.g. using *WordStat*, *Diction*, etc.) also covered minimally in this course
- ▶ BUT: much content will be shaped by participant problems

## Several main approaches to text analysis

- ▶ Purely qualitative  
(qualitative)
- ▶ Human coded, quantitative summary  
(qualitative/quantitative)

# Human coded example: Comparative Manifesto Project

## Enterprise & Jobs

Our programme of infrastructure investment through the Scottish Trust for Public Investments will give Scots businesses improved access to world markets through a modern and reliable road, rail, sea and air network. We will ensure Scotland does not get by-passed by the digital revolution by ensuring that Scotland has direct access to the internet and broadband capacity throughout the country. And our focus on reskilling Scotland will work to ensure that one of the key ingredients of a successful economy, a highly educated, flexible and skilled workforce, is in place to allow both the growth of indigenous enterprises, but also to encourage the relocation of high-skill, value-added international investors to our country.

Economic development agencies must become more focused and less bureaucratic. They must be more accessible and less regulatory. Their aim is to facilitate and add value to indigenous and incoming business. They should stimulate not suffocate.

Finally, because we believe in Scotland, because we stand for Scotland, we will be best placed to sell Scotland as a marketplace, as a holiday destination and as a key export partner. We will ensure that Scotland's businesses get better and wider representation across the world, and that every effort is made to promote Scotland as a world beating business and tourist centre. To this end, we will bring the tourist agency into Scotland's enterprise network.

411  
402  
602 401 601  
401 401  
401  
402 402  
303  
201  
303 402  
402  
601  
402 402  
402 402  
602 402  
402

## Several main approaches to text analysis (continued)

- ▶ **Purely machine processed**  
(quantitative with human decision elements)
- ▶ **Text as data approaches**  
(purely quantitative with minimal to no human decision elements)

# Key feature of quantitative text analysis

1. **Conversion** of texts into a common electronic format.
2. **(Sometimes) Pre-processing of texts.** e.g. stemming
3. **Conversion of textual features into a quantitative matrix.**  
Features can mean:
  - ▶ words  $\times$  documents
  - ▶ words  $\times$  some variable
  - ▶ word counts  $\times$  documents/variables
  - ▶ linguistic features  $\times$  documents
  - ▶ abstracted concepts  $\times$  other abstracted concepts
4. A **quantitative or statistical procedure** to extract information from the quantitative matrix
5. **Summary** and interpretation of the quantitative results

# LOGISTICS



## Short Course Schedule

Day	Date	Topic(s)	Details
Mon	22 July	Introduction and Issues in text analysis	Course goals; logistics; software overview; conceptual foundations; quantitative text analysis as a field; objectives; examples.
Tue	23 July	Forms of textual data, units of analysis, and definitions of features	Where to obtain textual data; formatting and working with text files; indexing and meta-data; units of analysis; and definitions of features and measures commonly extracted from texts, including stemming, stop-words, and feature weighting.
Wed	24 July	Research Strategies in Quantitative Text Analysis	This section addresses the main challenges involved in quantitative text research, reliability, measurement validity, scaling approaches and basic diagnostics, and sampling strategies.
Thu	25 July	Quantitative methods for comparing texts	Quantitative methods for comparing texts, such as characterizing texts through concordances, co-occurrences, and keywords in context; complexity and readability measures; dissimilarity measure; and an in-depth discussion of text types, tokens, and equivalencies.
Fri	26 July	Quantitative content analysis	How to convert text into quantitative matrixes for analysis using human coding; reliability checks; how to develop and test a coding scheme.

---

Mon	29 July	Dictionary methods	How to convert text into quantitative matrixes using dictionary approaches, including guidelines for constructing, testing, and refining dictionaries. Covers commonly used dictionaries such as LIWC, RID, and the Harvard IV-4, with applications.
Tue	30 July	Document classification	Statistical methods for classifying documents into categories, the nature of category systems, and special issues arising from using words as data. The topic also introduces validation and reporting methods for classifiers and discusses where these methods are applicable.
Wed	31 July	Non-parametric scaling models for text	The “Wordscores” approach to scaling latent traits using a Naïve Bayes foundation; Correspondence Analysis applied to texts.
Thurs	1 August	Parametric scaling models for text	Poisson scaling models (aka “wordfish”) of latent word and document traits, and their applications.
Fri	2 August	Working with Big Text Data: Twitter	An introduction to applying quantitative text analysis to Twitter Data, with examples.

---

## Software requirements for this course

- ▶ A text editor you know and love
  - ▶ Recommendations: Sublime Text 2, Emacs, TextMate (Mac), Notepad++ (Windows)
  - ▶ Many others available: see [http://en.wikipedia.org/wiki/List\\_of\\_text\\_editors](http://en.wikipedia.org/wiki/List_of_text_editors)
  - ▶ The key is that you know the difference between text editors and (e.g.) Microsoft Word
- ▶ Some familiarity with the **command line** is highly desirable
- ▶ Prior experience with a statistical package – we will use R in this course

## Software requirements (cont.)

- ▶ Any prior use of a computerized content analysis tool is helpful (but not essential) — we will learn/use QDAMiner/Wordstat
- ▶ Some of the software is home-grown: QUANTEDA (<http://github.com/kbenoit/quanteda>)
- ▶ Our exercises using software will be guided, with explicit instructions
- ▶ Lots of work with real texts and applications

# Who I am

- ▶ Instructor: Ken Benoit, London School of Economics  
kbenoit@lse.ac.uk
- ▶ TAs:
  - ▶ Paul Nulty, p.nulty@lse.ac.uk
  - ▶ Petra Martina Baumann, PetraMartina.Baumann@sbg.ac.at
- ▶ Course homepage:  
<http://www.kenbenoit.net/essex2013qta>
- ▶ *Introductions ...*

## Course resources

- ▶ **Syllabus:** describes class, lists readings, links to reading, and links to exercises and datasets
- ▶ **Web page** on <http://www.kenbenoit.net/essex2013qta>
  - ▶ Contains course handout
  - ▶ Slides from class
  - ▶ In-class exercises and supporting materials
  - ▶ Texts for analysis
  - ▶ (links to) Software tools and instructions for use
- ▶ **Main readings**
  - ▶ Krippendorff
  - ▶ Lots of articles
  - ▶ Some other texts or on-line articles linked to the course handout (downloadable online)

# EXAMPLES

# You have already done QTA!

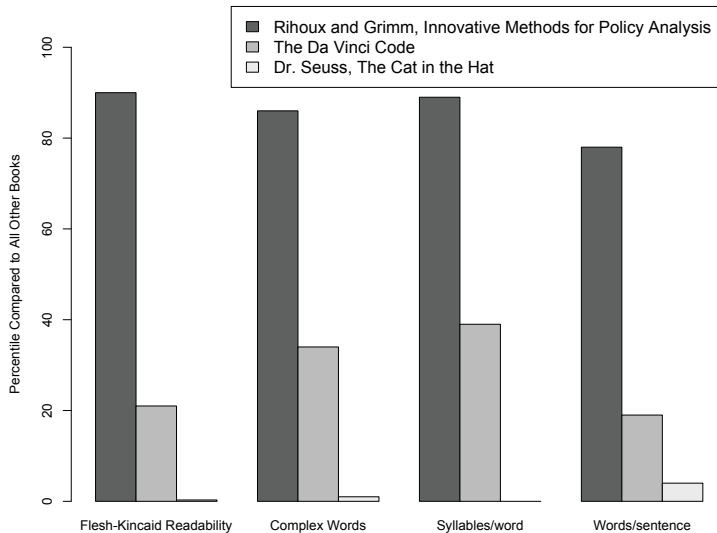
- ▶ Probably every day: Google searches (and many other Google products)
- ▶ Amazon.com does interesting text statistics:

Here is an analysis of the text of Dan Brown's *Da Vinci Code*:

<b>Readability</b> ( <a href="#">learn more</a> )		<b>Compared with other books</b>		
Fog Index:	8.8	20% are easier	▼	80% are harder
Flesch Index:	65.2	25% are easier	▼	75% are harder
Flesch-Kincaid Index:	6.9	21% are easier	▼	79% are harder
<b>Complexity</b> ( <a href="#">learn more</a> )				
Complex Words:	11%	34% have fewer	▼	66% have more
Syllables per Word:	1.5	39% have fewer	▼	61% have more
Words per Sentence:	11.0	19% have fewer	▼	81% have more
<b>Number of</b>				
Characters:	823,633	85% have fewer	▼	15% have more
Words:	138,843	88% have fewer	▼	12% have more
Sentences:	12,647	94% have fewer	▼	6% have more



# Comparing Texts on the Basis of Quantitative Information



## But Political Texts are More Interesting

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

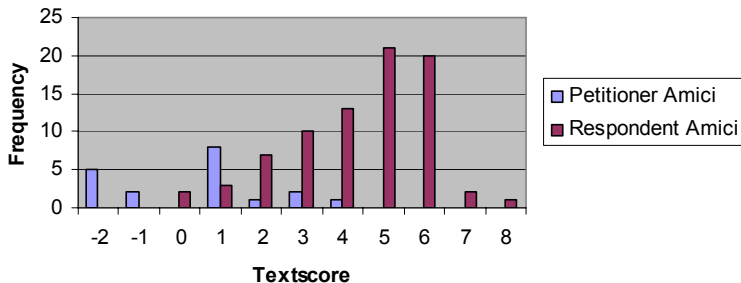


## Legal document scaling: “Wordscores”

### Amicus Curiae Textscores by Party

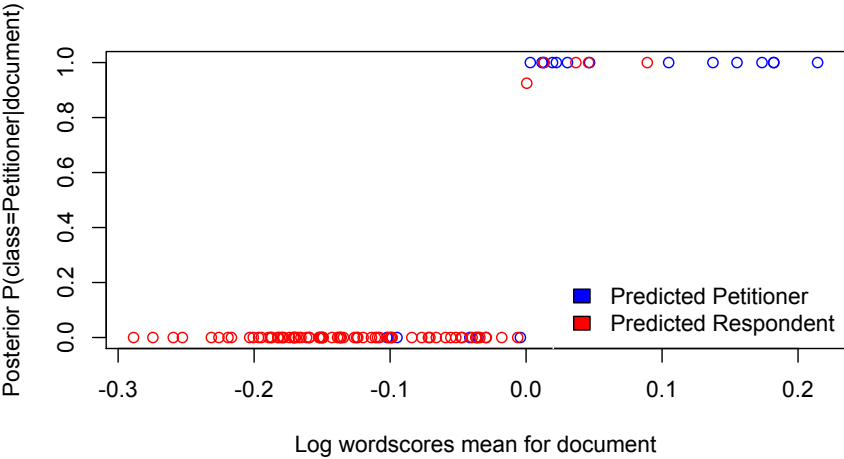
Using Litigants' Briefs as Reference Texts

(Set Dimension: *Petitioners = 1, Respondents = 5*)



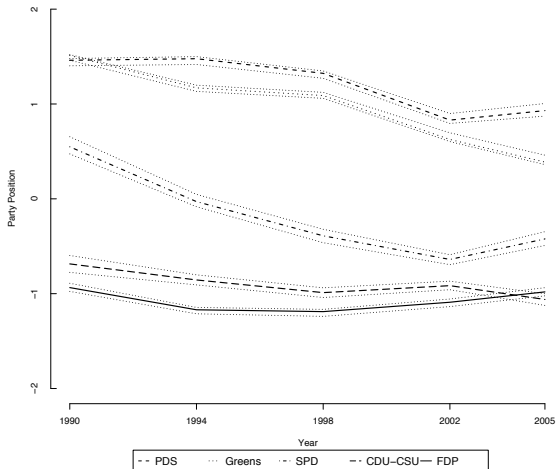
(from Evans et. al. 2007)

# Document classification: "Naive Bayes" classifier



# Party Manifestos: Poisson scaling

Left-Right Positions in Germany, 1990-2005  
including 95% confidence intervals



(from Slapin and Proksch, *AJPS* 2008)

# Party Manifestos: More scaling with Wordscores

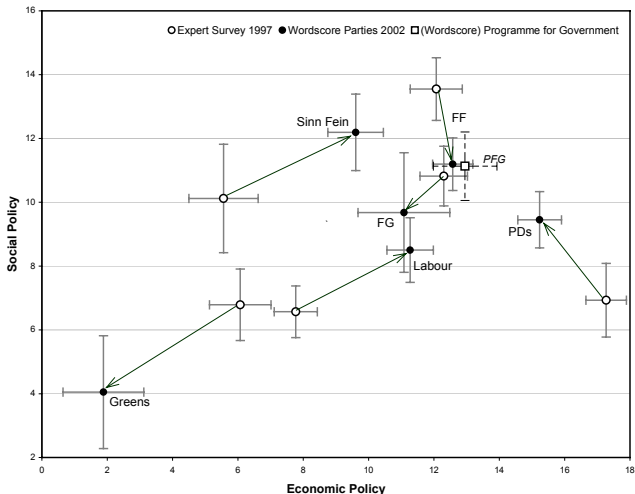
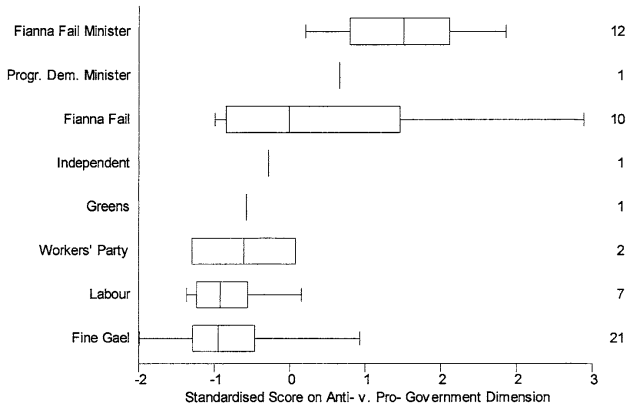


Figure 1. Movement from 1997 Positions on Economic and Social Policy, based on Wordscores Estimates. Bars indicate two standard errors on each scale.

(from Benoit and Laver, *Irish Political Studies* 2003)

# 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**

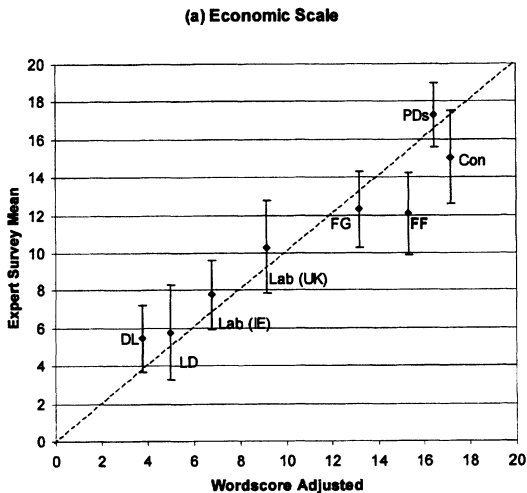


(from Benoit and Laver, *Irish Political Studies* 2002)



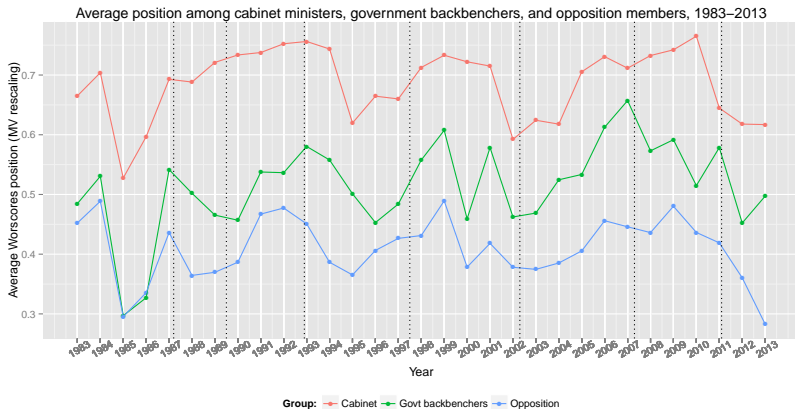
## Text scaling versus human experts

**FIGURE 2. Agreement Between Word Score Estimates and Expert Survey Results, Ireland and United Kingdom, 1997, for (a) Economic and (b) Social Scales**



(from Laver, Benoit and Carry, *ABSP* 2002)

# Government v. Opposition in yearly budget debates



(from Herzog and Benoit EPSA 2013)

## Published examples on reading list

- ▶ Schonhardt-Bailey (2008)
- ▶ Gebauer et al. (2007)