Day 1: Quantitative Text Analysis Overview and Fundamentals

Kenneth Benoit

Essex Summer School 2014

July 21, 2014

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 required
  - quantitative methods (intermediate statistics)
  - familiarity with R
  - ability to use a text editor
  - (optional) ability to process text files in a programming language such as Python

What is Quantitative Text Analysis?

- A variant of content analysis that is expressly quantititative, 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

docs \ words	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_ocaolain_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_odonnell_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_bruton_fg	1	10	6	4	4	3	0	6	16	5	3

Descriptive statistics on words

Scaling documents

Classifying documents

Extraction of topics

Vocabulary analysis

Sentiment analysis

This requires assumptions

- That texts represent an observable implication of some underlying characteristic of interest (usually an attribute of the author)
- That texts can be represented through extracting their *features*
    - most common is the bag of words assumption
    - many other possible definitions of "features"
- A document-feature matrix can be analyzed using quantitative methods to produce meaningful and valid estimates of the underlying characteristic of interest

ISSUES

Relationship to “content analysis”

Classical (quantitative) content analysis consists of applying explicit coding rules to classify content, then summarizing these numerically. Examples:

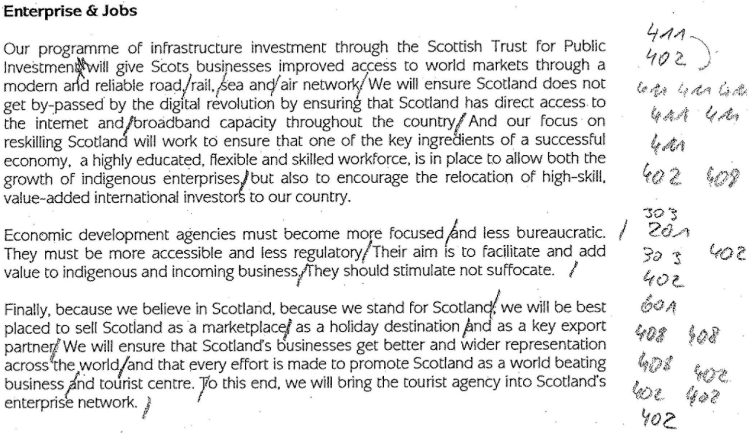
- Frequency analysis of article types in an academic journal (this is content analysis at the unit of the *article*)
- Determination of different forms of affect in sets of speeches, for instance positive or negative evaluations in free-form text responses on surveys, by applying a dictionary
- Machine coding of texts using dictionaries and complicated rules sets (e.g. using *WordStat*, *Diction*, etc.) also covered minimally in this course

Krippendorff book is 90% focused on this form of analysis, but still provides a good foundation

What role for "qualitative" analysis in QTA?

- 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 qualitiative 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 document-feature matrix

Human coded example: Comparative Manifesto Project



Enterprise & Jobs

Our programme of infrastructure investment through the Scottish Trust for Public Investment 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
411 411 411
411 411
411
402 408
303
201
303 402
402
601
408 408
408 402
402 402
402

Content analysis: uses human coding to define and select the features

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

	words										
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_ocaolain_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_odonnell_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_bruton_fg	1	10	6	4	4	3	0	6	16	5	3

Descriptive statistics on words

Scaling documents

Classifying documents

Extraction of topics

Vocabulary analysis

Sentiment analysis

Key feature of quantitative text analysis

- Selecting texts: Defining the *corpus*
- Conversion of texts into a common electronic format
- Defining documents: deciding what will be the doumentary unit of analysis

Key feature of quantitative text analysis (cont.)

- Defining features, for instance
  - words
  - word stems or lemmas: this is a form of defining *equivalence classes* for word features
  - word segments, espe-
    cially for languages using compound words, such as German, e.g.
    Rindfleischetikettierungsberwachungsaufgabenbertragungsgesetz
    (the law concerning the delegation of duties for the supervision of cattle marking and the labelling of beef)
    Saunauntensitzer
  - "word" sequences, especially when inter-word delimiters (usually white space) are not commonly used, as in Chinese

    莎拉波娃现在居住在美国东南部的佛罗里达。今年 4 月 9 日，莎拉波娃在美国第一大城市纽约度过了 1 8 岁生日。生日派对上，莎拉波娃露出了甜美的微笑。
  - linguistic features, such as parts of speech
  - (if qualitative coding is used) coded or annotated text segments

Key feature of quantitative text analysis (cont.)

- Conversion of textual features into a quantitative matrix. Features can mean:
- A quantitative or statistical procedure to extract information from the quantitative matrix
- Summary and interpretation of the quantitative results

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

	words										
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_ocaolain_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_odonnell_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_bruton_fg	1	10	6	4	4	3	0	6	16	5	3

Descriptive statistics on words

Scaling documents

Classifying documents

Extraction of topics

Vocabulary analysis

Sentiment analysis

LOGISTICS

Day	Date	Topic(s)	Details
Mon	21 July	Quantitative text analysis overview and fundamentals	Course goals; logistics; software overview; conceptual foundations; quantitative text analysis as a field; objectives; examples.
Tue	22 July	Working with texts, defining documents and features, weighting	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; identifying collocations.
Wed	23 July	Descriptive statistical methods for textual analysis	Quantitative methods for describing texts, such as characterizing texts through concordances, co-occurrences, and keywords in context; complexity and readability measures; and an in-depth discussion of text types, tokens, and equivalencies.
Thu	24 July	Quantitative methods for comparing texts	Quantitative methods for comparing texts, such as keyword identification, dissimilarity measures, association models, vector space models.
Fri	25 July	Automated 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.

Mon	28 July	Document classifiers	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.
Tue	29 July	Unsupervised models for scaling texts	The "Wordscores" approach to scaling latent traits using a Naïve Bayes foundation; Correspondence Analysis applied to texts.
Wed	30 July	Supervised models for scaling texts	Poisson scaling models (aka "wordfish") of latent word and document traits, and their applications.
Thurs	31 August	Clustering methods and topic models	An introduction to hierarchical clustering for textual data, including parametric topic models such as Latent Dirichlet Allocation (LDA)
Fri	1 August	Mining Social Media: An application to textual analysis of Twitter data.	Methods for extracting text and meta-data from Twitter feeds and applying sentiment analysis to these feeds.

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`
  - 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), e.g. QDAMiner/Wordstat
- Our software is home-grown: `quanteda` (`http://github.com/kbenoit/quanteda`)
- Our exercises using software will be guided, with explicit instricutions
- 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
- Course homepage:
  `http://www.kenbenoit.net/essex2014qta`
- *Introductions . . .*

Course resources

- Syllabus: describes class, lists readings, links to reading, and links to exercises and datasets
- Web page on `http://www.kenbenoit.net/essex2014qta`
  - 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 book
  - Lots of articles
  - Some other texts or on-line articles linked to the course handout (downloadable online)

EXAMPLES

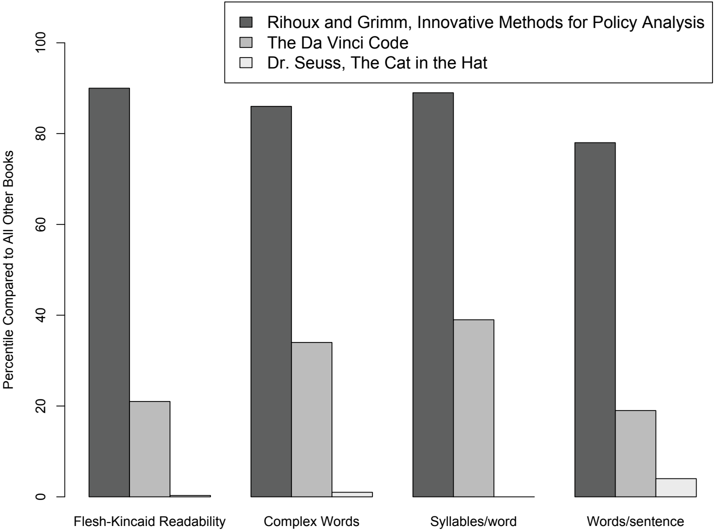
You have already done QTA!

- 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*:

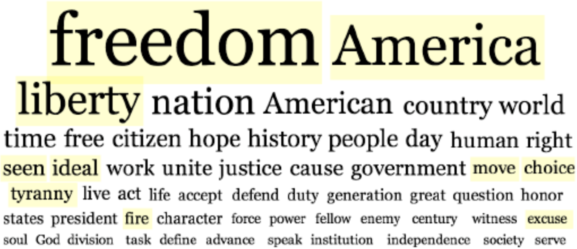
Readability (learn more)		**Compared with other books**	
Fog Index:	8.8	20% are easier	80% are harder
Flesch Index:	65.2	25% are easier	75% are harder
Flesch-Kincaid Index:	6.9	21% are easier	79% are harder
Complexity (learn more)			
Complex Words:	11%	34% have fewer	66% have more
Syllables per Word:	1.5	39% have fewer	61% have more
Words per Sentence:	11.0	19% have fewer	81% have more
Number of			
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

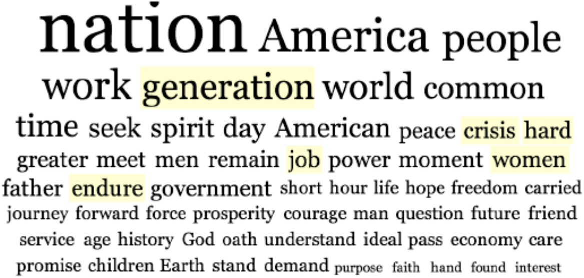


But Political Texts are More Interesting

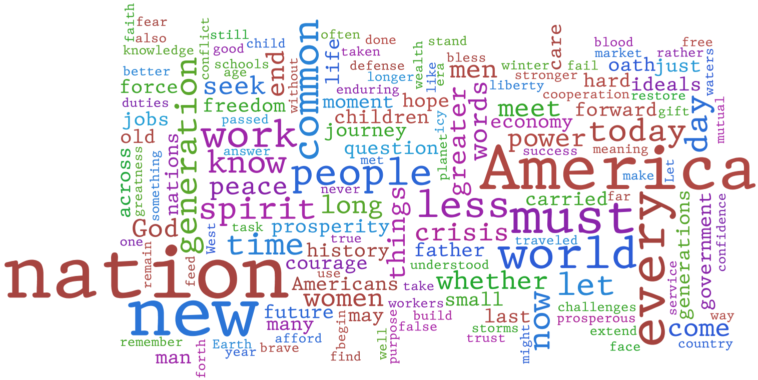
Bush's second inaugural address:



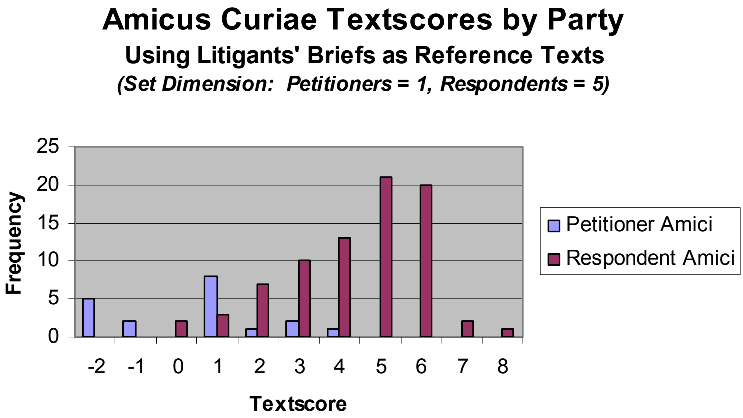
Obama's inaugural address:



Obama's Inaugural Speech in Wordle

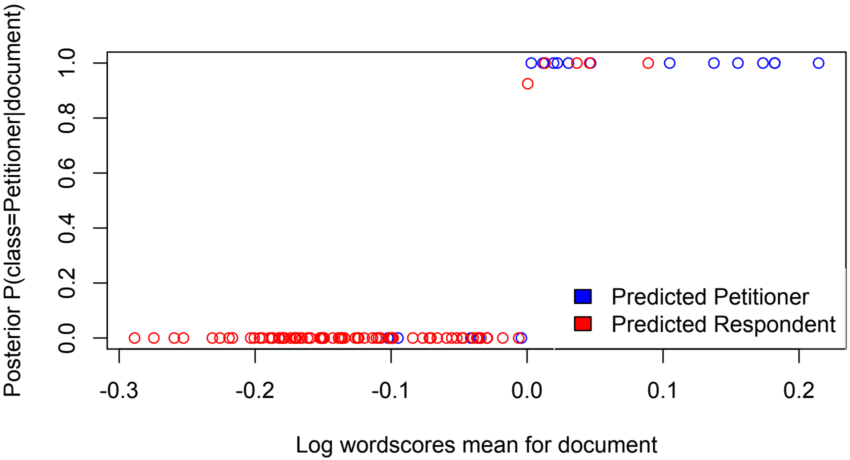


Legal document scaling: "Wordscores"

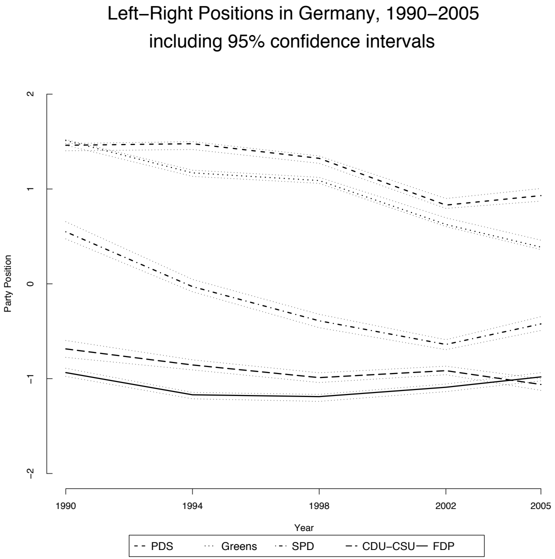


(from Evans et. al. 2007)

Document classification: "Naive Bayes" classifier



Party Manifestos: Poisson scaling



(from Slapin and Proksch, *AJPS* 2008)

Party Manifestos: More scaling with Wordscores

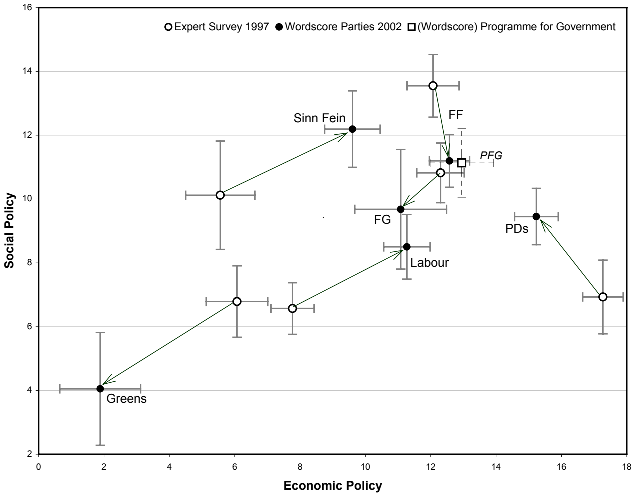
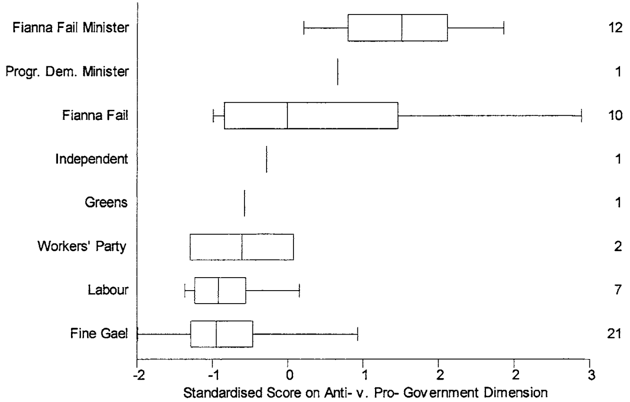


Figure 1. Movement from 1997 Positions on Economic and Social Policy, based on Wordscores Estimates. Bars indicate two standard errors on each scale.

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

No confidence debate speeches (Wordscores)

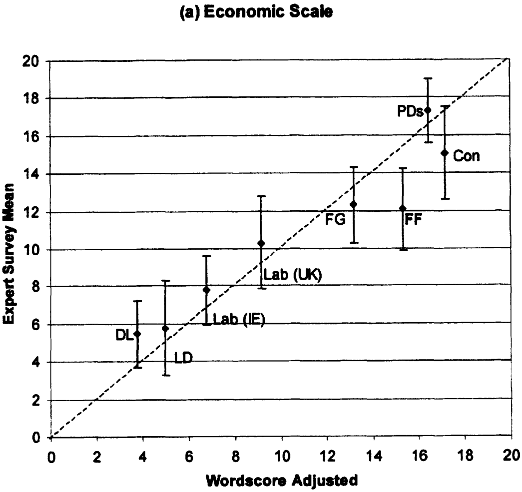
FIGURE 3. Box Plot of Standardized Scores of Speakers in 1991 Confidence Debate on "Pro- versus Antigovernment" Dimension, by Category of Legislator



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

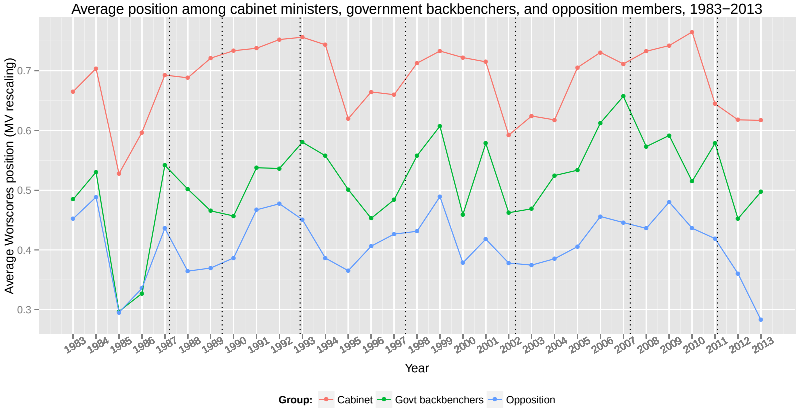
Text scaling versus human experts

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 Garry, *APSR* 2003)

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)