Day 1: Introduction and Issues in Quantitative Text Analysis

Kenneth Benoit

CEU April 14-21 2011

April 14, 2011

Today's Basic Outline

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

Class schedule: Typical day

9:00–10:30 Lecture

10:45–11:20 Focus on Examples

11:35–12:40 In-class exercises

MOTIVATION

Motivation

- Whom this class is for
- Learning objectives
- Prior knowledge
  - (very) basic quantitative methods
  - familiarity with some sort of quantitative analysis software
  - ability and willingness to try to learn a QTA software package
  - ability to use a text editor

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)

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
- 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
- Computer-assisted text analysis is not exclusively quantitative, but aids greatly even in conversion of qualitative text analysis into quantitative summaries — and typically CTA means QTA

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 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.

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)

LOGISTICS

Detailed Class Schedule

	Date	Topic(s)	Details
Thu	14 April	Introduction and Issues in text analysis	Course goals; logistics; software overview; Conceptual foundations; content analysis; objectives; examples.
Mon	18 April	Descriptive inference in text	Co-occurence, concordances, keywords in context; complexity and readability measures; also issues concerning sampling, validity, reliability, agreement in text analysis.
Mon	18 April	Classical quantitative content analysis	Manual unitization and coding approaches, including the CMP, Policy Agendas Project, and self-constructed themes. Software will use MaxQDA.
Tues	19 April	Automated dictionary-based approaches	Dictionary construction, and methods for automatically indexing texts for compiling scales of substantive quantities of interest. Also covers a variety of statistical issues surrounding text types, tokens, and equivalencies, including stemming, lemmatization, and trimming of texts based on word frequencies and *tf-idf*.
Wed	20 April	Words as Data approaches	Automatic "word indexing" and scoring using "Wordscores"; scaling models using parametric (Poisson) and non-parametric (correspondence analysis) methods.
Thur	21 April	Document Scaling	Purely statistical text analysis to recover political ideal points represented in text; some classification methods if we have time; and a brief course review.

Software requirements for this course

- A text editor you know and love
  - Nothing beats Emacs
  - 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
- Any prior use of a computerized content analysis tool is helpful (but not essential)
- Some of the software is home-grown
- Our exercises using software will be gentle and "assisted"

Who I am

- Ken Benoit, London School of Economics
  kbenoit@tlse.ac.uk
- Head of Methodology Institute
- http://www.kenbenoit.net/ceu2011cta
- *Introductions . . .*

Course resources

- Syllabus: describes class, lists readings, links to reading, and links to exercises and datasets
- Web page on `http://www.kenbenoit.net/ceu2011cta`
  - 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
  - Neuendorf
  - Other texts, esp. articles, are linked to the course handout and can be downloaded online

Prior probabilities and updating

A test is devised to automatically flag racist news stories.

- 1% of news stories in general have racist messages
- 80% of racist news stories will be flagged by the test
- 10% of non-racist stories will also be flagged

We run the test on a new news story, and it is *flagged as* racist.

Question: What is probability that the story is *actually* racist?

Any guesses?

Prior probabilities and updating

- What about without the test?
  - Imagine we run 1,000 news stories through the test
  - We expect that 10 will be racist
- With the test, we expect:
  - Of the 10 found to be racist, 8 should be flagged as racist
  - Of the 990 non-racist stories, 99 will be wrongly flagged as racist
  - That's a total of 107 stories flagged as racist
- So: the updated probability of a story being racist, conditional on being flagged as racist, is $\frac{8}{107} = 0.075$
- The *prior* probability of 0.01 is updated to only 0.075 by the positive test result
- This is an example of Bayes' Rule,

$$\Pr(R = 1|T = 1) = \frac{\Pr(T{=}1|R{=}1)\Pr(R{=}1)}{\Pr(T{=}1)}$$

EXAMPLES

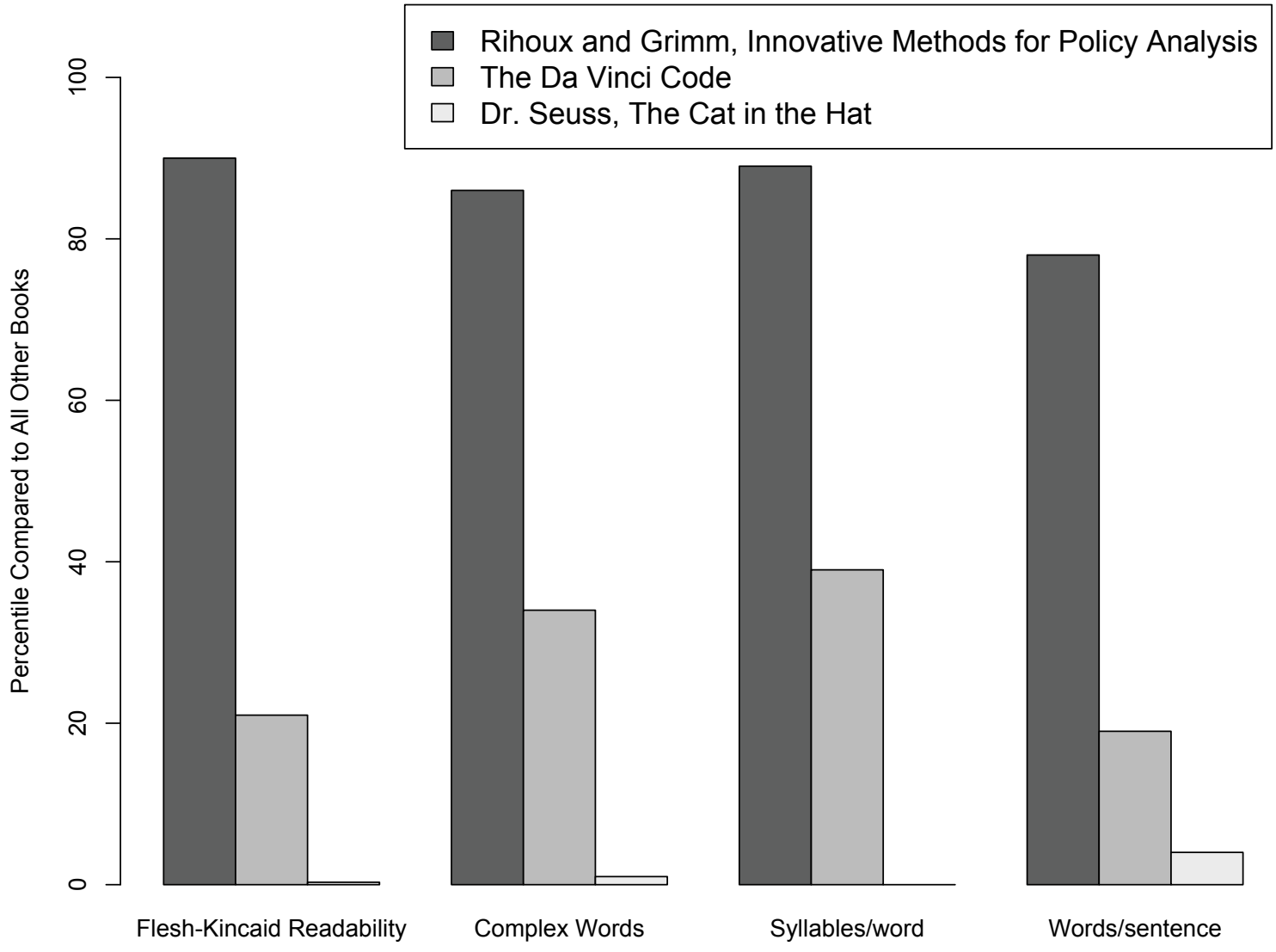
You have already done QTA!

- Unless you are one of five people on the planet who has never used an Internet search engine...
- `Amazon.com` also 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

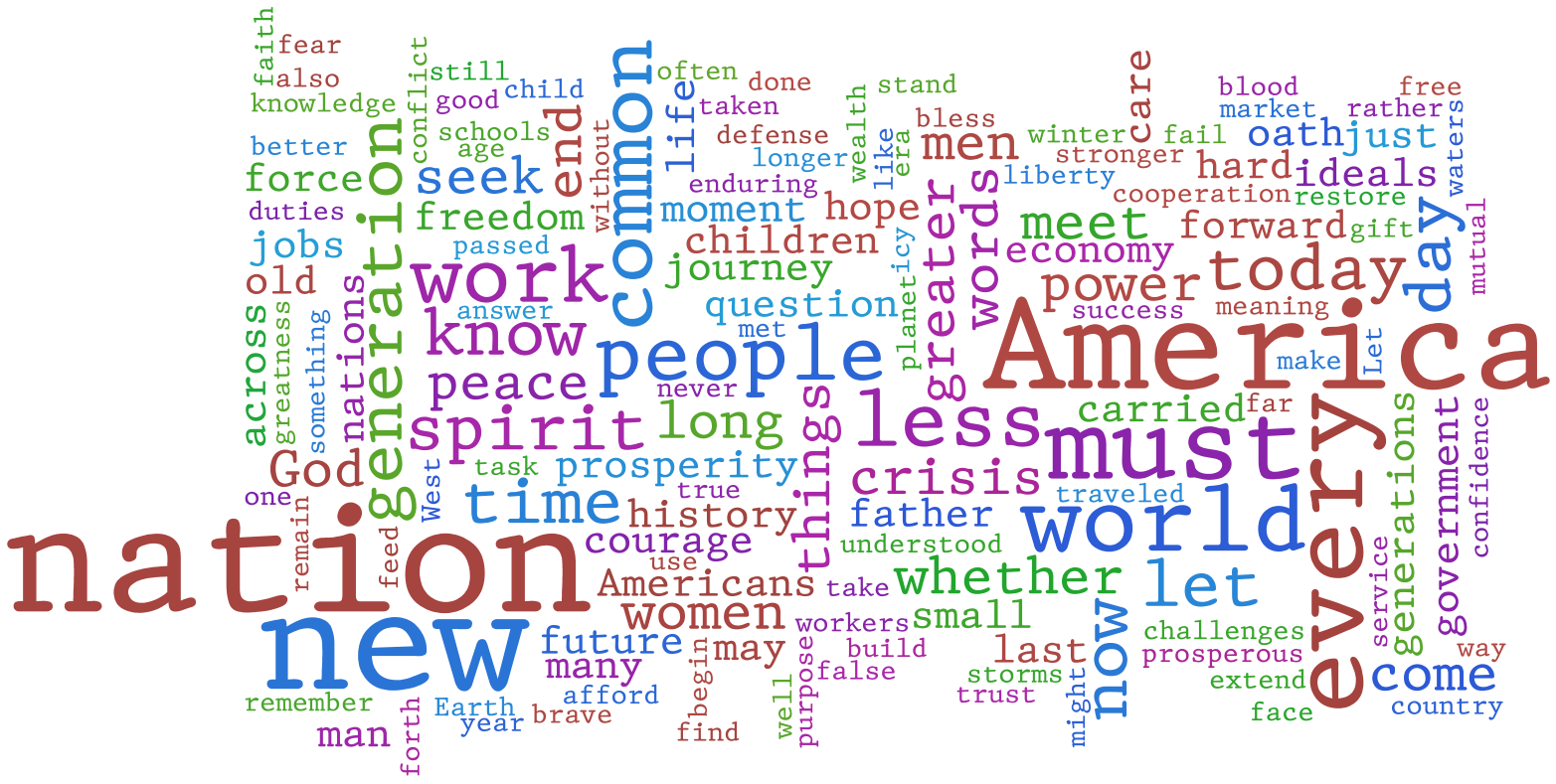
Bush's second inaugural address:

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

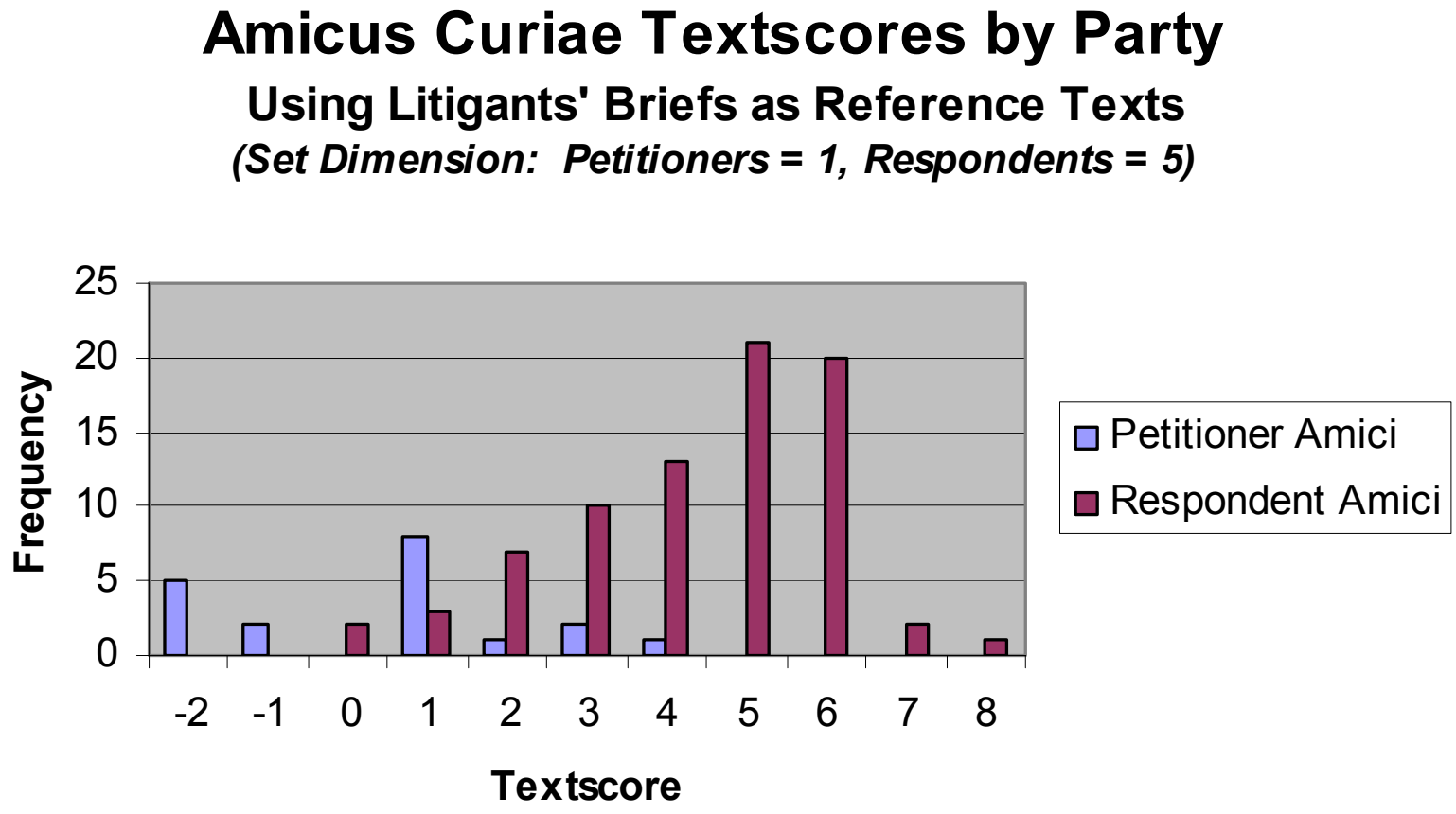
Obama's inaugural address:

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

Obama's Inaugural Speech in Wordle

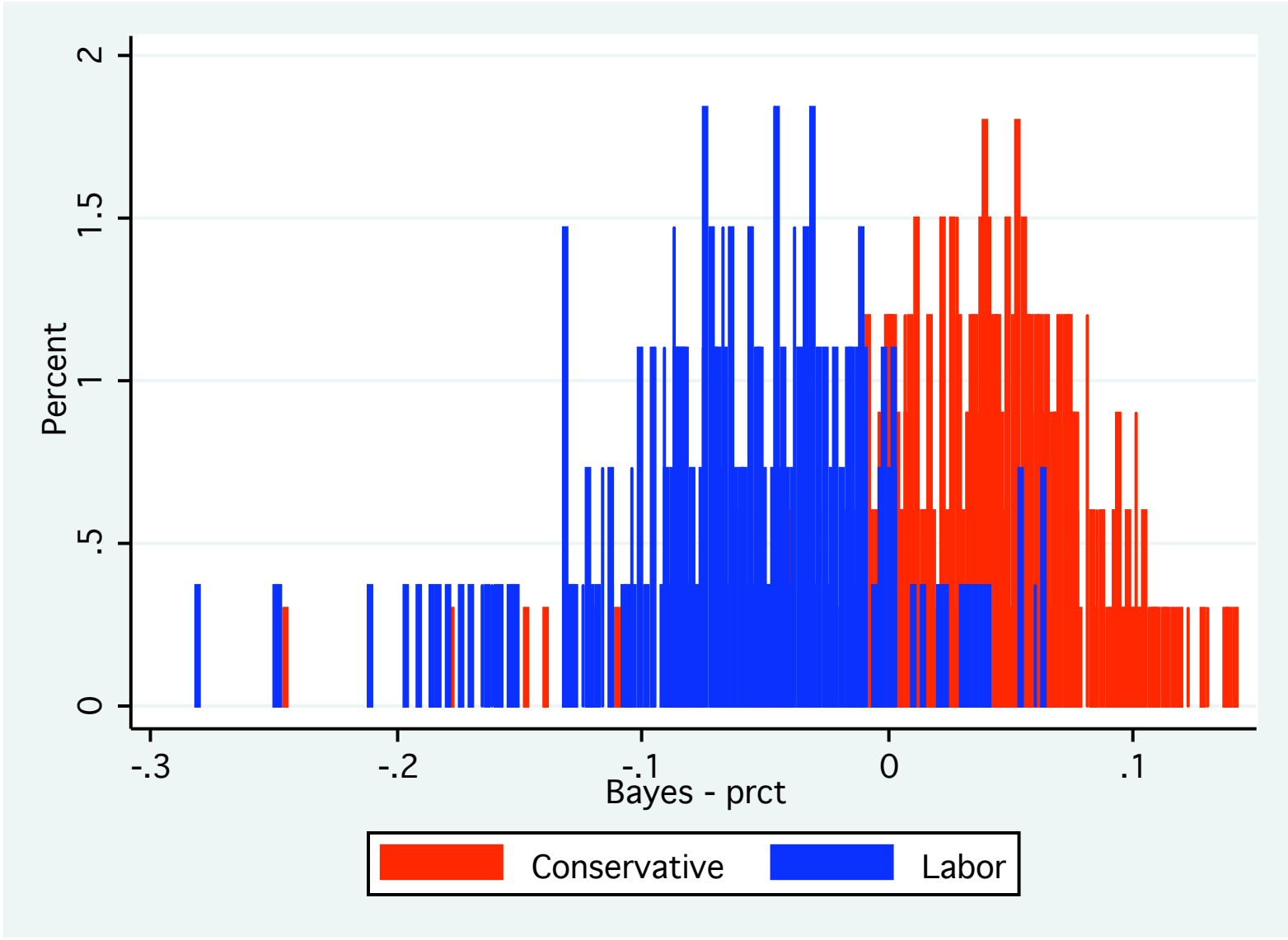


Legal document scaling: "Wordscores"



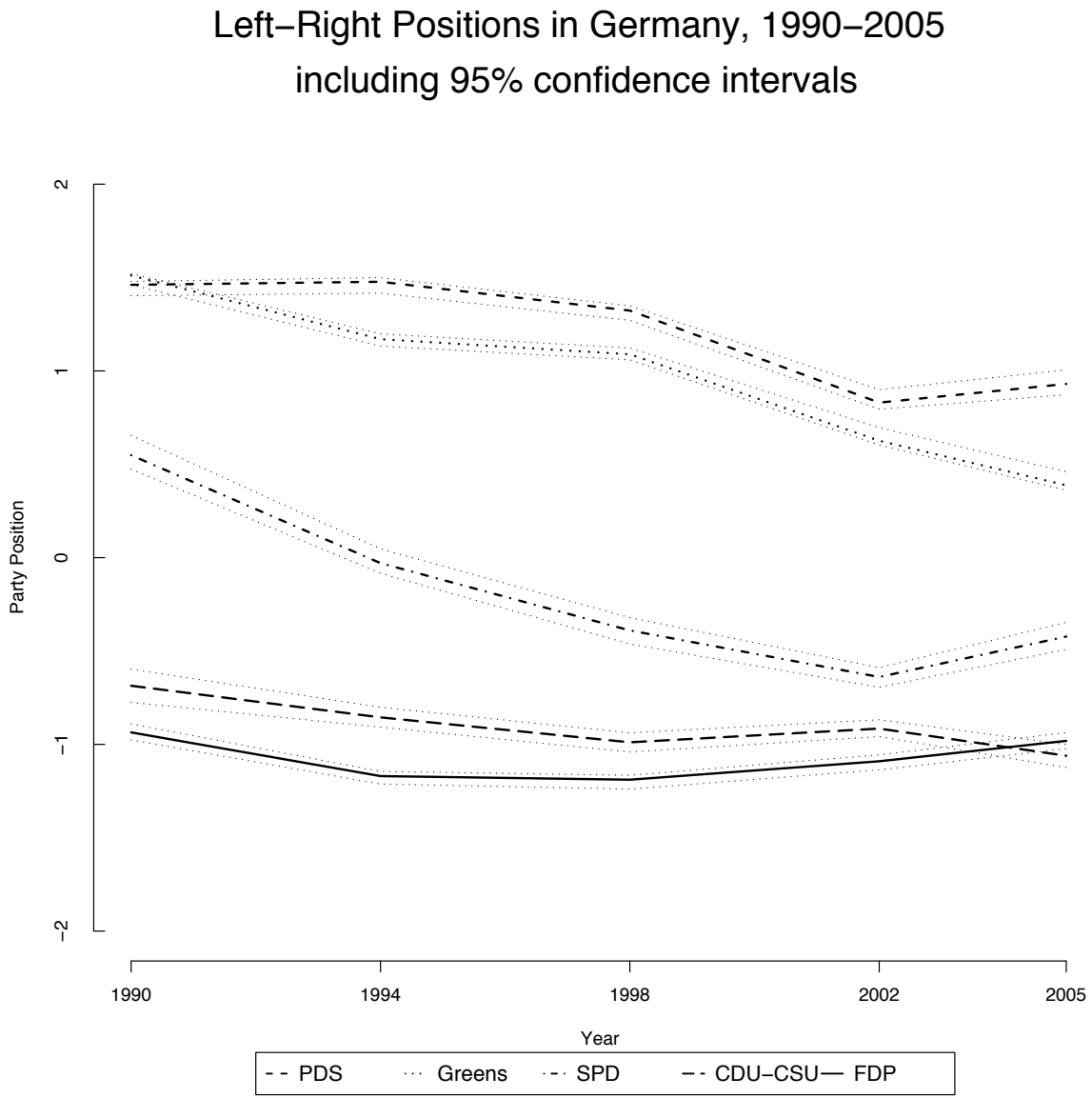
(from *JELS*, Evans et. al. 2008)

Legislative speeches: "Naive Bayes" classifier



(from work in progress by Nicolas Beauchamp)

Party Manifestos: Poisson scaling



(from Slapin and Proksch, forthcoming *AJPS* 2008)

Party Manifestos: More scaling with Wordscores

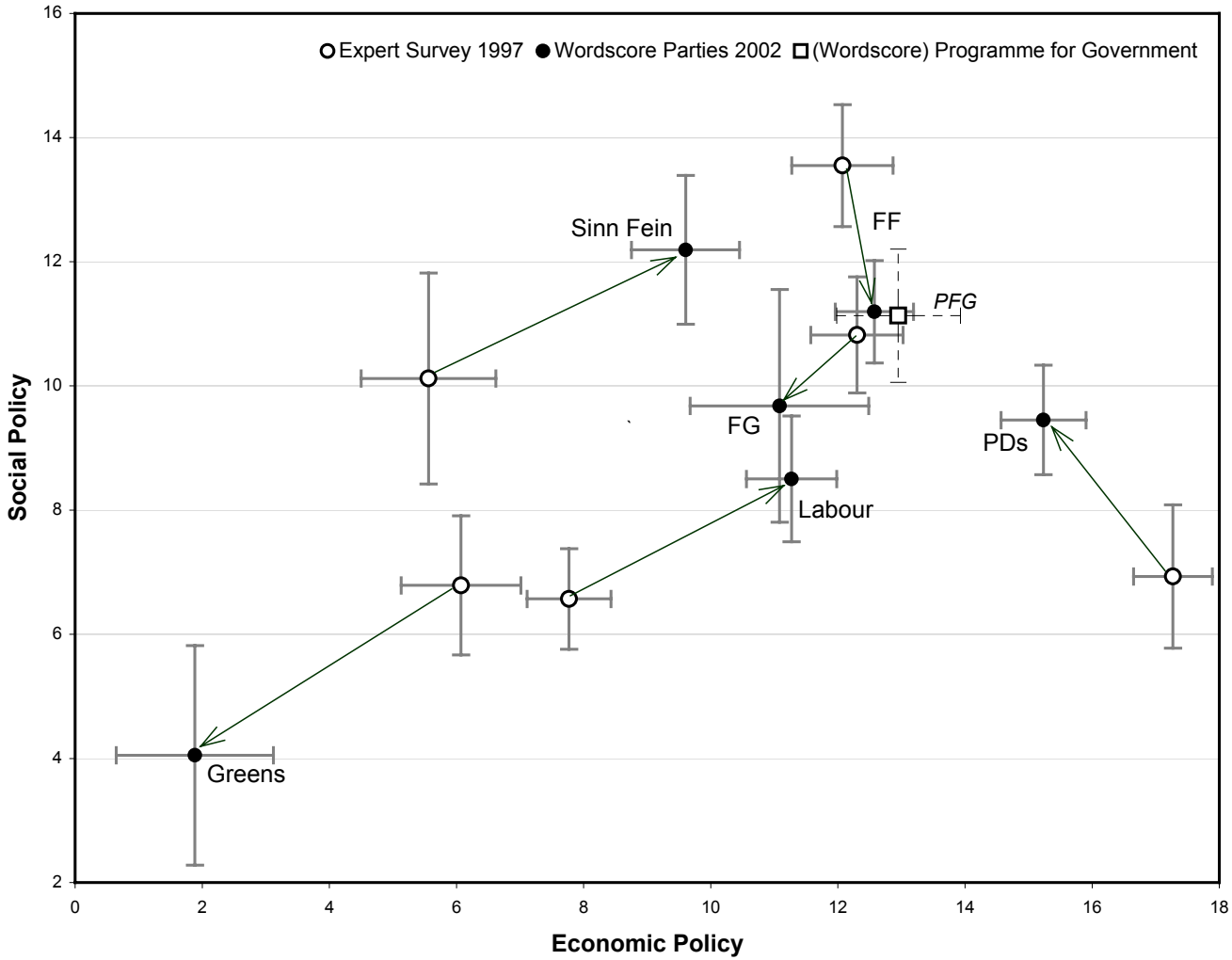
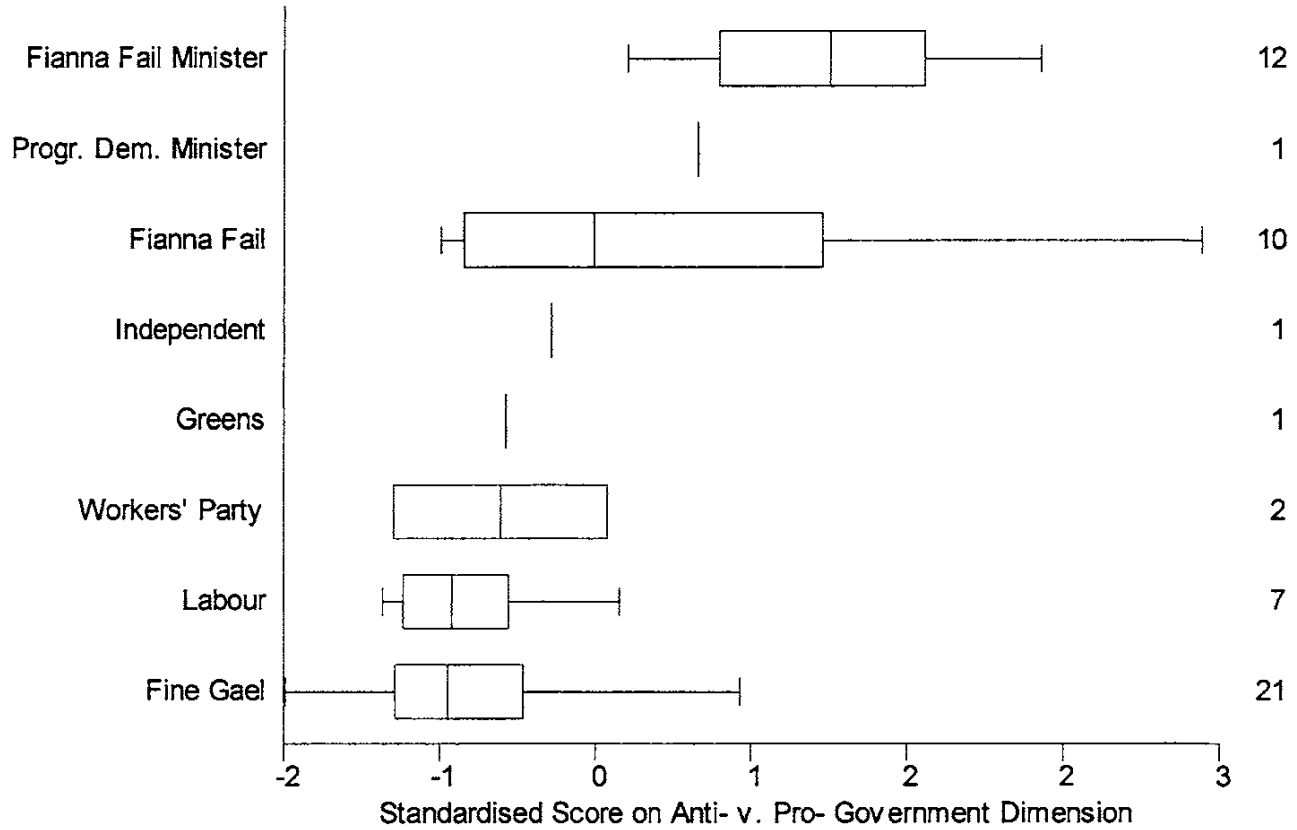


Figure 1. Movement from 1997 Positions on Economic and Social Policy, based on Wordscores Estimates. Bars indicate two standard errors on each scale.

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

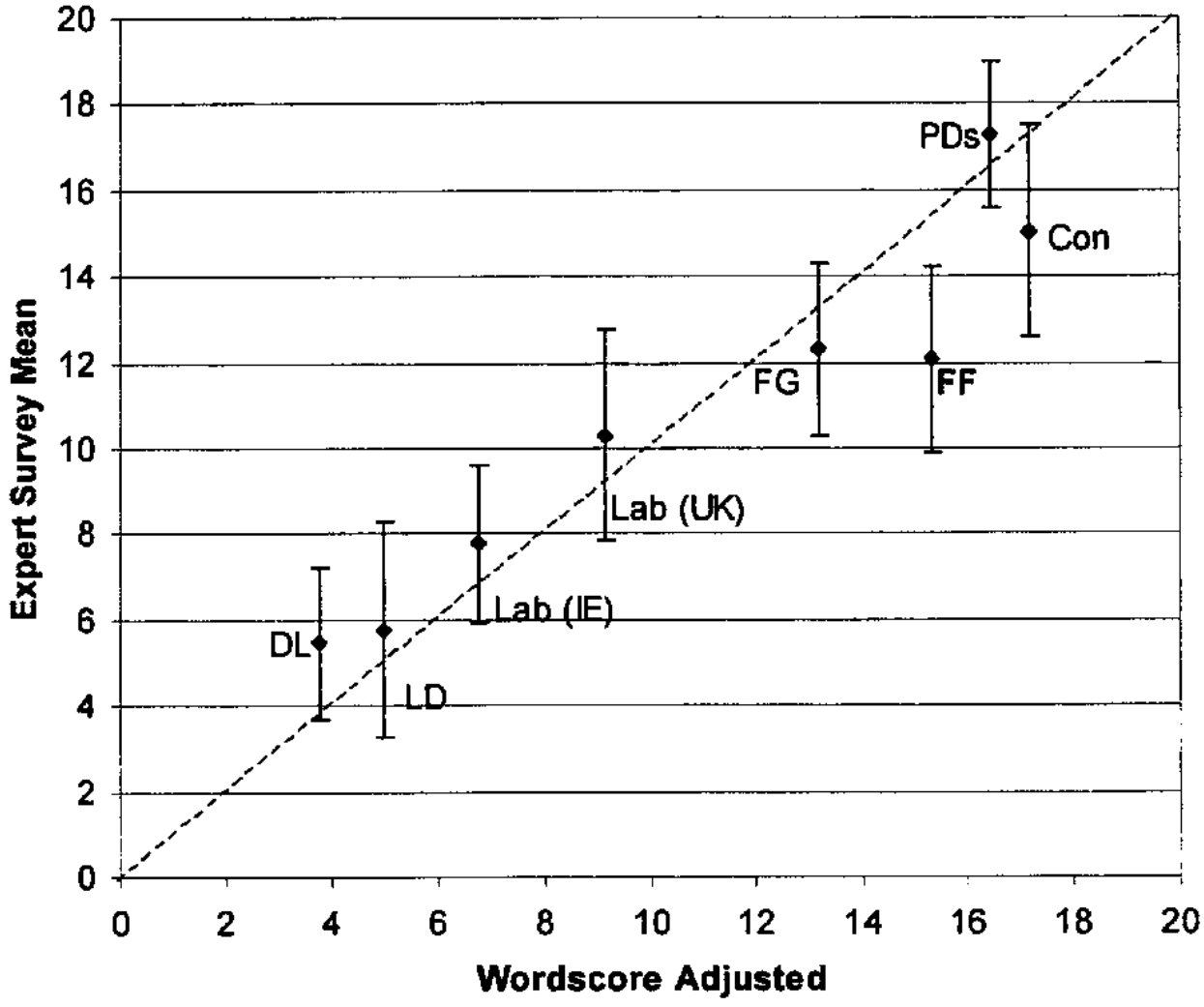
No confidence debate speeches (Wordscores)

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



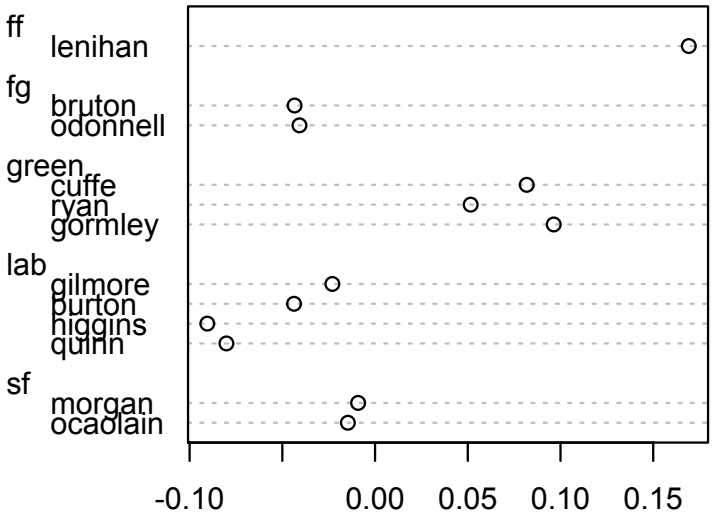
(from Benoit and Laver, *Irish Political Studies* 2002)

Text scaling versus human experts

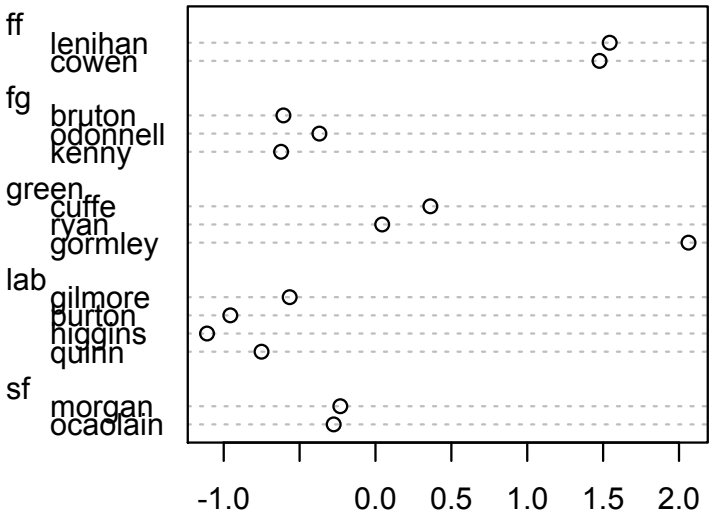


(from Laver, Benoit and Garry, *APSR* 2003)

NAMA and budget debates



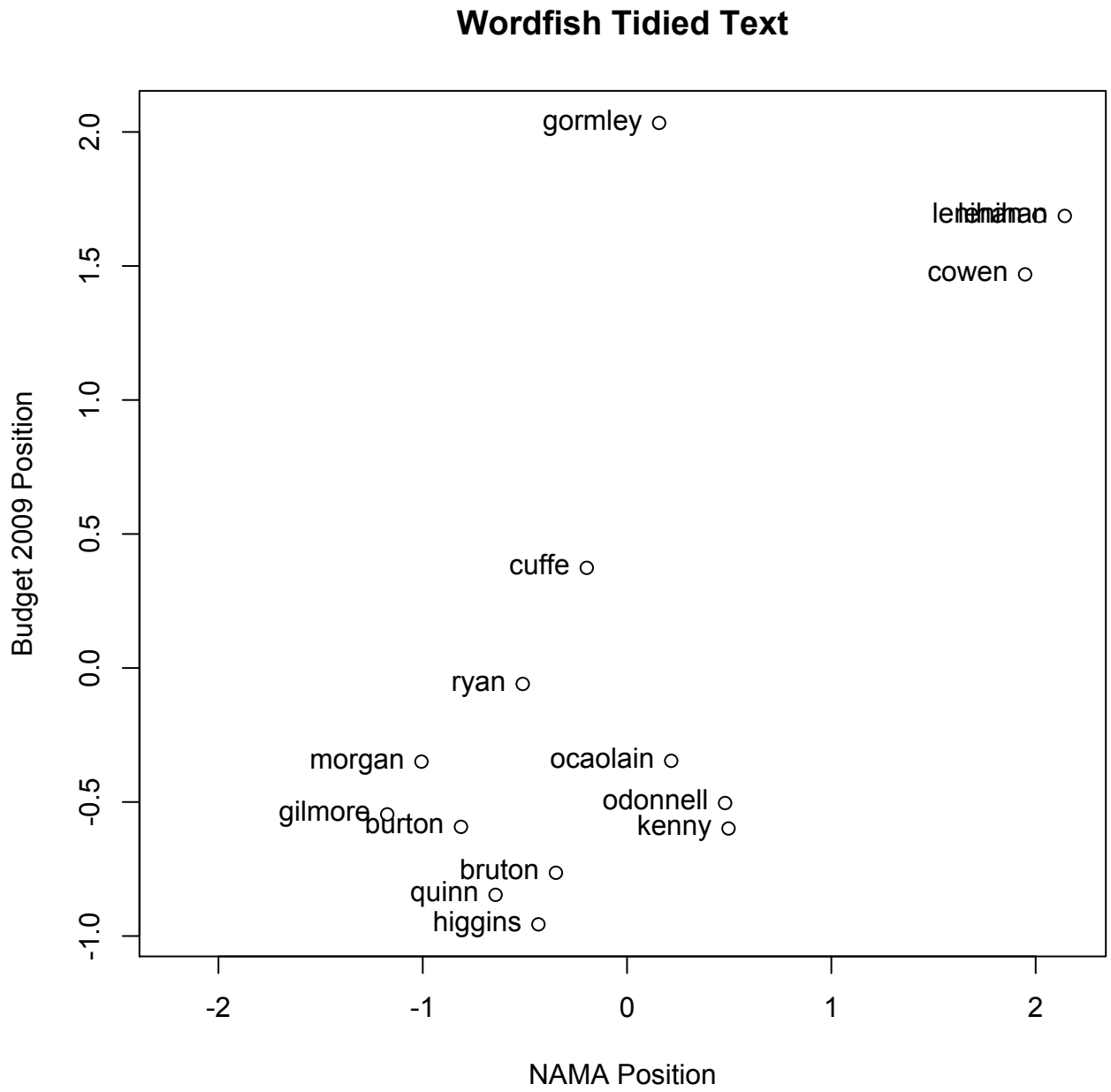
Wordscores LBG Position on Budget 2009



Normalized CA Position on Budget 2009

(from Lowe and Benoit Midwest 2010)

NAMA and budget debates 2



Published examples on reading list

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