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)
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 incentivise. It has the words made because had into get some through next where many irish

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 13 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
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 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 document-feature matrix.
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.
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 incentivise. It has the

<table>
<thead>
<tr>
<th>words</th>
<th>docs</th>
<th>made</th>
<th>because</th>
<th>had</th>
<th>into</th>
<th>get some</th>
<th>through</th>
<th>next</th>
<th>where</th>
<th>many</th>
<th>irish</th>
</tr>
</thead>
<tbody>
<tr>
<td>t06_kenny_fg</td>
<td>12</td>
<td>11</td>
<td>5</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>t05_cowen_ff</td>
<td>9</td>
<td>4</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>14</td>
<td>13</td>
<td>4</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>t14_ocaolain_sf</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>t01_lenihan_ff</td>
<td>12</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>11</td>
<td>9</td>
<td>16</td>
<td>14</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>t11_gormley_green</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>t04_morgan_sf</td>
<td>11</td>
<td>8</td>
<td>7</td>
<td>15</td>
<td>8</td>
<td>19</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>t12_ryan_green</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>t10_quinn_lab</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>t07_odonnell_fg</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>t09_higgins_lab</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>t03_burton_lab</td>
<td>4</td>
<td>8</td>
<td>12</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>8</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>t13_cuffe_green</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>16</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>t08_gilmore_lab</td>
<td>4</td>
<td>8</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>t02_bruton_fg</td>
<td>1</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>6</td>
<td>16</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>
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 documentary 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, especially 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18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 incentivise. It has the

Descriptive statistics on words
Scaling documents
Extraction of topics
Classifying documents
Sentiment analysis
Vocabulary analysis
LOGISTICS
<table>
<thead>
<tr>
<th>Day</th>
<th>Date</th>
<th>Topic(s)</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon</td>
<td>21 July</td>
<td>Quantitative text analysis overview and fundamentals</td>
<td>Course goals; logistics; software overview; conceptual foundations; quantitative text analysis as a field; objectives; examples.</td>
</tr>
<tr>
<td>Tue</td>
<td>22 July</td>
<td>Working with texts, defining documents and features, weighting</td>
<td>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.</td>
</tr>
<tr>
<td>Wed</td>
<td>23 July</td>
<td>Descriptive statistical methods for textual analysis</td>
<td>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.</td>
</tr>
<tr>
<td>Thu</td>
<td>24 July</td>
<td>Quantitative methods for comparing texts</td>
<td>Quantitative methods for comparing texts, such as keyword identification, dissimilarity measures, association models, vector space models.</td>
</tr>
<tr>
<td>Fri</td>
<td>25 July</td>
<td>Automated dictionary methods</td>
<td>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.</td>
</tr>
<tr>
<td>Day</td>
<td>Date</td>
<td>Topic(s)</td>
<td>Details</td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>---------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Mon</td>
<td>21 July</td>
<td>Quantitative text analysis</td>
<td>Overview and fundamentals: Course goals; logistics; software overview; conceptual foundations; quantitative text analysis as a field; objectives; examples.</td>
</tr>
<tr>
<td>Tue</td>
<td>22 July</td>
<td>Working with texts, defining</td>
<td>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.</td>
</tr>
<tr>
<td>Wed</td>
<td>23 July</td>
<td>Descriptive statistical methods</td>
<td>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.</td>
</tr>
<tr>
<td>Thu</td>
<td>24 July</td>
<td>Quantitative methods for comparing texts</td>
<td>Quantitative methods for comparing texts, such as keyword identification, dissimilarity measures, association models, vector space models.</td>
</tr>
<tr>
<td>Fri</td>
<td>25 July</td>
<td>Automated dictionary methods</td>
<td>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.</td>
</tr>
<tr>
<td>Mon</td>
<td>28 July</td>
<td>Document classifiers</td>
<td>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.</td>
</tr>
<tr>
<td>Tue</td>
<td>29 July</td>
<td>Unsupervised models for scaling</td>
<td>The “Wordscores” approach to scaling latent traits using a Naïve Bayes foundation; Correspondence Analysis applied to texts.</td>
</tr>
<tr>
<td>Wed</td>
<td>30 July</td>
<td>Supervised models for scaling</td>
<td>Poisson scaling models (aka “wordfish”) of latent word and document traits, and their applications.</td>
</tr>
<tr>
<td>Thurs</td>
<td>31 August</td>
<td>Clustering methods and topic</td>
<td>An introduction to hierarchical clustering for textual data, including parametric topic models such as Latent Dirichlet Allocation (LDA).</td>
</tr>
<tr>
<td>Fri</td>
<td>1 August</td>
<td>Mining Social Media: An</td>
<td>Methods for extracting text and meta-data from Twitter feeds and applying sentiment analysis to these feeds.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>application to textual analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>of Twitter data.</td>
<td></td>
</tr>
</tbody>
</table>
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
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 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

- 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](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*:

<table>
<thead>
<tr>
<th>Readability (learn more)</th>
<th>Compared with other books</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fog Index: 8.8</td>
<td>20% are easier ▼</td>
</tr>
<tr>
<td>Flesch Index: 65.2</td>
<td>25% are easier ▼</td>
</tr>
<tr>
<td>Flesch-Kincaid Index: 6.9</td>
<td>21% are easier ▼</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Complexity (learn more)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex Words: 11%</td>
<td>34% have fewer ▼</td>
</tr>
<tr>
<td>Syllables per Word: 1.5</td>
<td>39% have fewer ▼</td>
</tr>
<tr>
<td>Words per Sentence: 11.0</td>
<td>19% have fewer ▼</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Characters: 823,633</td>
<td>85% have fewer ▼</td>
</tr>
<tr>
<td>Words: 138,843</td>
<td>88% have fewer ▼</td>
</tr>
<tr>
<td>Sentences: 12,647</td>
<td>94% have fewer ▼</td>
</tr>
</tbody>
</table>
Comparing Texts on the Basis of Quantitative Information

- Flesh-Kincaid Readability
- Complex Words
- Syllables/word
- Words/sentence

Percentile Compared to All Other Books

Rihoux and Grimm, Innovative Methods for Policy Analysis
The Da Vinci Code
Dr. Seuss, The Cat in the Hat
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
tyrranny 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”

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

Log wordscores mean for document

Posterior $P(\text{class} = \text{Petitioner}|\text{document})$

Predicted Petitioner

Predicted Respondent

-0.3 -0.2 -0.1 0.0 0.1 0.2

0.0 0.2 0.4 0.6 0.8 1.0

Log wordscores mean for document

Posterior $P(\text{class} = \text{Petitioner}|\text{document})$

Predicted Petitioner

Predicted Respondent
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)
Fianna Fail ministers were overwhelmingly the most progovernment speakers in the debate, with Fianna Fail TDs (members of parliament) on average less pro-government in their speeches. At the other end of the scale, Labour, Fine Gael, and Workers’ Party TDs were the most systematically antigovernment in their speeches, closely followed by the sole Green TD. Not only does the word scoring plausibly locate the party groupings, but also it yields interesting information about individual legislators, whose scores may be compared to those of the various groupings. The position of government minister and PD leader Des O’Malley, for instance (the sole PD minister in Table 7), was less staunchly progovernment than that of his typical Fianna Fail ministerial colleagues. This may be evidence of the impending rift in the coalition, since in 1991 the PDs were shortly to leave the coalition with Fianna Fail.

We already noted that the word scoring of relatively short speeches may generate estimates of a higher uncertainty than those for relatively longer party manifestos. This is because our approach treats words as data and reflects the greater uncertainty that arises from having fewer data. In the point estimates of the 55 individual speeches we coded as virgin texts (not shown), greater uncertainty about the scoring of a virgin text was directly represented by its associated standard error. For the raw scores (with a minimum of $-0.41$ and a maximum of $-0.25$), the standard errors of the estimates derived from speeches ranged from $0.020$, for the shortest speech of 625 words, to $0.006$, for the longest speech of 6,396 words, delivered by the Labour Party leader Dick Spring. These errors are indeed larger than those arising in our manifesto analyses. However, substantively interesting distinctions between speakers are nonetheless possible on the basis of the resulting confidence intervals. Considering policy differences within Fine Gael, for example, the raw estimates (and 95% confidence intervals) of the positions of former FG Taoiseach Garrett FitzGerald were $-0.283$ ($-0.294$, $-0.272$), while those of future party leader Enda Kenny were $-0.344$ ($-0.361$, $-0.327$). This allows us to conclude with some confidence that Kenny was setting out a more robustly antigovernment position in the debate than party colleague Fitzgerald. Thus even when speeches are short, our method can detect strong variations in underlying positions and permit discrimination between texts, allowing us to infer how much of the difference between two estimates is due to chance and how much to underlying patterns in the data.

Overall we consider the use of word scoring beyond the analysis of party manifestos to be a considerable success, reproducing party positions in a no-confidence debate using no more than the relative word frequencies in speeches. This also demonstrates three important features of the word scoring technique. First, in a context where independent estimates of reference scores are not available, assuming reference text positions using substantive local knowledge may yield promising and sensible results. Second, we demonstrate that our method quickly and effortlessly handles a large number of texts that would have presented a daunting task using traditional methods. Third, we see that the method works even when texts are relatively short and provides estimates of the increased uncertainty arising from having less data.
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

Average position among cabinet ministers, government backbenchers, and opposition members, 1983–2013

(from Herzog and Benoit EPSA 2013)
Published examples on reading list

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