Day 1: Introduction and Issues in Quantitative Text Analysis

Kenneth Benoit

Essex Summer School 2012

July 9, 2012
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
  - (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 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.

- 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)
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)
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
## Detailed Class Schedule

<table>
<thead>
<tr>
<th>Day</th>
<th>Date</th>
<th>Topic(s)</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon</td>
<td>9 July</td>
<td>Introduction and Issues in text analysis</td>
<td>Course goals; logistics; software overview; conceptual foundations; content analysis; objectives; examples.</td>
</tr>
<tr>
<td>Tue</td>
<td>10 July</td>
<td>Textual Data, Sampling, and Working with a Text Corpus</td>
<td>Where to obtain textual data; formatting and working with text files; indexing and meta-data; sampling concerns with textual data.</td>
</tr>
<tr>
<td>Wed</td>
<td>11 July</td>
<td>Descriptive inference from text</td>
<td>Methods of summarizing texts and features of texts in order to characterize their properties. It covers many basic quantitative textual measures.</td>
</tr>
<tr>
<td>Thu</td>
<td>12 July</td>
<td>Research Design issues in textual studies</td>
<td>Reliability and validity and their role in designing and evaluating content-analysis based research; measures of reliability.</td>
</tr>
<tr>
<td>Fri</td>
<td>13 July</td>
<td>Thematic analysis, key words in context</td>
<td>Computer-assisted methods for developing themes from texts, examining key words in context, applying codes to texts.</td>
</tr>
<tr>
<td>Mon</td>
<td>16 July</td>
<td>Classical quantitative content analysis</td>
<td>Manual unitization and coding approaches, including the CMP, Policy Agendas Project, and self-constructed themes. The exercise will consist of an on-line quantitative coding experiment.</td>
</tr>
<tr>
<td>Tue</td>
<td>17 July</td>
<td>Automated dictionary-based approaches</td>
<td>Dictionary construction, and methods for automatically indexing texts for compiling scales of substantive quantities of interest.</td>
</tr>
<tr>
<td>Thurs</td>
<td>19 July</td>
<td>Classifiers: Introduction to the Naive Bayes Classifier</td>
<td>Extends “wordscores” into classification and scaling.</td>
</tr>
<tr>
<td>Fri</td>
<td>20 July</td>
<td>Document Scaling: Parametric models</td>
<td>Continues text scaling using completely automated methods based on parametric (Poisson) scaling.</td>
</tr>
</tbody>
</table>
Software requirements for this course

- A text editor you know and love
  - Nothing beats Emacs
  - 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/essex2012cta](http://www.kenbenoit.net/essex2012cta)
- 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
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*:

<table>
<thead>
<tr>
<th><strong>Readability</strong> (learn more)</th>
<th><strong>Compared with other books</strong></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><strong>Complexity</strong> (learn more)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex Words: 11%</td>
</tr>
<tr>
<td>Syllables per Word: 1.5</td>
</tr>
<tr>
<td>Words per Sentence: 11.0</td>
</tr>
</tbody>
</table>

**Number of**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Characters: 823,633</td>
<td>85% have fewer ▼</td>
<td>15% have more</td>
</tr>
<tr>
<td>Words: 138,843</td>
<td>88% have fewer ▼</td>
<td>12% have more</td>
</tr>
<tr>
<td>Sentences: 12,647</td>
<td>94% have fewer ▼</td>
<td>6% have more</td>
</tr>
</tbody>
</table>
Comparing Texts on the Basis of Quantitative Information

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

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

Percentile Compared to All Other Books

- 0
- 20
- 40
- 60
- 80
- 100
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”

Amicus Curiae Textscores by Party
Using Litigants' Briefs as Reference Texts
(Set Dimension: Petitioners = 1, Respondents = 5)

(from JELS, Evans et. al. 2008)
Legislative speeches: “Naive Bayes” classifier

(from work in progress by Nicolas Beauchamp)
Party Manifestos: Poisson scaling

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

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

Fianna Fail Minister
12

Progr. Dem. Minister
1

Fianna Fail
10

Independent
1

Greens
1

Workers’ Party
2

Labour
7

Fine Gael
21

(Standardised Score on Anti- v. Pro- Government Dimension)

(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

Note: The diagonal dashed line shows the axis of perfect agreement. Vertical bars represent one standard deviation of the expert scores (Ireland, N = 30; UK, N = 117).

Contextual judgments made by experts about the "real" position of Fianna Fáil, rather than of error in the computer analysis of the actual text of the party manifesto. Put in a slightly different way, the technique we propose performed, in just about every case, equivalently to a typical expert—which we take to be a clear confirmation of the external validity of our technique's ability to extract meaningful estimates of policy positions from political texts.

Coding non-English-language texts

Thus far we have been coding English-language texts, but since our approach is language-blind it should work equally well in other languages. We now apply it to German-language texts, analyzing these using no knowledge of German. Our research design is essentially similar to that we used for Britain and Ireland. As reference texts for Germany in the 1990s, we take the 1990 manifestos of four German political parties— the Greens, Social Democratic Party (SPD), Christian Democrats (CDU), and Free Democrats (FDP). Our estimates of the a priori positions of these texts on economic and social policy dimensions derive from an expert survey conducted in 1989 by Laver and Hunt (1992). Having calculated German word scores for both economic and social policy dimensions in precisely the same way as before, we move on to analyze six virgin texts. These are the manifestos of the same four parties in 1994, as well as manifestos for the former Communists (PDS) in both 1990 and 1994. Since no expert survey scores were collected for the PDS in 1990, or for any German party in 1994, we are forced to rely in our evaluation upon the face validity of our estimated policy positions for the virgin texts. However, the corpus of virgin texts presents us with an interesting and taxing new challenge. This is to locate the PDS on both economic and social policy dimensions, even though no PDS reference text was used to calculate the German word scores. We are thus using German word scores, calculated using no knowledge of German, to locate the policy positions of the PDS, using no information whatsoever about the PDS other than the words in its manifestos, which we did not and indeed could not read ourselves. The top panel in Table 6 summarizes the results of our analysis.

The first row in Table 6 reports our rescaled computer estimates of the economic policy positions of the six virgin texts. The main substantive pattern for the economic policy dimension is a drift of all established parties to the right, with a sharp rightward shift by the SDP. Though this party remains between the position of the Greens and that of the CDU, it has moved to a position significantly closer to the CDU. The face validity of this seems very plausible. Our estimated economic policy positions of the 1990 and 1994 PDS manifestos locate these firmly on the left of the manifestos of the other four parties, which has excellent face validity. The rescaled standard errors show that the PDS is indeed significantly to the left of the other parties but that there is no statistically significant difference between the 1990 and the 1994 PDS manifestos. In other words, using only word scores derived from the other four party manifestos in 1990 and no knowledge of German, the manifestos of the former Communists were estimated in both 1990 and 1994 to be on the far left of the German party system. We consider this to be an extraordinarily good result for our technique.

The third row in Table 6 reports our estimates of the social policy positions of the virgin texts. As in the

(from Laver, Benoit and Garry, APSR 2003)
NAMA and budget debates

(From Lowe and Benoit Midwest 2010)
NAMA and budget debates 2

Wordfish Tidied Text

NAMA Position

Budget 2009 Position

bruton
burton
cowen
cuffe
gilmore
gormley
higgins
kenny
lenihan
lenihan
morgan
ocaolain
odonnell
quinn
ryan

(from Lowe and Benoit Midwest 2010)
Published examples on reading list

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