Day 5: Quantitative content analysis

Kenneth Benoit

July 26, 2013

Hand-coding: "Classic" content analysis

- Key feature: use of "human" coders to implement a pre-defined coding scheme, by reading and coding texts

- Human decision-making is the central feature of coding decisions, not a computer or other mechanized tool

- Differs from thematic analysis in that the coding scheme is *fixed*

- Alternative 1: (somewhat more automated) is a dictionary approach

- Alternative 2: (entirely "automated") is inductive scaling of texts from the quantitative matrix

Hand-coding': "Classic" content analysis

- Validity is usually the objective, rather than reliability

- Another motivating factor could be ease of use, or the difficulty of implementing an automated procedure

- May be *computer-assisted*, especially for unitization

- Many common "CATA" or "CACA" tools exist – e.g. QDAMiner

Components of classical content analysis designs

Unitizing
: The systematic distinguishing of segments of text that are of interest to the analysis.

Sampling
: Choice (and justification of the choice) of text units to sample, from population of possible text units.

Coding
: Classifying each coded unit of text from the sample according to the pre-defined category scheme.

Summarizing
: Reducing the coded data to summary quantities of interest.

Inference and reporting
: The final steps wherein the analyzed results are used to generalize about social world, and communicating these results to others.

Coding Text Units

- ▶ The key step in transforming raw texts into representations that can be analyzed

- ▶ Involves reducing and quantifying the data into discrete categories

- ▶ Requires a pre-defined scheme with rules for how these should be applied

- ▶ Question in designing the scheme is to maximize on the precision-accuracy-reliability frontier

- ▶ This can only be done through an iterative process of design, with *human-involved reliability tests at each step*

- ▶ The Big Problem: the dilemma of maintaining backwards-compatibility versus achieving optimal design

Summarizing

- Involves characterizing the coded text units using additional quantification
- Most typical is a frequency distribution
- Scaling is also common: indexes constructed from summary information

Summarizing: Scale Example

- A very simple example comes from the CMP, using PER110 "European Union: Positive Mentions" and PER108 "European Union: Negative Mentions"

- The overall pro- versus anti- EU-ness can be assessed as PER110 - PER108. Theoretical range is $[-100, 100]$.

- A more complicated example is the CMP's famous "rile" index, which adds 13 categories of the "right" and subtracts from this the sum of 13 categories of the "left".

Inference and Reporting

- **Describing** category frequencies involves description of the coded texts as they are

- **Inference** involves generalization of quantities found in the texts to attributes of the authors who produced the texts

- **Sampling** issues are paramount in inference, but not in description

Coding Text Examples: The Comparative Manifestos Project

- 3,000+ party programmes
- 1948-2000
- 650+ parties
- 52 countries
- three books; hundreds of articles use it
- Procedure
    1. Human coders unitize the text into CMP text units are called quasi-sentences
    2. Coders assign each text unit a category
    3. Category percentages of the total text are are used to measure policy

Coding Text Examples: Newspaper content

- Smetko and Valkenburg 2000: Analyzed 2,601 news stories, 1,522 TV news stories, in the period surrounding the Amsterdam meetings of the heads of state in 1997
- Objective was to investigate framing: how the press framed European politics in terms of different modes
- Different frames were:
    - attribution of responsibility
    - human interest
    - conflict
    - morality
    - economics
- Used principal components analysis to link specific content categories to different factors, interepreted as the above five frames

Unitizing Texts

- Briefly read the CMP Coder Instructions in Appendix 2 of Mapping Policy Preferences II (filename `MPP2appendix2.pdf`)
- To unitize the text on the next slide.

Unitize this

We believe that continued double-figure inflation will destroy the basis of the New Zealand economy and cause untold misery. The fight against increases in the cost of living is the most important single issue in economic management.

People without jobs represent waste of productive effort: National supports a policy of full employment and the dignity of labour. We do not accept unemployment as a balancing factor in economic management.

Finally, the National Development Council will be restored and consultation resumed between Government departments, academic specialists and private industry, including farming and organised labour.

A Test: How many of you said seven?

We believe that continued double-figure inflation will destroy the basis of the New Zealand economy and cause untold misery. **/** *The fight against increases in the cost of living is the most important single issue in economic management.* **/** *People without jobs represent waste of productive effort:* **/** *National supports a policy of full employment* **/** *and the dignity of labour.* **/** *We do not accept unemployment as a balancing factor in economic management.* **/** *Finally, the National Development Council will be restored and consultation resumed between Government departments, academic specialists and private industry, including farming and organised labour.*

Unitizing Texts

- What were our experiences unitizing the CMP reliability test document?

- What were your impressions of this unitization scheme?

- What alternatives exist?
    - physical distinctions: time, length, size, volume
    - syntactical distinctions: words, sentences, paragraphs, chapters, articles, etc.
    - categorical distinctions: units defined by membership in a class or category – references to a particular (pre-defined) topic
    - propositional distinctions: constructions from structure of the language, e.g. separating clauses. A version of this forms the basis for the CMP's "quasi-sentence" scheme
    - thematic distinctions

- Some methods exist for *assessing the reliability of unitization* but these are not simple to compute

And now try to *code* it

We believe that continued double-figure inflation will destroy the basis of the New Zealand economy and cause untold misery. **/** *The fight against increases in the cost of living is the most important single issue in economic management.* **/** *People without jobs represent waste of productive effort:* **/** *National supports a policy of full employment* **/** *and the dignity of labour.* **/** *We do not accept unemployment as a balancing factor in economic management.* **/** *Finally, the National Development Council will be restored and consultation resumed between Government departments, academic specialists and private industry, including farming and organised labour.*

And the ("gold standard") answer is:

We believe that continued double-figure inflation will destroy the basis of the New Zealand economy and cause untold misery. ▰ The fight against increases in the cost of living is the most important single issue in economic management. ▰

414
414

People without jobs represent waste of productive effort: ▰ National supports a policy of full employment ▰ and the dignity of labour. ▰ We do not accept unemployment as a balancing factor in economic management. ▰

410
408
701
701

Finally, the National Development Council will be restored and consultation resumed between Government departments, academic specialists and private industry, including farming and organised labour. ▰

405

414 "Economic Orthodoxy: Positive"
410 "Productivity: Positive"
408 "Economic Goals"
701 "Labour Groups: Positive"
405 "Corporatism: Positive"

Unitization empirical results from CMP tests


Master N is dashed red line

SD = 19
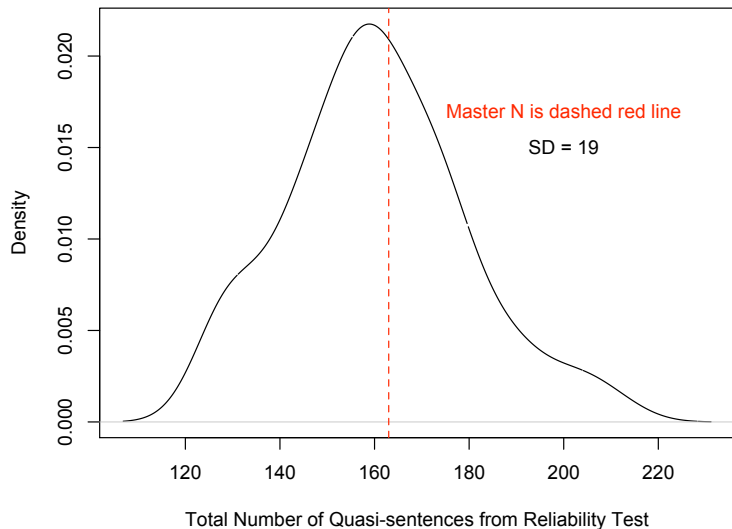
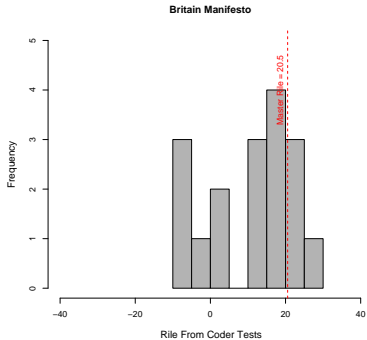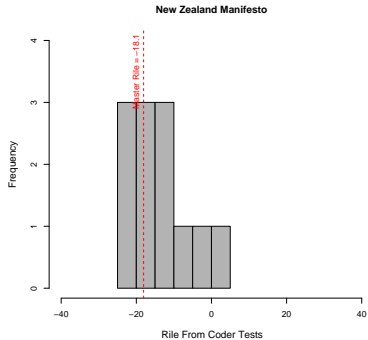# Empirical results from Mikhaylov and Benoit 2010

Caveats before I show you some compromising pictures:

- ▶ We are not out to smear mud on the CMP! We actually like and respect the CMP and believe in the usefulness of their objective.

- ▶ *At the same time*, no research project should be immune from improvement

- ▶ There are weaknesses in the data and these are worth knowing

- ▶ The structure of the tests: Ask trained coders used by the CMP to code CMP manifestos to complete a recoding test online, for a test that was used as an example in the CMP coding instructions. Text was pre-unitized.
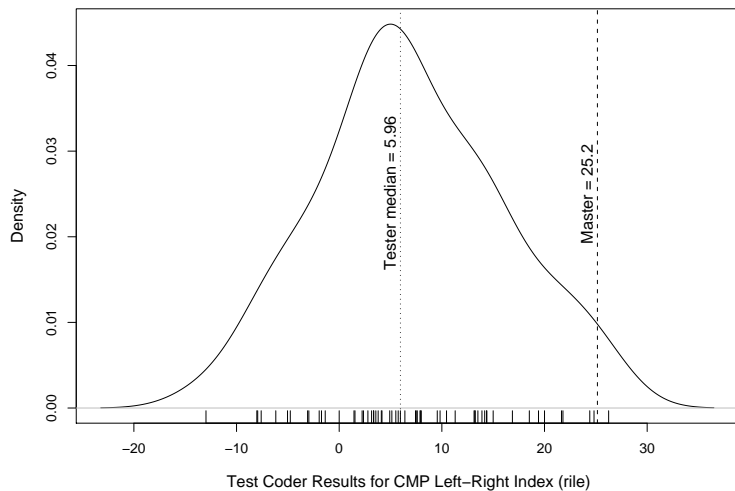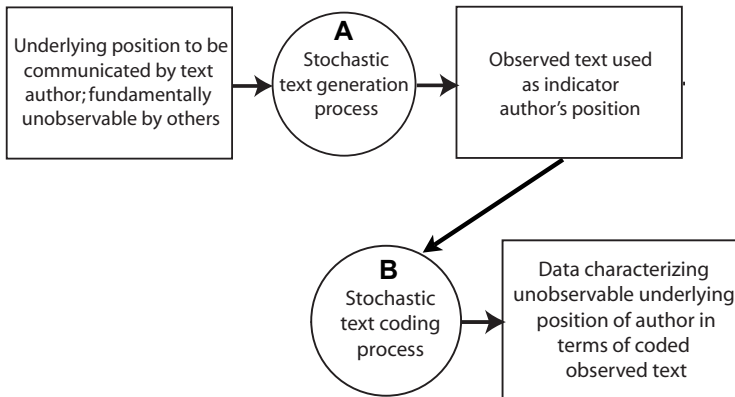
Empirical results from CMP reliability tests
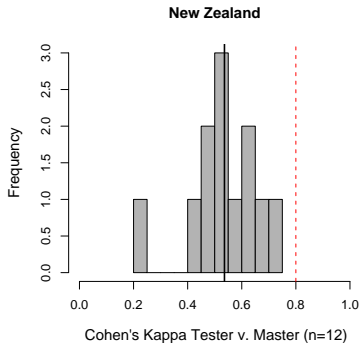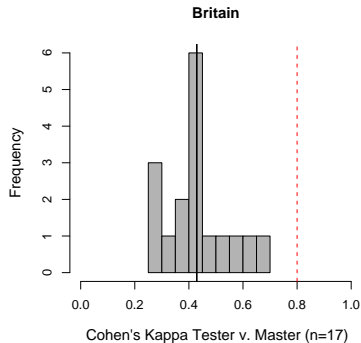
Empirical results from CMP reliability tests

Empirical results from CMP reliability tests

The Big Picture

Underlying position to be communicated by text author; fundamentally unobservable by others

A Stochastic text generation process

Observed text used as indicator author's position

B Stochastic text coding process

Data characterizing unobservable underlying position of author in terms of coded observed text

Scaling Issues

- ▶ Scaling becomes a major issue when we wish to construct quantities of interest from quantitative content analyses
- ▶ Simple example: Proportion of content of a given type (e.g. anti-Lisbon treaty)
- ▶ Complex example: Left-right policy positions (e.g. CMP "Rile")
- ▶ Are the metrics "natural"?
- ▶ Does the output metric resemble the input metric (if any)?
- ▶ What properties should the scale have, such as boundaries, type of increase, etc?
- ▶ How can uncertainty be characterized for the given scale?

Logit scale for left-right

- The Comparative Manifesto Project scales policy positions as absolute porportional difference, measured by proportion of "Right" mentions less proportion of "Left" mentions: $\frac{(R-L)}{N}$

- Problems:
    - Addition of irrelevant content shifts the scale toward zero
    - Assumes the additional mentions increase emphasis in a linear scale

- The alternative is to scale $\frac{(R-L)}{(R+L)}$ (Kim and Fording 2002; Laver and Garry 2000), but this too has problems:
    - Still linear shift in position for increase in repetition
    - Quickly maxes out at the extremes

- Lowe, Benoit, Mikhaylov and Laver (2010) propose using a logistic odds-ratio scale $\log\frac{R}{L}$
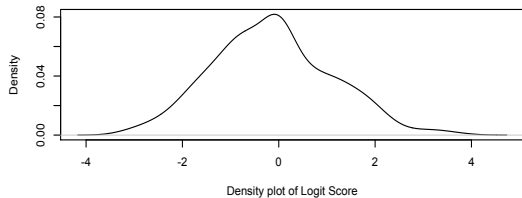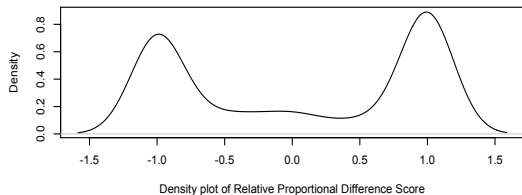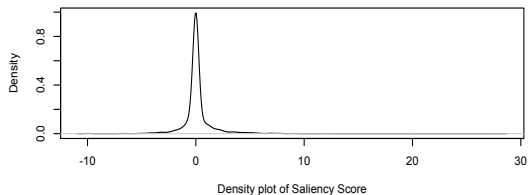
Comparing scales:

$\hat{\theta}^{(S)}$ v. $\hat{\theta}^{(R)}$

Protectionism
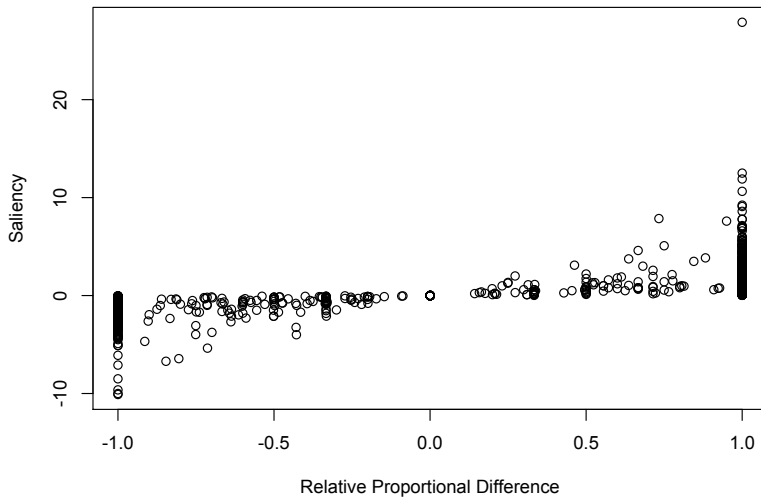
Comparing scales

Protectionism

distributions

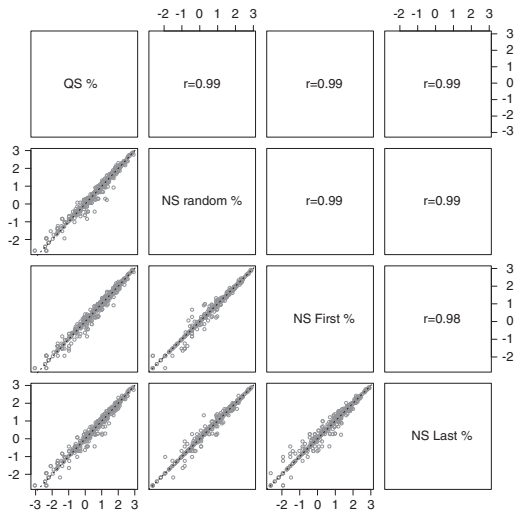# And some results from natural v. quasi-sentences

Fig. 2. Comparing quasi-sentence aggregate category percentages to natural sentence recodings
Notes: Three rules are compared: randomly assign the code based on constituent QSs; take the first QS code for the natural sentence; and take the last QS code for the natural sentences. Total manifestos analysed: 15.