Day 5: Classical quantitative content analysis

Kenneth Benoit

Essex Summer School 2011

July 15, 2011

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
- Alternatives are the purely statistical analysis of text as data, where human decisions are minimal or non-existent, and statistical methods are used to scale quantities from texts
- Other alternatives could be purely descriptive approaches to word frequency analysis

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
- Common "CATA" or "CACA" tools:
    - MaxQDA
    - T-Lab
    - Atlas-TI (formerly NUD*IST)
    - WordStat
    - TextPack
    - Diction
    - General Inquirer
    - many others

Components of manual coding approaches

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.

Sampling Texts

- (Mainly we have already covered this on Days 3–4)
- In hand-coded schemes, sampling may be more deliberate
- For the Comparative Manifesto Project, the case study for this topic, "sampling" consists of selecting all texts of a particular class

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
- Examples

  Category frequencies: Coded category frequency measures, such as the proportion of times "economy" is mentioned in a speech, or the proportion of mentions of the environment

  Type/token measures: Frequency tabulations of token types and their frequencies

  Range/variance: Here we might be interested in the total number or the spread or variance of categories used in particular documents or by particular speakers
- May also involve scales or indexes constructed from summary information

Summarizing: Example

Democratic	Republican
iraq	consent
administration	ask
year	unanimous
health	bill
families	committee
program	senate
care	30
debt	2006
women	border
veterans	senator
help	vote
americans	law
country	hearing
children	authorized
new	further
education	states
funding	proceed
workers	order
programs	session
disaster	time

Top 20 Democratic and Republican words from the 2006 US Senate (source: Nicholas Beauchamp 2008)

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 26 categories of the "right" and subtracts from this the sum of 13 categories of the "left".

Inference and Reporting

- This involves drawing conclusions from the research, and these conclusions will depend on the *validity* established by the research design

- Reporting means communicating the results in a clear and relevant fashion. (This can be challenging – see for instance the Schonhardt-Bailey article.)

- No iron-clad rules here – use your discretion as applied to a particular case

Unitizing Texts

- Briefly read the CMP Coder Instructions in Appendix 2 of Mapping Policy Preferences II (on the web page for Day 2).
- 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

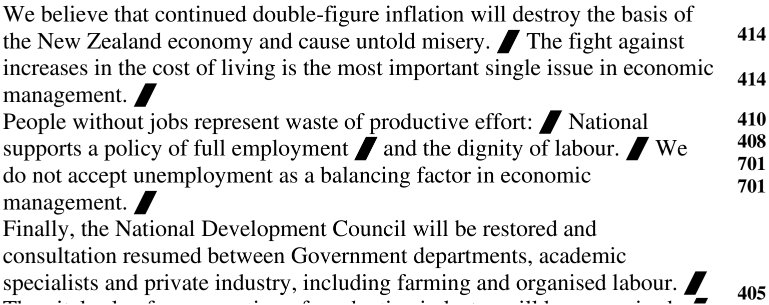
And now try to *code* it

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 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.

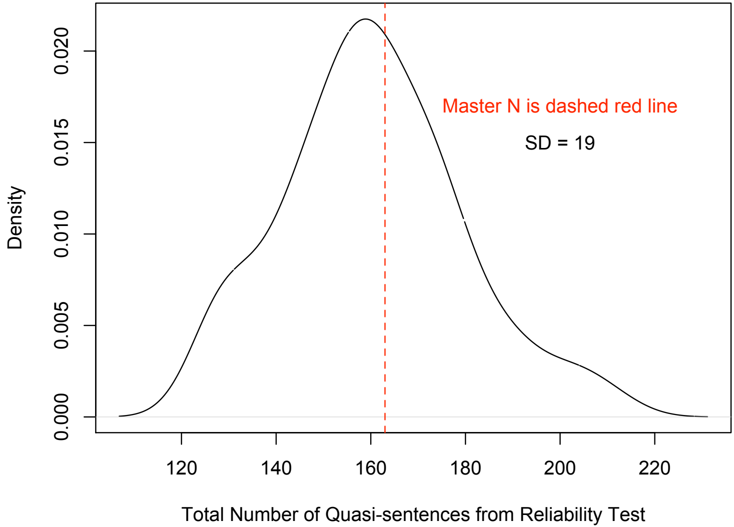
And the ("gold standard") answer is:



We believe that continued double-figure inflation will destroy the basis of
the New Zealand economy and cause untold misery. / The fight against 414
increases in the cost of living is the most important single issue in economic 414
management. /
People without jobs represent waste of productive effort: / National 410
supports a policy of full employment / and the dignity of labour. / We 408
do not accept unemployment as a balancing factor in economic 701
management. / 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

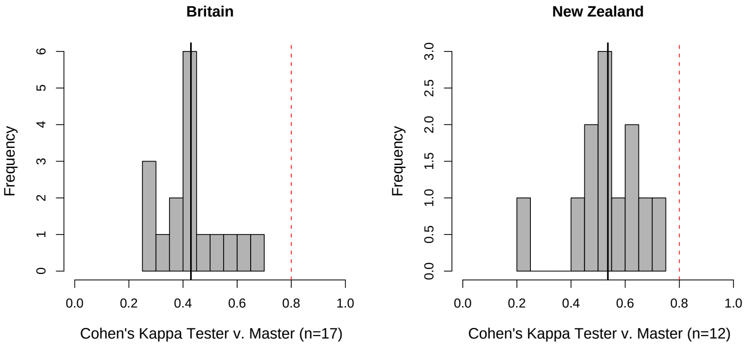


Empirical results from Mikhaylov and Benoit 2010

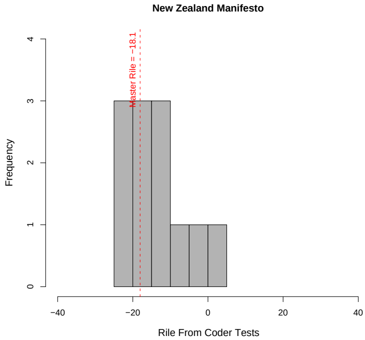
Caveats before I show you some compromising pictures:

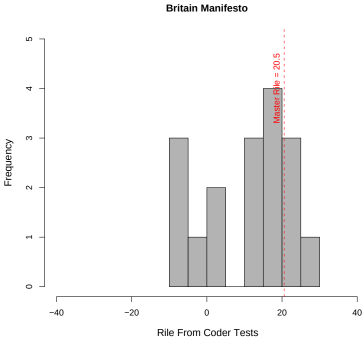
- We are not out to smear mud on the CMP! We actually like and respect the CMP and believe in the usefulness of their objective.
- *At the same time*, no research project should be immune from improvement
- There are weaknesses in the data and these are worth knowing
- The structure of the tests: Ask 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

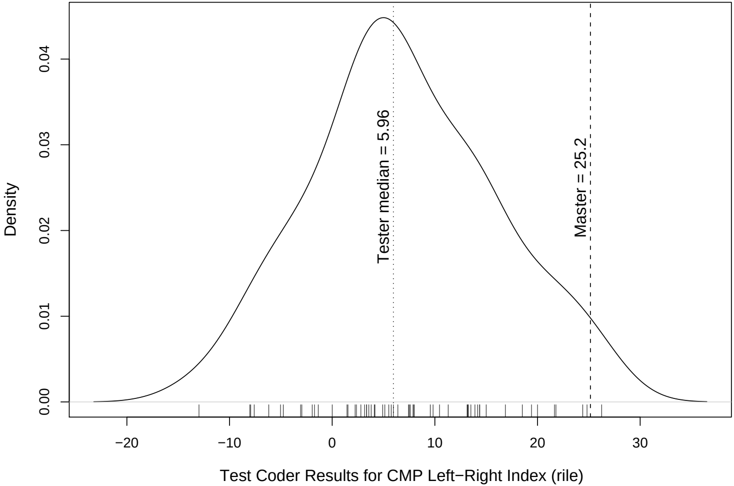


Empirical results from CMP reliability tests

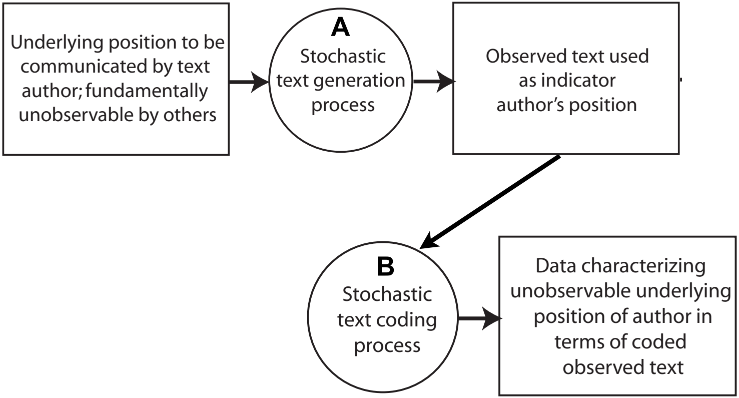

New Zealand Manifesto


Britain Manifesto

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

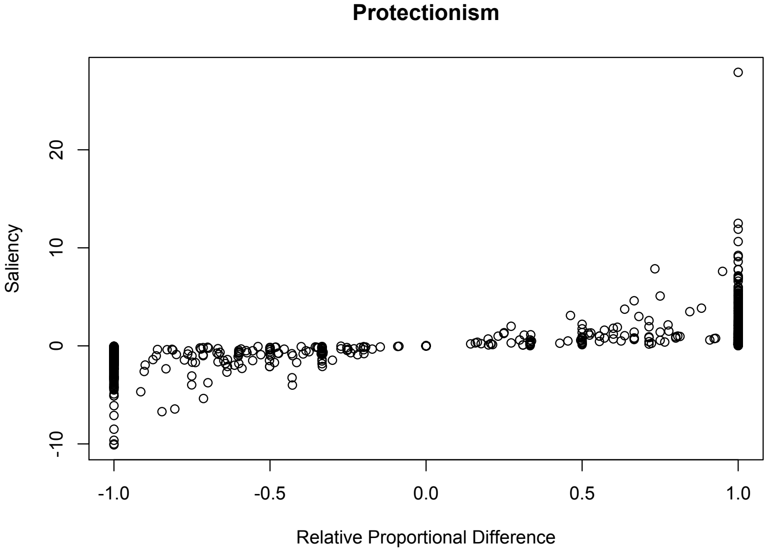
- Scaling becomes a major issue when we wish to construct quantities of interest from quantitative content analyses
- Simple example: Proportion of content of a given type (e.g. anti-Lisbon treaty)
- Complex example: Left-right policy positions (e.g. CMP "Rile")
- Are the metrics "natural"?
- Does the 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:

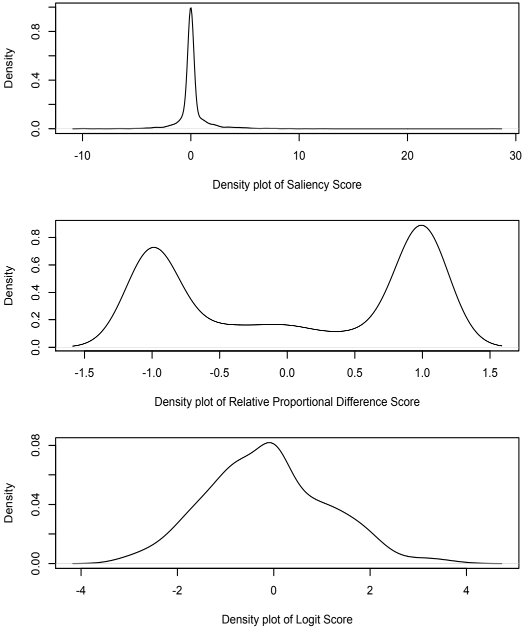
$\hat{\theta}^{(S)}$ v. $\hat{\theta}^{(R)}$



Comparing scales

Protectionism

distributions



Content Analysis Programs

Yoshikoder (Hamlet, Diction, Textpack, Wordstat, etc.)
LIWC (Linguistic Inquiry and Word Count, Pennebacker)
General Inquirer (Stone et al.)
Alceste (Image corp.)
See Lowe's review and also Alexa and Zuell (2000).

Content Analysis Programs

Yoshikoder is one of many classical content analysis programs having a basic handful of functions:

- Category building
- Concordance construction
- Frequency reports

Not as fancy as Wordstat but. . .

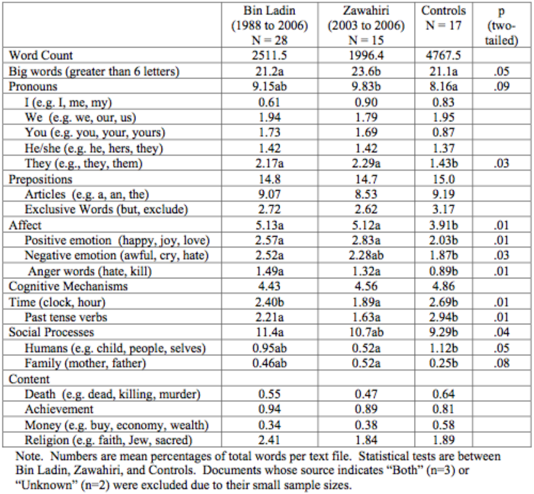
- free!
- works with non-english text
- works on all operating systems

Content Analysis Programs

LIWC is both a dictionary and a program (english only)
(one form of this dictionary is translated into Yoshikoder format and available from www.yoshikoder.org) Mostly used for social psychology applications
Has an online version
Example:

- Zawahiri vs. bin Laden vs. the world. . . (Pennebaker and Chung)

bin Laden vs. Zawahiri vs. Controls



	Bin Ladin (1988 to 2006) N = 28	Zawahiri (2003 to 2006) N = 15	Controls N = 17	p (two-tailed)
Word Count	2511.5	1996.4	4767.5	
Big words (greater than 6 letters)	21.2a	23.6b	21.1a	.05
Pronouns	9.15ab	9.83b	8.16a	.09
I (e.g. I, me, my)	0.61	0.90	0.83	
We (e.g. we, our, us)	1.94	1.79	1.95	
You (e.g. you, your, yours)	1.73	1.69	0.87	
He/she (e.g. he, hers, they)	1.42	1.42	1.37	
They (e.g., they, them)	2.17a	2.29a	1.43b	.03
Prepositions	14.8	14.7	15.0	
Articles (e.g. a, an, the)	9.07	8.53	9.19	
Exclusive Words (but, exclude)	2.72	2.62	3.17	
Affect	5.13a	5.12a	3.91b	.01
Positive emotion (happy, joy, love)	2.57a	2.83a	2.03b	.01
Negative emotion (awful, cry, hate)	2.52a	2.28ab	1.87b	.03
Anger words (hate, kill)	1.49a	1.32a	0.89b	.01
Cognitive Mechanisms	4.43	4.56	4.86	
Time (clock, hour)	2.40b	1.89a	2.69b	.01
Past tense verbs	2.21a	1.63a	2.94b	.01
Social Processes	11.4a	10.7ab	9.29b	.04
Humans (e.g. child, people, selves)	0.95ab	0.52a	1.12b	.05
Family (mother, father)	0.46ab	0.52a	0.25b	.08
Content				
Death (e.g. dead, killing, murder)	0.55	0.47	0.64	
Achievement	0.94	0.89	0.81	
Money (e.g. buy, economy, wealth)	0.34	0.38	0.58	
Religion (e.g. faith, Jew, sacred)	2.41	1.84	1.89	

Note. Numbers are mean percentages of total words per text file. Statistical tests are between Bin Ladin, Zawahiri, and Controls. Documents whose source indicates "Both" (n=3) or "Unknown" (n=2) were excluded due to their small sample sizes.

Content Analysis Programs

The General Inquirer is perhaps the oldest content analysis program still in existence (1967)
13000 words (and 6336 word sense disambiguation rules)
An online version is available at Maryland
Example:

- speeches from US presidential candidates (2000)