Quantitative Content Analysis: Lecture 7

Matthias Haber

29 March 2017

Today's Outline

- Assignment 1 solution
- CMP: Validity and reliability issues
- Assignment 2

Reliability & Validity

Reliability

- Are measures that are derived from text analysis be stable when repeated?

Validity

- Does the text analysis measure what it is supposed to measure

CMP: Reliability within

Misclassification problem

- Ambuiguities and overlap in the Coding scheme:
  - '401: Free enterprise: Favorable mentions of free enterprise capitalismm; superiority of individual enterprise over state control systems. . .' +'402: Incentives: Need for wage and tax policies to induce enterprise' +'501: Environmental protection'
- CMP data are coded only once
- No estimate for the uncertainty resulting from the coding process

CMP: Reliability within (II)

The CMP's argument

- Reliability exists for the data set as a whole
  +i.e. for the RiLe scale or where substantial sub-sets of the data are input together the data has high reliability and validity
- Misclassification should 'cancel out'
  +Miclassification within left/right/uncoded codes occur
  +Misclassification between the groups of codes should be rare

Are misclassifications self cancelling?

Mickhaylov et al. 2010

- Texts used in CMP-training;
  +Manifestos for UK and NZ +'Gold-Standard'/'Master coding' gives 'true' coding
- Actual CMP-coders code manifestos
  +'Worst' 25% of coders are excluded in evaluation +(GB: 17, NZ: 12)
- Coding of the two manifestos only uses ~20 of the 57 categories
- Correct unitizing is given for coders

1. Intercoder-reliability

Intercoder-reliability

- Agreement between coders
- Calculated at
  +Full 56 categories level +RiLe-relevant level, i.e. with right, left and other as categories, only
- .35 - .36 for GB and .4 - .47 for NZ
  +Remember Krippendorff's suggestion from the first slide...

2. Coder-Master reliability

Coder-Master reliability - Agreement between coders and gold standard - Suggested kappa: .8
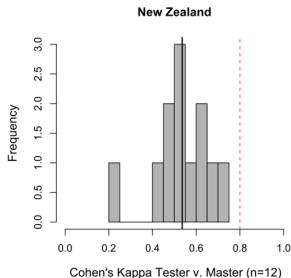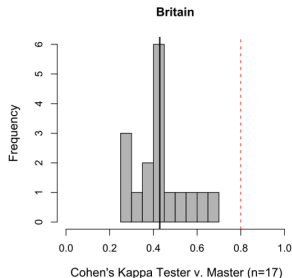
2. Coder-Master reliability (II)

- CMP argument: Disagreement cancels out
  +Data set as a whole is relable and valid +Misclassifications occur, but do not bias RiLe +i.e. occur between codes pertaining to the same side
- Misclassifications:
  +Gold standard gives true classification +Are misclassification within right/left/other or across? +Measured on the RiLe level, i.e. three categories

3. True vs. observed RiLe classification

		True Rile Category			Total
		Left	Right	None	
	Left	430	100	188	718
		0.59	0.11	0.19	
Coded	Right	41	650	115	806
Rile		0.06	**0.69**	0.11	
	None	254	193	712	1,159
		0.35	0.20	**0.70**	
	Total	725	943	1,015	1,668
	"False negative" rate	.41	.31	.30	
	"False positive" rate	.15	.09	.27	

Frequently misclassified categories

- Look at the probability for a code to be coded as left, right or other
- Compare with true coding (gold standard)
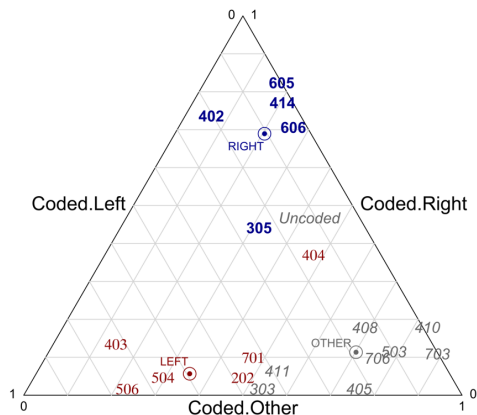
Frequently misclassified categories (II)

- Coders assign truly left/right codes to the opposite categories (305 & 404)
- Truly uncoded text (headings etc.) are more likely to be coded left or right than other
- Additional noise?
- Bias in the RiLe scale?

Assignment 2

1. Register with the Manifesto Project and download the full CMP data set (v 2016b) directly into R using the manifestoR package.

2. Chose a country of your interest and create an overview for the positions provided by the CMP (RiLe, Planeco, etc.) for each party. Display your results graphically.

3. Analyze an entire party family (e.g. green parties) across multiple countries (>2) and choose an appropriate graphical presentation.

4. Calculate an alternative right-left measures from Week 6 for **either** 2. and 3. and describe how results change.

5. Exclude some of the most unreliable categories described in Mickhaylov et al. 2012 for **either** 2. or 3., recalculate and describe possible changes.

Assignment 2 (II)

- Assignment is due on April 11
- Submit code via email
- Hand in short report in class

Next Session

- NO CLASS NEXT WEEK!
- Challenges involved in working with CMP data