Content Analysis Dictionaries

William Lowe

Hertie School of Governance

22nd September 2020

Classical content analysis

Content is, or is constructed from, *categories* e.g.

→ human rights, welfare state, national security

Substantively these often have *valence*, e.g.

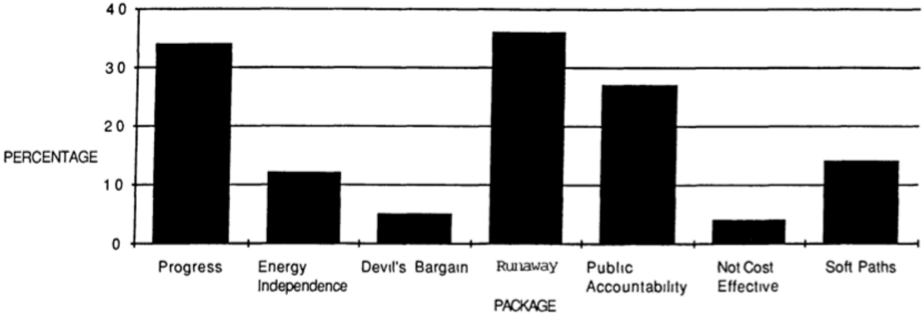
→ pro-welfare state vs. anti-welfare state, lots of CMP categories

But they are invariably treated as *nominal level* variables

We are typically interested in them for

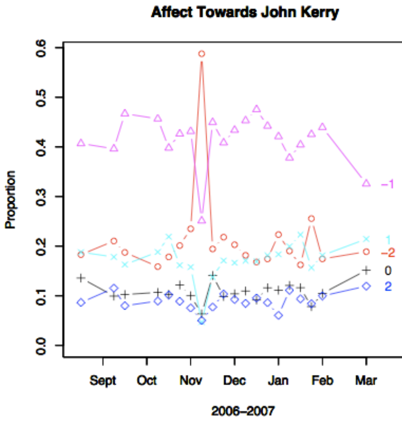
→ simple descriptions, making comparisons, tracing temporal dynamics

Talking like a newspaper



From Gamson and Modigliani (1989)

Talking like a presidential candidate



From **Hopkin.King2010**

Talking like a terrorist

	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	

Talking about drugs

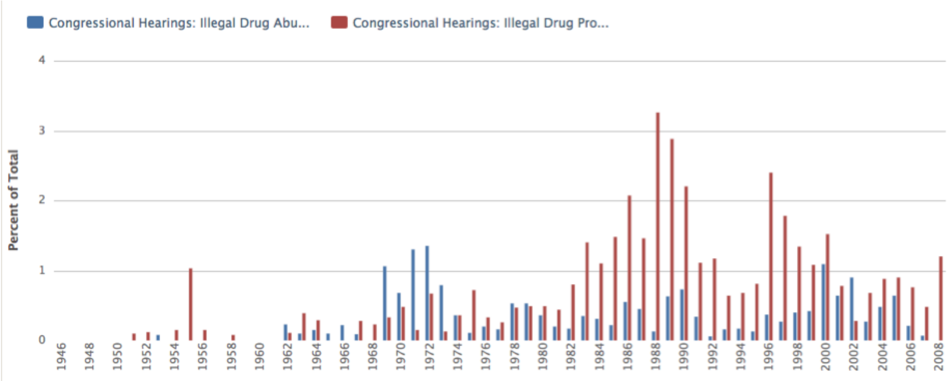
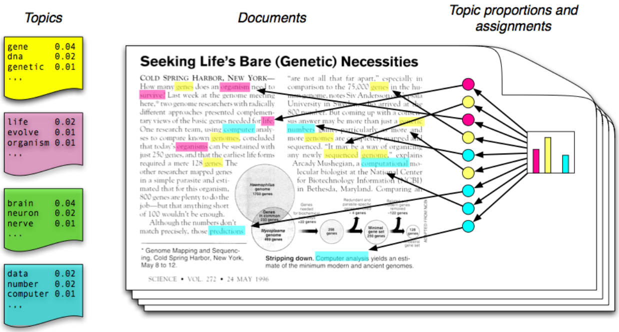


Figure: Congressional Bills Project website (retrieved 2010)

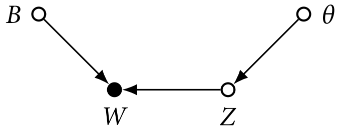
Classical content analysis

Categories are

- → equivalence classes over words
- → representable as assignments of a K-valued category membership variable Z to each word

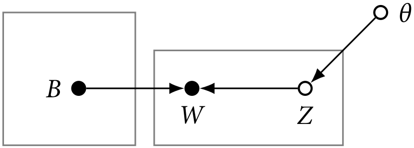


Topics



W_i is the i-th word in the document
Z_i is true topic of W_i
$\theta_k = P(Z = k)$ in this document
B_k is the distribution $P(W \mid Z = k)$

Now let's claim we *know* some things



Content analysis dictionary

ECONOMY	
state reg	**market econ** assets bid
accommodation age	choice* compet*
ambulance assist	constrain* …
benefit …	

from Laver and Garry's (2000) dictionary

As a posterior: $P(Z \mid W)$

Dictionary is an explicit and very *certain* statement of $P(Z \mid W)$

W	P(Z = state reg | W)	P(Z = market econ | W)
age	1	0
benefit	1	0
…	…	…
assets	0	1
bid	0	1
…	…	…

...from a underspecified likelihood

The *only* way this could be true is if the data had been generated like

$P(W \mid Z)$

	state reg	market econ
P(W = "age" | Z)	a	**0**
P(W = "benefit" | Z)	b	**0**
…	…	…
P(W = "assets" | Z)	**0**	c
P(W = "bid" | Z)	**0**	d
…	…	…

...leading to a posterior over content

Define the category *counts*

$$Z_k = \sum_i^N P(Z = k \mid W_i)$$

and estimate category posterior probabilities, a.k.a. relative *proportions* using

$$\hat{\theta}_k = \frac{Z_k}{\sum_j^K Z_j}$$

...leading to a posterior over content

Define the category *counts*

$$Z_k = \sum_i^N P(Z = k \mid W_i)$$

and estimate category posterior probabilities, a.k.a. relative *proportions* using

$$\hat{\theta}_k = \frac{Z_k}{\sum_j^K Z_j}$$

When θ is a set of multinomial parameters, *and the model assumptions are correct*, this could be a reasonable estimator.

Reconstruction

Dictionary-based content analysis was *not* developed this way

→ Originally (e.g. **???**) there was no probability model

Reconstruction

Dictionary-based content analysis was *not* developed this way

→ Originally (e.g. **???**) there was no probability model

We are reconstructing it to compare and contrast with topic models which make the *same* structural assumptions but operate in exploratory, not confirmatory mode

Connecting CCA content to politics

We're usually interested in category proportions per unit (usually document), e.g.

→ *How much* of this document is about national defense?

→ What is the *difference* of aggregated left and aggregated right categories (RILE)

→ How does the *balance* of human rights and national defense change over time?

Inference About content

Statistically speaking, we are just dealing with proportions of various kinds

- → a proportion
- → a difference of proportions
- → a ratio of proportions

Under certain sampling assumptions we can make inferences about a population

Simple inference about proportions

Example: in the 2001 Labour manifesto there are 872 matches to Laver and Garry's *state reg* category

- → 0.029 (nearly 3%) of the document's words
- → 0.066 (about 6%) of words that matched *any* categories

The document has 30157 words, so the *first* proportion is estimated as

$$\hat{\theta}_{state\ reg} = 0.029 \ \ [0.027, 0.030]$$

What does this mean?

Inference about proportions

Think of the party headquarters repeatedly *drafting* this manifesto

The true proportion – the one suitable to the party's policies – is fixed but every draft is slightly different

The confidence interval reflects the fact that we expect long manifestos to have more precise information about policy

Inference about proportions

Think of the party headquarters repeatedly *drafting* this manifesto

The true proportion – the one suitable to the party's policies – is fixed but every draft is slightly different

The confidence interval reflects the fact that we expect long manifestos to have more precise information about policy

This interval is computed as if

- → every word was a new independent piece of information
- → we're never wrong about word categories

Ratios: How new was `New Labour'?

Was the Conservative party in 1992 more or less for state intervention than ‘New’ Labour in 1997?

Compare instances of *state reg* and *market econ* in the manifestos

party	*state reg*	*market econ*
Conservative	320	643
Labour	396	268

Quantities of interest: Risk ratios

Compute two *risk ratios*:

$$
\begin{aligned}
RR_{state\ reg} &= \frac{P(state\ reg \mid \text{cons})}{P(state\ reg \mid \text{lab})} \\
RR_{market\ econ} &= \frac{P(market\ econ \mid \text{cons})}{P(market\ econ \mid \text{lab})}
\end{aligned}
$$

and 95% confidence intervals

Interpreting risk ratios

If $RR = 1$ then the category occurs at the same rate in labour and conservative manifestos

If $RR = 2$ then the conservative manifesto contains *twice* as much *state reg* language as the labour manifesto

If $RR = .5$ then the conservative manifesto contains *half* as much *state reg* language as the labour manifesto

If the confidence interval for RR contains 1 then we *no evidence* that *state reg* and *market econ* occur at different rates

Risk ratios

	Risk Ratio
market econ	1.45 [1.26, 1.67]
state reg	0.49 [0.42, 0.57]

Conservative manifesto generates *market econ* words 45% more often

→ 45% = 100(1.45 - 1)%

Conservative manifesto only generates 49% as many *state reg* words as Labour. Equivalently Labour generates them about *twice* as often

Log ratios

It's often more useful to work with log ratios

$$\log(2) \approx 0.69$$
$$\log(0.5) \approx -0.69$$

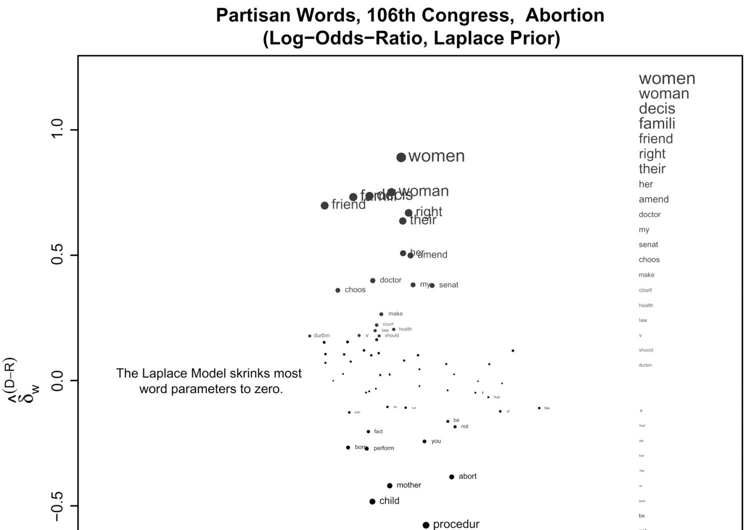
which are

→ symmetric, with an interpretable 0

→ proportional (percentage increase/decreases)

Log ratios as forensics

Log ratios of words: keyness



Ratios, ratios everywhere

party	*state reg*	*market econ*
Conservative	320	643
Labour	396	268

Looking forward a little, there are two separate sorts of information in tables like these

Marginal information:

→ e.g. state regulation is mentioned 320+396=716 times, and market economy 643+268=911 times.

Ratios, ratios everywhere

party	*state reg*	*market econ*
Conservative	320	643
Labour	396	268

Association information:

- → conservatives mention state regulation 320/643 = about 50% as much as market economy
- → labour mentions it 396/268 = about 50% more than market economy.

So the odds ratio (0.5 / 1.5) = about 0.33.

This, plus the marginal information, *completely characterizes* this table.

A psychological aside

Are people *really* sensitive to these sorts of associational statistics?

A psychological aside

Are people *really* sensitive to these sorts of associational statistics?

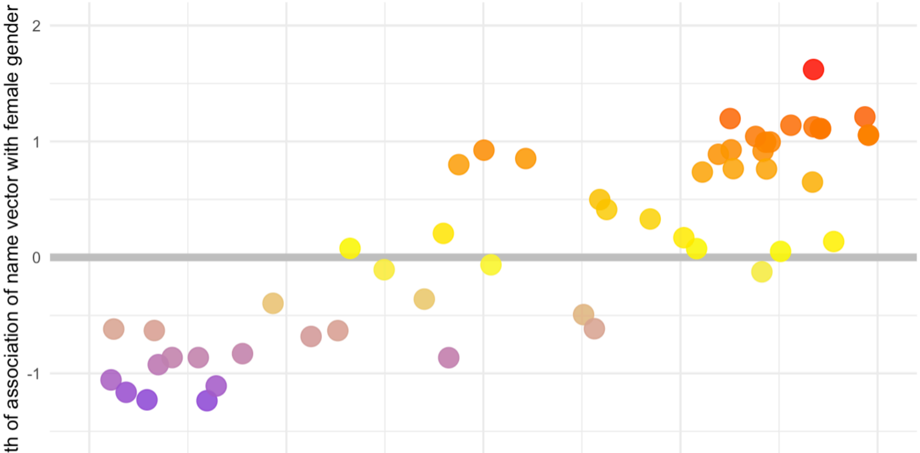
It seems they are:

→ Even infants track conditional probabilities (**???**)

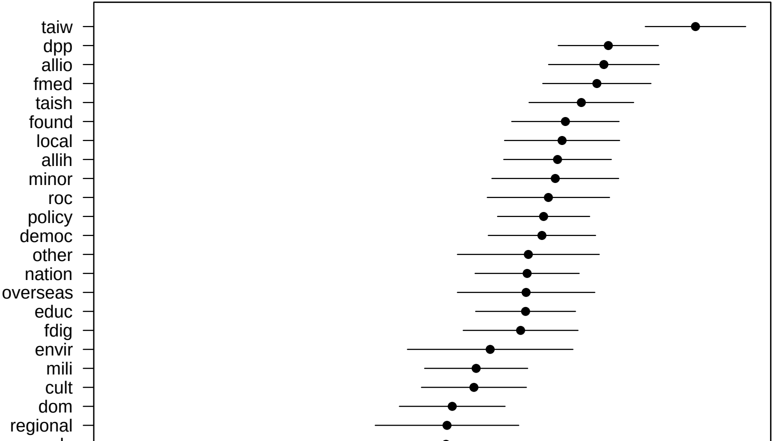
→ Purely statistical textual measures recover Implicit Association Test biases (**???**)

Word embeddings

Contextual similarity tracks real relations (as it must!)



Category count as a dependent variable



Category counts as a dependent variable

District vs party focus in speeches

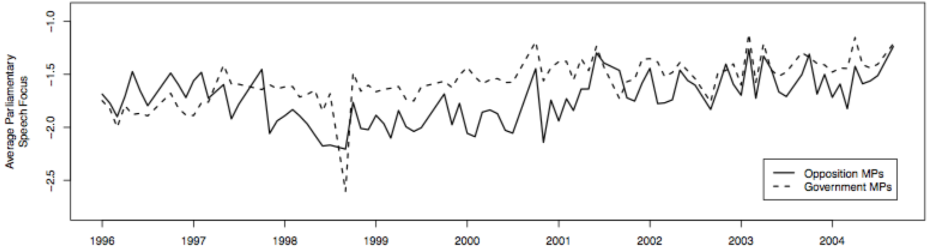


Figure: (Kellerman and Proksch, MS)

Data: [district words, party words]

Category counts as a dependent variable

Logit reminder:

→ when you are modeling two category counts as a function of covariates the linear predictor is a smoothed version of their log ratio

$$\text{district words, party words} \sim Multinomial(\pi^{\text{district}}_{,} N_i)$$

$$\log \frac{\pi_i^{\text{district}}}{(1 - \pi_i^{\text{district}}}) = \log \frac{\pi_i^{\text{district}}}{\pi_i^{\text{party}}} = \ldots$$

OK, how do I make such a dictionary?

Find a suitable tool

→ Maximise measurement validity

→ Minimise *measurement error*

OK, how do I make such a dictionary?

Find a suitable tool

→ Maximise measurement validity

→ Minimise *measurement error*

(Sell high, buy low)

Find a suitable tool

Wordstat

LIWC (maybe don’t)

Hamlet

Atlas-ti (?)

Yoshikoder

What's to go wrong?

The source of measurement error

Measurement error in classical content analysis is primarily failure of *this* assumption:

W	P(Z = state reg | W)	P(Z = market econ | W)
age	1	0
benefit	1	0
…	…	…
assets	0	1
bid	0	1
…	…	…

Consequences of measurement error

What are the effects of measurement error in category counts?

Being directly wrong, e.g.

- → Estimated rates are too *low* (bias)
- → Some of estimates are more biased than others

Being *indirectly* wrong, e.g.

- → Subtractive or ratio left-right measures are too *centrist*

Measurement error: example

Assume

- → a vocabulary of only two words 'benefit' and 'assets'
- → a *subtractive* measure of position (Laver and Garry):

$$\frac{Z_{\text{market econ}} - Z_{statereg}}{Z_{\text{market econ}} + Z_{statereg}}$$

Then we hope that the posterior over categories is:

	state reg	*market econ*	
"benefit"	1	0	1
"assets"	0	1	1

Measurement error: example

but if word generation happened like this…

	state reg	*market econ*
“benefit”	0.7	0.2
“assets”	0.3	0.8
total	1	1

then

$$P(W = \text{"asset"} | \mid Z = \text{state reg}) > 0$$

so, e.g.

$$P(Z = \text{state reg} \mid W = \text{"asset"}) < 1$$

Measurement error: example

Assume

- → $Z_{market\ econ} = 10$
- → $Z_{state\ reg} = 20$

Then the *true* difference is

$$\frac{(10-20)}{(10+20)} = -0.33$$

Under perfect measurement we would expect

- → 20 'benefit's
- → 10 'assets's

Measurement error: example

Under *imperfect* measurement we expect

→ 16 'benefit' (14 from *state reg* but 2 from *market econ*)

→ 14 'assets' (8 from *market econ* but 6 from *state reg*)

Measurement error: example

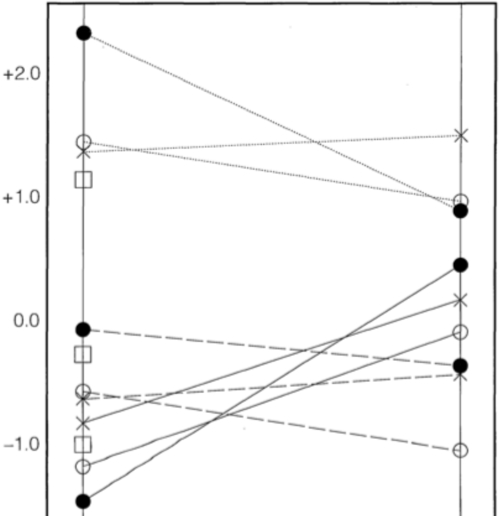
The proportional difference measure is now

$$\frac{(14 - 16)}{(14 + 16)} = -0.07$$

Apparently much closer to the centre, but only because of measurement error

All relative measures will have this problem (and all kinds of text analyzers)

In action (Laver and Garry 2000)



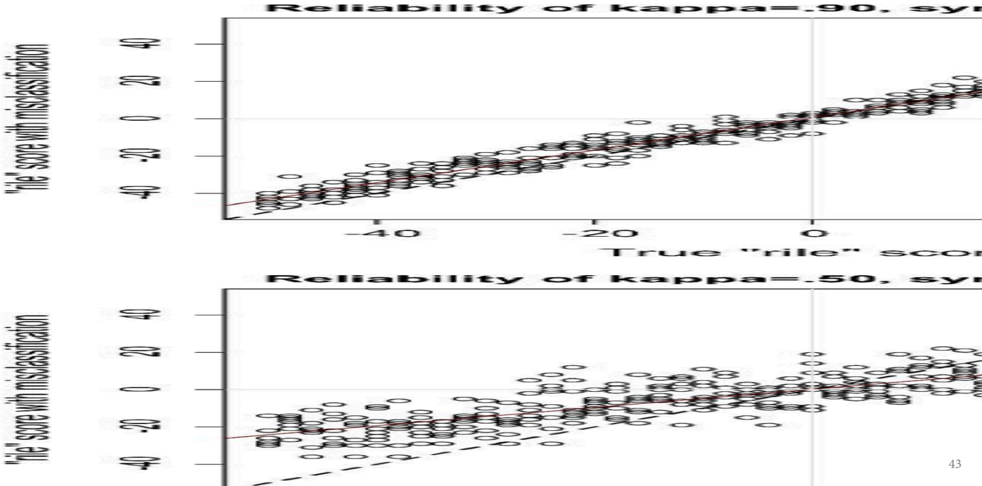
In action with people, not dictionaries

Table 3 Misclassification matrix for true versus observed Rile

		True Rile category			
		Left	*None*	*Right*	*Total*
	Left	430	188	100	718
		0.59	0.19	0.11	
Coded	None	254	712	193	1159
Rile		0.35	**0.70**	0.20	
	Right	41	115	650	806
		0.06	0.11	**0.69**	
	Total	725	1015	943	1668
	False negative rate	0.41	0.30	0.31	
	False positive rate	0.15	0.27	0.09	

Note. The top figure in each cell is the raw count; the bottom figure is the column proportion. The figures are empirically computed from combined British and New Zealand manifesto tests. The false negative rate is 1—sensitivity, whereas the false positive rate is 1—specificity.

Attenuation (Mikhaylov et al. 2011)



Solutions: Some theological approaches

Solutions: Some theological approaches

Solutions: Do not sin in the first place

"Beatings will continue until morale improves"

Solutions: Do not sin in the first place

“Beatings will continue until morale improves”

An often non-obvious fact about content dictionaries:

- → *precision*: proportion of words used the way your dictionary assumes
- → *recall*: proportion of words used that way that are in your dictionary

always trade-off…

Sins of ommission vs sins of commission

Every field reinvents this distinction:

- → precision and recall
- → specificity and sensitivity
- → users and producer’s accuracy
- → type 1 and type 2 error

Humility and self-examination

Keyword in context analyses (KWIC) allow you to scan all contexts of a word

→ How many of them are the sense or usage you want?

KWIC: benefit*

	pre	keyword	post
1	also keep all the other	benefits	that pensioners currently receive ,
2	regulation will have to have	benefits	exceeding costs , and regulations
3	and Controlled Immigration Britain has	benefited	from immigration . We all
4	positive contribution But if those	benefits	are to continue to flow
5	Nor ther n Ireland brings	benefits	to all parts of our
6	their home , will also	benefit	first-time buyers . Empowering individuals
7	you help yourself ; you	benefit	and the country benefits .
8	you benefit and the country	benefits	. So now , I
9	result of our tax and	benefit	measures compared to 1997 .
10	result of personal tax and	benefit	measures introduced since 1997 ,
11	, the savings on unemployment	benefits	will go towards investing more
12	trebled the number on incapacity	benefits	. We will help 17
13	Work programme and reform Incapacity	Benefit	, with the main elements
14	main elements of the new	benefit	regime in place from 2008
15	stronger penalties . To the	benefit	of business and household consumers
16	effective directive to provide real	benefits	to consumers and new opportunities
17	better.We are examining the potential	benefits	of a parallel Expressway on
18	ways to lock in the	benefit	of new capacity . We
19	are determined to spread the	benefits	of enterprise to every community
20	to get ahead , to	benefit	from improving public services ,
21	of the school workforce is	benefiting	staff and helping to tailor
22	teachers and pupils get the	benefit	of the range of support

Last week

References

Gamson, W. A. & Modigliani, A. (1989). 'Media discourse and public opinion on nuclear power: A constructionist approach'. *American Journal of Sociology*, *95*(1), 1–37.