Content Analysis Dictionaries

William Lowe

Hertie School of Governance

23rd September 2020

Plan

Dictionary based content analysis

The underlying measurement model

How to read a dictionary

Using the output

How to do it

How not to do it

Measurement error and its consequences

Classical dictionary content analysis

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

→ human rights, welfare state, national security

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

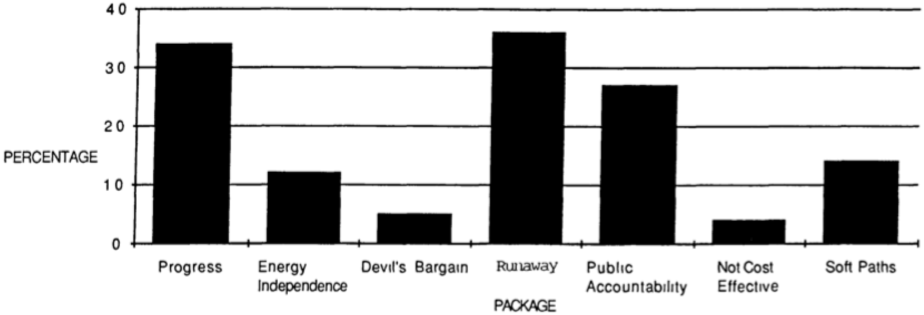
→ pro-welfare state vs. anti-welfare state, lots of CMP categories

But they are invariably treated as *nominal level* variables

We are typically interested in them for

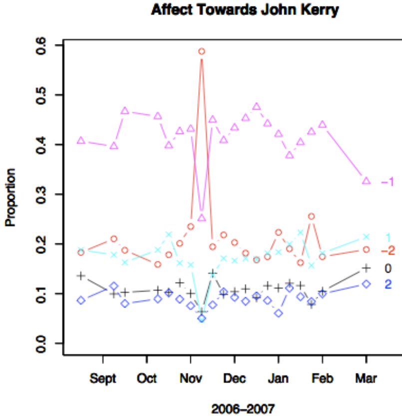
→ simple descriptions, making comparisons, tracing temporal dynamics

Talking like a newspaper



From Gamson and Modigliani (1989)

Talking like a presidential candidate



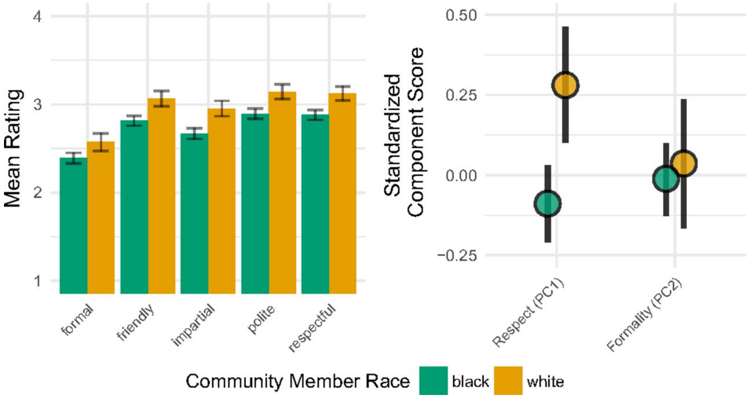
From Hopkins and King (2010)

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	
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

Talking to police

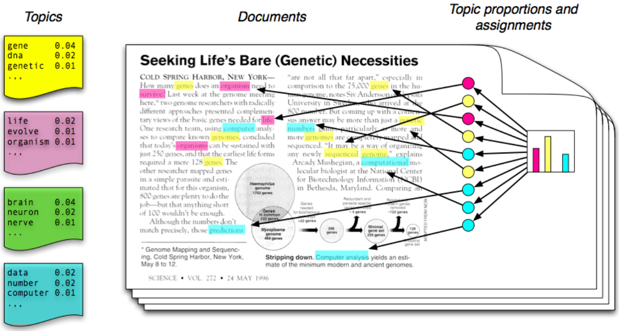


From Voigt et al. (2017)

Classical content analysis

Categories are

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



Topics

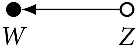
W_i is the i-th word in the document

Z_i is true topic of W_i

$\theta_k = P(Z = k)$ in this document

β_k in B is the distribution $P(W \mid Z = k)$

Topics



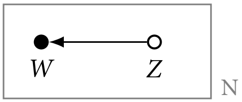
W_i is the i-th word in the document

Z_i is true topic of W_i

$\theta_k = P(Z = k)$ in this document

β_k in B is the distribution $P(W \mid Z = k)$

Topics



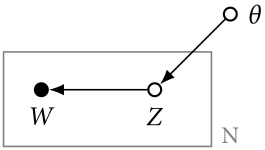
W_i is the i-th word in the document

Z_i is true topic of W_i

$\theta_k = P(Z = k)$ in this document

β_k in B is the distribution $P(W \mid Z = k)$

Topics



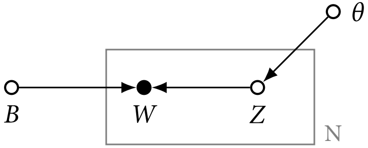
W_i is the i-th word in the document

Z_i is true topic of W_i

$\theta_k = P(Z = k)$ in this document

β_k in B is the distribution $P(W \mid Z = k)$

Topics



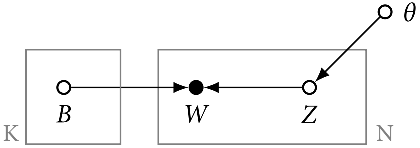
W_i is the i-th word in the document

Z_i is true topic of W_i

$\theta_k = P(Z = k)$ in this document

β_k in B is the distribution $P(W \mid Z = k)$

Topics



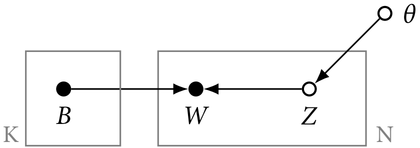
W_i is the i-th word in the document

Z_i is true topic of W_i

$\theta_k = P(Z = k)$ in this document

β_k in B is the distribution $P(W \mid Z = k)$

Topics

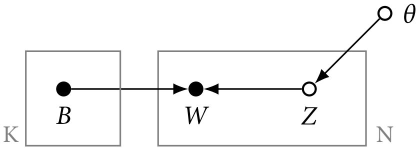


B is a 'dictionary' that explains how often each word should be generated in each of the K topics

→ An entry in the dictionary B is a vector of word generation probabilities β_k

This week we will assume we *know* B

Topics



B is a 'dictionary' that explains how often each word should be generated in each of the K topics

→ An entry in the dictionary B is a vector of word generation probabilities β_k

This week we will assume we *know* B

Strictly speaking we will assume that we know enough about its *inverse* $W \longrightarrow Z$ to say

→ for each word W, what its Z is.

Dictionary B

Here's a excerpt from the Economy section of the dictionary in Laver and Garry (2000)

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

How to read it

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
…	…	…

If we're so sure about Z...

Then estimating the proportion $Z = k$ in a document is easy.

First count up all the 'hits', where $Z = k$

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

then divide by the sum

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

and that's our estimate of the document content

Discrimination

Stating $P(Z \mid W)$ is the *discrimination* direction from last week

→ we didn't even learn it from data, we just asserted it!

So what must the generation process have looked like?

Discrimination

Stating $P(Z \mid W)$ is the *discrimination* direction from last week

→ we didn't even learn it from data, we just asserted it!

So what must the generation process have looked like?

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

	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
…	…	…

Reconstruction

Why do we seem to be going about this backwards?

- → Because we are reconstructing old practice as a measurement process
- → which allows us to learn from data things we previously only asserted
- → and understand exactly where and when things go wrong

But let's see how to work with the *output* of the process before looking into its quirks

Connecting 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 (Robinson 2016)

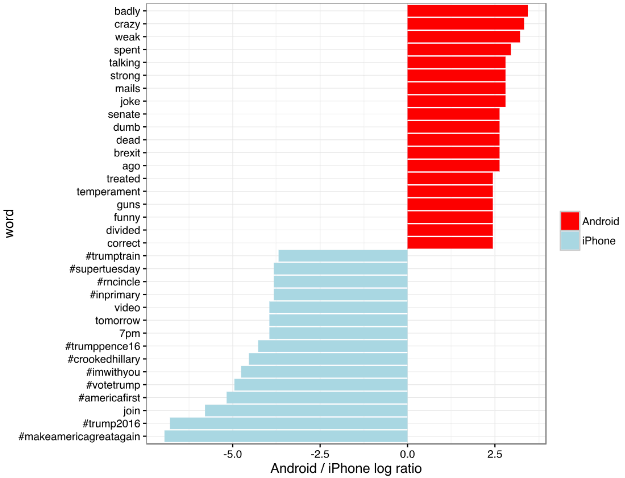
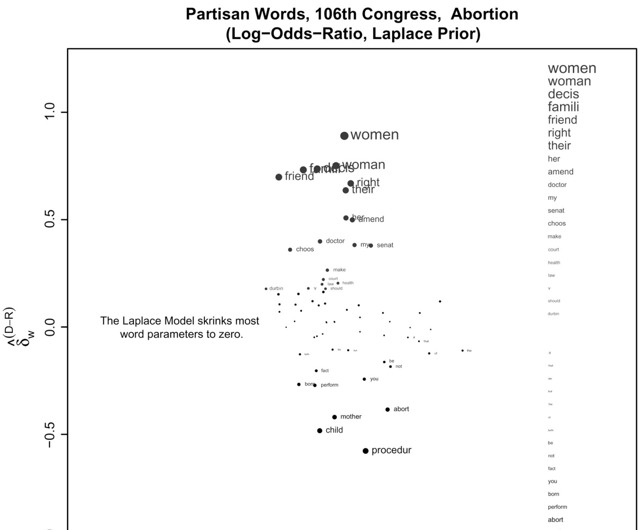
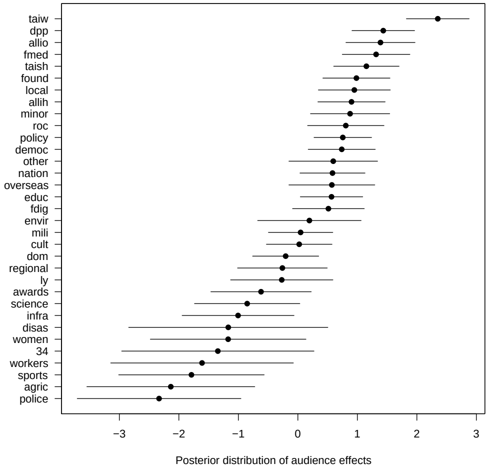


Figure: David Robinson (2016)

Log ratios of words: (Monroe et al. 2008)

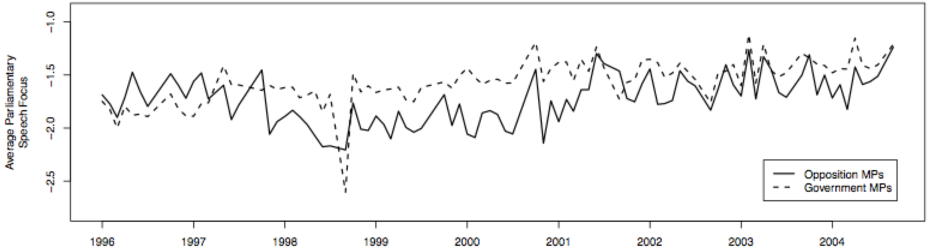


Category count as a dependent variable



Category counts as a dependent variable

District vs party focus in speeches



From Kellerman and Proksch, MS

OK, how do I make such a dictionary?

Find a suitable tool

- → Wordstat
- → LIWC (maybe don't)
- → Hamlet
- → Atlas-ti (?)
- → Yoshikoder

OK, how do I make such a dictionary?

Find a suitable tool

→ Wordstat

→ LIWC (maybe don't)

→ Hamlet

→ Atlas-ti (?)

→ Yoshikoder

Then assign words to

→ maximise measurement validity

→ minimise *measurement error*

OK, how do I make such a dictionary?

Find a suitable tool

→ Wordstat

→ LIWC (maybe don't)

→ Hamlet

→ Atlas-ti (?)

→ Yoshikoder

Then assign words to

→ maximise measurement validity

→ minimise *measurement error*

(Sell high, buy low)

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):

$$\frac{Z_{\text{market econ}} - Z_{\text{state reg}}}{Z_{\text{market econ}} + Z_{\text{state reg}}}$$

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*)

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

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*)

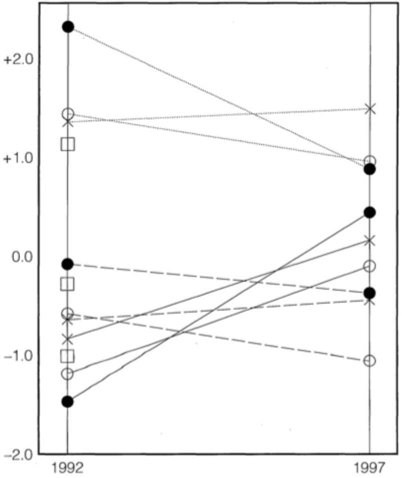
The proportional difference measure is now

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

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.

From Mikhaylov et al. (2011)

So what to do now?

That's for next 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.

Hopkins, D. J. & King, G. (2010). 'A method of automated nonparametric content analysis for social science'. *American Journal of Political Science*, *54*(1), 229–247.

Laver, M. & Garry, J. (2000). 'Estimating policy positions from political texts'. *American Journal of Political Science*, *44*(3), 619–634.

Mikhaylov, S., Laver, M. & Benoit, K. R. (2011). 'Coder reliability and misclassification in the human coding of party manifestos'. *Political Analysis*, *20*(1), 78–91.

Monroe, B. L., Colaresi, M. & Quinn, K. M. (2008). 'Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict'. *Political Analysis*, *16*(4), 372–403.

References

Pennebaker, J. W. & Chung, C. K. (2008). Computerized text analysis of al-qaeda transcripts. In K. Krippendorff & M. A. Bock (Eds.), *The content analysis reader*. Sage.

Robinson, D. (2016, August 9). *Text analysis of trump's tweets confirms he writes only the (angrier) android half*. http://varianceexplained.org/r/trump-tweets/

Sullivan, J. & Lowe, W. (2010). 'Chen shui-bian: On independence'. *China Quarterly*, *203*, 619–638.

Voigt, R., Camp, N. P., Prabhakaran, V., Hamilton, W. L., Hetey, R. C., Griffiths, C. M., Jurgens, D., Jurafsky, D. & Eberhardt, J. L. (2017). 'Language from police body camera footage shows racial disparities in officer respect'. *Proceedings of the National Academy of Sciences*, *114*(25), 6521–6526.