Quantitative Text Analysis

William Lowe

Hertie School of Governance

7th September 2020

Text as data

Transcendental question

What are the *conditions for the possibility* of learning about these things by counting words?

In plainer language:

→ How could this possibly work?

Big picture

There is a *message* or *content* that cannot be directly observed, e.g.

→ the topic of this lecture
→ my position on some political issue
→ the importance of defence issues to a some political party

and *behaviour*, including *linguistic behaviour*, e.g.

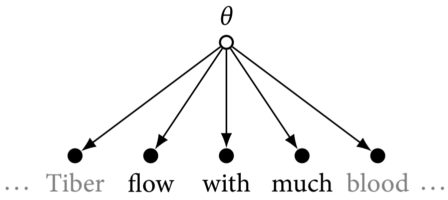
→ yelling
→ writing
→ lecturing

which *can* be directly observed.

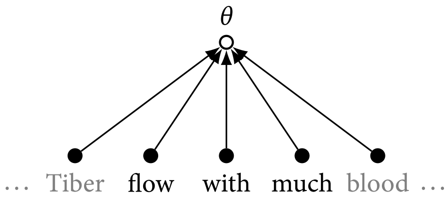
Although language can do things directly (Austin, 1962), we'll focus on the *expressed message* and the *words*…

Communication

To *communicate* a message θ – to inform, persuade, demand, or threaten – a producer (the speaker or writer) *generates* words of different kinds in different quantities



To *understand* a message the consumer (the hearer, reader, coder) uses those words to *reconstruct* the message



Communication

This process is

- → stable (Grice, 1993; Searle, 1995)
- → conventional (Lewis, 2011)
- → disruptible (Riker et al., 1996)
- → empirically underdetermined (Davidson, 1985; Quine, 1960)

How to model this without having to solve the problems of linguistics (psychology, politics) first?

Rely on:

- → instrumentality
- → reflexivity
- → randomness

Communication: Instrumentality

Instrumentality from 'them': Language use is a form of action (Austin, 1962; Krebs & Dawkins, 1984; Wittgenstein, 1958)

Note the distinction between

X means *Y* | *X* is used to mean *Y*

Instrumentality from us:

→ we aren't actually interested in words themselves; that's are for linguists

→ we aren't actually interested in what's in the head; that's for psychologists

Except as they help explain things we are interested in. Text is just data

Communication: Reflexivity

Politicians are often nice enough to talk as if they really do communicate this way

> *My theme here has, as it were, four heads. [...] The first is articulated by the word "opportunity" [...] the second is expressed by the word "choice" [...] the third theme is summed up by the word "strength" [and] my fourth theme is expressed well by the word "renewal".*

(Note however, these words occur 2, 7, 2, and 8 times in 4431 words)

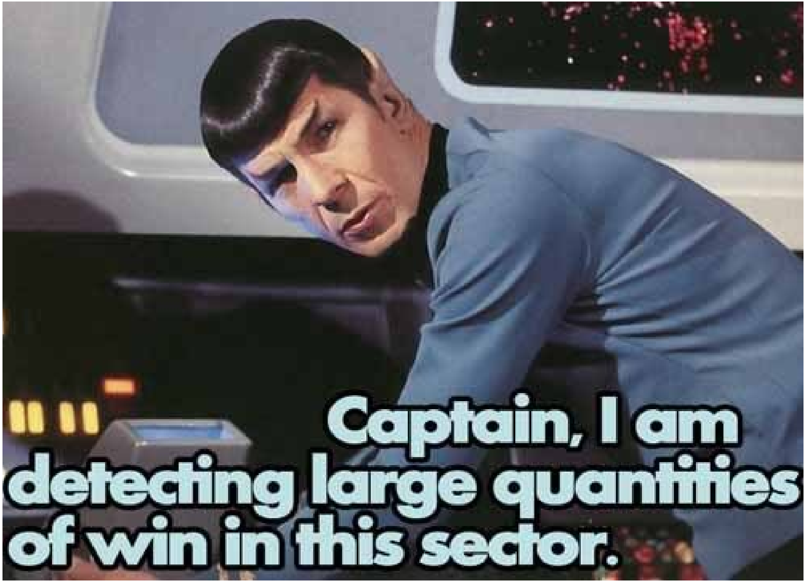
Communication: Reflexivity

Politicians are often nice enough to talk as if they really do communicate this way

> *My theme here has, as it were, four heads. […] The first is articulated by the word "opportunity" […] the second is expressed by the word "choice" […] the third theme is summed up by the word "strength" [and] my fourth theme is expressed well by the word "renewal".*

(Note however, these words occur 2, 7, 2, and 8 times in 4431 words)

> *A couple months ago we weren't expected to win this one, you know that, right? We weren't…Of course if you listen to the pundits, we weren't expected to win too much. And now we're winning, winning, winning the country – and soon the country is going to start winning, winning, winning.*


Captain, I am
detecting large quantities
of win in this sector.

Communication and comparability

Quantitative text analysis works best when language usage is stable, conventionalized, and instrumental.

Implicitly, that means *institutional language*, e.g.

- → courts
- → legislatures
- → op-eds
- → financial reporting

Institution-specificity inevitably creates a *comparability* problem, e.g.

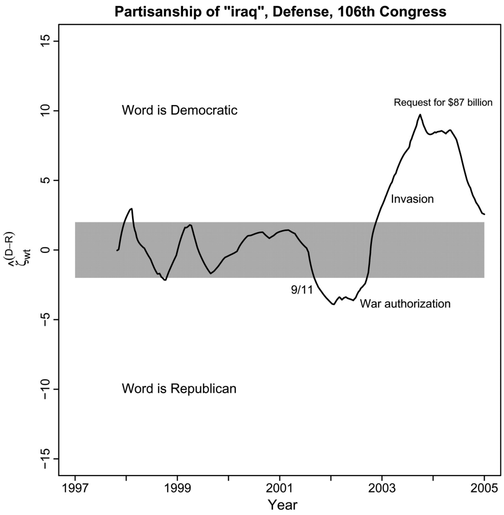
- → upper vs lower chamber vs parliamentary hearings
- → bureaucracy vs lobby groups (Klüver, 2009)
- → European languages (Proksch et al., 2019)

Instability

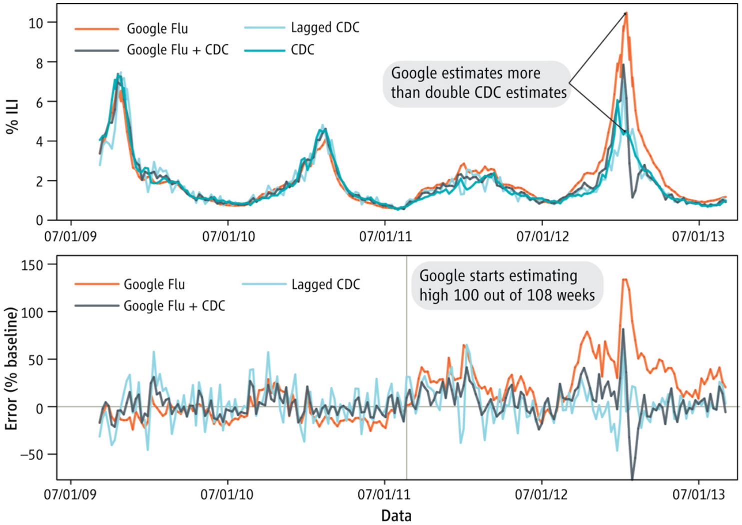
We are going to design instruments to measure θ and are going to assume that the $\theta \longrightarrow W$ relationships are institutionally stable

What if they aren't?

Rhetorical instability



Algorithmic instability



Reflexive solutions

Sometimes these actors are happy to solve comparability problems for us, e.g.

→ Lower court opinions (Corley et al., 2011) or Amicus briefs (Collins et al., 2015) *embedded in* Supreme Court opinions

→ ALEC model bills *embedded in* state bills (Garrett & Jansa, 2015)

A perfect jobs for *text-reuse* algorithms…

Communication: Randomness

Why randomness?

You almost never *say exactly the same words twice*, even when you haven't changed your mind about the message.

Hence words are the result of some kind of *sampling process*.

We model this process as random because we don't know or care about all the causes of variation (and because we're all secretly Bayesians)

Note: this is randomness *conditional on the institution*

Words as data

What do we know about words as data?

They are *difficult*

- → High dimensional
- → Sparsely distributed (with skew)
- → Not equally informative

Difficult words

Example: Conservative party 2017 manifesto compared to other parties in two elections:

High dimensional	3784 word types (adult native english speakers know 20-35,000)
Sparse	Of these, the Conservatives only used 23.5)
Skewed	Of these 10.8% words appear exactly once

Difficult words

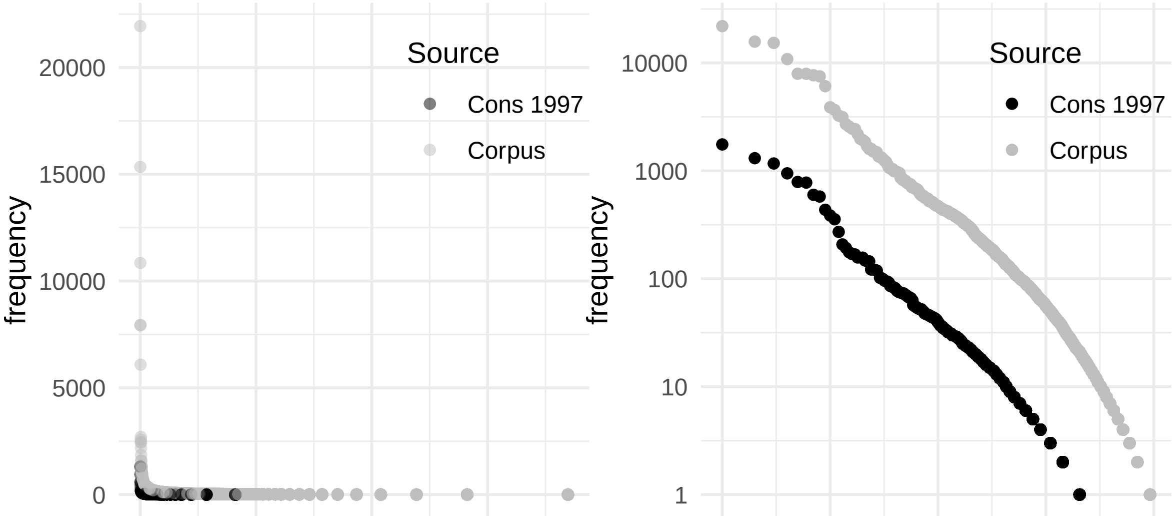
More generally: the Zipf-Mandelbrot law (Mandelbrot, 1966; Zipf, 1932)

$$F(W_i) \propto 1/\mathrm{rank}(W_i)^{\alpha}$$

where rank(.) is the frequency *rank* of a word in the vocabulary and $\alpha \approx 1$

(This is a Pareto distribution in disguise)

Difficult at all scales



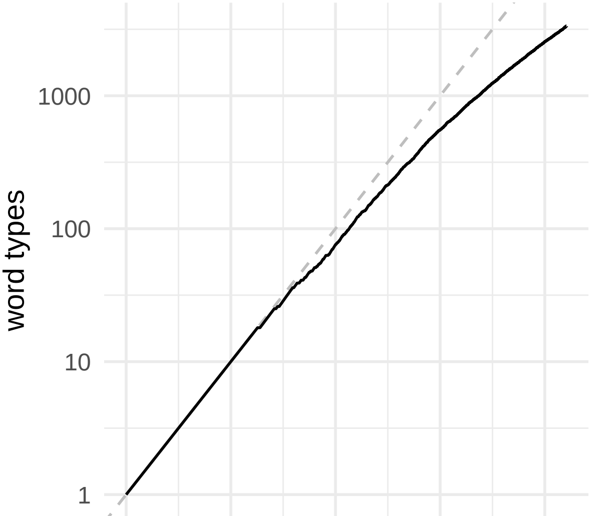
Types and Tokens

More generally: the Heaps-Herdan Law states that the number of word types appearing for the first time after n tokens is

$$D(n) = Kn^{\beta}$$

where K is between 10 and 100 and $\beta \approx 0.5$ for English.

Here's the Conservative party manifesto



Frequency and interestingness

Frequency is inversely proportional to substantive interestingness

	Word	Freq.
1	the	21939
2	and	15747
3	to	15347
4	of	10850
5	we	7943
6	will	7930

Top 10

	Word	Freq.
16078	1.83	1
16079	2.20	1
16080	1.35	1
16081	33.34	1
16082	1.71	1
16083	rigination	1

Bottom ten

	Word	Freq.
20	people	1929
26	new	1507
27	government	1493
33	support	1212
34	work	1143
36	uk	1058

Top ten minus *stopwords*

Dealing with difficult words

Removing stopwords, while standard in computer science, is not necessarily better…

Example:

- → Standard collections contain, ‘him’, ‘his’, ‘her’ and ‘she’.
- → Words you’d want to keep when analyzing an abortion debates.

Dealing with difficult words

For large amounts of text summaries are not enough.

We need a *model* to provide assumptions about

→ *equivalence*

→ *exchangeability*

Text as data started off making most use of equivalence, and ended up with increasingly sophisticated versions of exchangeability

Since ontogeny recapitulates phylogeny, let's walk through some standard text processing steps, asserting equivalences along the way…

Punctuation invariance

As I look ahead I am filled with foreboding. Like the Roman I seem to see 'the river Tiber flowing with much blood'…"
(E. Powell, 1968)

Punctuation invariance

As I look ahead I am filled with foreboding. Like the Roman I seem to see 'the river Tiber flowing with much blood'…"
(E. Powell, 1968)

index	token
1	as
2	i
3	look
4	ahead
5	i
6	am
7	…

index	token
1	like
2	the
3	roman
4	i
5	seem
6	to
7	…

Lexical univocality

type	count
as	1
i	2
look	1
ahead	1
am	1
…	…

token	count
like	1
the	1
roman	1
i	1
seem	1
to	1
…	…

Order invariance

		unit	
		'doc' 1	'doc' 2
type	ahead	1	0
	am	1	0
	as	1	0
	i	2	1
	like	0	1
	look	1	0
	roman	0	1
	seem	0	1
	the	0	1
	to	0	1
	…	…	…

Count data

We have turned a corpus into a *contingency table*.

→ (Or a term-document / document-term / document-feature matrix, in the lingo)

Count data

We have turned a corpus into a *contingency table*.

→ (Or a term-document / document-term / document-feature matrix, in the lingo)

Everything you learned in your categorical data analysis course applies

→ except that the variables of interest: θ are *not observed*

Count data

We have turned a corpus into a *contingency table*.

→ (Or a term-document / document-term / document-feature matrix, in the lingo)

	ahead	am	i	like	look		
doc 1	1	1	2	0	1	...	θ_{doc1}
doc 2	0	0	1	1	0	...	θ_{doc2}
	β_{ahead}	β_{am}	β_{i}	β_{like}	β_{look}		

Everything you learned in your categorical data analysis course applies

→ except that the variables of interest: θ are *not observed*

Statistical assumptions about words

Word counts/rates are conditionally *Poisson*:

$$W_j \sim \text{Poisson}(\lambda_j)$$

Curiously

$$\text{E}[W] = \text{Var}[W] = \lambda$$

Rate models are naturally *multiplicative*.

→ Rates increase / decrease by X%

Model assumptions are will turn on how λ is related to θ

Statistical assumptions about words

That means that for fixed document lengths, counts are conditionally *Multinomial*:

$$W_{i1} \ldots W_{iV} \sim \text{Mult}(W_{i1} \ldots W_{iV} \mid \pi_1 \ldots \pi_V, N_i)$$

Here

$$\text{E}[W] = N\pi$$

and

$$\text{Cov}[W_i, W_j] = -N\pi_i\pi_j$$

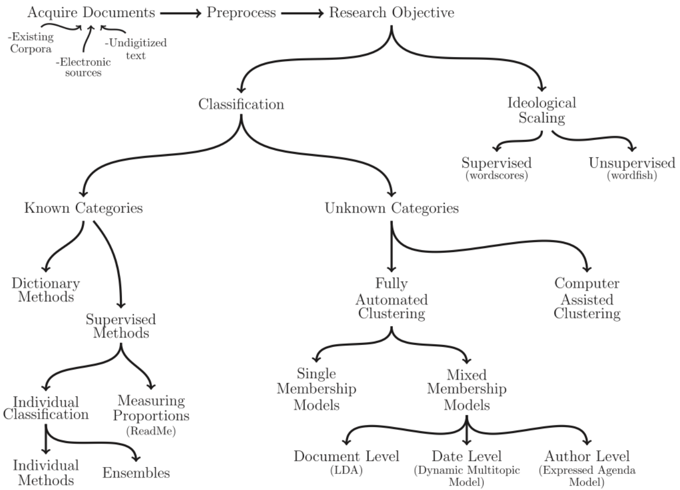
Negative covariance is due to the 'budget constraint'

Modelling decisions

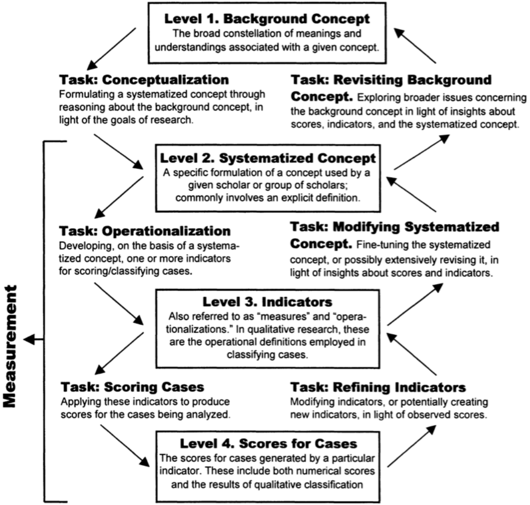
For each research problem involving content analysis we need to ask:

- → What counts as a *word*?
- → What counts as a *document*?
- → What *structure* does θ has
- → What is *observed*, what is *assumed*, and what is *inferred*?
- → What the *relationship* is between θ and the words? The model

The space of models



In the general measurement problem



Exchangeability and the 'bag of words'

The marginal challenge

The original:

> *Reasonable relief of the working middle class and debt reduction do not exclude each other but are complementary. That proved to be true during the past four years. We oppose higher taxes on citizens and businesses. They prevent growth and kill jobs, thus putting at risk the very existence of countless workers and their families.*

Given the *marginal distribution of word types* a.k.a. the ‘bag of words’ how different could the meaning of the resulting document be?

Slightly different...

A possible reconstruction

> *We oppose higher taxes because they prevent growth and kill jobs, risking the very existence of countless families and businesses. Debt reduction is complementary to reasonable relief of the working middle class; they do not exclude each other. That proved to be true in the last four years.*

But much the same sense

Looser constraints

We can make this a more semantic challenge by only demanding the substantically interesting word margins are maintained.

> *businesses reasonable existence countless reduction families prevent risking working exclude oppose higher growth relief middle proved taxes class years kill jobs debt true last four complementary*

Removing stopwords mostly just removes *grammatical* constraints

But negation!

If we can add grammatical functors at will, could we make the opposite meaning by negating everything?

In principle (and in practice for some discursive forms) yes, yes we could.

And the bag of words assumption would fail

But negation!

If we can add grammatical functors at will, could we make the opposite meaning by negating everything?

In principle (and in practice for some discursive forms) yes, yes we could.

And the bag of words assumption would fail

An interesting *empirical fact* about political discourse is that actors do not tend to disagree by negation but by redirection or diversion.

- → Simple version: You talk about the environment, I talk about economic growth
- → Sophisticated version: The ‘heresthetic’ (Riker et al., 1996)

References

Austin, J. L. (1962). 'How to do things with words'. Clarendon Press.

Chater, N. & Brown, G. D. A. (1999). 'Scale-invariance as a unifying psychological principle.' *Cognition*, *69*(3), B17–24.

Collins, P. M., Corley, P. C. & Hamner, J. (2015). 'The influence of amicus curiae briefs on u.s. supreme court opinion content: The influence of amicus curiae'. *Law & Society Review*, *49*(4), 917–944.

Corley, P. C., Collins, P. M. & Calvin, B. (2011). 'Lower court influence on u.s. supreme court opinion content'. *The Journal of Politics*, *73*(1), 31–44.

Davidson, D. (1985). 'Inquiries into truth and interpretation'. Clarendon Press.

Garrett, K. N. & Jansa, J. M. (2015). 'Interest group influence in policy diffusion networks'. *State Politics & Policy Quarterly*, *15*(3), 387–417.

Grice, P. (1993). 'Studies in the way of words' (3. print). Harvard Univ. Press.

References

Klüver, H. (2009). 'Measuring interest group influence using quantitative text analysis'. *European Union Politics*, *10*(4), 535–549.

Krebs, J. & Dawkins, R. (1984). Animal signals: Mind-reading and manipulation. In J. Krebs & N. B. Davies (Eds.), *Behavioural ecology: An evolutionary approach* (2nd ed., pp. 380–402). Blackwell Science.

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014). 'The parable of google flu: Traps in big data analysis'. *Science*, *343*(6176), 1203–1205.

Lewis, D. K. (2011). 'Convention: A philosophical study' (Nachdr.). Blackwell.

Mandelbrot, B. (1966). Information theory and psycholinguistics: A theory of word frequencies. In P. Lazarsfeld & N. Henry (Eds.), *Readings in mathematical social science*. MIT Press.

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

Proksch, S.-O., Lowe, W., Wäckerle, J. & Soroka, S. (2019). 'Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches'. *Legislative Studies Quarterly*, *44*(1), 97–131.

Quine, W. v. O. (1960). 'Word and object'. MIT Press.

Riker, W. H., Calvert, R. L., Mueller, J. E. & Wilson, R. K. (1996). 'The strategy of rhetoric: Campaigning for the american constitution'. Yale University Press.

Searle, J. R. (1995). 'The construction of social reality'. Free Press.

Wittgenstein, L. (1958). 'Philosophical investigations' (G. E. M. Anscombe, Trans.). Blackwell.

Zipf, G. K. (1932). 'Selected studies of the principle of relative frequency in language'. Oxford University Press.