Week 5: Quantitative Text Analysis

Kenneth Benoit

TCD HT 2017

20 March 2017

Week 5 Outline

- Key features of QTA
  - Quantitative text analysis workflow
  - Key basic concepts
- Documents and features
  - Strategies for selecting documents
  - Defining features
  - Parts of speech
  - Filtering features
  - "stopwords"
- Descriptive text analysis
  - Key words in context
- Dictionary analysis
- The Wordfish Scaling Model

Key features of QTA

Basic QTA Process: Texts → Feature matrix → Analysis

When I presented the supplementary budget to this House last April, I said we could work our way through this period of severe economic distress. Today, I can report that notwithstanding the difficulties of the past eight months, we are now on the road to economic recovery.

In this next phase of the Government's plan we must stabilise the deficit in a fair way, safeguard those worst hit by the recession, and stimulate crucial sectors of our economy to sustain and create jobs. The worst is over.

This Government has the moral authority and the well-grounded optimism rather than the cynicism of the Opposition. It has the imagination to create the new jobs in energy, agriculture, transport and construction that this green budget will

docs	words made	because	had	into	get	some	through	next	where	many	irish
t06_kenny_fg	12	11	5	4	8	4	3	4	5	7	10
t05_cowen_ff	9	4	8	5	5	5	14	13	4	9	8
t14_ocaolain_sf	3	3	3	4	7	3	7	2	3	5	6
t01_lenihan_ff	12	1	5	4	2	11	9	16	14	6	9
t11_gormley_green	0	0	0	3	0	2	0	3	1	1	2
t04_morgan_sf	11	8	7	15	8	19	6	5	3	6	6
t12_ryan_green	2	2	3	7	0	3	0	1	6	0	0
t10_quinn_lab	1	4	4	2	8	4	1	0	1	2	0
t07_odonnell_fg	5	4	2	1	5	0	1	1	0	3	0
t09_higgins_lab	2	2	5	4	0	1	0	0	2	0	0
t03_burton_lab	4	8	12	10	5	5	4	5	8	15	8
t13_cuffe_green	1	2	0	0	11	0	16	3	0	3	1
t08_gilmore_lab	4	8	7	4	3	6	4	5	1	2	11
t02_bruton_fg	1	10	6	4	4	3	0	6	16	5	3

Descriptive statistics on words

Scaling documents

Classifying documents

Extraction of topics

Vocabulary analysis

Sentiment analysis

Key feature of quantitative text analysis

1. Selecting texts: Defining the *corpus*
2. Conversion of texts into a common electronic format
3. Defining documents: deciding what will be the doumentary unit of analysis

Key feature of quantitative text analysis (cont.)

4. Defining features. These can take a variety of forms, including tokens, equivalence classes of tokens (dictionaries), selected phrases, human-coded segments (of possibly variable length), linguistic features, and more.
5. Conversion of textual features into a quantitative matrix
6. A quantitative or statistical procedure to extract information from the quantitative matrix
7. Summary and interpretation of the quantitative results

```
When I presented the
supplementary budget to
this House last April, I
said we could work our
way through this period
of severe economic
distress. Today, I can
report that
notwithstanding the
difficulties of the past
eight months, we are now
on the road to economic
recovery.
```

```
In this next phase of the
Government's plan we must
stabilise the deficit in
a fair way, safeguard
those worst hit by the
recession, and stimulate
crucial sectors of our
economy to sustain and
create jobs. The worst is
over.
```

```
This Government has the
moral authority and the
well-grounded optimism
rather than the cynicism
of the Opposition. It has
the imagination to create
the new jobs in energy,
agriculture, transport
and construction that
this green budget will
```

docs	words made	because	had	into	get	some	through	next	where	many	irish
t06_kenny_fg	12	11	5	4	8	4	3	4	5	7	10
t05_cowen_ff	9	4	8	5	5	5	14	13	4	9	8
t14_ocaolain_sf	3	3	3	4	7	3	7	2	3	5	6
t01_lenihan_ff	12	1	5	4	2	11	9	16	14	6	9
t11_gormley_green	0	0	0	3	0	2	0	3	1	1	2
t04_morgan_sf	11	8	7	15	8	19	6	5	3	6	6
t12_ryan_green	2	2	3	7	0	3	0	1	6	0	0
t10_quinn_lab	1	4	4	2	8	4	1	0	1	2	0
t07_odonnell_fg	5	4	2	1	5	0	1	1	0	3	0
t09_higgins_lab	2	2	5	4	0	1	0	0	2	0	0
t03_burton_lab	4	8	12	10	5	5	4	5	8	15	8
t13_cuffe_green	1	2	0	0	11	0	16	3	0	3	1
t08_gilmore_lab	4	8	7	4	3	6	4	5	1	2	11
t02_bruton_fg	1	10	6	4	4	3	0	6	16	5	3

Descriptive statistics on words

Scaling documents

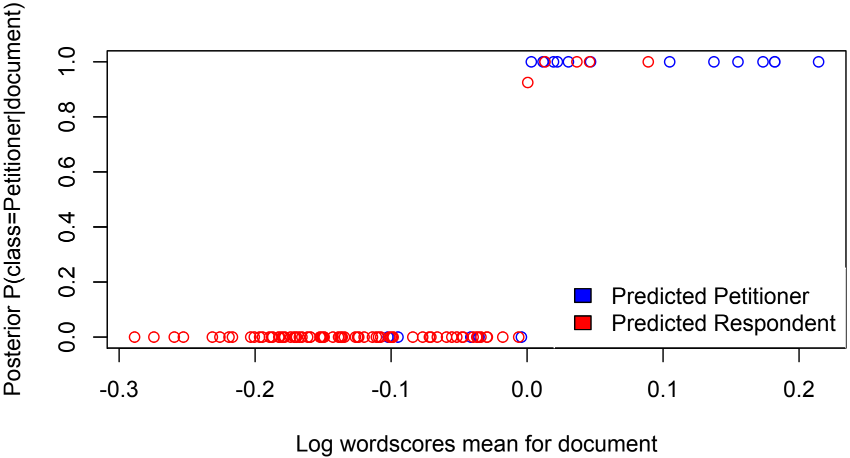
Classifying documents

Extraction of topics

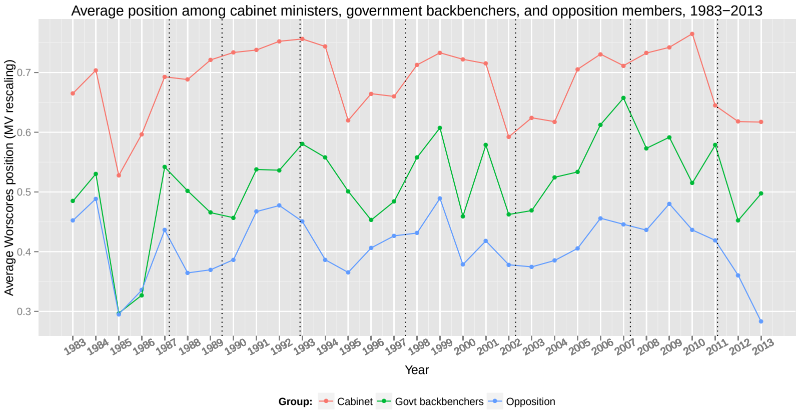
Vocabulary analysis

Sentiment analysis

Example: Document classification using the "Naive Bayes" classifier

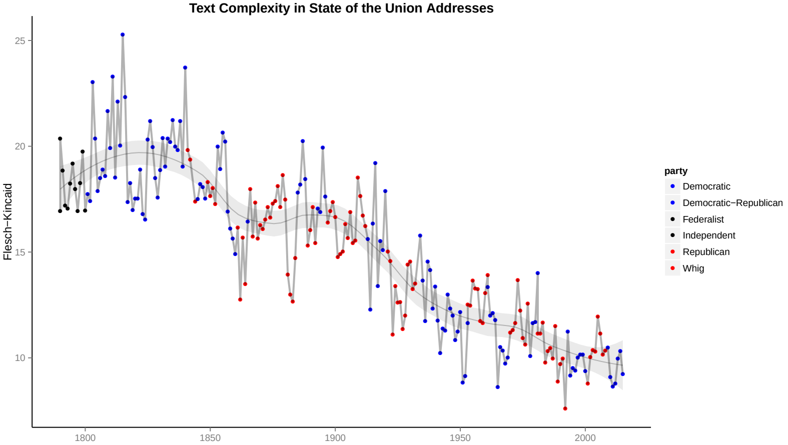


Government v. Opposition in yearly budget debates



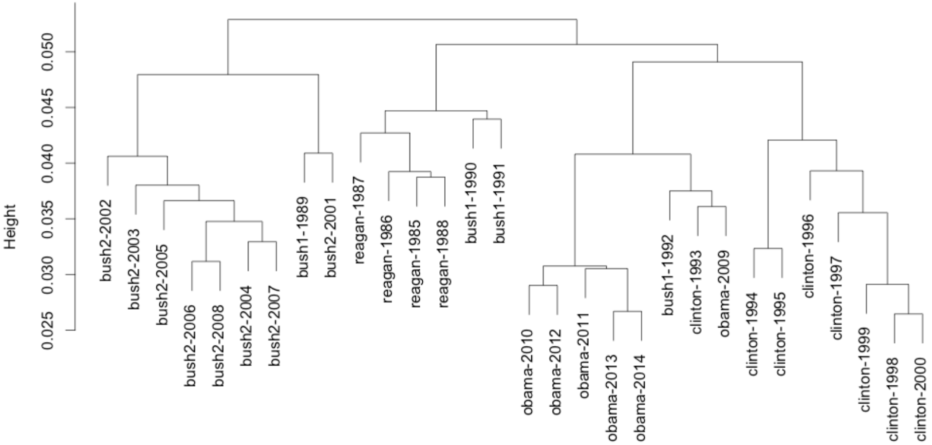
(from Herzog and Benoit EPSA 2013)

Reading level of US State-of-the-Union addresses over time

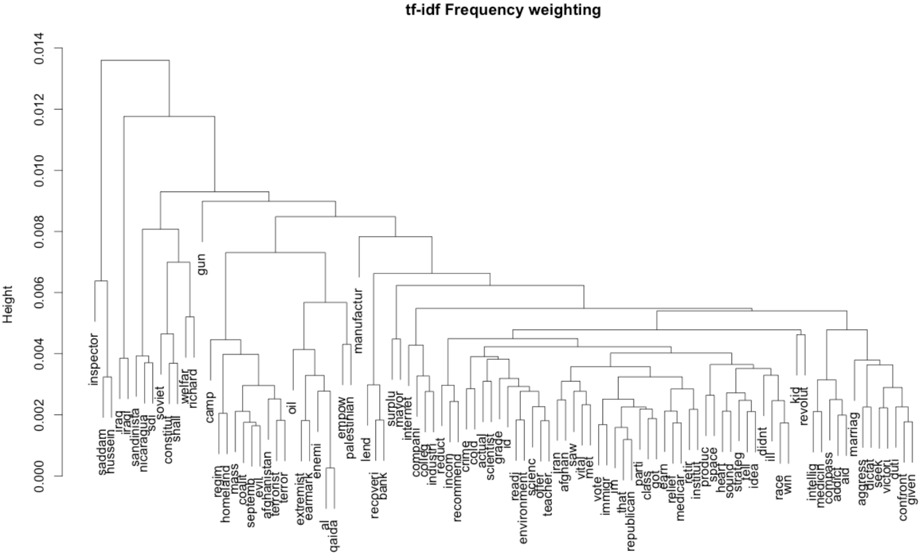


Wordcloud of Tweets from 2014 EP campaign, by list-leading candidate

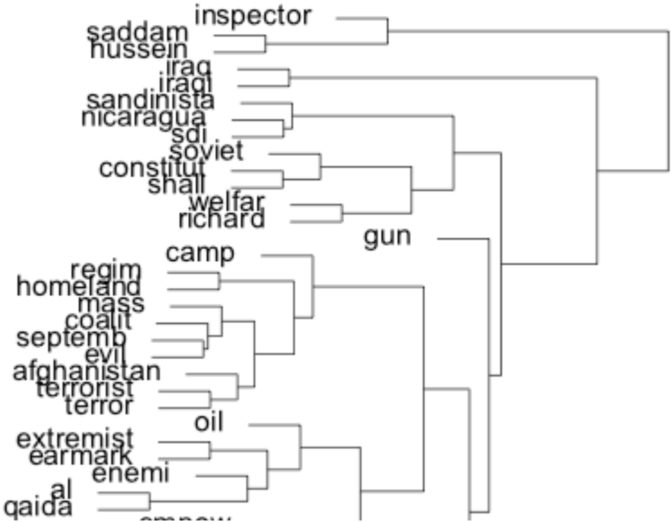
Hierachical clustering: Presidential State of the Union addresses



Dendrogram: Presidential State of the Union addresses



Dendrogram: Presidential State of the Union addresses



Assumptions

- That texts represent an observable implication of some underlying characteristic of interest (usually an attribute of the author)
- That texts can be represented through extracting their *features*
  - most common is the bag of words assumption
  - many other possible definitions of "features"
- A document-feature matrix can be analyzed using quantitative methods to produce meaningful and valid estimates of the underlying characteristic of interest

Some key basic concepts

(text) corpus
: a large and structured set of texts for analysis

types
: for our purposes, a unique word

tokens
: any word – so token count is total words

stems
: words with suffixes removed

lemmas
: canonical word form (the base form of a word that has the same meaning even when different suffixes (or prefixes) are attached)

keys
: such as dictionary entries, where the user defines a set of equivalence classes that group different word types

parts of speech
: grammatical types such as nouns, verbs, etc.

Some more key basic concepts

"key" words Words selected because of special attributes, meanings, or rates of occurrence

stop words Words that are designated for exclusion from any analysis of a text

readability provides estimates of the readability of a text based on word length, syllable length, etc.

complexity A word is considered "complex" if it contains three syllables or more

diversity (lexical diversity) A measure of how many types occur per fixed word rate (a normalized vocabulary measure)

Documents and Features

Strategies for selecting units of textual analysis

- Words
- *n*-word sequences
- pages
- paragraphs
- Themes
- Natural units (a speech, a poem, a manifesto)
- Key: depends on the research design

Defining Features

- words
- word stems or lemmas: this is a form of defining *equivalence classes* for word features
- word segments, especially for languages using compound words, such as German, e.g.
  Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz
  (the law concerning the delegation of duties for the supervision of cattle marking and the labelling of beef)
  Saunauntensitzer

Defining Features (cont.)

- "word" sequences, especially when inter-word delimiters (usually white space) are not commonly used, as in Chinese

  莎拉波娃现在居住在美国东南部的佛罗里达。今年４月９日，莎拉波娃在美国第一大城市纽约度过了１８岁生日。生日派对上，莎拉波娃露出了甜美的微笑。

- linguistic features, such as parts of speech
- (if qualitative coding is used) coded or annotated text segments
- linguistic features: parts of speech

Parts of speech

- the Penn "Treebank" is the standard scheme for tagging POS

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Example using **spacyr**

```
require(spacyr)

## Loading required package:  spacyr

spacy_initialize()
spacy_parse("Pierre Vinken, 61 years old, will join the board as a nonexecutive
Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.")

## Error in if (type == "str" || type == "unicode")
return(char_to_R(var)):  missing value where TRUE/FALSE needed
```

Strategies for feature selection

- document frequency How many documents in which a term appears
- term frequency How many times does the term appear in the corpus
- deliberate disregard Use of "stop words": words excluded because they represent linguistic connectors of no substantive content
- purposive selection Use of a *dictionary* of words or phrases
- declared equivalency classes Non-exclusive synonyms, what I call a *thesaurus* (lots more on these on Day 4)

Common English stop words

a, able, about, across, after, all, almost, also, am, among, an, and, any, are, as, at, be, because, been, but, by, can, cannot, could, dear, did, do, does, either, else, ever, every, for, from, get, got, had, has, have, he, her, hers, him, his, how, however, I, if, in, into, is, it, its, just, least, let, like, likely, may, me, might, most, must, my, neither, no, nor, not, of, off, often, on, only, or, other, our, own, rather, said, say, says, she, should, since, so, some, than, that, the, their, them, then, there, these, they, this, tis, to, too, twas, us, wants, was, we, were, what, when, where, which, while, who, whom, why, will, with, would, yet, you, your

- But no list should be considered universal

A more comprehensive list of stop words

as, able, about, above, according, accordingly, across, actually, after, afterwards, again, against, aint, all, allow, allows, almost, alone, along, already, also, although, always, am, among, amongst, an, and, another, any, anybody, anyhow, anyone, anything, anyway, anyways, anywhere, apart, appear, appreciate, appropriate, are, arent, around, as, aside, ask, asking, associated, at, available, away, awfully, be, became, because, become, becomes, becoming, been, before, beforehand, behind, being, believe, below, beside, besides, best, better, between, beyond, both, brief, but, by, cmon, cs, came, can, cant, cannot, cant, cause, causes, certain, certainly, changes, clearly, co, com, come, comes, concerning, consequently, consider, considering, contain, containing, contains, corresponding, could, couldnt, course, currently, definitely, described, despite, did, didnt, different, do, does, doesnt, doing, dont, done, down, downwards, during, each, edu, eg, eight, either, else, elsewhere, enough, entirely, especially, et, etc, even, ever, every, everybody, everyone, everything, everywhere, ex, exactly, example, except, far, few, fifth, first, five, followed, following, follows, for, former, formerly, forth, four, from, further, furthermore, get, gets, getting, given, gives, go, goes, going, gone, got, gotten, greetings, had, hadnt, happens, hardly, has, hasnt, have, havent, having, he, hes, hello, help, hence, her, here, heres, hereafter, hereby, herein, hereupon, hers, herself, hi, him, himself, his, hither, hopefully, how, howbeit, however, id, ill, im, ive, ie, if, ignored, immediate, in, inasmuch, inc, indeed, indicate, indicated, indicates, inner, insofar, instead, into, inward, is, isnt, it, itd, itll, its, its, itself, just, keep, keeps, kept, know, knows, known, last, lately, later, latter, latterly, least, less, lest, let, lets, like, liked, likely, little, look, looking, looks, ltd, mainly, many, may, maybe, me, mean, meanwhile, merely, might, more, moreover, most, mostly, much, must, my, myself, name, namely, nd, near, nearly, necessary, need, needs, neither, never, nevertheless, new, next, nine, no, nobody, non, none, noone, nor, normally, not, nothing, novel, now, nowhere, obviously, of, off, often, oh, ok, okay, old, on, once, one, ones, only, onto, or, other, others, otherwise, ought, our, ours, ourselves, out, outside, over, overall, own, particular, particularly, per, perhaps, placed, please, plus, possible, presumably, probably, provides, que, quite, qv, rather, rd, re, really, reasonably, regarding, regardless, regards, relatively, respectively, right, said, same, saw, say, saying, says, second, secondly, see, seeing, seem, seemed, seeming, seems, seen, self, selves, sensible, sent, serious, seriously, seven, several, shall, she, should, shouldnt, since, six, so, some, somebody,

Stemming words

Lemmatization refers to the algorithmic process of converting words to their lemma forms.

stemming the process for reducing inflected (or sometimes derived) words to their stem, base or root form. Different from *lemmatization* in that stemmers operate on single words without knowledge of the context.

both convert the morphological variants into stem or root terms

example: `produc` from `production`, `producer`, `produce`, `produces`, `produced`

Descriptive text analysis

Exploring Texts: Key Words in Context

KWIC *Key words in context* Refers to the most common format for concordance lines. A KWIC index is formed by sorting and aligning the words within an article title to allow each word (except the stop words) in titles to be searchable alphabetically in the index.

```
kwic(data_corpus_inaugural, "nuclear* *", window = 3)
```

```
##
## [1973-Nixon, 428:429]       the limitation of |    nuclear arms     |
## [1977-Carter, 1103:1104]   elimination of all |   nuclear weapons   |
## [1985-Reagan, 2208:2209]  further increase of |   nuclear weapons   |
## [1985-Reagan, 2229:2230]           one day of |   nuclear weapons   |
## [1985-Reagan, 2264:2265]           the use of |   nuclear weapons   |
## [1985-Reagan, 2334:2335]   that would destroy |  nuclear missiles   |
## [1985-Reagan, 2369:2370]      It would render |   nuclear weapons   |
## [1985-Reagan, 2396:2397]        the threat of | nuclear destruction |
## [1997-Clinton, 1668:1669]       the threat of |      nuclear ,      |
## [2009-Obama, 1604:1605]       to lessen the |   nuclear threat    |
##
## [1973-Nixon, 428:429]     , and to
## [1977-Carter, 1103:1104]  from this Earth
```

Finding "key" differential words

- *"keyness"* can also refer to the extent to which specific words occur at differential rates across classes or categories of a variable
- Common methods for forming this association are χ^2 and G^2 (likelihood ratio) statistics
- Often a useful starting point for finding words for forming a *dictionary*

Keyness example: χ^2

```r
period <- ifelse(docvars(data_corpus_inaugural, "Year") < 1945,
                 "pre-war", "post-war")
# compare Trump 2017 to other post-war presidents
pwdfm <- dfm(corpus_subset(data_corpus_inaugural, period == "post-war"))
head(textstat_keyness(pwdfm, target = "2017-Trump"), 10)
```

```
##                 chi2            p
## protected   76.79339 0.000000e+00
## will        49.67662 1.812883e-12
## while       48.33243 3.597567e-12
## obama       47.94909 4.374279e-12
## we've       47.94909 4.374279e-12
## america     29.11681 6.814327e-08
## again       27.88512 1.287361e-07
## everyone    27.73848 1.388726e-07
## your        26.75528 2.309197e-07
## transferring 25.59527 4.210694e-07
```

Keyness example: [2]

```
# using the likelihood ratio method
head(textstat_keyness(dfm_smooth(pwdfm), measure = "lr", target = "2017-Trump")

##                   G            p
## will      24.653370 6.862464e-07
## america   14.064544 1.766425e-04
## your      11.662713 6.376529e-04
## while     11.166438 8.329039e-04
## again     11.083498 8.709934e-04
## protected 10.591417 1.136138e-03
## american   9.993537 1.570906e-03
## back       8.209112 4.168054e-03
## dreams     7.075173 7.815929e-03
## country    6.800841 9.111495e-03
```

Dictionary analysis

Rationale for dictionaries

- Rather than count words that occur, pre-define words associated with specific meanings
- Two components:

  key the label for the equivalence class for the concept or canonical term

  values (multiple) terms or patterns that are declared equivalent occurences of the key class
- Frequently involves lemmatization: transformation of all inflected word forms to their "dictionary look-up form" — more powerful than stemming

“Dictionary”: a misnomer?

- A *dictionary* is really a thesaurus: a canonical term or concept (a “key”) associated with a list of equivalent synonyms
- But dictionaries tend to be exclusive: they single out features defined as keys, selecting the terms or patterns linked to each key
- An alternative is a “thesaurus” concept: a tag of key equivalency for an associated set of terms, but non-exclusive
  - WC = `wc, toilet, restroom, bathroom, jack, loo`
  - vote = `poll, suffrage, franchis*, ballot*, ^vot$`

Bridging qualitative and quantitative text analysis

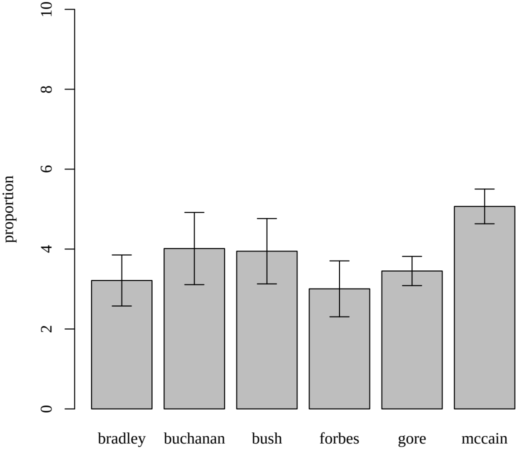
- A hybrid procedure between qualitative and quantitative classification the fully automated end of the text analysis spectrum
- "Qualitiative" since it involves identification of the concepts and associated keys/categories, and the textual features associated with each key/category
- Dictionary construction involves a lot of contextual interpretation and qualitative judgment
- Perfect reliability because there is no human decision making as part of the text analysis procedure

Well-known dictionaries: General Inquirer

- General Inquirer (Stone et al 1966)
- Example: self = *I*, *me*, *my*, *mine*, *myself*
  selves = *we*, *us*, *our*, *ours*, *ourselves*
- Latest version contains 182 categories – the "Harvard IV-4" dictionary, the "Lasswell" dictionary, and five categories based on the social cognition work of Semin and Fiedler
- Examples: "self references", containing mostly pronouns; "negatives", the largest category with 2291 entries
- Also uses disambiguation, for example to distinguishes between *race* as a contest, *race* as moving rapidly, *race* as a group of people of common descent, and *race* in the idiom "rat race"
- Output example:
  `http://www.wjh.harvard.edu/~inquirer/Spreadsheet.html`

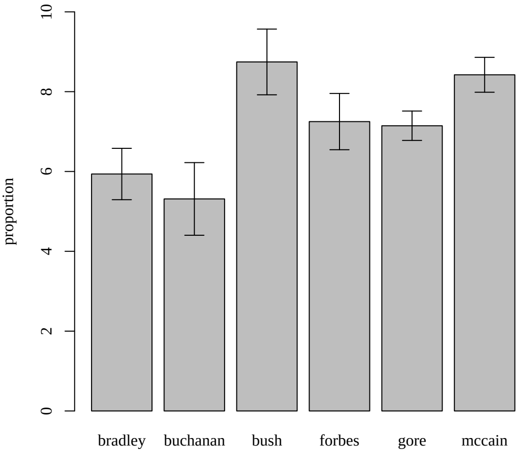
General Inquirer Applied to US Presidential Candidate Speeches (2000)

Negative language



General Inquirer Applied to US Presidential Candidate Speeches (2000)

Positive language



Well-known dictionaries: Regressive Imagery Dictionary

- Consists of about 3,200 words and roots, assigned to 29 categories of primary process cognition, 7 categories of secondary process cognition, and 7 categories of emotions
- designed to measure primordial vs. conceptual thinking
  - Conceptual thought is abstract, logical, reality oriented, and aimed at problem solving
  - Primordial thought is associative, concrete, and takes little account of reality – the type of thinking found in fantasy, reverie, and dreams
- Categories were derived from the theoretical and empirical literature on regressive thought by Martindale (1975, 1990)

Regressive Imagery Dictionary categories

- Full listing of categories

1 orality	21 brink-passage	41 aggression	62 novelty
2 anality	22 narcissism	42 expressive behaviour	63 negation
3 sex	23 concreteness	43 glory	64 triviality
4 touch	24 ascend	44 female role	65 transmute
5 taste	25 height	45 male fole	
6 odour	26 descent	46 self	
7 general sensation	27 depth	47 related others	
8 sound	28 fire	48 diabolic	
9 vision	29 water	49 aspiration	
10 cold	30 abstract thought	50 angelic	
11 hard	31 social behaviour	51 flowers	
12 soft	32 instrumental behaviour	52 synthesize	
13 passivity	33 restraint	53 streight	
14 voyage	34 order	54 weakness	
15 random movement	35 temporal references	55 good	
16 diffusion	36 moral imperative	56 bad	
17 chaos	37 positive affect	57 activity	
18 unknown	38 anxiety	58 being	
19 timelessness	39 sadness	59 analogy	
20 counscious	40 affection	61 integrative con	

- More on categories:
`http://www.kovcomp.co.uk/wordstat/RID.html`

Linquistic Inquiry and Word Count

- Created by Pennebaker et al — see `http://www.liwc.net`
- uses a dictionary to calculate the percentage of words in the text that match each of up to 82 language dimensions
- Consists of about 4,500 words and word stems, each defining one or more word categories or subdictionaries
- For example, the word *cried* is part of five word categories: sadness, negative emotion, overall affect, verb, and past tense verb. So observing the token *cried* causes each of these five subdictionary scale scores to be incremented
- Hierarchical: so "anger" are part of an *emotion* category and a *negative emotion* subcategory
- You can buy it here: `http://www.liwc.net/descriptiontable1.php`

Example: Terrorist speech

	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.

Example: Laver and Garry (2000)

- A *hierarchical* set of categories to distinguish policy domains and policy positions – similar in spirit to the CMP
- Five domains at the top level of hierarchy
  - economy
  - political system
  - social system
  - external relations
  - a " 'general' domain that has to do with the cut and thurst of specific party competition as well as uncodable pap and waffle"
- Looked for word occurences within "word strings with an average length of ten words"
- Built the dictionary on a set of specific UK manifestos

Example: Laver and Garry (2000): Economy

TABLE 1 Abridged Section of Revised Manifesto Coding Scheme

1 ECONOMY
Role of state in economy

- 1 1 ECONOMY/+State+
  Increase role of state
  - 1 1 1 ECONOMY/+State+/Budget
    Budget
    - 1 1 1 1 ECONOMY/+State+/Budget/Spending
      Increase public spending
      - 1 1 1 1 1 ECONOMY/+State+/Budget/Spending/Health
      - 1 1 1 1 2 ECONOMY/+State+/Budget/Spending/Educ. and training
      - 1 1 1 1 3 ECONOMY/+State+/Budget/Spending/Housing
      - 1 1 1 1 4 ECONOMY/+State+/Budget/Spending/Transport
      - 1 1 1 1 5 ECONOMY/+State+/Budget/Spending/Infrastructure
      - 1 1 1 1 6 ECONOMY/+State+/Budget/Spending/Welfare
      - 1 1 1 1 7 ECONOMY/+State+/Budget/Spending/Police
      - 1 1 1 1 8 ECONOMY/+State+/Budget/Spending/Defense
      - 1 1 1 1 9 ECONOMY/+State+/Budget/Spending/Culture
    - 1 1 1 2 ECONOMY/+State+/Budget/Taxes
      Increase taxes
      - 1 1 1 2 1 ECONOMY/+State+/Budget/Taxes/Income
      - 1 1 1 2 2 ECONOMY/+State+/Budget/Taxes/Payroll
      - 1 1 1 2 3 ECONOMY/+State+/Budget/Taxes/Company
      - 1 1 1 2 4 ECONOMY/+State+/Budget/Taxes/Sales
      - 1 1 1 2 5 ECONOMY/+State+/Budget/Taxes/Capital
      - 1 1 1 2 6 ECONOMY/+State+/Budget/Taxes/Capital gains
    - 1 1 1 3 ECONOMY/+State+/Budget/Deficit
      Increase budget deficit
      - 1 1 1 3 1 ECONOMY/+State+/Budget/Deficit/Borrow
      - 1 1 1 3 2 ECONOMY/+State+/Budget/Deficit/Inflation

Example: Laver and Garry (2000)

```
ECONOMY / +STATE
    accommodation
    age
    ambulance
    assist
    ...

ECONOMY / -STATE
    choice*
    compet*
    constrain*
    ...
```

Advantage: Multi-lingual

APPENDIX B
DICTIONARY OF THE COMPUTER-BASED CONTENT ANALYSIS

	NL	UK	GE	IT
Core	elit*	elit*	elit*	elit*
	consensus*	consensus*	konsens*	consens*
	ondemocratisch* ondemokratisch*	undemocratic*	undemokratisch*	antidemocratic*
	referend*	referend*	referend*	referend*
	corrupt*	corrupt*	korrupt*	corrot*
	propagand*	propagand*	propagand*	propagand*
	politici*	politici*	politiker*	politici*
	bedrog *bedrieg*	*deceit* *deceiv*	täusch* betrüg* betrug*	ingann*
	verraa *verrad*	*betray*	*verrat*	tradi*
	schaam*	shame*	scham* schäm*	vergogn*
	schand*	scandal*	skandal*	scandal*
	waarheid*	truth*	wahrheit*	verità
	oneerlijk*	dishonest*	unfair* unehrlich*	disonest*
Context	establishm*	establishm*	establishm*	partitocrazia
	heersend*	ruling*	*herrsch*	
	capitul* kapitul*			
	kaste*			
	leugen* lieg*		lüge*	menzogn* mentir*

(from Rooduijn and Pauwels 2011)

Disdvantage: Highly specific to context

- Example: Loughran and McDonald used the Harvard-IV-4 TagNeg (H4N) file to classify sentiment for a corpus of 50,115 firm-year 10-K filings from 1994–2008
- found that almost three-fourths of the "negative" words of H4N were typically not negative in a financial context e.g. *mine* or *cancer*, or *tax*, *cost*, *capital*, *board*, *liability*, *foreign*, and *vice*
- Problem: polysemes – words that have multiple meanings
- Another problem: dictionary lacked important negative financial words, such as *felony*, *litigation*, *restated*, *misstatement*, and *unanticipated*

Different dictionary formats

- General Inquirer: see `http://www.wjh.harvard.edu/~inquirer/inqdict.txt`
- WordStat: see `http://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/`
- LIWC: for an example see the Moral Foundations dictionary at `http://www.moralfoundations.org/othermaterials`
- quanteda (see demo code)

A quick introduction to regular expressions

- an expanded version of the "glob" matching implemented in most command line interpreters, i.e.
  - `*` matches zero or more characters
  - `?` matches any one character (and in some environments, zero trailing characters)
  - `[]` may match any characters within a range inside the brackets
- a much more powerful version are regular expressions, which also exist in several (slightly) different versions
- R has both the POSIX 1003.2 and the Perl Compatible Regular Expressions implemented, see `?regex`
- Additional materials:
  - great cheat sheet
  - useful tutorial and reference

Wordfish

The Poisson distribution

$$
\begin{aligned}
f_{Poisson}(y_i|\lambda) &= \begin{cases} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} & \forall\ \lambda > 0 \text{ and } y_i = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} \\
Pr(Y|\lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \\
\lambda &= e^{X_i\beta} \\
\mathrm{E}(y_i) &= \lambda \\
\mathrm{Var}(y_i) &= \lambda
\end{aligned}
$$

The Poisson scaling "wordfish" model

Data:

- Y is N (speaker) $\times$ V (word) term document matrix
  $V \gg N$

Model:

$$
\begin{aligned}
P(Y_i \mid \theta) &= \prod_{j=1}^{V} P(Y_{ij} \mid \theta_i) \\
Y_{ij} &\sim \text{Poisson}(\lambda_{ij}) \qquad (1) \\
\log \lambda_{ij} &= \alpha_i + \theta_i \beta_j + \psi_j
\end{aligned}
$$

Estimation:

- Easy to fit for large V (V Poisson regressions with α offsets)

Model components and notation

$$\log \lambda_{ij} = \alpha_i + \theta_i \beta_j + \psi_j$$

Element	*Meaning*
i	indexes documents
j	indexes word types
θ_i	the unobservable "position" of document i
β_j	word parameters on θ – the relationship of word j to document position
ψ_j	word "fixed effect" (function of the frequency of word j)
α_i	document "fixed effects" (a function of (log) document length to allow estimation in Poisson of an essentially multinomial process)

"Features" of the parametric scaling approach

- Standard (statistical) inference about parameters
- Uncertainty accounting for parameters
- Distributional assumptions are made explicit (as part of the data generating process motivating the choice of stochastic distribution)
  - *conditional independence*
  - *stochastic process* (e.g. $\mathrm{E}(Y_{ij}) = \mathrm{Var}(Y_{ij}) = \lambda_{ij}$)
- Permits hierarchical reparameterization (to add covariates)
- Generative model: given the estimated parameters, we could generate a document for any specified length

Some reasons why this model is wrong

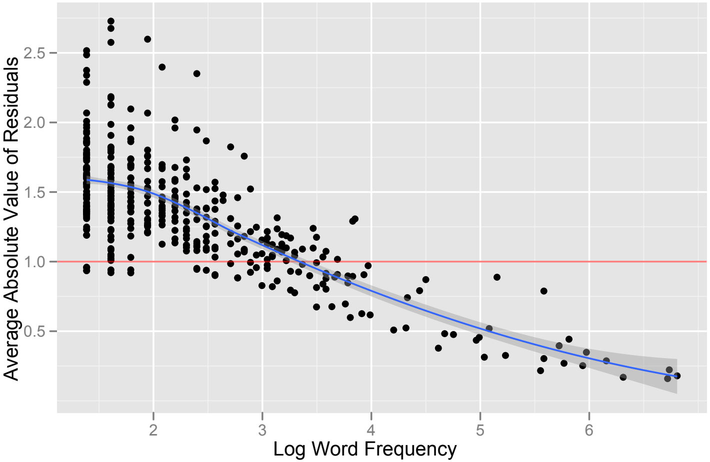
- Words occur in order - violates positional independence
- Words occur in combinations (as collocations) "carbon tax" / "income tax" / "inhertiance tax" / "capital gains tax" /"bank tax"
- Sentences (and topics) occur in sequence (extreme serial correlation)
- Style may mean means we are likely to use synonyms
- Rhetoric may lead to repetition. ("Yes we can!") – anaphora

Assumptions of the model (cont.)

- Poisson assumes $\mathrm{Var}(Y_{ij}) = \mathrm{E}(Y_{ij}) = \lambda_{ij}$
- For many reasons, we are likely to encounter overdispersion or underdispersion
  - overdispersion when "informative" words tend to cluster together
  - underdispersion could (possibly) occur when words of high frequency are uninformative and have relatively low between-text variation (once length is considered)
- This should be a *word*-level parameter

Overdispersion in German manifesto data

(data taken from Slapin and Proksch 2008)



One solution: Model overdispersion

Lo, Proksch, and Slapin:

$$\mathrm{Poisson}(\lambda) = \lim_{r \to \infty} \mathrm{NB}\left(r, \frac{\lambda}{\lambda + r}\right)$$

$$Y_{ij} \sim \mathrm{NB}\left(r, \frac{\lambda_{ij}}{\lambda_{ij} + r}\right)$$

where the variance inflation parameter r varies across *documents*:

$$Y_{ij} \sim \mathrm{NB}\left(r_i, \frac{\lambda_{ij}}{\lambda_{ij} + r_i}\right)$$

Relationship to multinomial

If each feature count Y_{ij} is an independent Poisson random variable with mean μ_{ij}, then we can formulate this as the following log-linear model:

$$\log \mu_{ij} = \lambda + \alpha_i + \psi_j^* + \theta_i \beta_j^* \tag{2}$$

where the log-odds that a generated token will fall into feature category j relative to the last feature J is:

$$\log \frac{\mu_{ij}}{\mu_{iJ}} = (\psi_j^* - \psi_J^*) + \theta_i(\beta_j^* - \beta_J^*) \tag{3}$$

which is the formula for multinomial logistic

Poisson/multinomial process as a DAG

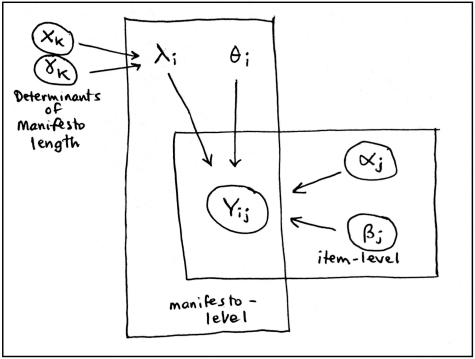


Figure 2: Directed acyclic graph of the one-dimensional Poisson IRT for document and item parameters to category counts Y_{ij}

How to estimate this model

Iterative maximimum likelihood estimation:

- If we knew Ψ and β (the word parameters) then we have a Poisson regression model
- If we knew α and θ (the party / politician / document parameters) then we have a Poisson regression model too!
- So we alternate them and hope to converge to reasonable estimates for both
- Implemented in the `austin` package as `wordfish`

An alternative is MCMC with a Bayesian formulation

Marginal maximum likelihood for wordfish

Start by guessing the parameters

Algorithm:

- Assume the current party parameters are correct and fit as a Poisson regression model
- Assume the current word parameters are correct and fit as a Poisson regression model
- Normalize θs to mean 0 and variance 1

Repeat

Identification

The *scale* and *direction* of θ is undetermined — like most models with latent variables

To identify the model in Wordfish

- Fix one α to zero to specify the left-right direction (Wordfish option 1)
- Fix the $\hat{\theta}$s to mean 0 and variance 1 to specify the scale (Wordfish option 2)
- Fix two $\hat{\theta}$s to specify the direction and scale (Wordfish option 3 and Wordscores)

Note: Fixing two reference scores does not specify the policy domain, it just identifies the model

Or: Use non-parametric methods

- Non-parametric methods are algorithmic, involving no "parameters" in the procedure that are estimated
- Hence there is no uncertainty accounting given distributional theory
- Advantage: don't have to make assumptions
- Disadvantages:
  - cannot leverage probability conclusions given distribtional assumptions and statistical theory
  - results highly fit to the data
  - not really assumption-free, if we are honest

Correspondence Analysis

- CA is like factor analysis for categorical data
- Following normalization of the marginals, it uses Singular Value Decomposition to reduce the dimensionality of the word-by-text matrix
- This allows projection of the positioning of the words as well as the texts into multi-dimensional space
- The number of dimensions – as in factor analysis – can be decided based on the eigenvalues from the SVD

Singular Value Decomposition

- A matrix $\underset{i \times j}{\mathbf{X}}$ can be represented in a dimensionality equal to its rank k as:

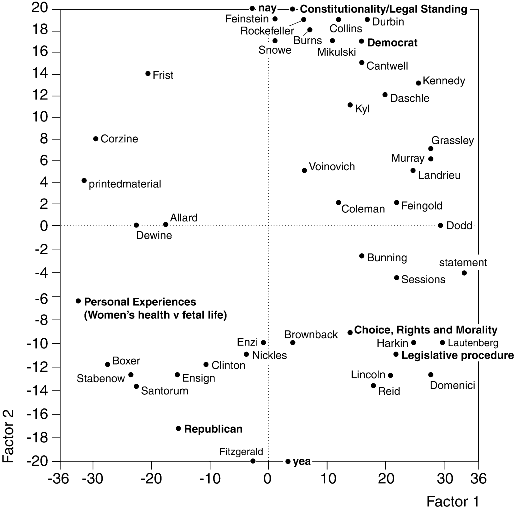
$$\underset{i \times j}{\mathbf{X}} = \underset{i \times k}{\mathbf{U}} \underset{k \times k}{\mathbf{d}} \underset{j \times k}{\mathbf{V}'} \tag{4}$$

- The **U**, **d**, and **V** matrixes "relocate" the elements of **X** onto new coordinate vectors in n-dimensional Euclidean space
- Row variables of **X** become points on the **U** column coordinates, and the column variables of **X** become points on the **V** column coordinates
- The coordinate vectors are perpendicular (*orthogonal*) to each other and are normalized to unit length

Correspondence Analysis and SVD

- Divide each value of $\mathbf{X}$ by the geometric mean of the corresponding marginal totals (square root of the product of row and column totals for each cell)
  - Conceptually similar to subtracting out the χ^2 expected cell values from the observed cell values
- Perform an SVD on this transformed matrix
  - This yields singular values $\mathbf{d}$ (with first always 1.0)
- Rescale the row ($\mathbf{U}$) and column ($\mathbf{V}$) vectors to obtain canonical scores (rescaled as $U_i\sqrt{f_{..}/f_{i.}}$ and $V_j\sqrt{f_{..}/f_{j.}}$)

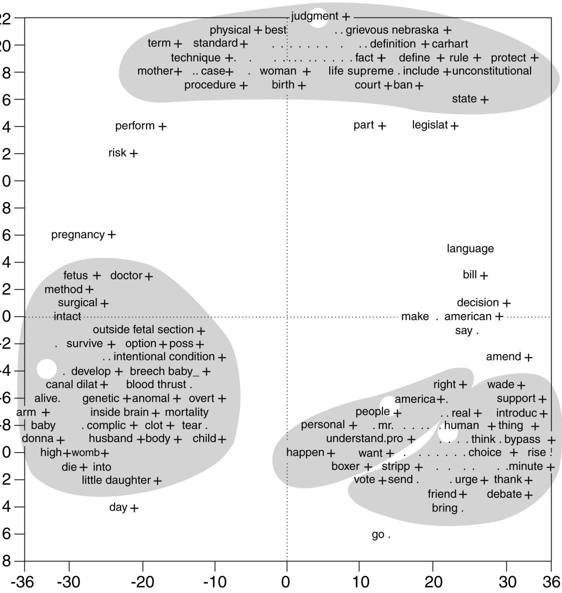
Example: Schonhardt-Bailey (2008) - speakers



	Eigenvalue	% Association	% Cumulative
Factor 1	0.30	44.4	44.4
Factor 2	0.22	32.9	77.3

Fig. 3. Correspondence analysis of classes and tags from Senate debates on Partial-Birth Abortion Ban Act

Example: Schonhardt-Bailey (2008) - words



How to get confidence intervals for CA

- There are problems with bootstrapping: (Milan and Whittaker 2004)
  - rotation of the principal components
  - inversion of singular values
  - reflection in an axis

How to account for uncertainty

- Ignore the problem and hope it will go away
  - SVD-based methods (e.g. correspondence analysis) typically do not present errors
  - and traditionally, point estimates based on other methods have not either

How to account for uncertainty

- Analytical derivatives
  - Using the multinomial formulation of the Poisson model, we can compute a Hessian for the log-likelihood function
  - The standard errors on the θ_i parameters can be computed from the covariance matrix from the log-likelihood estimation (square roots of the diagonal)
  - The covariance matrix is (asymptotically) the inverse of the negative of the Hessian
    (where the negative Hessian is the observed Fisher information matrix, a.ka. the second derivative of the log-likelihood evaluated at the maximum likelihood estimates)
  - Problem: These are *too small*

How to account for uncertainty

- Parametric bootstrapping (Slapin and Proksch, Lewis and Poole)
  Assume the distribution of the parameters, and generate data after drawing new parameters from these distributions.
  Issues:
  - slow
  - relies heavily (twice now) on parametric assumptions
  - requires some choices to be made with respect to data generation in simulations
- Non-parametric bootstrapping
- (and yes of course) Posterior sampling from MCMC

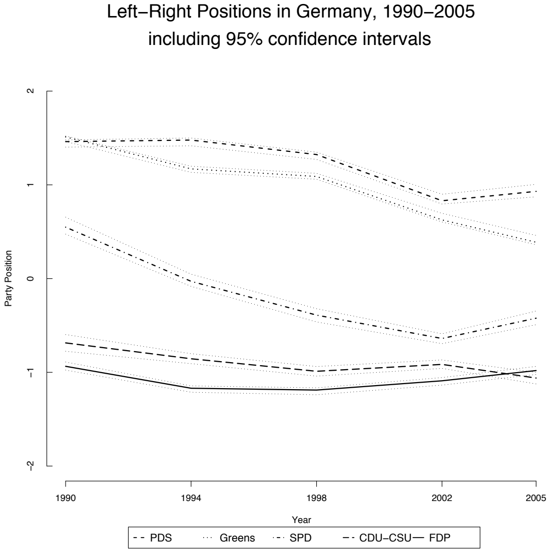
How to account for uncertainty

- Non-parametric bootstrapping
  - draw new versions of the texts, refit the model, save the parameters, average over the parameters
  - slow
  - not clear how the texts should be resampled

How to account for uncertainty

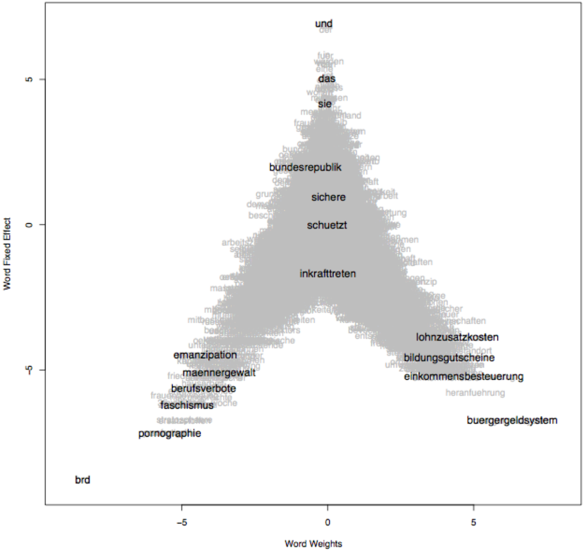
- For MCMC: from the distribution of posterior samples

Parametric Bootstrapping and analytical derivatives yield "errors" that are too small



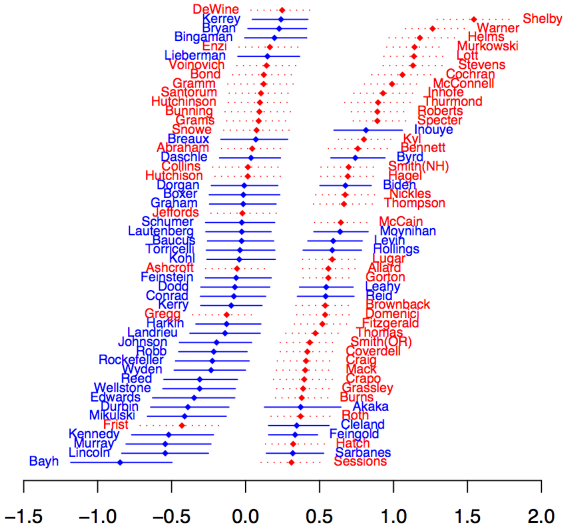
Frequency and informativeness

Ψ and β (frequency and informativeness) tend to trade-off



Plotting θ

Plotting θ (the ideal points) gives estimated positions. Here is Monroe and Maeda's (essentially identical) model of legislator positions:



Interpreting multiple dimensions

To get one dimension for each policy area, split up the document by hand and use the subparts as documents (the Slapin and Proksch method)

There is currently *no* implementation of Wordscores or Wordfish that extracts two or more dimensions at once

- But since Wordfish is a type of factor analysis model, there is no reason in principle why it could not

Interpreting scaled dimensions

- Another (better) option: compare them other known descriptive variables
- Hopefully also *validate* the scale results with some human judgments
- This is necessary even for single-dimensional scaling
- And just as applicable for non-parametric methods (e.g. correspondence analysis) as for the Poisson scaling model