Computer-Assisted Content Analysis: Documents in space

Will Lowe (University of Mannheim)

Menu

Session 1: Dictionary-based 'classical' content analysis and topic models

Session 2: Classification and evaluation

Session 3:

- Scaling Models
  - Scaling models
    Dimensionality
    Uncertainty
    as a visualisation tool

The bag of words

	ahead	am	am	i	like	look		content
doc 1	1		1	1	2	0	1	...	θ_1
doc 2	0		0	0	1	1	0	...	θ_2

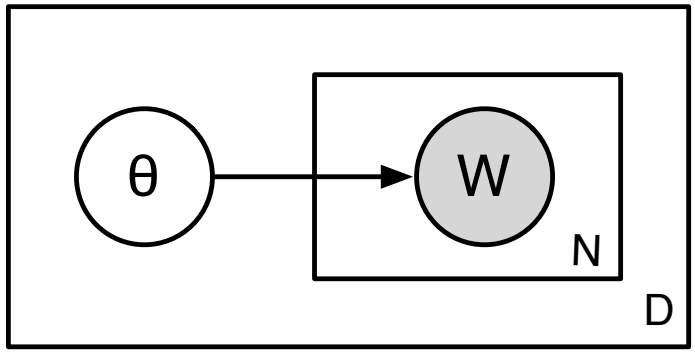
For each research problem involving content analysis we need to ask:

- What *structure* θ has
- What *modeling strategy* to take
- What the *relationship* is between θ and the words (i.e. the model)

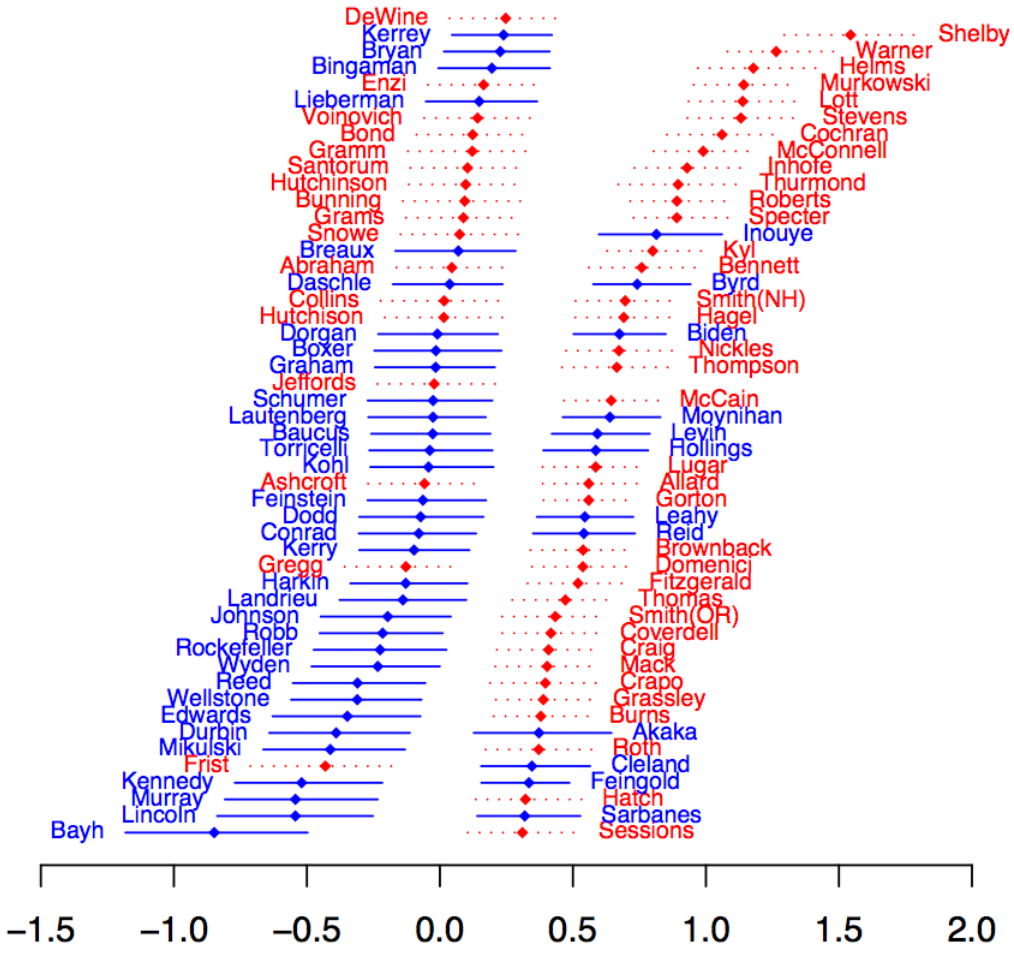
Answers

1. θ is a (possibly multidimensional) position

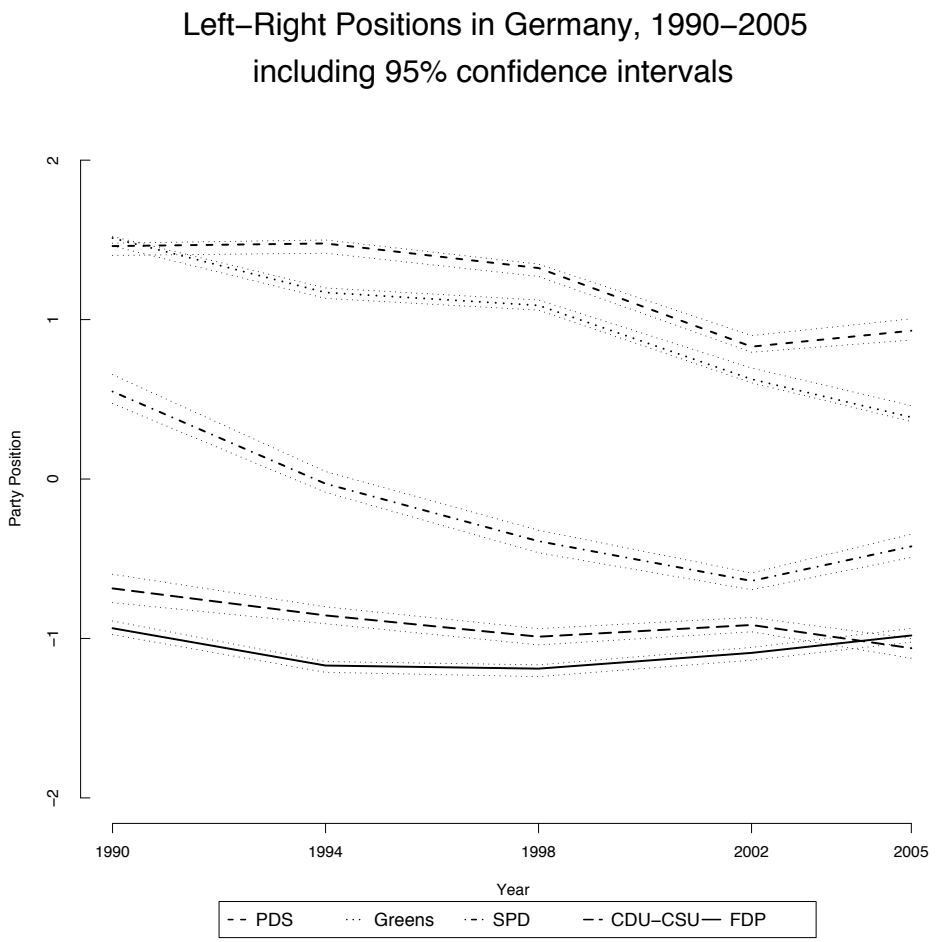
2. We want a generative model of words: indirect approach

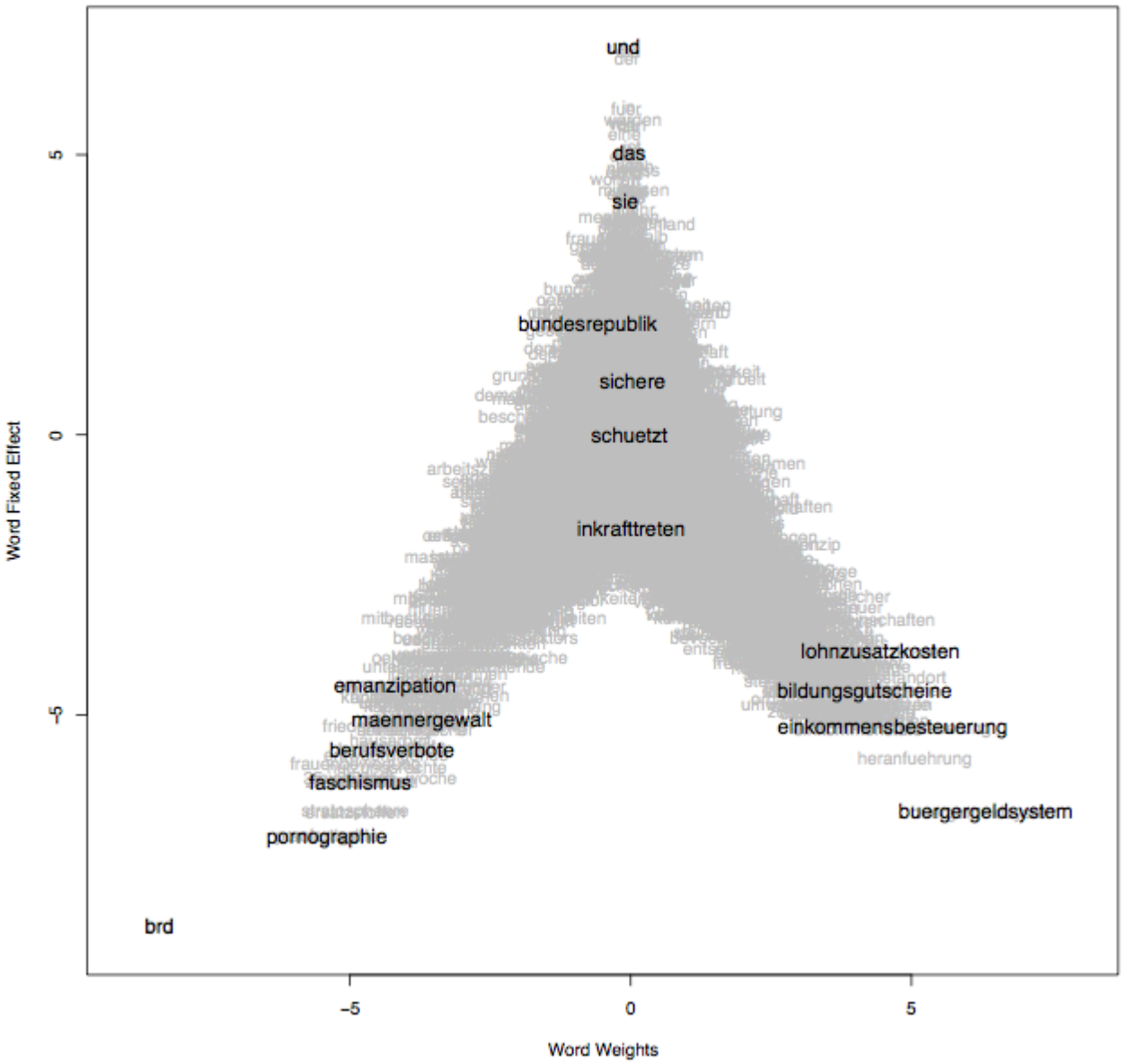


3. θ has a proximity ('ideal point') structure connecting document and word positions


DeWine
Kerrey
Bryan
Bingaman
Enzi
Lieberman
Voinovich
Bond
Gramm
Santorum
Hutchinson
Bunning
Grams
Snowe
Breaux
Abraham
Daschle
Collins
Hutchison
Dorgan
Boxer
Graham
Jeffords
Schumer
Lautenberg
Baucus
Torricelli
Kohl
Ashcroft
Feinstein
Dodd
Conrad
Kerry
Gregg
Harkin
Landrieu
Johnson
Robb
Rockefeller
Wyden
Reed
Wellstone
Edwards
Durbin
Mikulski
Frist
Kennedy
Murray
Lincoln
Bayh
Shelby
Warner
Helms
Murkowski
Lott
Stevens
Cochran
McConnell
Inhofe
Thurmond
Roberts
Specter
Inouye
Kyl
Bennett
Byrd
Smith(NH)
Hagel
Biden
Nickles
Thompson
McCain
Moynihan
Levin
Hollings
Lugar
Allard
Gorton
Leahy
Reid
Brownback
Domenici
Fitzgerald
Thomas
Smith(OR)
Coverdell
Craig
Mack
Crapo
Grassley
Burns
Akaka
Roth
Cleland
Feingold
Hatch
Sarbanes
Sessions
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

German parties 1990-2005




und
das
sie
bundesrepublik
sichere
schuetzt
inkrafttreten
lohnzusatzkosten
bildungsgutscheine
einkommensbesteuerung
heranfuehrung
buergergeldsystem
emanzipation
maennergewalt
berufsverbote
faschismus
pornographie
brd
Word Fixed Effect
5
0
−5
−5
0
5
Word Weights

Structure

Typically scaling models assume that

- relative word usage is reflective of political ideology
- Positions are unidimensional in W
- Positions drive word counts stochastically according to a particular form for $P(W_j \mid \theta)$
- Bag of words: counts of W_j are conditionally independent given θ

$$P(W_1 \ldots W_V) \;=\; \prod_j^V P(W_j \mid \theta)\; P(\theta)$$

Existing Models

We focus on:

Wordfish: (Slapin and Proksch, 2008)

(Also Laver et al. 2003, Monroe and Maeda, 2004; Beauchamp, 2008; Pennings and Keman, 2002; König and Luig, 2009, Goodman 1979, etc.)

Wordfish

The position word relationship is

$$\begin{aligned} W_{ij} &\sim \text{Poisson}(\mu_{ij}) \\ \log \mu_{ij} &= \psi_j + \beta_j \theta_i + \alpha_i \end{aligned}$$

Each word is a Poisson Process (stochastic component) driven by word and document parameters (the systematic component)

Word parameters:

- β how fast counts increase or decrease with changes in position
- ψ how frequent words are irrespective of position

Wordfish

The position word relationship is

$$\begin{aligned} W_{ij} &\sim \text{Poisson}(\mu_{ij}) \\ \log \mu_{ij} &= \psi_j + \beta_j \theta_i + \alpha_i \end{aligned}$$

Each word is a Poisson Process driven by word and document parameters

Document parameters:

- θ the position being expressed
- α a fixed effects for documents controlling for length...

An equivalent view

If document *length* is uninformative about position then we can condition on it.

In fact Wordfish already does

$$
\begin{aligned}
P(W_i \mid \theta_i, N) &= \text{Multinomial}(\boldsymbol{\pi}, N) \\
\log \frac{\pi_{ij}}{\pi_{i1}} &= \psi_j^* + \beta_j^* \theta_i
\end{aligned}
$$

where $\psi_j^* \leftarrow (\psi_j - \psi_1)$, $\beta_j^* \leftarrow (\beta_j - \beta_1)$ and α_i cancels

Wordfish is secretly a Multinomial Response Model

Estimation

Wordfish models are fit using Conditional Maximum Likelihood (regression without independent variables)

Iterate:

- Fix document parameters (α and θ) and maximize word parameters (β and ψ)
- Fix new word parameters (β and ψ) and maximize document parameters (α and θ)

This can be quite slow depending on the size of your dataset. . .

Identification

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

As is the case with all scaling models (e.g. NOMINATE), the likelihood function is not identified.

Without fixing some parameters, there are infinite combinations of θ and β, which could provide the same likelihood (we would not arrive at a unique solution).

Solution: fix mean of document positions θ to 0 and St. Dev. to 1. Set one document fixed effect to 0. Set directionality of scale. This means that you cannot

directly compare estimates ACROSS different estimations.

What about differences across languages?

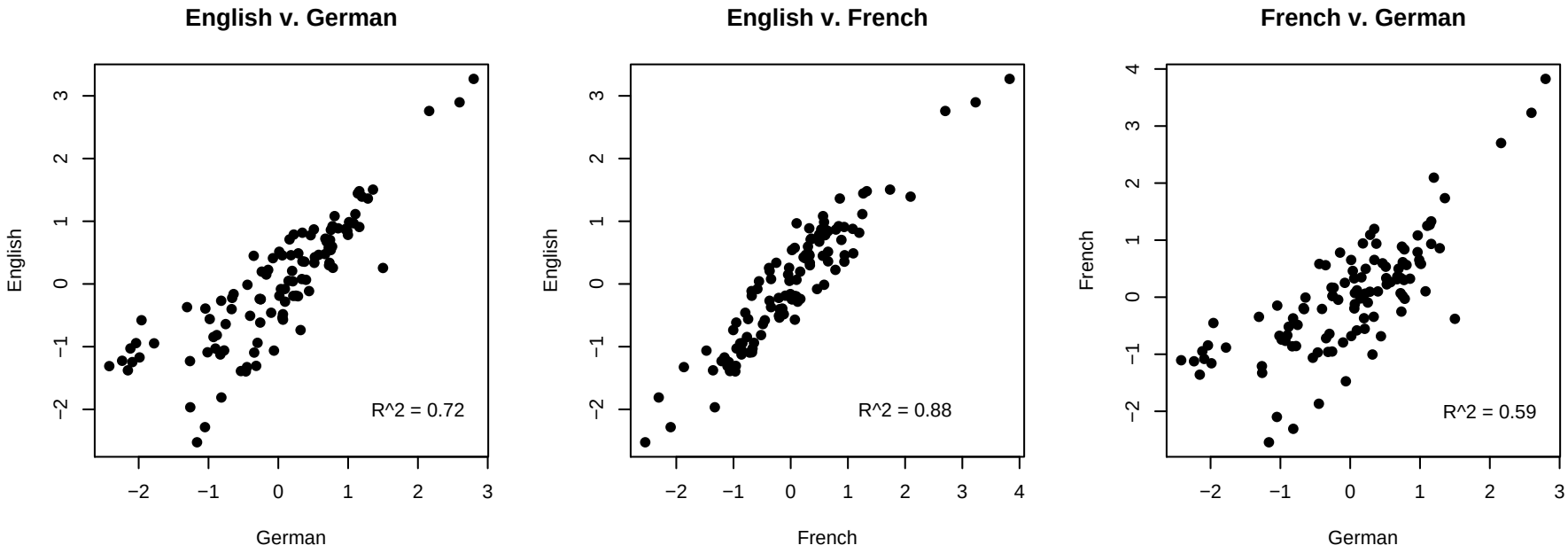
Ideal case: get the exact same political texts in high quality translations

Estimate Wordfish and compare across different languages

This is possible: European Parliament speeches (translated into all official languages of the EU)

What about differences across languages?

Positions of National Parties in the European Parliament



Dimension Issues

What the heck is θ?

How do we know that positions on only one dimension are being expressed in the text?

Dimension Issues

What the heck is θ?

Whatever maximizes the Likelihood

Approximately the first principal component of $\log W$

Like all scaling techniques (e.g. NOMINATE), Wordfish is effectively *exploratory* – you have to figure out what the dimension really is. This is the reason why you need to think about your data before applying the method.

Dimension Issues

How do we know that positions on only one dimension are being expressed? How do we get positions on a specific policy issue?

Force the assumption to be true by including substantive knowledge about your texts:

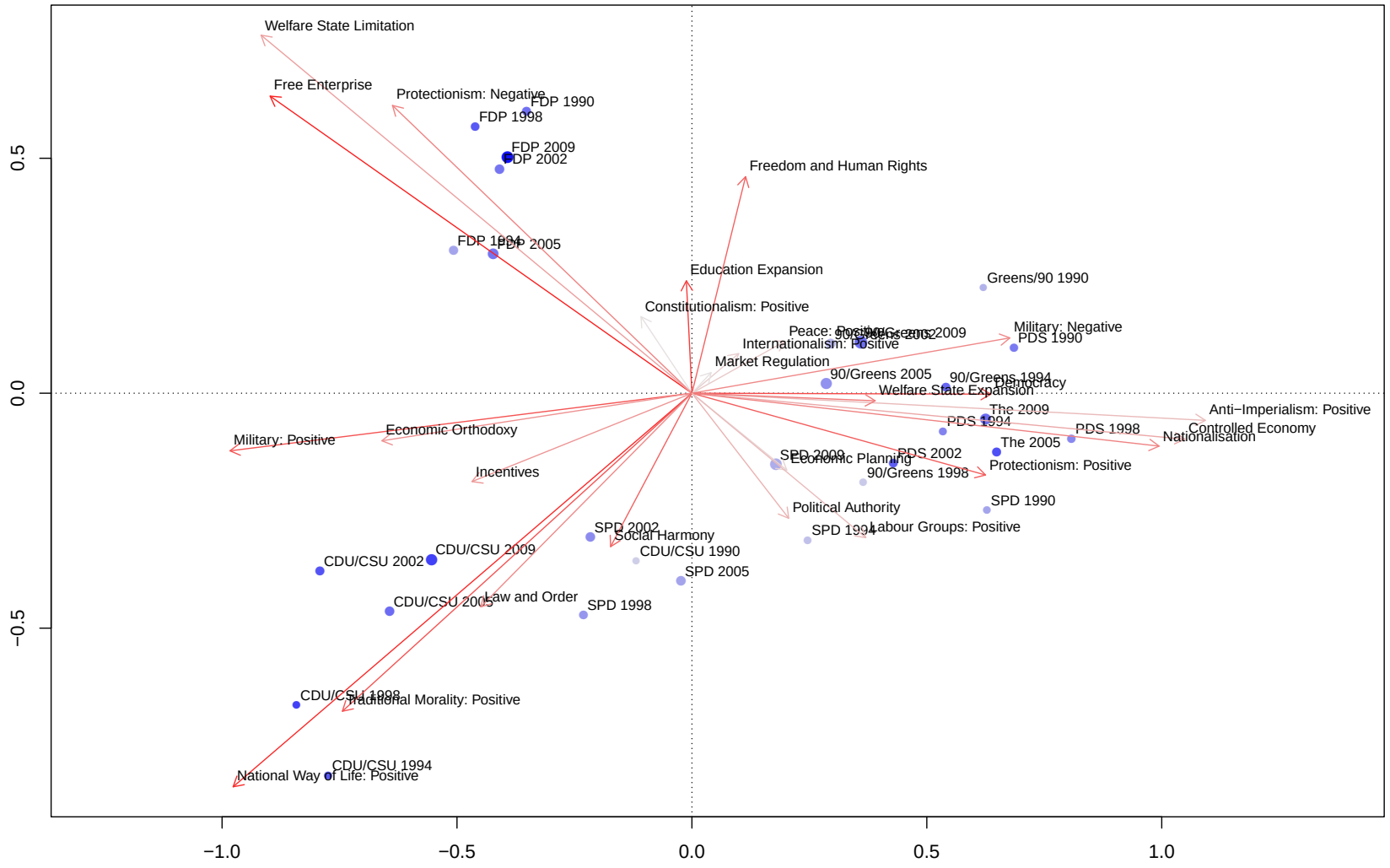
> Use only those texts (or sections thereof) that are guaranteed to be on the same topic and scale them separately
>
> E.g. speeches from a specific debate or policy sections of manifestos (e.g. Slapin and Proksch 2008)

Monroe and Quinn (endlessly forthcoming) use a topic model

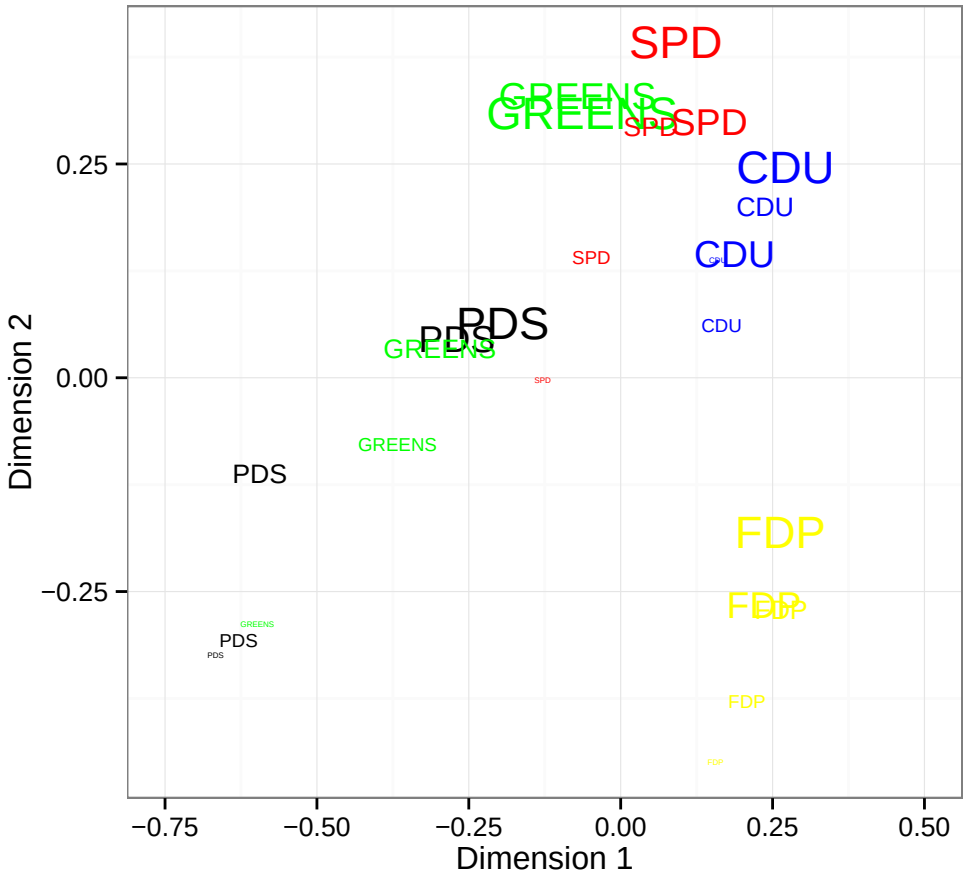
Advantage: Heroic assumptions are (closer to being) true

Orthogonality?

Substantively meaningful dimensions do not always line up with the axes of θ...


Welfare State Limitation
Free Enterprise
Protectionism: Negative
FDP 1990
FDP 1998
FDP 2009
FDP 2002
FDP 1994
FDP 2005
Freedom and Human Rights
Education Expansion
Constitutionalism: Positive
Greens/90 1990
Peace: Positive
Internationalism: Positive
Military: Negative
PDS 1990
Market Regulation
90/Greens 2005
90/Greens 1994
Democracy
Welfare State Expansion
The 2009
PDS 1994
PDS 1998
Anti−Imperialism: Positive
Controlled Economy
Nationalisation
Military: Positive
Economic Orthodoxy
The 2005
PDS 2002
Economic Planning
Protectionism: Positive
Incentives
90/Greens 1998
SPD 1990
Political Authority
Labour Groups: Positive
SPD 1994
SPD 2002
Social Harmony
CDU/CSU 2009
CDU/CSU 1990
CDU/CSU 2002
SPD 2005
CDU/CSU 2005
Law and Order
SPD 1998
CDU/CSU 1998
Traditional Morality: Positive
CDU/CSU 1994
National Way of Life: Positive
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
1.0

Multiple Dimension Issues



(In)Stability of the Political Lexicon

What if the political lexicon is unstable over time?

New issues appear, old issues disappear

If this happens frequently, then scaling algorithms will pick up shifts in the policy agenda – rather than shifts in party positions.

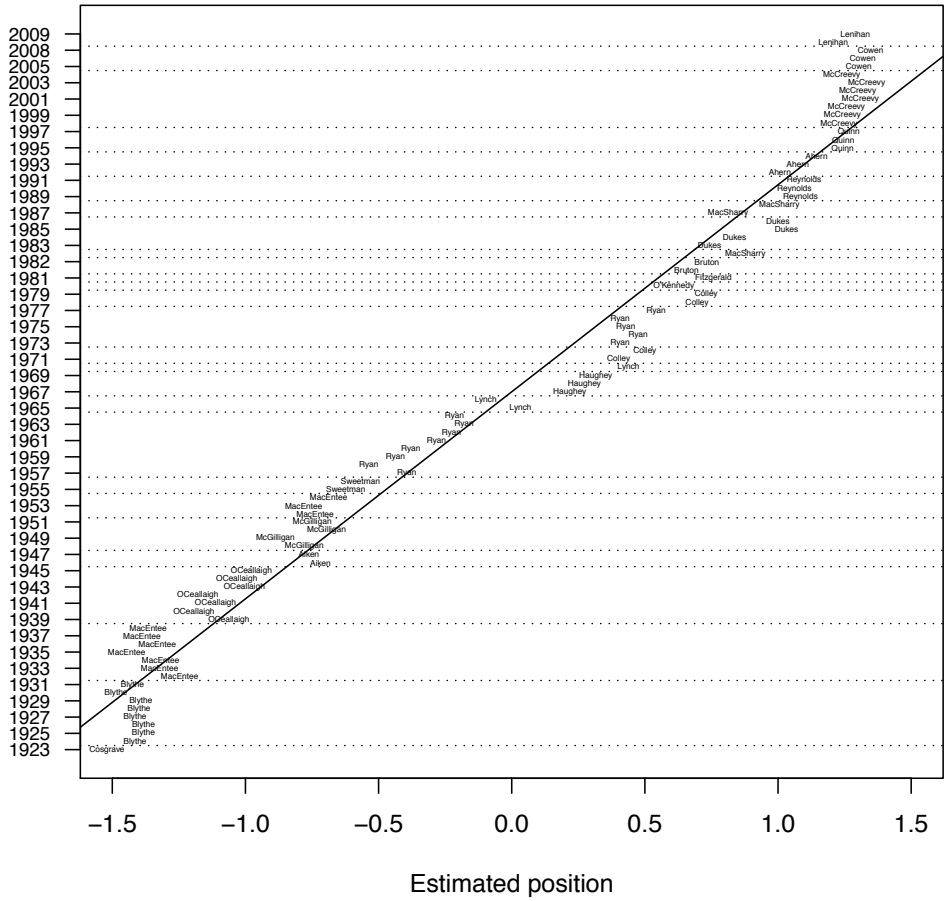
In fact, this is one assumption: that word usage reflects ideology.

For example, it becomes seriously problematic when all parties start talking about the "issue" of the day.

Then we can distinguish between elections, but not very well between parties

We can (try to) get around this by focusing on those words that remain in the political vocabulary across time.

Worst Case Scenario



Policy Dimensionality

Emerging consensus:

Identifying policy dimensions in text is about vocabulary choice: substantive issue.

Different routes on how to get there are possible.

Should be theory driven; validity is a big concern.

Validation

Do these models extract valid positions?

- multiple dimensions, drifting language, and topics
- non-dimensional structure
- selection bias and party control of the floor

Validating positions and uncertainty

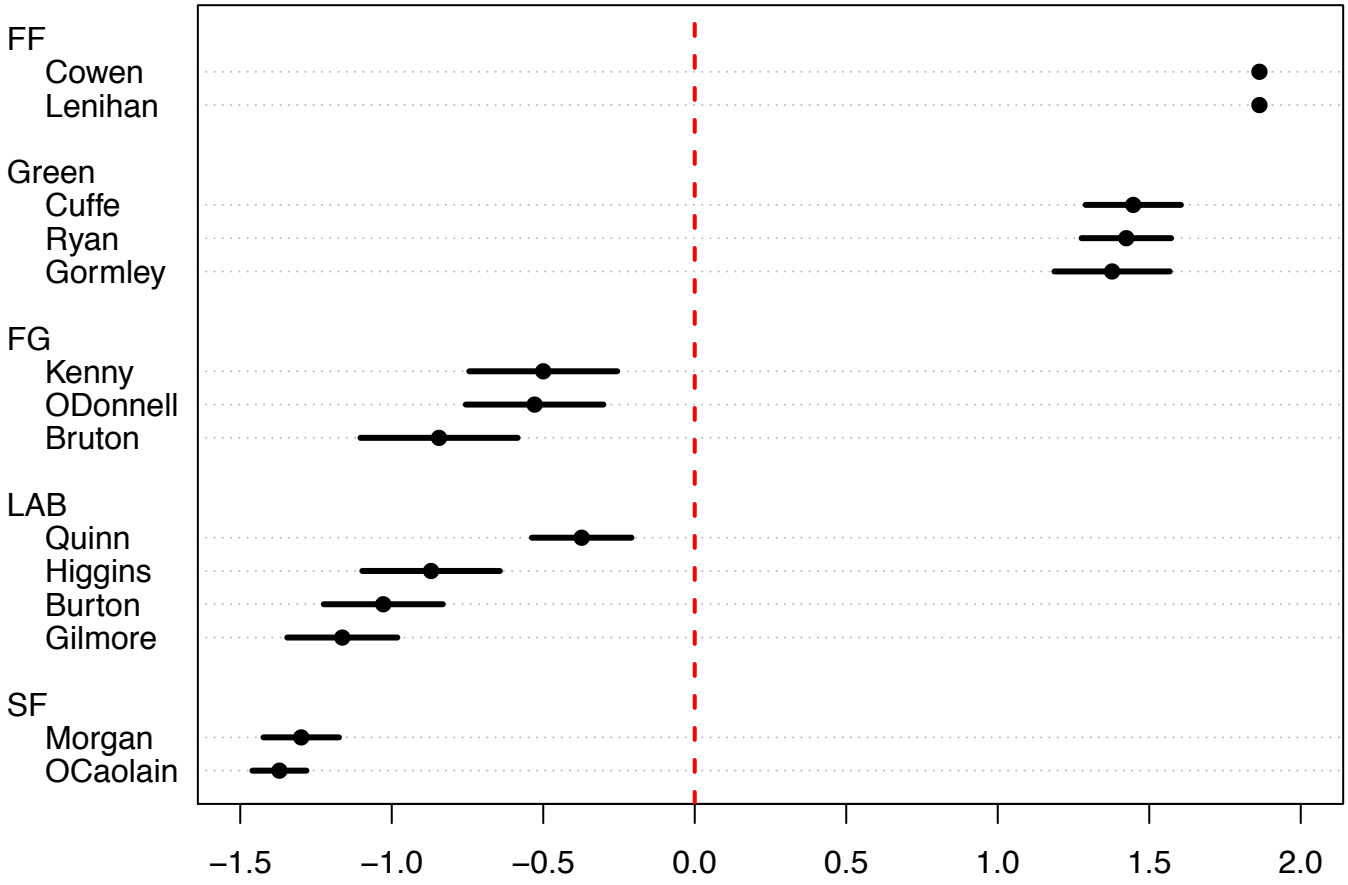
14 speeches from the debate on Ireland's 2010 budget coded by 18 PhD students (LSE and TCD)

Task: Identify *speaker positions*, twice

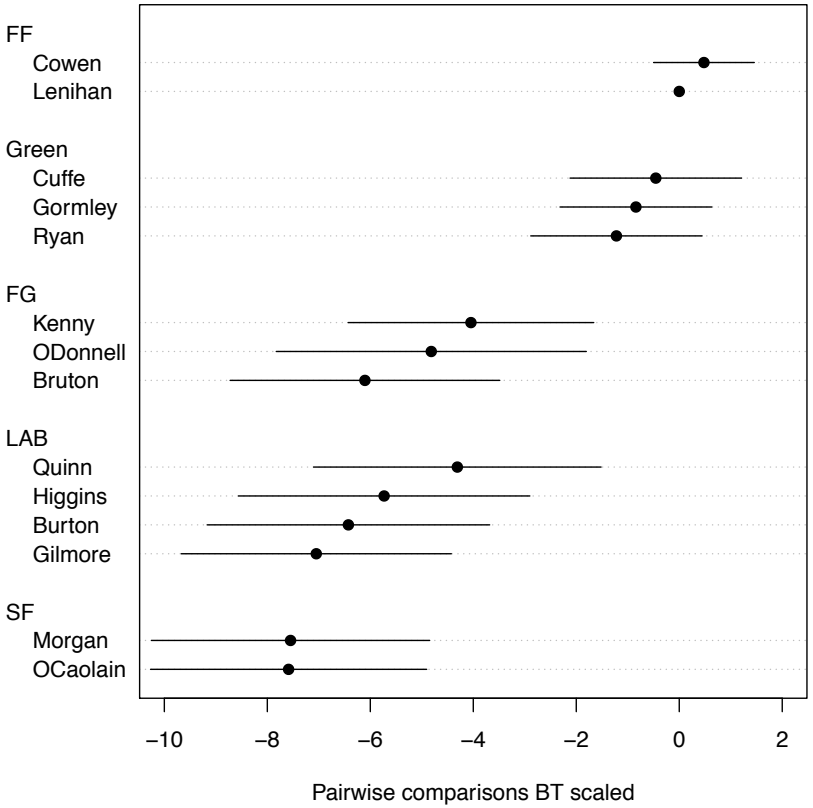
directly with uncertainty limits
by pairwise comparison (and indicate uncertainty)

(Lowe & Benoit 2014)

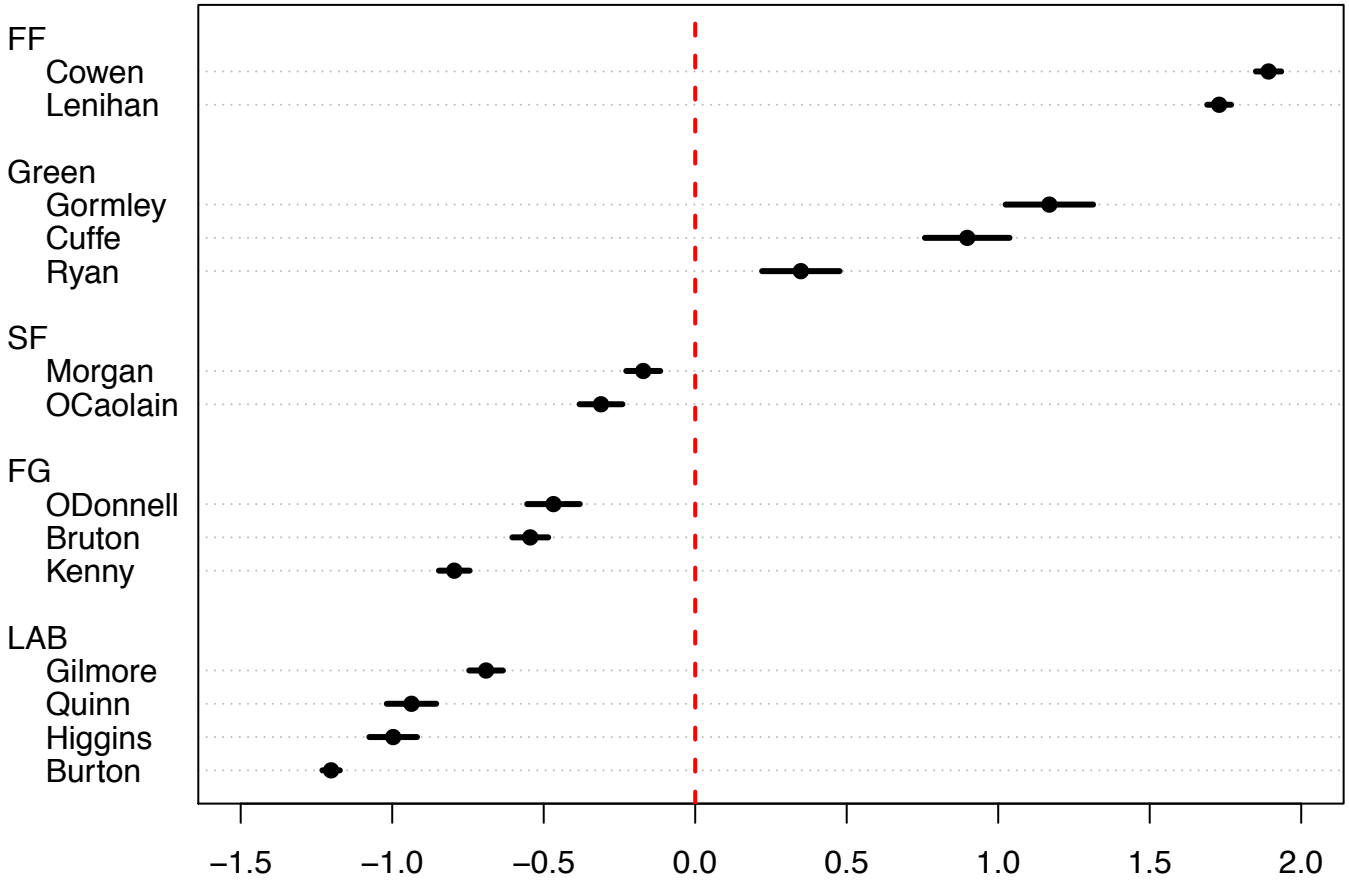
Respondents' positions: direct

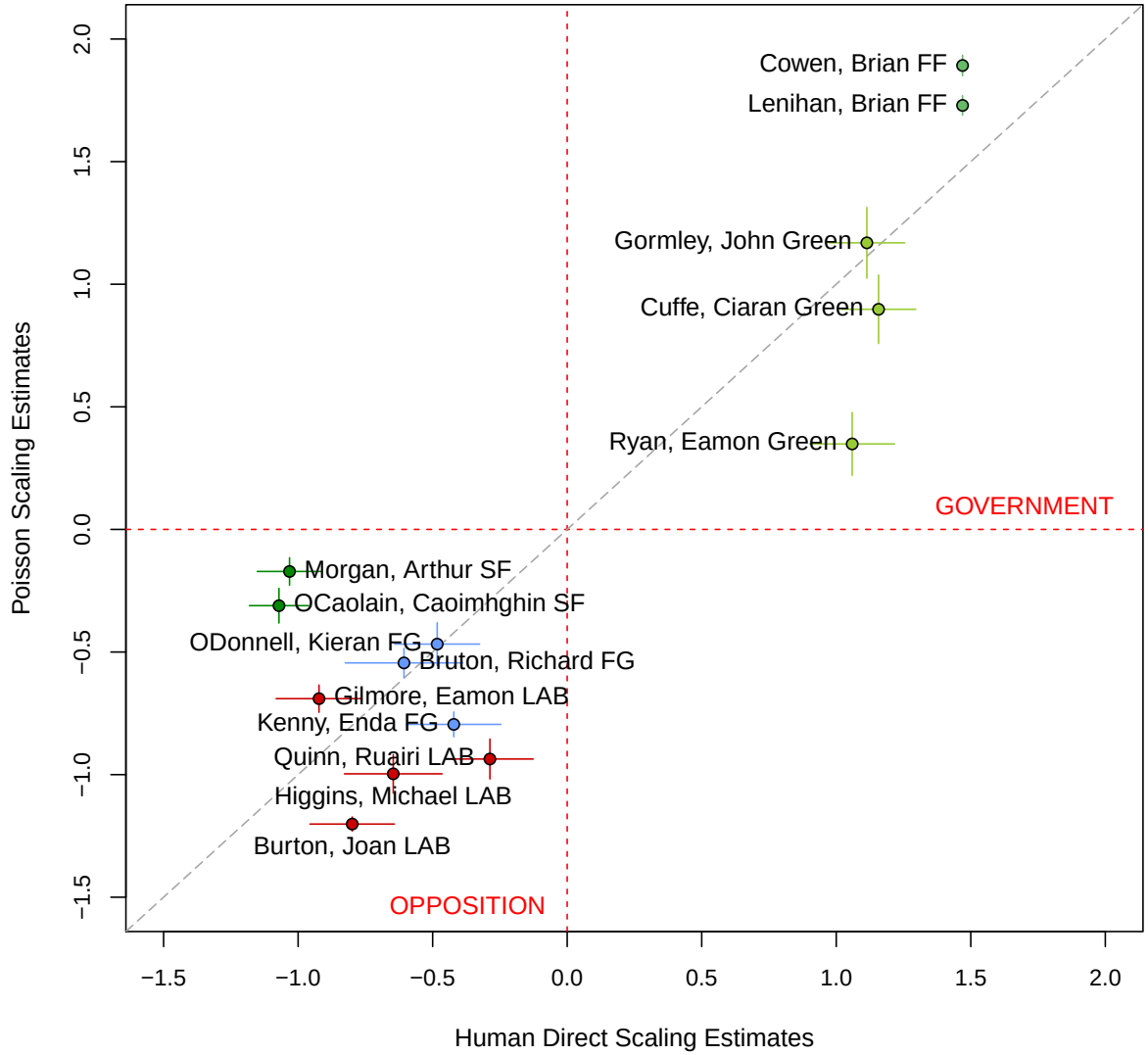


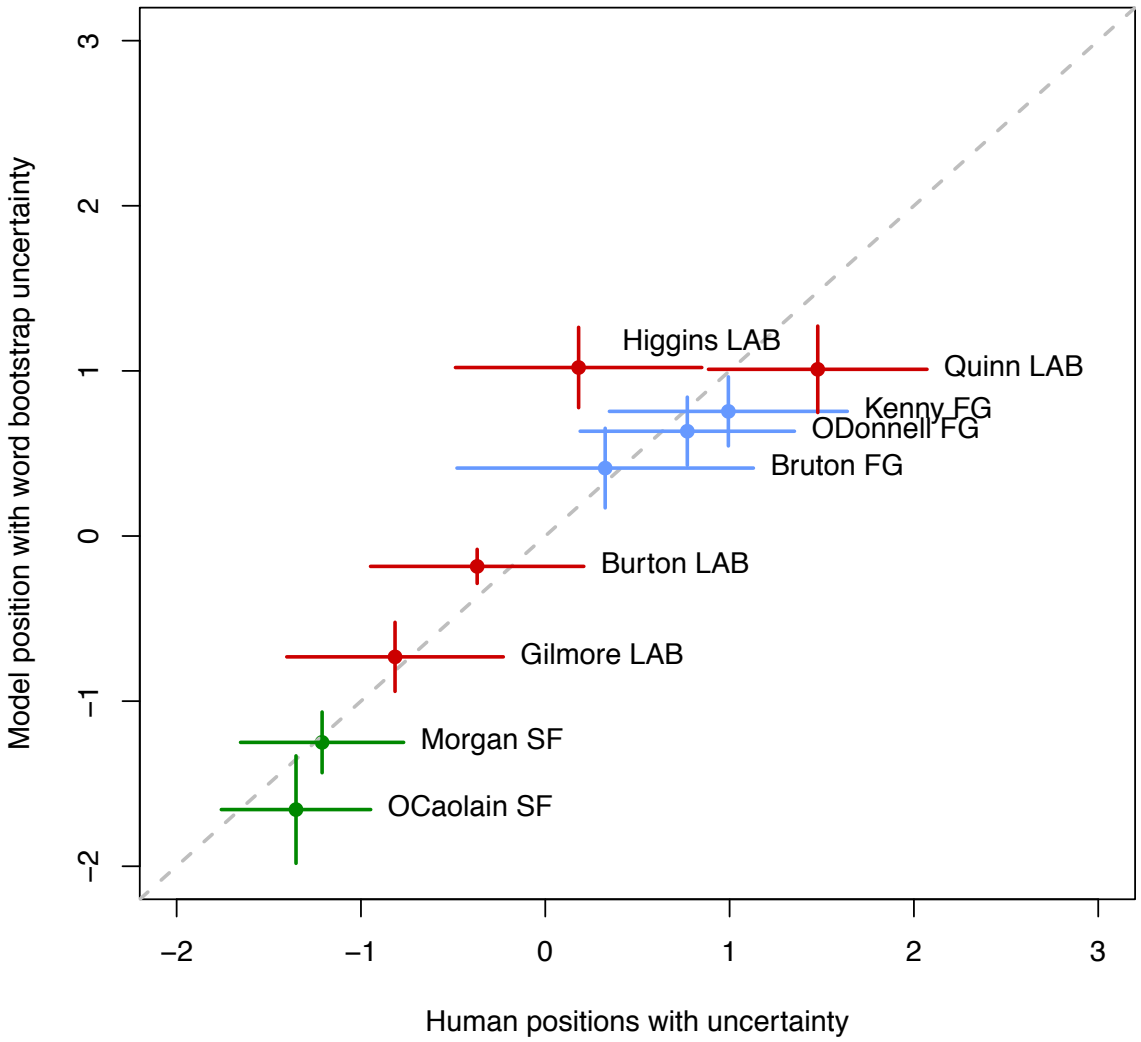
Respondents' positions: pairwise



Model's positions




Cowen, Brian FF
Lenihan, Brian FF
Gormley, John Green
Cuffe, Ciaran Green
Ryan, Eamon Green
GOVERNMENT
Morgan, Arthur SF
OCaolain, Caoimhghin SF
ODonnell, Kieran FG
Bruton, Richard FG
Gilmore, Eamon LAB
Kenny, Enda FG
Quinn, Ruairi LAB
Higgins, Michael LAB
Burton, Joan LAB
OPPOSITION
Poisson Scaling Estimates
Human Direct Scaling Estimates
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


Higgins LAB
Quinn LAB
Kenny FG
ODonnell FG
Bruton FG
Burton LAB
Gilmore LAB
Morgan SF
OCaolain SF
Model position with word bootstrap uncertainty
Human positions with uncertainty
−2
−1
0
1
2
3

Validation: positioning

These models

- distinguish *government and opposition* less than coders
- correctly *order parties*, except for Sinn Féin
- identify more *variation* in the Green party

Individual speaker positions are apparently difficult to assign

Validation: uncertainty (five methods)

Uncertainty measures try to answer the question:

How different could this speech have been?
(while still expressing the same position)

Three ways to be uncertain (in a model)

ML: assume the model is correct & *word parameters* are well estimated
(Lowe & Benoit 2010)

Bayes: assume the model is correct
(Monroe & Maeda 2004, Lo et al. 2011)

Bootstrap: assume that expected word rate μ_{ij} is correct
(Lowe & Benoit 2011)

Bootstrapping text

Resampling ‘writes’ the speeches that could have been given, but weren’t. . .

Bootstrapping text

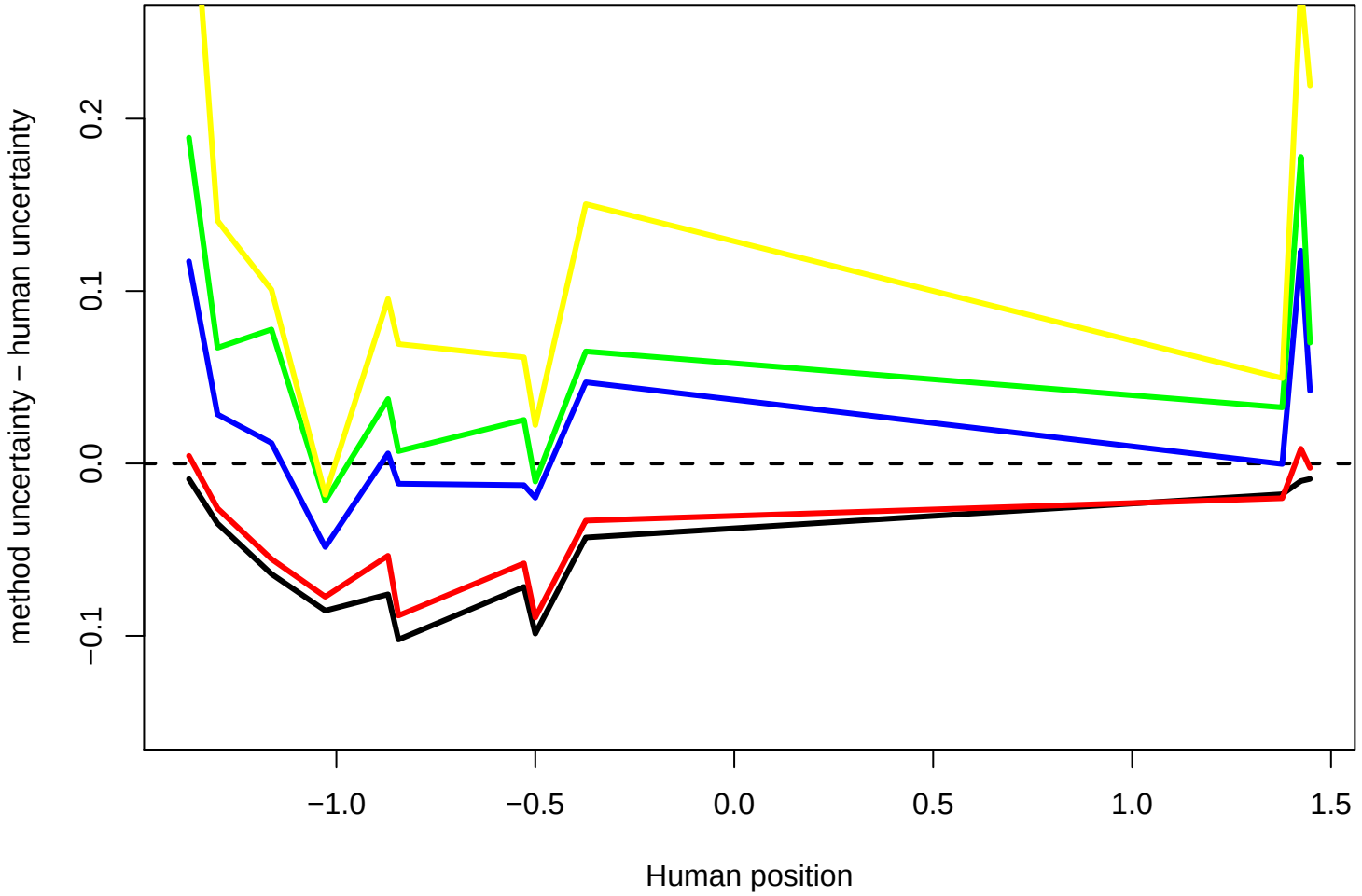
Resampling 'writes' the speeches that could have been given, but weren't. . .

parametric: Resample from the *fitted word counts* and refit

word: Resample *individual words* and refit

sentence: Resample *natural sentences* and refit

block: Resample overlapping length K word sequences


method uncertainty – human uncertainty
0.2
0.1
0.0
−0.1
−1.0
−0.5
0.0
0.5
1.0
1.5
Human position

Conclusions from validation

Substantively these scaling models induce dimensional structure from patterns of relative emphasis

- This is some, but not all of the structure reported by experts and coders

They do mostly they recover expert-assigned political positions, but can get fooled by extra-dimensional structure

Conclusions from validation

Substantively these scaling models induce dimensional structure from patterns of relative emphasis

- This is some, but not all of the structure reported by experts and coders

They do mostly they recover expert-assigned political positions, but can get fooled by extra-dimensional structure

They are *over-confident* about position, but this can be *partially corrected*.

Visualisation

But I *don't really care* about the microfoundations of political position-taking using text. . .

An alternative derivation

Wordfish is a model intermediate in complexity between:

$$\log \mu_{ij} = \psi_j + \alpha_i$$

and

$$\log \mu_{ij} = \psi_j + \alpha_i + \eta_{ij}$$

by projecting the variation in η_{ij} into a lower dimensional space

$$\eta_{ij} \approx \beta_j \theta_i$$

Visualising relative emphasis

6000 manually-coded Tweets from `#OccupyWallStreet`, `#15m`, `#greekrevolution`.

Visualising relative emphasis

6000 manually-coded Tweets from `#OccupyWallStreet`, `#15m`, `#greekrevolution`.

Practical problem: how to visualise cross-tabular content, like

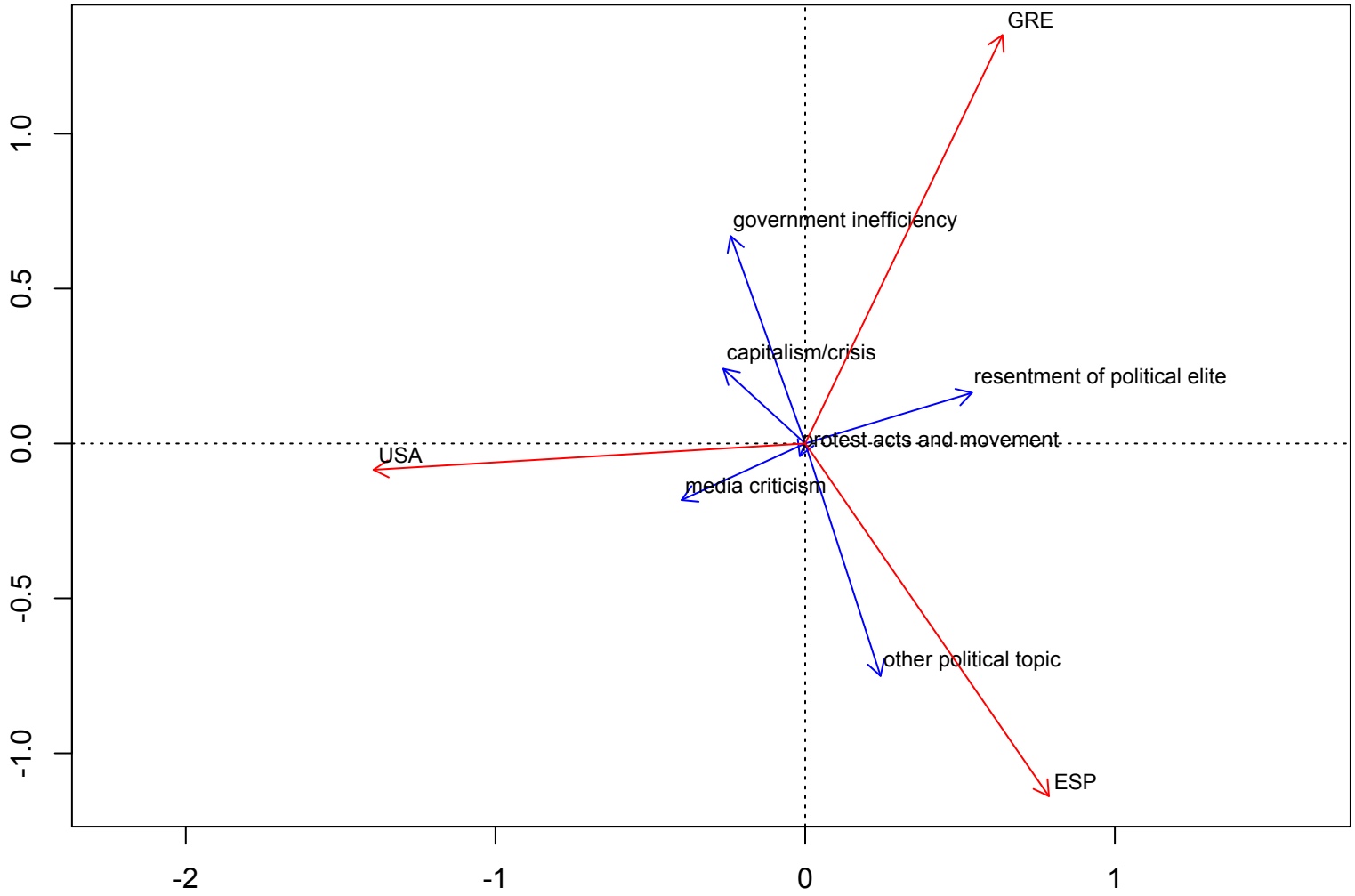
	ESP	GRE	USA
capitalism/crisis	33	68	85
government inefficiency	1	33	26
media criticism	42	22	73
other political topic	40	3	14
protest acts and movement	487	409	479
resentment of political elite	101	118	19

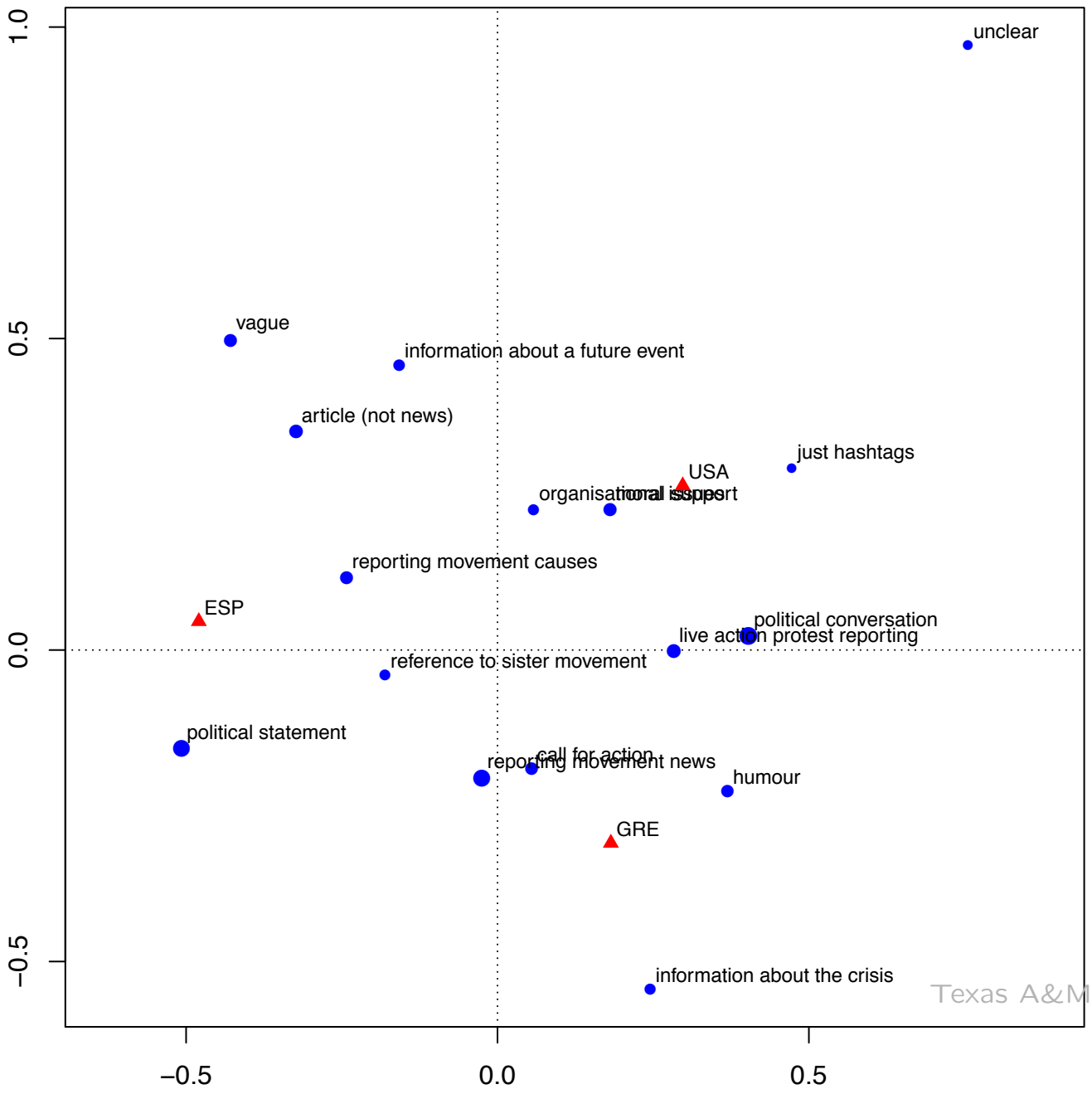
(Theocharis et al., 2013)

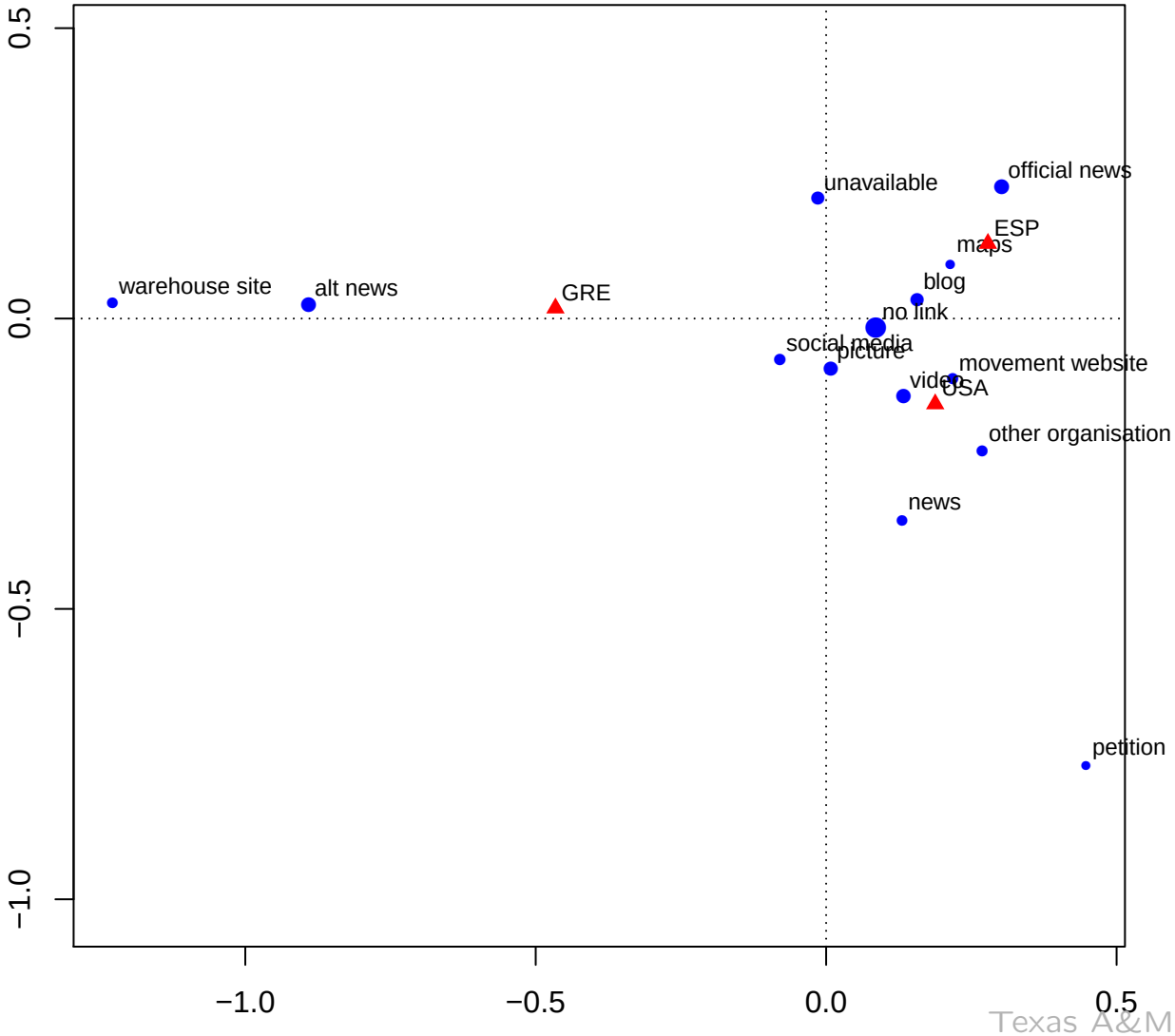
Visualising relative emphasis

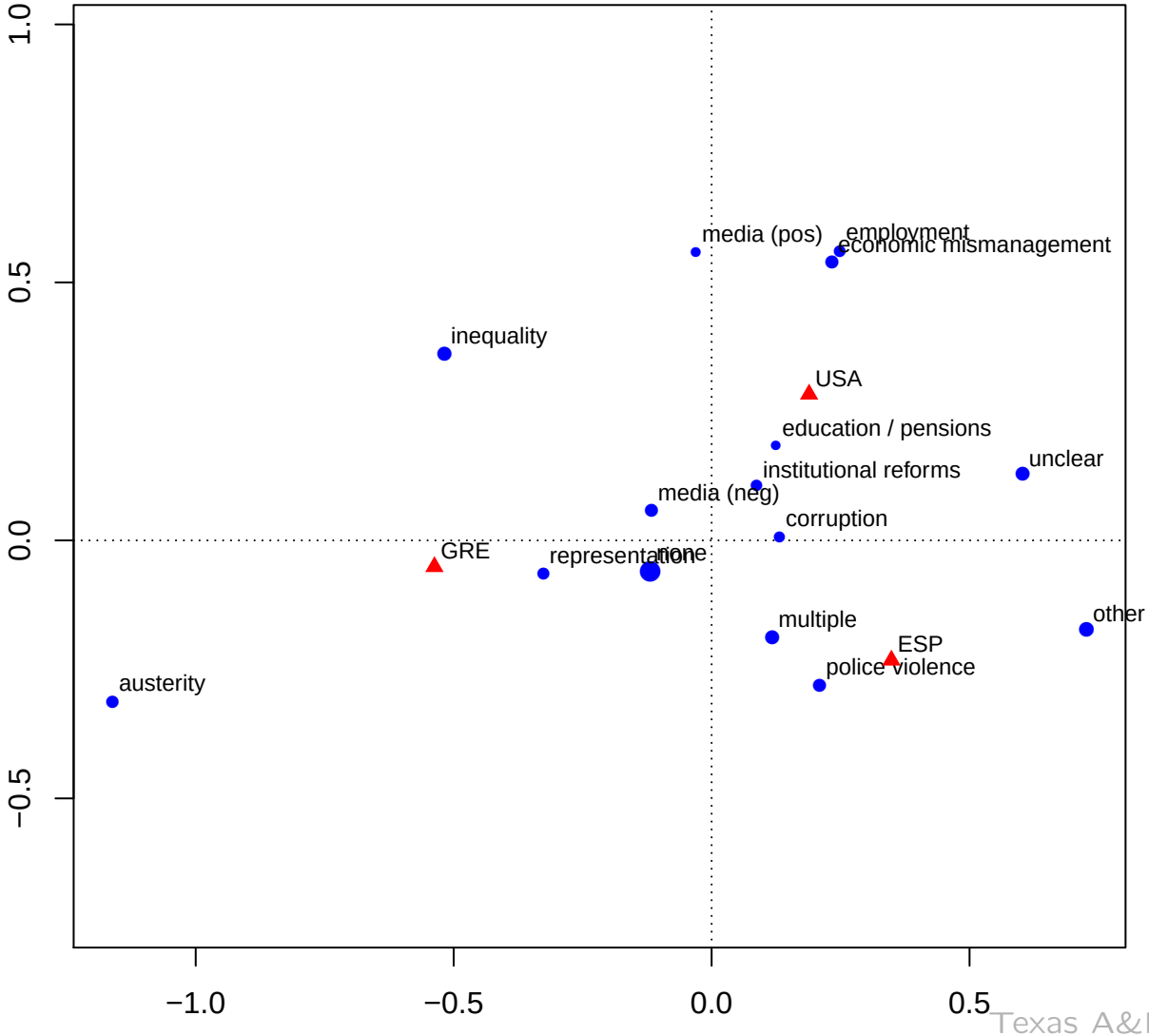
Fit a *scaling model* for countries θ and topics β

Visualise using a *biplot* (Gabriel 1971, Greenacre 2010)


GRE
government inefficiency
capitalism/crisis
resentment of political elite
protest acts and movement
USA
media criticism
other political topic
ESP
1.0
0.5
0.0
-0.5
-1.0
-2
-1
0
1


unclear
vague
information about a future event
article (not news)
just hashtags
USA
organisational issues
moral support
reporting movement causes
ESP
political conversation
live action protest reporting
reference to sister movement
political statement
call for action
reporting movement news
humour
GRE
information about the crisis
1.0
0.5
0.0
−0.5
−0.5
0.0
0.5


0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
warehouse site
alt news
GRE
unavailable
official news
ESP
maps
blog
no link
social media
picture
movement website
video
USA
other organisation
news
petition


media (pos)
employment
economic mismanagement
inequality
USA
education / pensions
unclear
institutional reforms
media (neg)
corruption
GRE
representation
none
multiple
other
ESP
police violence
austerity
1.0
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5

Texas A&M Jan 2015

A surprising application to IR

Russian artillery south of the Chechen capital Grozny blasted Chechen positions overnight before falling silent at dawn, witnesses said on Tuesday.

Israel said on Tuesday it sent humanitarian aid to Colombia where a massive earthquake last week killed at least 938 people and injured 400.

Event data extraction

Russian artillery^{S} south of the Chechen capital Grozny blasted^{223} Chechen positions^{T} overnight before falling silent at dawn, witnesses said on Tuesday.

Israel^{S} said on Tuesday it sent humanitarian aid^{073} to Colombia^{T} where a massive earthquake^{S} last week killed^{222} at least 938 people^{T} and injured 400.

Event data extraction

Russian artillery^{S} south of the Chechen capital Grozny blasted^{223} Chechen positions^{T} overnight before falling silent at dawn, witnesses said on Tuesday.

Israel^{S} said on Tuesday it sent humanitarian aid^{073} to Colombia^{T} where a massive earthquake^{S} last week killed^{222} at least 938 people^{T} and injured 400.

```
20010901 RUS CHE 223
20020804 ISR COL 073
20020804 -- COL 222
```

Dyadic event data (Serbia-Bosnia)

Week	Code	Description
1995-07-11	211	SEIZE POSSESSION
	212	ARREST PERSON
	223	MILITARY ENGAGEMENT
1995-07-12	211	SEIZE POSSESSION
	223	MILITARY ENGAGEMENT
	173	SPECIF THREAT
	191	CANCEL EVENT
	211	SEIZE POSSESSION
	095	PLEAD
	111	TURN DOWN
	212	ARREST PERSON
	081	MAKE AGREEMENT
	023	NEUTRAL COMMENT
	032	VISIT
	031	MEET

Scaled dyadic event data

Week	Code	Score [-10,10)
1995-07-11	211	-9.2
	212	-9.0
	223	-10.0
1995-07-12	211	-9.2
	223	-10.0
	173	-7.0
	191	-2.2
	211	-9.2
	095	1.2
	111	-4.0
	212	-9.0
	081	6.5
	023	-0.2
	032	1.9
	031	1.0

Summed scaled dyadic event data

The human elements

Coders read newswire and extract events
(e.g. GEDS projects, Swisspeace)

Experts assign scores to event types
(e.g. Goldstein 1995, Shellman 2004)

Analysts aggregate and infer conflict dynamics
(Goldstein & Pevehouse 1997, Pevehouse & Goldstein 1999)

Automating a conflict scale

Schrodt (2007) applied IRT models to event data

Motivation:

- International actions in a week are like roll-call votes in a parliament
- or like responses on a survey (with lots of questions)

Automating a conflict scale

Schrodt (2007) applied IRT models to event data

Motivation:

- International actions in a week are like roll-call votes in a parliament
- or like responses on a survey (with lots of questions)

But it didn't work

Automating a conflict scale

Schrodt (2007) applied IRT models to event data

Motivation:

- International actions in a week are like roll-call votes in a parliament
- or like responses on a survey (with lots of questions)

But it didn't work

There's a *substantive* reason for that

Two kinds of item structure

Dominance

IRT models used in voting, surveying, and testing applications assume that *perfect data* would have dominance structure

	'easy'							'hard'
	A	B	C	D	E	F	G	H
low θ	1	1	1	0	0	0	0	0
	1	1	1	1	0	0	0	0
	1	1	1	1	1	0	0	0
	1	1	1	1	1	1	0	0
	1	1	1	1	1	1	1	0
high θ	1	1	1	1	1	1	1	1

Two kinds of item structure

Proximity

Unfolding models used in ideal point estimation assume that *perfect data* would have a structure

	'left'							'right'
	A	B	C	D	E	F	G	H
'left' θ	1	1	1	0	0	0	0	0
	0	1	1	1	0	0	0	0
	0	0	1	1	1	0	0	0
	0	0	0	1	1	1	0	0
	0	0	0	0	1	1	1	0
'right' θ	0	0	0	0	0	1	1	1

Unfolding models for event data

Happily we have *several*, often used for text scaling:

$$\log E[Y_{ij}] = \alpha_i + \psi_j + \theta_i \beta_j \qquad \text{RC Model}$$

$$Y_{ij}/N = \alpha_i \psi_j (1 + \theta_i \beta_j) \qquad \text{CA}$$

Y_i is a table each week's events counts (a 'document')
Scale weeks θ_i and infer event types β_j

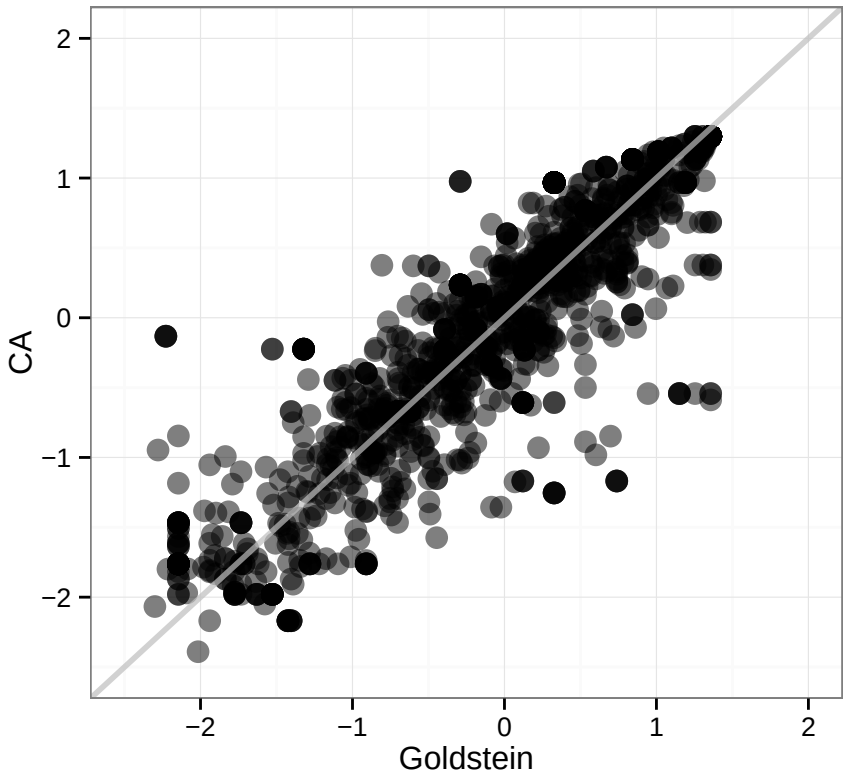
First example

Scaling CAMEO codes in the Middle East (ISR-PAL)

Aggregate event types up to 20 top level CAMEO categories

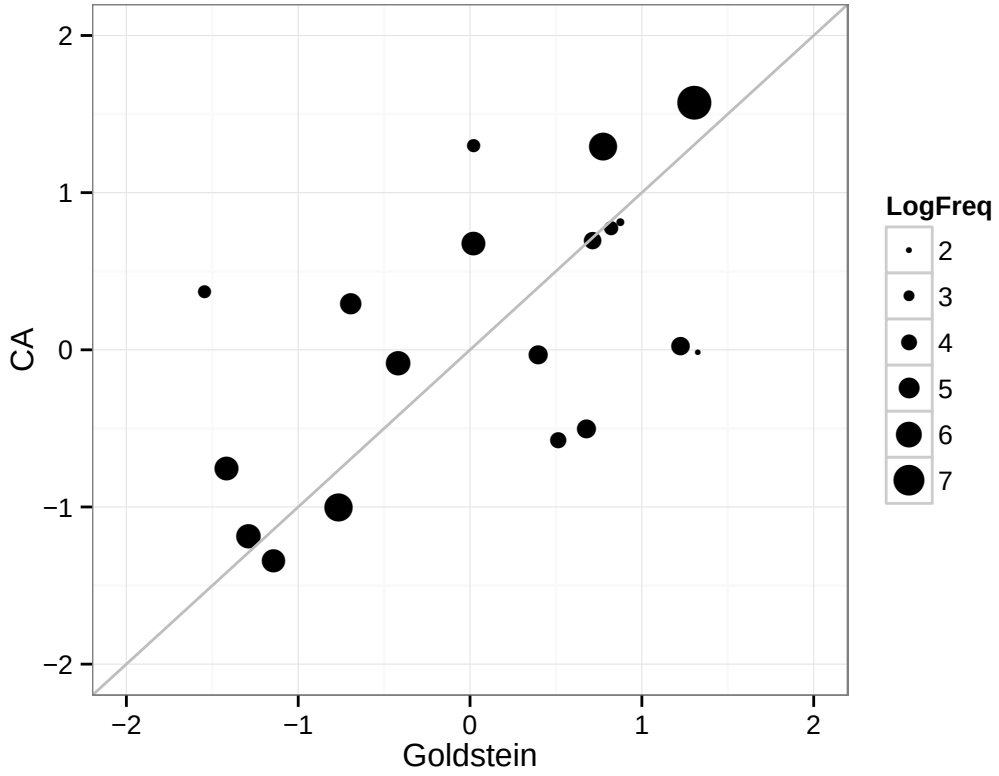
Compare to weekly averaged Goldstein scores (1979-2011)

Manual vs induced weekly conflict θ



$r = 0.89$

Item scores β for event types



Did we need all those categories?

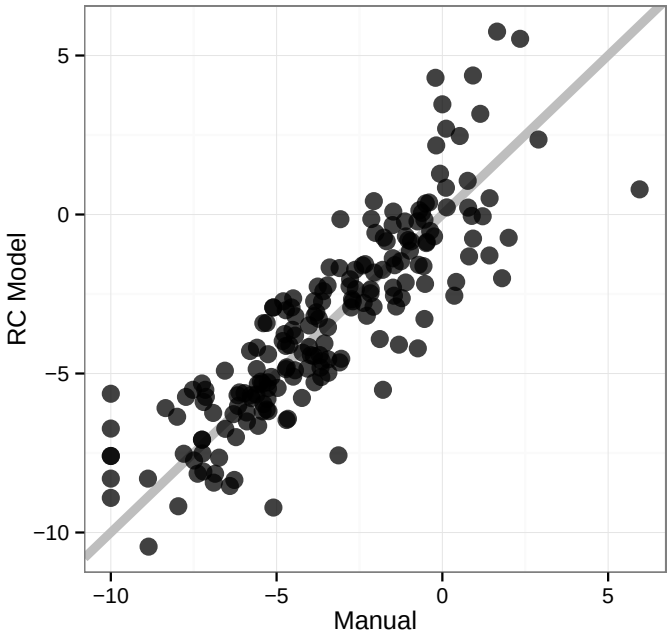
Scaling WEIS codes in Yugoslavia (SER-BOS)

Aggregate to 4 basic event categories (Schrodt, 2012)

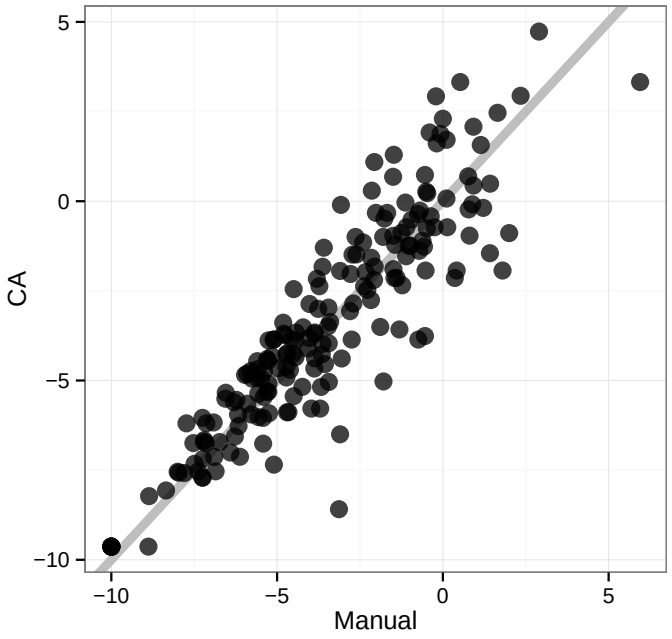
Week	Mat.Conf.	Mat.Coop.	Verb.Conf.	Verb.Coop.	N_t
⋮	⋮	⋮	⋮	⋮	⋮
1995-07-09	29	3	4	5	41
1995-07-16	21	4	12	13	50
1995-07-23	12	3	4	1	20
1995-07-30	4	2	2	4	12
⋮	⋮	⋮	⋮	⋮	⋮
	939	380	304	483	

Compare to weekly averaged Goldstein scores

Manual vs induced weekly conflict θ



r=0.84



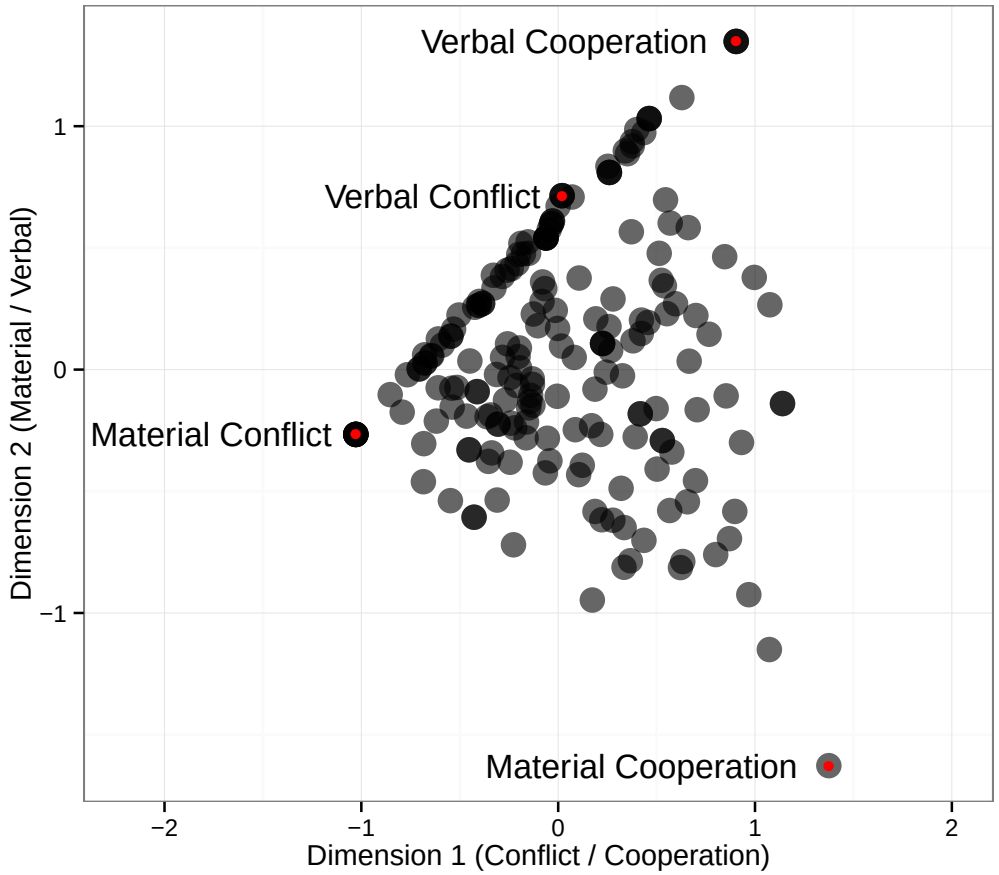
r=0.9

Item scores β for event types

This works because both models recover the type-averaged Goldstein scores

Event	Av. Goldstein	RC	CA
material conflict	-1.13	-1.42	-1.27
verbal conflict	-0.54	0.02	-0.28
verbal cooperation	0.72	0.60	0.55
material cooperation	0.95	0.80	1.00
		r=0.93	r=0.98

Item scores for event types



Automating a conflict scale

Those event coders and conflict experts were onto something. . .

and we understand it enough to automate it

Automating a conflict scale

Those event coders and conflict experts were onto something. . .

> and we understand it enough to automate it

Substantively

> expert-validated event scaling is now cheap and potentially important for bringing *context sensitivity* back into event data analysis

In conclusion

What are you going to scale today?

Texas A&M Jan 2015

CA as an (approximation to) unfolding

An unfolding type 'unimodal model' (ter Braak 1985)

$$\text{link}(\mu_{ki}) = a_k - \tfrac{1}{2}(x_i - u_k)^2/t_k^2,$$

CA transition eqns.

$$\lambda^{1-\alpha} x_i = \sum_k y_{ki} u_k / y_{+i} \quad (i = 1, \ldots, n),$$

$$\lambda^{\alpha} u_k = \sum_i y_{ki} x_i / y_{k+} \quad (k = 1, \ldots, m),$$

ML unimodal model update eqns.

$$x_i = \sum_k \frac{y_{ki} u_k}{t_k^2} \Big/ \sum_k \frac{y_{ki}}{t_k^2} - \left[\sum_k \frac{(x_i - u_k)\mu_{ki}}{t_k^2} \Big/ \sum_k \frac{y_{ki}}{t_k^2} \right],$$

$$u_k = \sum_i y_{ki} x_i / y_{k+} - \left[\sum_i (x_i - u_k)\mu_{ki} / y_{k+} \right].$$

see also (Heiser 1981, Polak et al. 2009)