Week 5: Scaling Latent Traits Using Texts

LSE MY459: Quantitative Text Analysis
https://lse-my459.github.io/

Ryan Hübert

Slides last updated: 17 February 2025

Outline for today

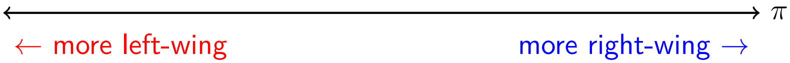
- ➜ What is scaling?
- ➜ Supervised scaling with Wordscores
- ➜ Unsupervised scaling with Wordfish and Wordshoal

What is scaling?

What is scaling?

Scaling is a set of quantitative tools for measuring **latent traits**, which are typically measured on continuous “scales”

- Classic example: political ideology (π) on “left-right” scale ($\mathbb{R}$)



π

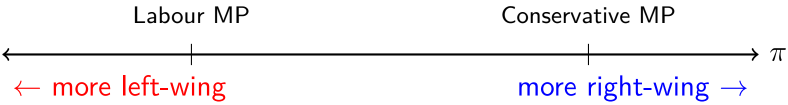
$\leftarrow$ more left-wing

more right-wing $\rightarrow$

What is scaling?

Scaling is a set of quantitative tools for measuring **latent traits**, which are typically measured on continuous "scales"

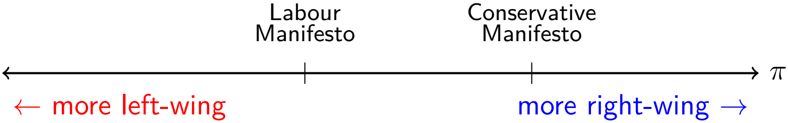
- ➜ Classic example: political ideology (π) on "left-right" scale ($\mathbb{R}$)
  - ➜ In political science: lots of research placing individual politicians on ideological scales



What is scaling?

Scaling is a set of quantitative tools for measuring **latent traits**, which are typically measured on continuous "scales"

- Classic example: political ideology (π) on "left-right" scale ($\mathbb{R}$)
  - Also—lots of research placing *documents* on ideological scales as well (Recall: texts are observable implications)



What is scaling?

But that's not all...

- ➜ Policy positions on economic vs social dimension
- ➜ Inter- and intra-party differences
- ➜ Soft news vs hard news
- ➜ Petitioner vs respondent in legal briefs
- ➜ ...and any other continuous scale

From a *methodological* perspective, scaling methods are quite similar to stuff we've learned so far

Supervised scaling with Wordscores

Supervised scaling methods

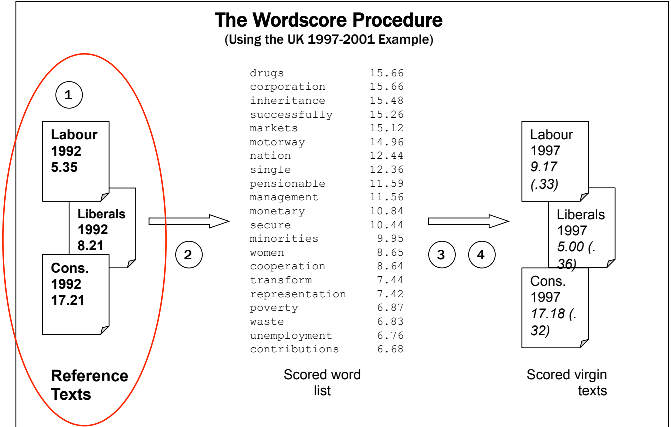
Wordscores method by Laver, Benoit and Garry (2003):

- Start with set of texts with *known* positions on one (or more) “left-right” scale
- Use these texts to measure the so-called *word scores* for each word used in these texts (this is the “supervision”)
- Use these word scores to create document-level scores for different set of texts with unknown positions

Very similar procedure to classification!

Source: https://doi.org/10.1017/S0003055403000698

The Wordscores procedure



The Wordscore Procedure

(Using the UK 1997-2001 Example)

1

Labour 1992 5.35

Liberals 1992 8.21

Cons. 1992 17.21

Reference Texts

2

drugs	15.66
corporation	15.66
inheritance	15.48
successfully	15.26
markets	15.12
motorway	14.96
nation	12.44
single	12.36
pensionable	11.59
management	11.56
monetary	10.84
secure	10.44
minorities	9.95
women	8.65
cooperation	8.64
transform	7.44
representation	7.42
poverty	6.87
waste	6.83
unemployment	6.76
contributions	6.68

Scored word list

3 4

Labour 1997 *9.17 (.33)*

Liberals 1997 *5.00 (. 36)*

Cons. 1997 *17.18 (. 32)*

Scored virgin texts

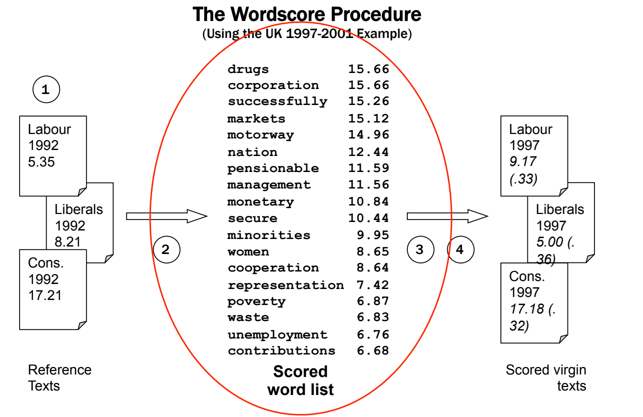
Step 1: Obtain reference texts with a priori known positions (`setref`)

Step 2: Generate word scores from reference texts (`wordscore`)

Step 3: Score each virgin text using word scores (`textscore`)

Step 4: (optional) Transform virgin text scores to original metric

The Wordscores procedure



The Wordscore Procedure

(Using the UK 1997-2001 Example)

drugs	15.66
corporation	15.66
successfully	15.26
markets	15.12
motorway	14.96
nation	12.44
pensionable	11.59
management	11.56
monetary	10.84
secure	10.44
minorities	9.95
women	8.65
cooperation	8.64
representation	7.42
poverty	6.87
waste	6.83
unemployment	6.76
contributions	6.68

Scored word list

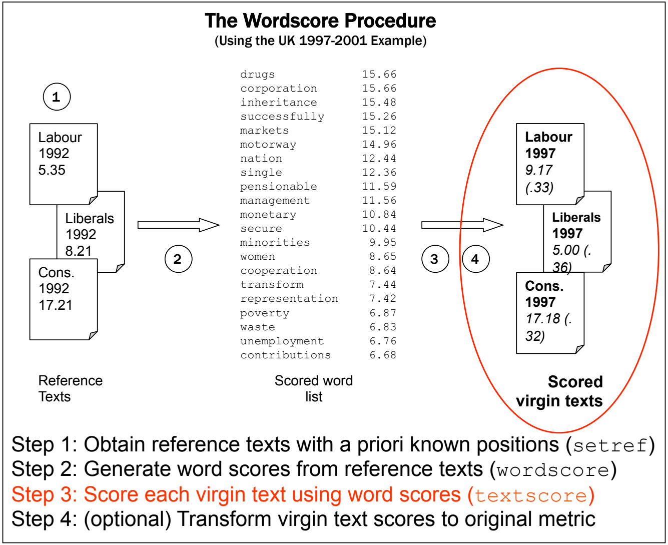
Step 1: Obtain reference texts with a priori known positions (`setref`)

Step 2: Generate word scores from reference texts (`wordscore`)

Step 3: Score each virgin text using word scores (`textscore`)

Step 4: (optional) Transform virgin text scores to original metric

The Wordscores procedure



The Wordscore Procedure

(Using the UK 1997-2001 Example)

1

Labour 1992 5.35

Liberals 1992 8.21

Cons. 1992 17.21

Reference Texts

2

drugs	15.66
corporation	15.66
inheritance	15.48
successfully	15.26
markets	15.12
motorway	14.96
nation	12.44
single	12.36
pensionable	11.59
management	11.56
monetary	10.84
secure	10.44
minorities	9.95
women	8.65
cooperation	8.64
transform	7.44
representation	7.42
poverty	6.87
waste	6.83
unemployment	6.76
contributions	6.68

Scored word list

3 4

Labour 1997 *9.17 (.33)*

Liberals 1997 *5.00 (. 36)*

Cons. 1997 *17.18 (. 32)*

Scored virgin texts

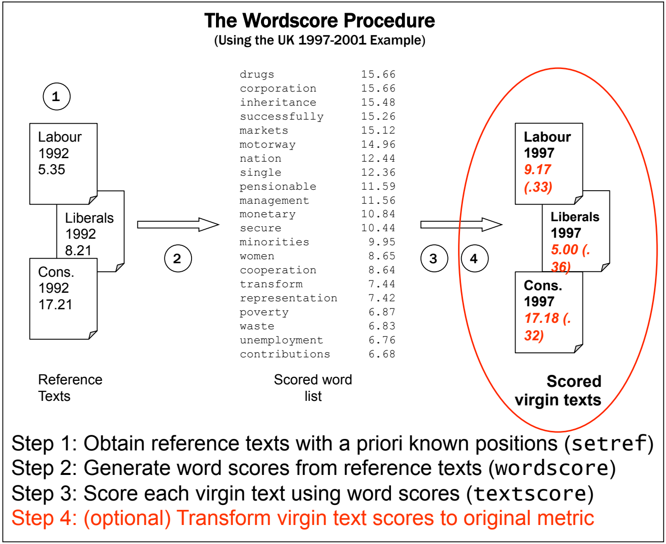
Step 1: Obtain reference texts with a priori known positions (`setref`)

Step 2: Generate word scores from reference texts (`wordscore`)

Step 3: Score each virgin text using word scores (`textscore`)

Step 4: (optional) Transform virgin text scores to original metric

The Wordscores procedure



The Wordscore Procedure

(Using the UK 1997-2001 Example)

1

Labour
1992
5.35

Liberals
1992
8.21

Cons.
1992
17.21

Reference
Texts

2

drugs	15.66
corporation	15.66
inheritance	15.48
successfully	15.26
markets	15.12
motorway	14.96
nation	12.44
single	12.36
pensionable	11.59
management	11.56
monetary	10.84
secure	10.44
minorities	9.95
women	8.65
cooperation	8.64
transform	7.44
representation	7.42
poverty	6.87
waste	6.83
unemployment	6.76
contributions	6.68

Scored word
list

3 4

Labour
1997
9.17
(.33)

Liberals
1997
5.00 (.
36)

Cons.
1997
17.18 (.
32)

Scored
virgin texts

Step 1: Obtain reference texts with a priori known positions (`setref`)
Step 2: Generate word scores from reference texts (`wordscore`)
Step 3: Score each virgin text using word scores (`textscore`)
Step 4: (optional) Transform virgin text scores to original metric

Calculating Wordscores: prerequisites

We begin with a set of texts divided into:

- The **labeled set** (or "reference texts" in LBG 2003)
- The **unlabeled set** (or "virgin texts" in LBG 2003)

Each text i in the labeled set must have a pre-defined "score," which we notate π_i, since it indicates a position on a scale

- It is a single number locating the text on a single dimension
- This can be on any scale, such as 1 to 20
- Can use arbitrary endpoints, such as –1 to 1

We do not know the scores of the unlabeled set, as that's what we're trying to learn!

Calculating Wordscores: prerequisites

Running example: corpus of UK manifestos available in `quanteda`

```
Corpus consisting of 6 documents and 6 docvars.
Con_1992 :
"Conservative Party 1992  The Best Future for Britain FOREWOR..."

Lab_1992 :
" THE LABOUR PARTY MANIFESTO TIME TO GET BRITAIN WORKING AGAI..."

LD_1992 :
"1992 CHANGING BRITAIN FOR GOOD AFTER THE HEADLINE  "BE WARNE..."

Con_1997 :
"THE CONSERVATIVE MANIFESTO 1997 FOREWORD THE CONSERVATIVE AD..."

Lab_1997 :
"BRITAIN WILL BE BETTER WITH NEW LABOUR 'OUR CASE IS SIMPLE: ..."

LD_1997 :
"make the difference The Liberal Democrat Manifesto 1997 £2.4..."
```

Calculating Wordscores: prerequisites

- We'll use the 1992 documents to create our word scores ("train")
- Then we'll apply estimated wordscores to 1997 documents to measure their ideology
- We'll assume these scores for the labeled set:
  - Conservative 1992: 17.21
  - Labour 1992: 5.35
  - Liberal Democrat 1992: 8.21
- Where do these come from? Expert coding. . . (more later)

Calculating Wordscores: pre-processing

Convert the *labeled* texts into a document-feature matrix

- ➜ You do not need to do stemming or remove stop words, but you can if you want (see Lowe and Benoit 2013)
- ➜ Just be consistent and do what makes sense for your task

We will use the usual notation we've developed so far:

- ➜ $\mathbf{W}$ is an $N \times J$ document feature matrix
- ➜ $\mathbf{W}_i$ is a specific document (row) of the DFM
- ➜ $M_i = \sum_j \mathbf{W}_i$ is the document length for document i

Keep in mind: for now, $\mathbf{W}$ is a DFM for the labeled documents only

Calculating Wordscores: pre-processing

```
Document-feature matrix of: 3 documents, 7,103 features
  (58.20% sparse) and 6 docvars.
          features
docs        conservative party best future britain
  Con_1992            22    14   38     35      76
  Lab_1992             6     6   10     13      36
  LD_1992              5     4   11     19      40
[ reached max_nfeat ... 7,098 more features ]
```

Calculating Wordscores: normalise DFM

Begin the estimation process by normalising the DFM

- ➜ Convert $\mathbf{W}$ into a **relative document-feature matrix** $\mathbf{F}$, by dividing each W_{ij} by its **word total marginals**
- ➜ In other words, divide each row by the sum of that row

So, each element of $\mathbf{F}$ is calculated by

$$F_{ij} = \frac{W_{ij}}{M_i} = \frac{W_{ij}}{\sum_j W_{ij}}$$

Calculating Wordscores: normalise DFM

```
Document-feature matrix of: 3 documents, 5,419 features
  (45.21% sparse) and 6 docvars.
          features
docs        conservative   party    best  future britain
  Con_1992       0.00113 0.00072 0.00195 0.00179 0.00389
  Lab_1992       0.00077 0.00077 0.00129 0.00168 0.00465
  LD_1992        0.00042 0.00033 0.00092 0.00159 0.00334
[ reached max_nfeat ... 5,414 more features ]
```

Calculating Wordscores: document probs. in labeled set

Using $\mathbf{F}$, next we compute the **relative document probabilities** for each of the labeled texts

Do this by creating an $N \times J$ matrix $\mathbf{P}$, where each element P_{ij} is

$$P_{ij} = \frac{F_{ij}}{F_j} = \frac{F_{ij}}{\sum_i F_{ij}}$$

➜ For each document i and word j, P_{ij} is the probability that we are reading a certain reference document i if it contains j

Calculating Wordscores: document probs. in labeled set

```
Document-feature matrix of: 3 documents, 5,419 features
  (45.21% sparse) and 6 docvars.
          features
docs        conservative   party    best  future britain
  Con_1992       0.48577 0.39264 0.46816 0.35427 0.32750
  Lab_1992       0.33395 0.42417 0.31055 0.33169 0.39104
  LD_1992        0.18028 0.18319 0.22129 0.31404 0.28146
[ reached max_nfeat ... 5,414 more features ]
```

Calculating Wordscores: calculate word scores

Using $\mathbf{P}$, we can compute a J-length vector S that gives the "word score" for each word

Specifically, each element in the vector is calculated as follows:

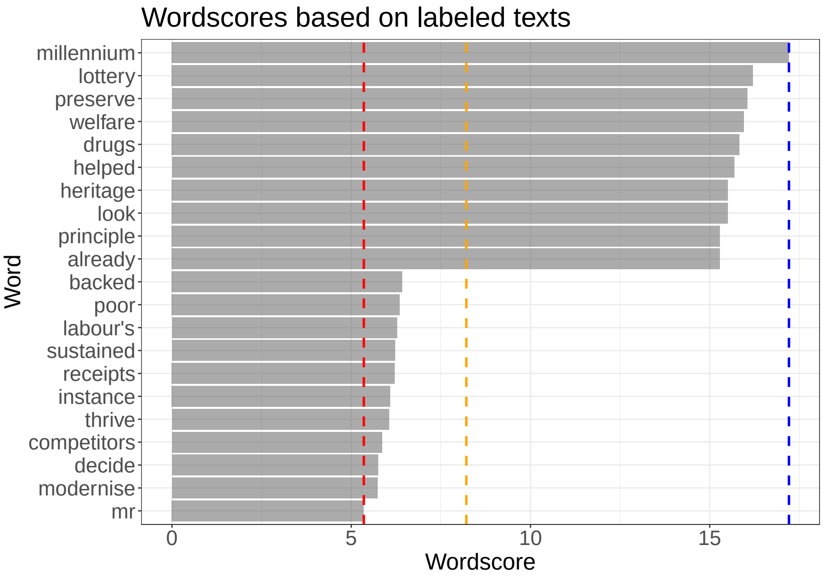
$$s_j = \sum_i (\pi_i \times P_{ij})$$

➜ s_j is the average of each document i's scores π_i, weighted by each word's P_{ij}

Calculating Wordscores: calculate word scores

```
# A tibble: 5,419 x 2
   word         word.score
   <chr>             <dbl>
 1 conservative      11.6
 2 party             10.5
 3 best              11.5
 4 future            10.4
 5 britain           10.0
 6 foreword           8.72
 7 end                8.87
 8 parliament         9.94
 9 new               10.1
10 millennium        17.2
# i 5,409 more rows
```

Calculating Wordscores: calculate word scores



Calculating Wordscores: apply to unlabeled texts

Now, we obtain a *single* score for each unlabeled text, relative to the labeled texts

For each unlabeled document: take the mean of the word scores of its words, weighted by their term frequency:

1. Create a matrix $\mathbf{F}^{\mathrm{u}}$ for the unlabeled documents the same way you did for the labeled documents
   - ➜ Important: only include features in $\mathbf{F}^{\mathrm{u}}$ that are also in the feature set for the labeled documents
   - ➜ Any new words that are not in the labeled set are ignored
2. Calculate the score of unlabeled document i as follows:

$$\hat{\pi}_i = \sum_j (F_{ij}^{\mathrm{u}} \times s_j)$$

Calculating Wordscores: apply to unlabeled texts

```
Con_1997 Lab_1997  LD_1997
10.88510 10.39209 10.35097
```

Calculating Wordscores: rescale scores for unlabeled set

Unfortunately, these scores aren't on same scale as labeled set

- Notice they are bunched together
- This is because there are a lot of "nondiscriminating" words pulling the estimates toward the middle

We rescale them in a couple ways:

- Original LBG paper proposes one way
- Martin and Vanberg (2007) propose another

Won't get into details, but can do both in `quanteda`

Calculating Wordscores: rescale scores for unlabeled set

No rescaling (option `rescaling="none"`):

```
Con_1997 Lab_1997  LD_1997
10.88510 10.39209 10.35097
```

LBG rescaling (option `rescaling="lbg"`):

```
 Con_1997  Lab_1997   LD_1997
17.672301  7.406080  6.549772
```

Martin-Vanberg rescaling (option `rescaling="mv"`)::

```
 Con_1997  Lab_1997   LD_1997
17.210000  5.350000  4.360755
```

Calculating Wordscores: other considerations

You can do a couple other things:

1. You will likely want to provide confidence intervals for your estimates on the unlabeled set
   - ➜ Can use an analytical approach (as in LBG 2003)
   - ➜ Or you can bootstrap (as in Lowe and Benoit 2013)
2. You may wish to smooth the labeled DFM to deal with zeroes

Getting labeled texts

What kind of considerations go into choosing labeled texts?

1. Labeled set should contain texts that clearly represent an a priori known dimension
2. Labeled texts should be as discriminating as possible, e.g. “extreme” texts
3. Labeled texts must contain lots of words
4. Labeled & unlabeled texts must come from same lexical universe

What about the labels (i.e., positions on scale)? Some latitude here, but most important consideration is use case:

- ➜ Choose labels on pre-existing scale for interpretabilty?
- ➜ Can always rescale, e.g. from –1 to 1

Linking Wordscores to a language model

There's a link between Naive Bayes classification and Wordscores

For a word w_j, define the following posterior probability of it being in document d_i as:

$$\Pr(d_i|w_j) = \frac{\Pr(w_j|d_i)\Pr(d_i)}{\Pr(w_j)}$$

Note that $\Pr(d_i|w_j)$ is just P_{ij} from the $\mathbf{P}$ matrix above

Ordering reference scores all N documents in the labeled set such that $\pi_1 < \pi_2 < \cdots < \pi_N$, then the word score for word w_j is:

$$s_j = \pi_1 P_{1j} + \pi_2 P_{2j} + \cdots + \pi_N P_{Nj} = \sum_{i=1}^{N-1} P_{ij}(\pi_i - \pi_N) + \pi_N$$

Linking Wordscores to a language model

The reference scores are on an arbitrary scale, so why not rescale the labeled documents so that $\pi_1 = -1.0$ and $\pi_N = 1.0$?

For word j, we get the "simple" word score of

$$s_j = (1 - 2P_{1j}) + \sum_{i=2}^{N-1} P_i(\pi_i - 1)$$

With just two labeled documents, this becomes:

$$s_j = 1 - 2P_{1j}$$

Implies: LBG's "word scores" come from a linear combination of class posterior probabilities from a Bayesian model!

Unsupervised scaling with Wordfish and Wordshoal

Unsupervised methods scale *distance*

Unsupervised scaling with DFMs measure *similarity* or *distance* in feature use

- ➜ We'll discuss this in more detail next lecture

Fundamental problem: *distance on which scale?*

- ➜ Ideally, something we care about, e.g. policy positions, ideology, preferences, sentiment
- ➜ But often other dimensions (language, rhetoric style, authorship) are more predictive
- ➜ First dimension in unsupervised scaling will capture main source of variation, whatever that is

Unlike supervised models, validation comes *after* estimation

Unsupervised scaling methods

Goal: unsupervised scaling of ideological positions

Two main approaches

- **Parametric methods** model feature occurrence according to some distribution, e.g. Poisson distribution
  - word effects and “positional” effects are unobserved parameters to be estimated
  - e.g. Wordfish (Slapin and Proksch 2008) and Wordshoal (Lauderdale and Herzog 2016)
- **Non-parametric methods** typically based on the Singular Value Decomposition of a matrix
  - For lack of time, we won’t cover these here

Wordfish

Slapin and Proksch (2008) introduce **Wordfish**

Built on a model of language (albeit a different one than we've seen)

Assume that the frequency with which politician i uses word j is drawn from a **Poisson distribution**

$$W_{ij} \sim \text{Poisson}(\lambda_{ij}),\ \text{where } \lambda_{ij} = \exp(\alpha_i + \psi_j + \mu_j \pi_i)$$

- ➜ α_i is "loquaciousness" of politician i (document fixed effect)
- ➜ ψ_j is frequency of word j (word fixed effect)
- ➜ μ_j is "discrimination parameter" of word j
- ➜ π_i is the politician's ideological position

Source: https://doi.org/10.1111/j.1540-5907.2008.00338.x

Wordfish

Recall the idea behind language models:

- ➜ They encode ideas about the data generating process for texts
- ➜ They're *always* simplifications, and to some extent "wrong"

Why not use multinomial model from before?

- ➜ Short answer: that's not what Slapin and Proksch used!
- ➜ We want to estimate "ideology" but the multinomial model doesn't have enough parameters—it's underspecified
- ➜ Poisson is easier to estimate in this context

Why Poisson specifically?

- ➜ It is a commonly used way to model the data generating process for *count data*
- ➜ DFMs are datasets of *word* count data!

How to estimate this model

Conditional maximum likelihood estimation:

➜ If we knew ψ and μ (the word parameters) then we have a Poisson regression model

➜ If we knew α and π (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 `quanteda` package as `textmodel_wordfish()`

(An alternative is MCMC with a Bayesian formulation or variational inference using an Expectation-Maximization algorithm (Imai et al. 2016))

After estimation, then identification

After estimation, you get a model

$$\hat{\lambda}_{ij} = \exp(\hat{\alpha}_i + \hat{\psi}_j + \hat{\mu}_j \hat{\pi}_i)$$

We care about the $\hat{\pi}_i$ parameters—we'll interpret as our latent trait!

But the *scale* and *direction* of π is undetermined

We need to "identify" the model to make it interpretable

- ➜ **"Simple" normalisation**: fix the scale by, for example, normalising $\hat{\pi}_i$s to have mean 0 and variance 1
- ➜ **Anchoring**: set the "left-right" direction by, for example fixing values of two $\hat{\pi}_i$s (e.g, to –1 and 1) and normalise remaining $\hat{\pi}_i$s on that scale

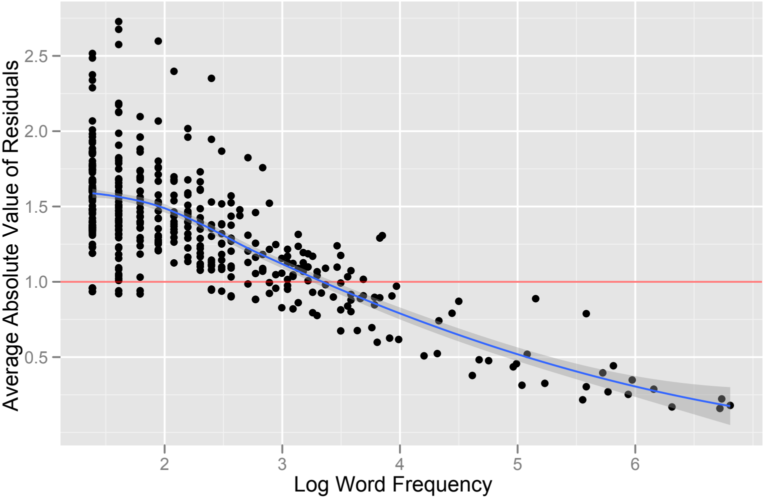
Some nice features of the parametric scaling approach

1. Can do standard (statistical) inference about parameters, e.g. estimates of uncertainty
2. The distributional assumptions are made explicit as part of the data generating process motivating the choice of stochastic distribution
3. Allows for hierarchical reparameterization, for example to add covariates
4. Generative model: given the estimated parameters, we could generate a document for any specified length

But some reasons why this model is "wrong"

1. Like all language models we've discussed: conditional independence of words is heroic!
   - ➜ But we will keep calm and carry on
2. Method produces heteroskedastic errors
   - ➜ Overdispersion – occurs when "informative" words tend to cluster together
   - ➜ Underdispersion – could (possibly) occur when high frequency words are uninformative and have relatively low between-text variation (once length is considered)

Overdispersion in German manifesto data



Data from Slapin and Proksch (2008)

One solution to model overdispersion

Lo, Proksch, and Slapin (2014) suggest using a negative binomial model (instead of a Poisson model):

$$W_{ij} \sim \mathrm{NB}\left(r_i, \frac{\lambda_{ij}}{\lambda_{ij} + r_i}\right), \text{ where } \lambda_{ij} = \exp(\alpha_i + \psi_j + \mu_j \pi_i)$$

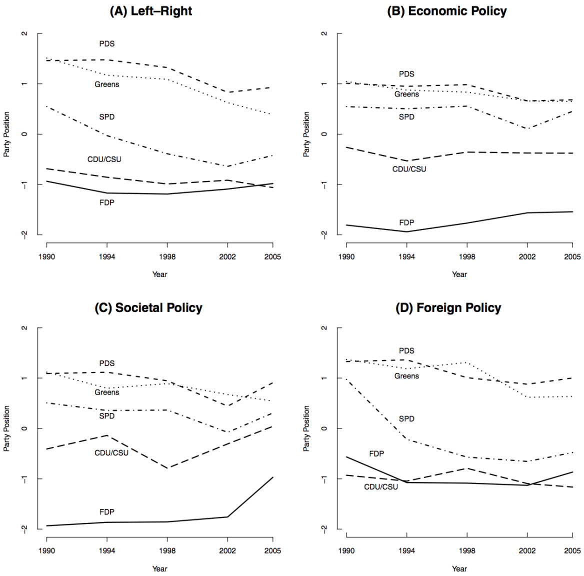
and where r_i is a variance inflation parameter varying across documents

Technical fix that can also have a substantive interpretation, e.g., “ideological ambiguity”

Source: https://doi.org/10.1017/S0007123414000192

Example from Slapin and Proksch (2008)

FIGURE 1 Estimated Party Positions in Germany, 1990–2005



(A) Left–Right

(B) Economic Policy

(C) Societal Policy

(D) Foreign Policy

Example from Slapin and Proksch (2008)

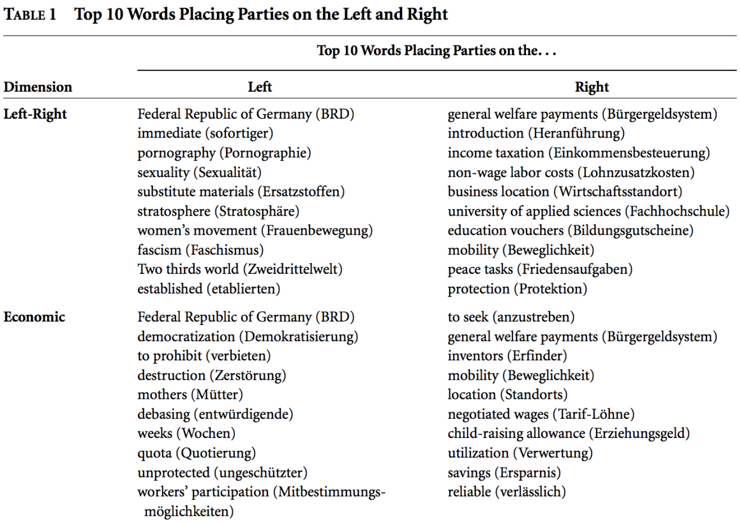


TABLE 1 Top 10 Words Placing Parties on the Left and Right

	Top 10 Words Placing Parties on the. . .	
Dimension	**Left**	**Right**
Left-Right	Federal Republic of Germany (BRD)	general welfare payments (Bürgergeldsystem)
	immediate (sofortiger)	introduction (Heranführung)
	pornography (Pornographie)	income taxation (Einkommensbesteuerung)
	sexuality (Sexualität)	non-wage labor costs (Lohnzusatzkosten)
	substitute materials (Ersatzstoffen)	business location (Wirtschaftsstandort)
	stratosphere (Stratosphäre)	university of applied sciences (Fachhochschule)
	women's movement (Frauenbewegung)	education vouchers (Bildungsgutscheine)
	fascism (Faschismus)	mobility (Beweglichkeit)
	Two thirds world (Zweidrittelwelt)	peace tasks (Friedensaufgaben)
	established (etablierten)	protection (Protektion)
Economic	Federal Republic of Germany (BRD)	to seek (anzustreben)
	democratization (Demokratisierung)	general welfare payments (Bürgergeldsystem)
	to prohibit (verbieten)	inventors (Erfinder)
	destruction (Zerstörung)	mobility (Beweglichkeit)
	mothers (Mütter)	location (Standorts)
	debasing (entwürdigende)	negotiated wages (Tarif-Löhne)
	weeks (Wochen)	child-raising allowance (Erziehungsgeld)
	quota (Quotierung)	utilization (Verwertung)
	unprotected (ungeschützter)	savings (Ersparnis)
	workers' participation (Mitbestimmungsmöglichkeiten)	reliable (verlässlich)

Example from Slapin and Proksch (2008)

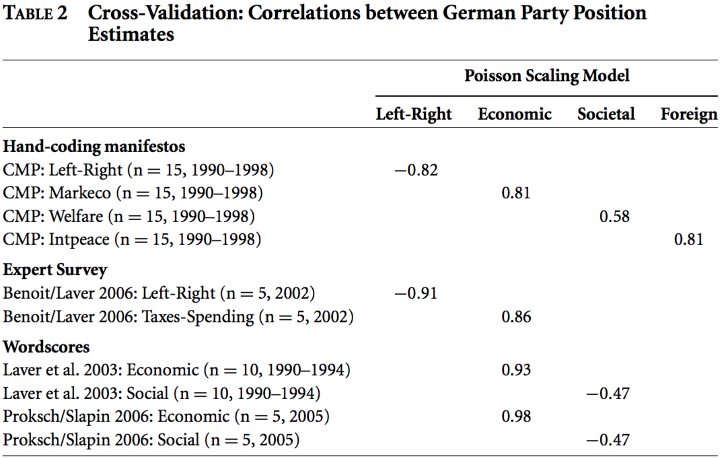


TABLE 2 Cross-Validation: Correlations between German Party Position Estimates

	Poisson Scaling Model			
	Left-Right	**Economic**	**Societal**	**Foreign**
Hand-coding manifestos				
CMP: Left-Right (n = 15, 1990–1998)	−0.82			
CMP: Markeco (n = 15, 1990–1998)		0.81		
CMP: Welfare (n = 15, 1990–1998)			0.58	
CMP: Intpeace (n = 15, 1990–1998)				0.81
Expert Survey				
Benoit/Laver 2006: Left-Right (n = 5, 2002)	−0.91			
Benoit/Laver 2006: Taxes-Spending (n = 5, 2002)		0.86		
Wordscores				
Laver et al. 2003: Economic (n = 10, 1990–1994)		0.93		
Laver et al. 2003: Social (n = 10, 1990–1994)			−0.47	
Proksch/Slapin 2006: Economic (n = 5, 2005)		0.98		
Proksch/Slapin 2006: Social (n = 5, 2005)			−0.47	

Wordshoal

Two key limitations of Wordfish applied to legislative text:

1. Word discrimination parameters assumed to be constant across debates (unrealistic, think e.g. “debt”)
2. May not capture left-right ideology but topic variation

Slapin and Proksch partially avoid these issues by scaling different types of debates separately

But resulting estimates are confined to set of speakers who spoke on each topic

Wordshoal by Lauderdale and Herzog (2016) suggests aggregate debate-specific ideal points into a reduced number of scales

Source: https://doi.org/10.1093/pan/mpw017

Wordshoal

The frequency with which politician i uses word j in debate k is drawn from a **Poisson distribution**:

$$w_{ijk} \sim \text{Poisson}(\lambda_{ijk}), \text{ where } \lambda_{ijk} = \exp(\alpha_{ik} + \psi_{jk} + \mu_{jk}\pi_{ik})$$
$$\text{and } \pi_{ik} \sim \mathcal{N}(\nu_k + \kappa_k\theta_i, \tau_i)$$

With parameters:

- ➜ α_{ik} is "loquaciousness" of politician i in debate k
- ➜ ψ_{jk} is frequency of word j in debate k
- ➜ μ_{jk} is discrimination parameter of word j in debate k
- ➜ π_{ik} is the politician's ideological position in debate k
- ➜ ν_k is baseline ideological position of debate k
- ➜ κ_k is correlation of debate k with common dimension
- ➜ θ_i is overall ideological position of politician i

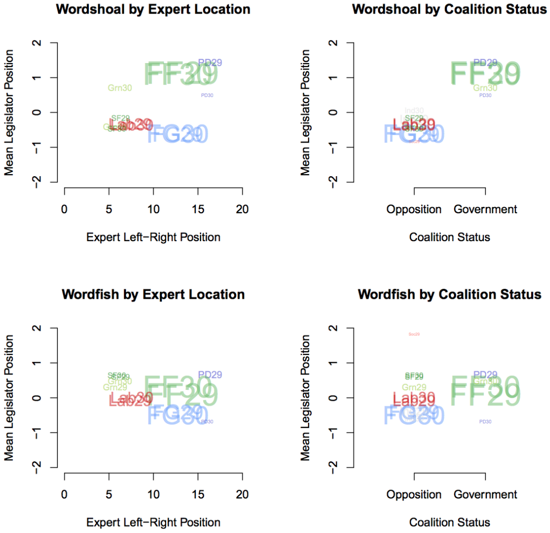
Intuition: debate-specific estimates are aggregated into a single position using dimensionality reduction

Wordshoal

New quantities of interest to estimate:

- Politicians' *overall* position versus their *debate-specific* positions
- Strength of association between ideological scale of debate and the general ideological scale
- Association of words with general scales, and stability of word discrimination parameters across debates

Example from Lauderdale and Herzog (2016)



Wordshoal by Expert Location

Wordshoal by Coalition Status

Wordfish by Expert Location

Wordfish by Coalition Status

Example from Lauderdale and Herzog (2016)

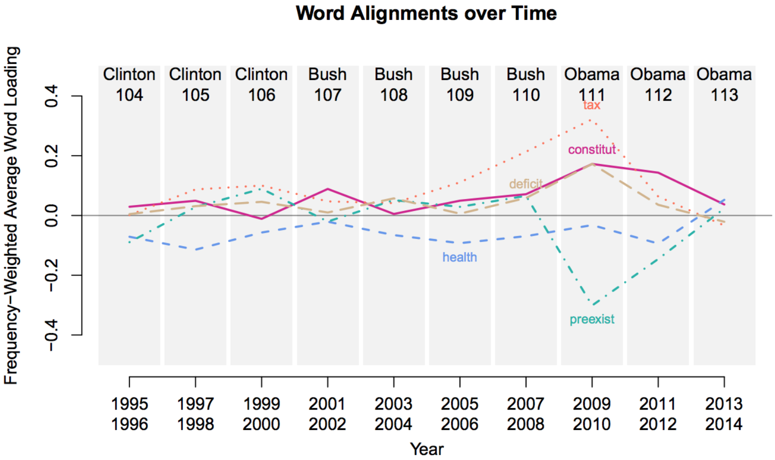
Reproducing Table 2 from Lauderdale and Herzog (2016)

- ➜ Five most and least “polarising” debates from the 30th Dáil

High government-opposition polarization	κ_k
Social Welfare and Pensions (No. 2) Bill 2009 (Second Stage)	0.942
Early Childhood Care and Education (Motion)	0.887
Private Members’ Business - Vaccination Programme (Motion)	0.824
Capitation Grants (Motion)	0.819
Confidence in Government (Motion)	0.814
Low government-opposition polarization	
Cancer Services Reports (Motion)	0.003
Finance (No. 2) Bill 2007 (Committee and Remaining Stages)	0.002
Finance Bill 2011 (Report and Final Stages)	0.002
Private Members’ Business - Mortgage Arrears (Motion)	0.002
Wildlife (Amendment) Bill 2010 (Committee and Remaining Stages)	0.001

Example from Lauderdale and Herzog (2016)

Word loading: “Association of word with general scale across debates”


Word Alignments over Time
Frequency−Weighted Average Word Loading
Clinton 104
Clinton 105
Clinton 106
Bush 107
Bush 108
Bush 109
Bush 110
Obama 111
Obama 112
Obama 113
tax
constitut
deficit
health
preexist
Year