Polimetrics

Lecture 5
Wordfish

From words to preferences

approach
evidence
collect
warehouse
derived
related
processed
describes
much
lives
scale
small
new
system
whether
smaller
social
created
term
challenges
set
storage
anything
enough
continues
often
insights
things
size
velocity
scope
analytics
sources
dataset
able
services
systems
something
otherwise
ethical
software
use
commerce
tools
excel
growth
opportunity
across
disparate
requires
many
questions
amounts
technologies
computational
different
aspect
later
reality
made
volume
possibilities
get
solve
main
meaning
computing
techniques
collecting
variety
capable
maybe
analysis
technology
value
decision
refers
diverse
really
machine
think
world
right
like
large
just
relative
relationships
using
now
understand
particular
need
problem
important
help
means
actually
technological
datasets
complex
deal
business
information
society
standard
advances
capturing
organizations
number
processing
changes
way
used
analytical
definition
processes
time
change
decisions
solutions
essential
lot
every
fit
making
tb
store
patterns
single
make
databases
storing
buzzword
traditional
products
internet
methods
mean
organization
behavior
people
memory
also
issues
tool
gather
better

Wordfish

Computer-based content analysis provides an alternative and systematic way to estimate **party positions** from **political texts**

In the last years, a growing increase of methods that saves to the researcher the huge cost of human coded content analysis

Here we will focus on **Wordfish**

Wordfish

The Wordfish technique treats ideology as a **latent variable**. This means that ideology is not something that the researcher can **directly observe**, rather it must be indirectly estimated based upon **observable actions** taken by parties and their members

The observable action we are most concerned with here is the writing of **documents** (such as election manifestos) and/or the **speeches** given by politicians in some given circumstances (such as legislative speeches)

Wordfish assumes that the **language** used by political parties expresses political ideology, that is…

…**Ideology** manifests itself in the **word choice** of politicians when writing party documents or saying something for example

Wordfish

More specifically, Wordfish assumes that parties' **relative word** usage within party documents conveys information about their positions in a policy space

To give an example, the technique assumes that if one party uses the word '**freedom**' more frequently than the word '**equality**' in a document on economic policy while another party uses 'equality' more often than 'freedom' in a similar document, these **two words** – 'equality' and 'freedom' – **provide information** about party ideology with regard to the economic policy dimension, and **discriminate** between the parties

Wordfish

The interpretation of the **estimated dimension** in Wordfish is left to the researcher (contrary to other quantitative position estimation techniques such as **Wordscores**)

In the previous example, Wordfish does not tell the researcher whether ‘equality’ is a ‘left-wing word’ while ‘freedom’ is a ‘right-wing word’

The algorithm will simply use the relative frequencies of these words as data to locate the documents on a scale, and it is up to the researcher to make an assessment about what constitutes ‘left’ and ‘right’ based upon her **knowledge of politics** (*a-posteriori* method!)

Wordfish

Critics of word frequency-based approaches are quick to point out that such algorithms are **ignorant of sentence structure and context**

For instance, the expressions '*We are against lowering taxes, and for tax increases*' and '*We are for lowering taxes, and against tax increases*' use the exact same words with the **same frequencies**, even though the meaning is reversed.

A word frequency approach used on only these statements, however, will provide **identical estimates**!

While this may indeed be cause for concern for **short statements**, this is not necessarily problematic for the analysis of **long texts** such as election manifestos

Wordfish

Critics of word frequency-based approaches are quick to point out that such algorithms are **ignorant of sentence structure and context**

For instance, the expressions ‘*We are against lowering taxes, and for tax increases*’ and ‘*We are for lowering taxes, and against tax increases*’ use the exact same words with the **same frequencies**, even though the meaning is reversed.

Furthermore, while the word tax won’t help to discriminate the two texts, other words can do that: e.g. *cut taxes to cut public debt* OR *cut taxes to raise welfare expenditure*

Wordfish Estimation Process

Wordfish uses a **bag-of-words approach**: a text is represented as a **vector of word counts** or **occurrences**

Multiple document vectors are then put together in a **term-document matrix**, where each **column** represents a document and each **row** represents a unique word, or term

The cells of the matrix contain the **number of times** the unique word (term) is mentioned in each document

Data matrix: an example

CASE_LBL	LDP	JSP	KOMEI	DSP
総理	0	38	29	23
国民	11	18	23	22
基地	0	3	47	3
政府	6	21	8	15
大学	7	4	1	38
わが国	10	2	12	7
政治	3	6	6	15
大臣	0	8	1	18
経済	8	9	2	6
安保	0	4	10	8
ゴルフ	0	0	21	0
社会	0	10	8	3
政策	2	8	1	7
物価	2	3	8	5
学生	1	2	1	13
本土	1	5	3	8
予算	1	11	2	3
考え	5	3	5	3
方式	0	0	0	16
公害	0	0	15	0

Wordfish Estimation Process

The **order of words** is lost and elements in the matrix simply represent the **term frequency**

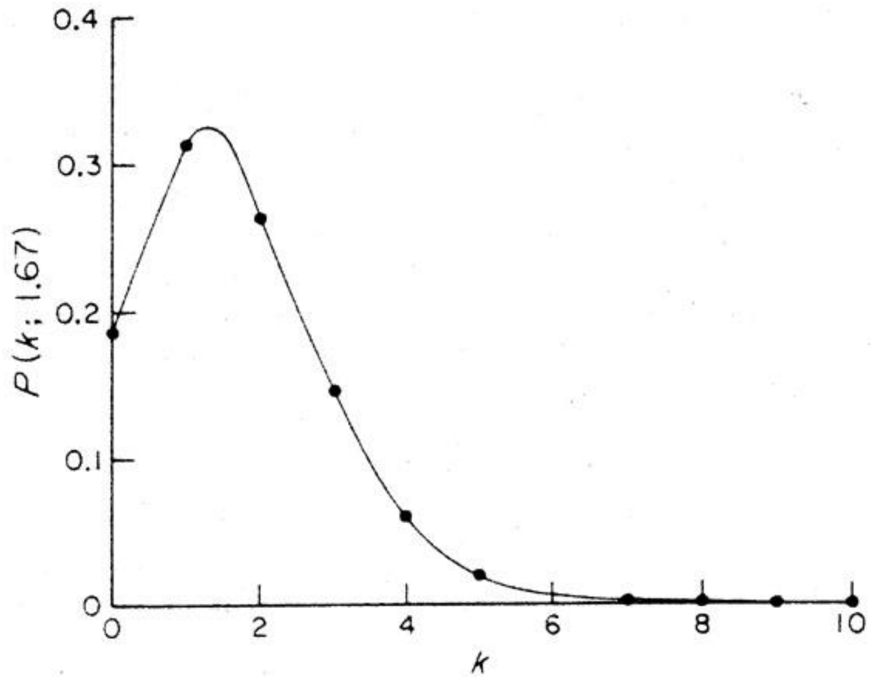
(and Wordfish assumes that each word is independent from the others… which is untrue, still…not an issue)

Therefore, this approach assumes **individual words** are distributed at **random throughout a text**

Wordfish Estimation Process

But which are the statistical distributions which most accurately approximate word usage?

Wordfish assumes that word frequencies (the number of times an actor *i* mentions word *j)* are generated by/drawn from a **Poisson process**, a distribution that is **heavily skewed**, as is the case of word usage



Wordfish Estimation Process

The **systematic component** of this process contains four parameters: word fixed effects, document (party) fixed effects, document (party) positions, word weights (discriminating parameters)

Word fixed effects are included to capture the fact that some words need to be used much more often in a language. Such words may serve a grammatical purpose but they have no substantive or ideological meaning, such as conjunctions or definite and indefinite articles

The **document fixed effect** parameters control for the possibility that some documents in the analysis may be significantly longer than others. When using manifestos to estimate party positions, for example, this can happen when some parties in some years write much longer manifestos

Wordfish Estimation Process

The **document positions parameters** tells us the positions of the parties relative to the other parties in the political space

The **word discrimination parameters** allow the researcher to analyze **which words help to differentiate party positions**

In previous example, ‘equality’ would have a high absolute value for its discrimination value and its usage would most likely be associated with left-wing parties. The word ‘freedom’ would also have a high absolute value but with the opposite sign because its usage would be associated with right-wing parties

This allows the researcher to estimate party positions and uncover the variations in political language that are responsible for placing parties on this dimension

More formally

Formally the functional form of the model is as follows:

$y_{ijt} \approx POISSON(\lambda_{ijt})$ where y_{ijt} is the count of word *j* in party *i*'s manifesto (or speech) at time *t*

The lambda parameter λ has the following systematic component:

$$\lambda_{ijt} = exp(\alpha_{it} + \psi_i + \beta_i * \omega_{it})$$

with α as a set of **document fixed effects at time t**, Ψ as a set of **word fixed effects**, β as estimates of **word specific weights** capturing the importance of word *j* in discriminating between documents (manifestoes or speeches), and ω as the estimate of party *i*'s **position** at time *t* (therefore *it* is indexing one specific document)

More formally

WORDFISH uses an expectation maximization (EM) algorithm to retrieve maximum likelihood estimates for all parameters (not so true… it’s a multinomial response model, still, not an issue)

The implementation of this algorithm entails an iterative process:

first party parameters are held fixed at a certain value while word parameters are estimated, **then** word parameters are held fixed at their new values while the party positions are estimated

This process is **repeated until the parameter estimates** reach an acceptable level of convergence

Conditions for using Wordfish

The model specification used by Wordfish works best as **more data is available**, meaning as **more documents** are used in the analysis and as those documents contain **more unique words**

If the documents do not contain a sufficient number of unique words, there will not be adequate information to estimate document parameters

Conditions for using Wordfish

The **Challenge of Dynamic Estimation**: Using text to estimate party positions **over time** creates an additional challenge. On the one hand, we would like to use as much information in the texts as possible. On the other hand, we would like to estimate position change over time. This is a trade-off

For example, if the **political debate changes and new vocabulary** enters the political lexicon in election *t*, then this will differentiate the texts at point *t* from those at point *t-1*

In fact, in this instance, we are likely to pick up a **policy agenda** shift in texts, whereas we are interested in party position change

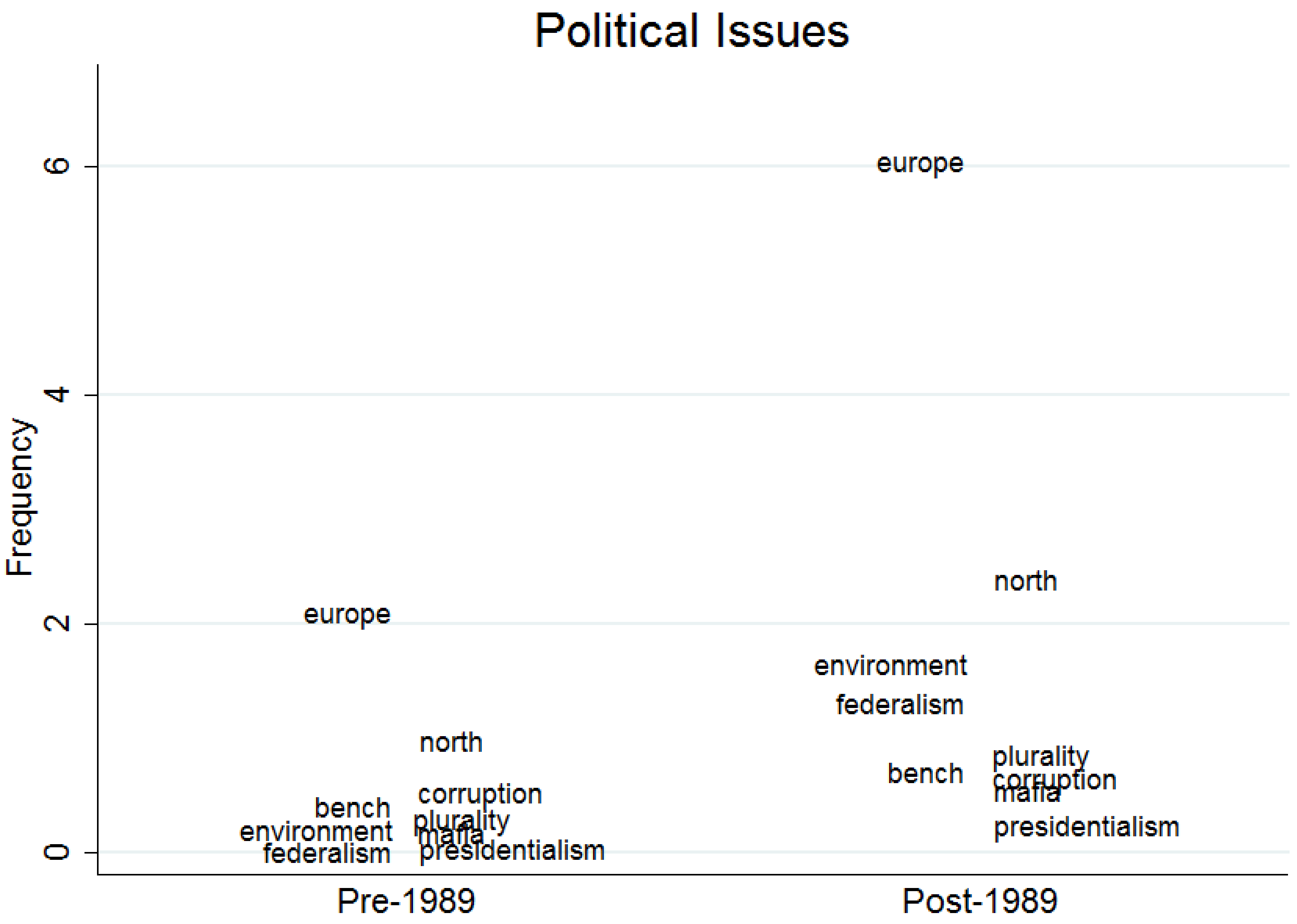
Conditions for using Wordfish

Potential route to addressing this issue: carefully select the **words** that enter the analysis!!!

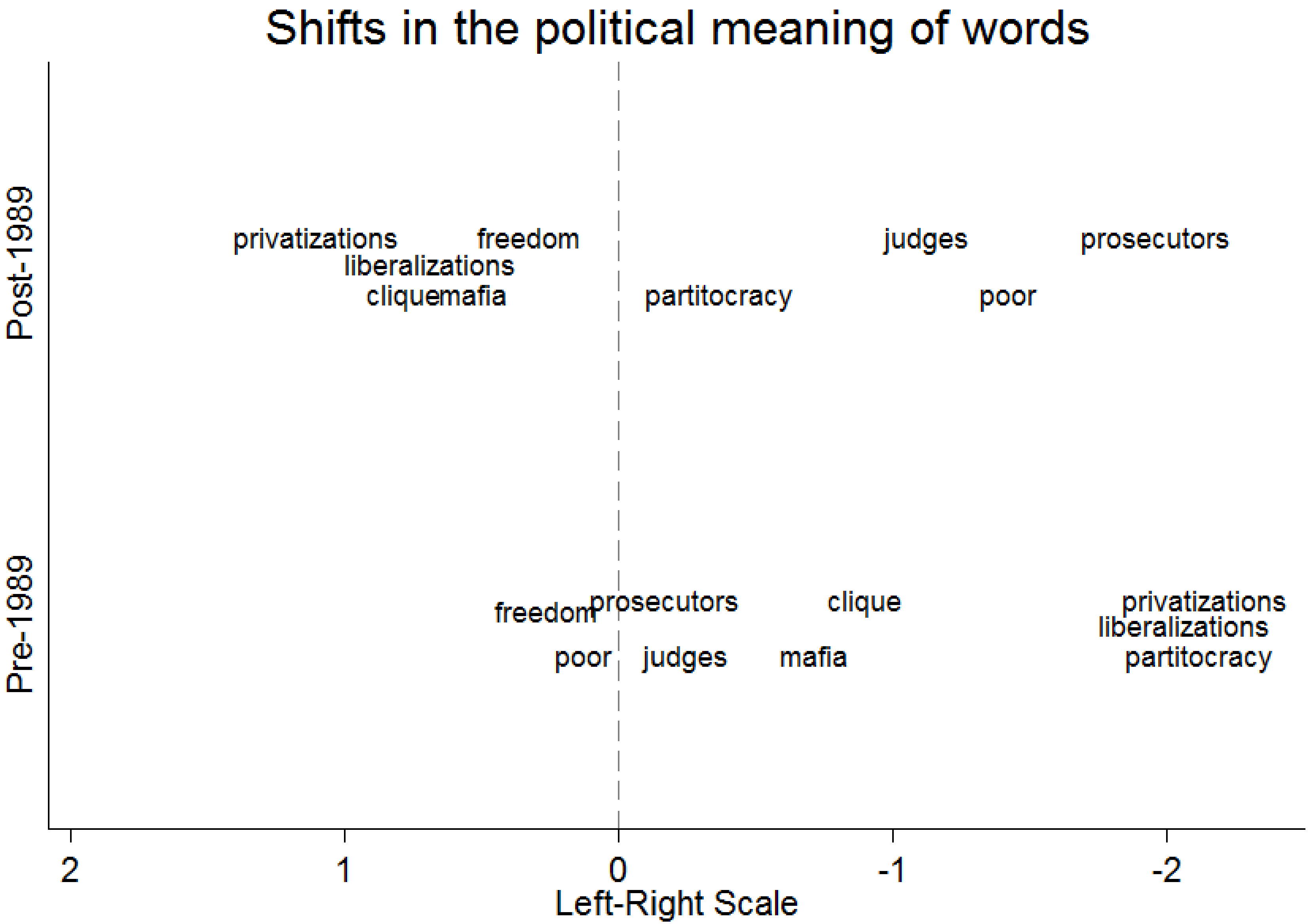
Thus, if there is movement of parties, it can only be due to **different word usage**

This requires that the **word data over time** must be comparable at a minimum level

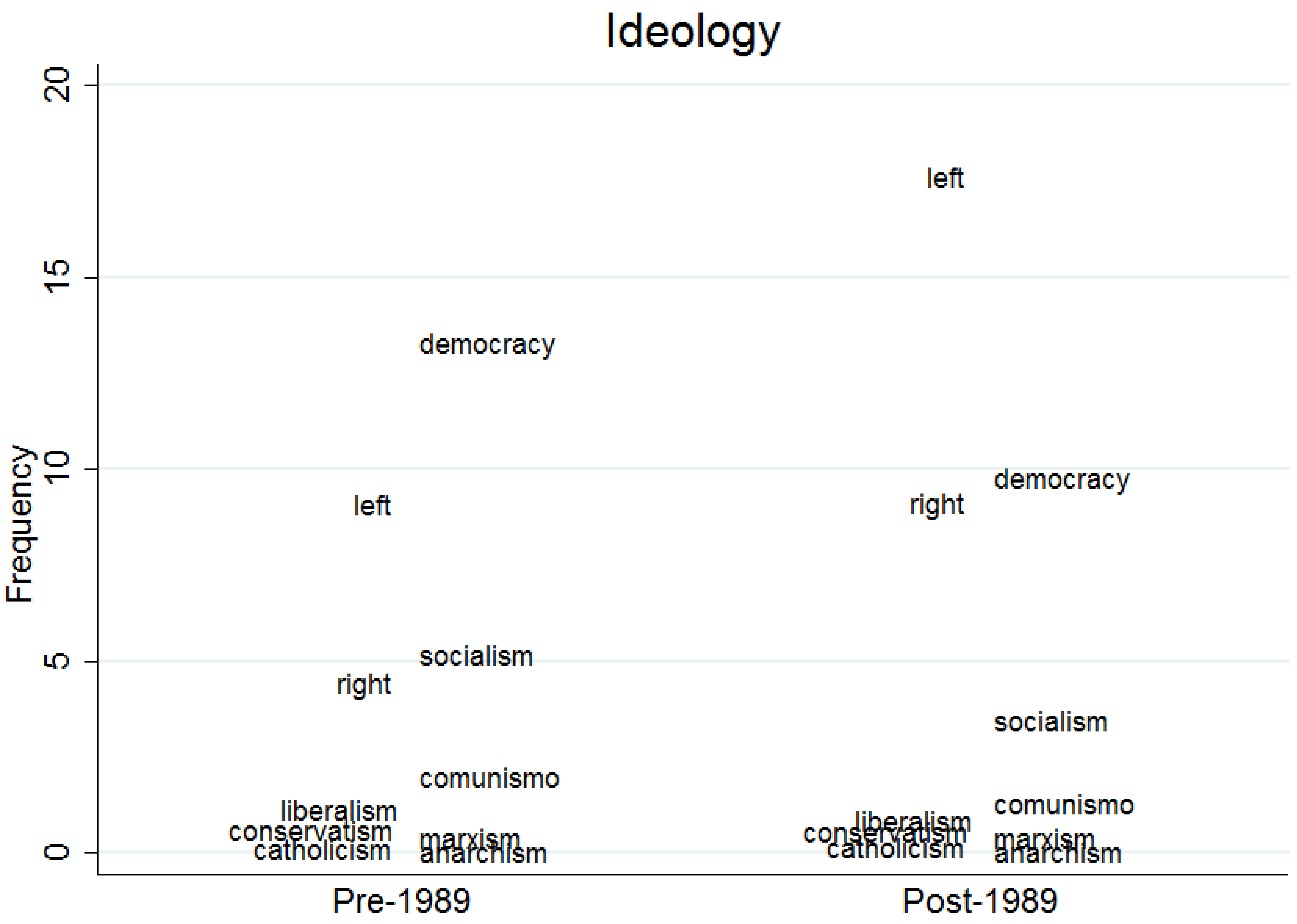
Evolution of Language



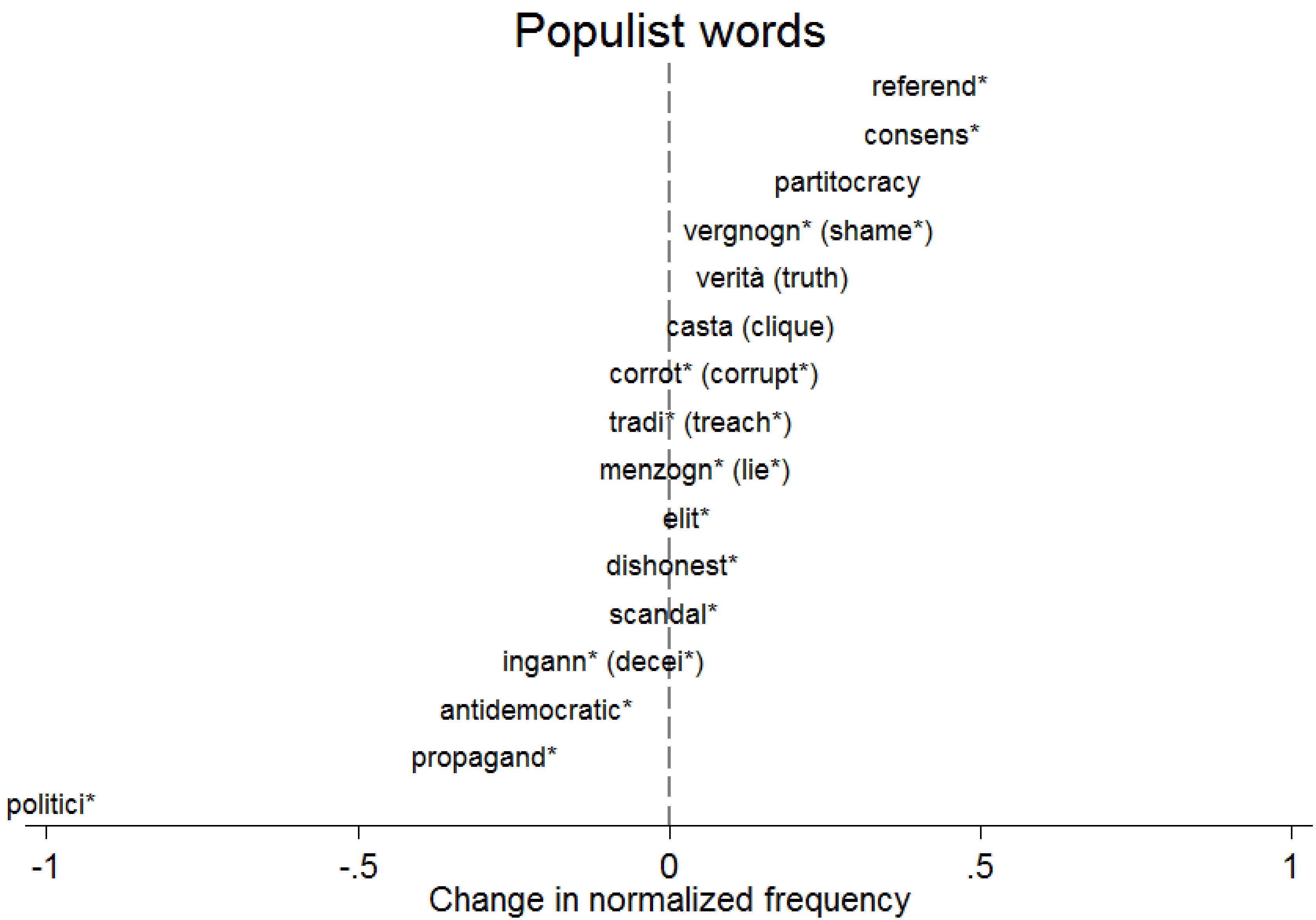
Bad news


Shifts in the political meaning of words
Post-1989
privatizations
freedom
liberalizations
cliquemafia
partitocracy
judges
prosecutors
poor
Pre-1989
freedom
prosecutors
clique
privatizations
liberalizations
partitocracy
poor
judges
mafia
2
1
0
-1
-2
Left-Right Scale

Good news


Ideology
Frequency
0
5
10
15
20
left
democracy
socialism
right
comunismo
liberalism
conservatism
catholicism
marxism
anarchism
left
democracy
right
socialism
comunismo
liberalism
conservatism
catholicism
marxism
anarchism
Pre-1989
Post-1989

Evolution of Language



Conditions for using Wordfish

Take the example of parties' manifestos in Germany since 1970 to 2005. Assume that you want to analyze such documents with Wordfish

Now assume that the political lexicon in the manifestos at election time *t* contains an issue that is no longer relevant at time *t+1*, e.g. official relations with the GDR (East Germany)

If all parties make a statement with regard to the GDR at point *t* but not at *t+1*, then the words **will not only distinguish** parties at point *t*, but also distinguish the elections

As a result, if all words are counted, even the rare ones, the parties are more likely to be **clustered by election**

Conditions for using Wordfish

The same is true if we have some changes in the **actual meaning** of some political words

Which word inclusion criteria? Include in the analysis only words that fulfill a minimum threshold criterion based on **non-informative and informative priors**

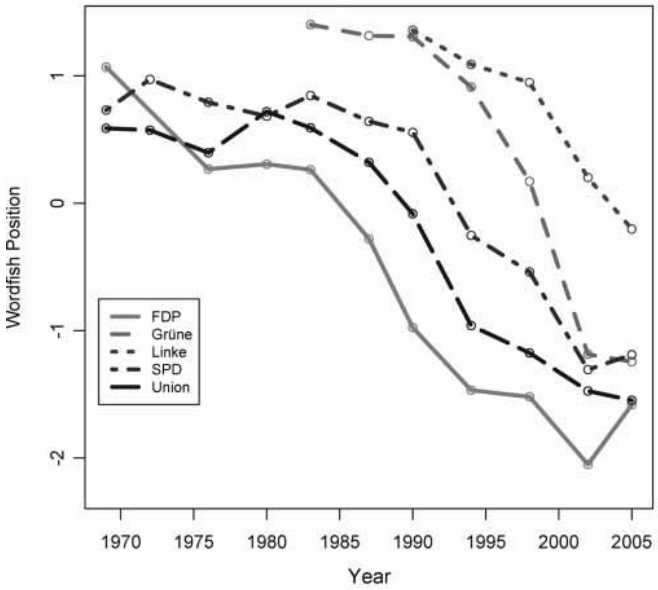
First alternative: in the term-document matrix includes words that are mentioned in a minimum number of documents (say, in at least 20%), thus essentially keeping words that are deemed important enough to be mentioned either over time by one party or by several parties

Conditions for using Wordfish

Second alternative: in the term-document matrix includes only those words that appear both pre- and post-1990, i.e., reunification added words to the German political lexicon that were not in it previously. Likewise, some words that were previously important likely fell out of use.

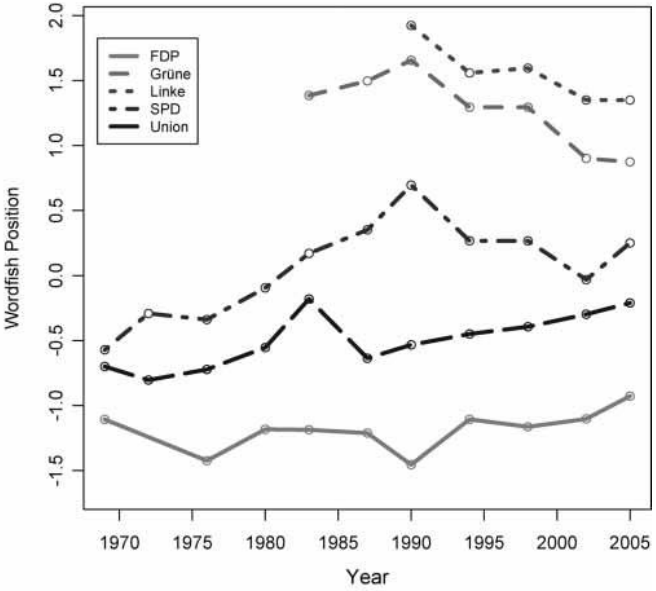
If we do not control for this fact, we would see a large jump in all parties around 1990 as they all shift their word usage to account for new political realities

German Party Position Estimates, 1969-2005
(Dataset A: all words)



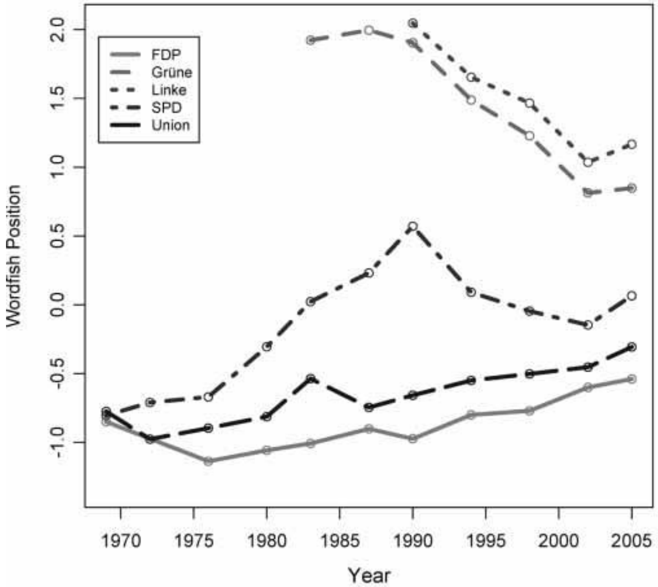
41,684 unique words, 44 documents.

German Party Position Estimates, 1969-2005
(Dataset B: stemmed words in at least 20% of all docs)



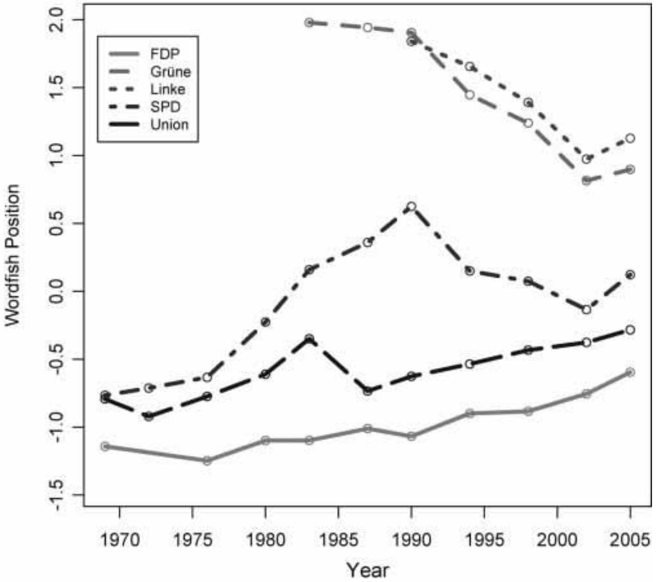
3,455 unique words, 44 documents.

German Party Position Estimates, 1969-2005
(Dataset C: words mentioned pre/post 1990)



11,273 unique words, 44 documents.

German Party Position Estimates, 1969-2005
(Dataset D: stemmed words mentioned pre/post 1990)



8,178 unique words, 44 documents.

Conditions for using Wordfish

The model is identified by fixing the mean position at 0 and the standard deviation at 1 and by constraining the FDP in 1990 to have a smaller value than the PDS in 1990

As suspected, agenda effects over time dominate the results when all words are used

Excluding rare words induces stability and the results are corroborated by their good face validity

Wordfish: a summary

The Wordfish method does not depend on documents with ex ante assigned reference scores. Position estimates derived using Wordfish are based only on the information in the texts.

This lack of an ex ante defined dimensionality is a double-edged sword: while Wordfish scales texts independently of prior information, it renders uncertain the exact nature of the dimension being estimated (as it happens in all inductive approaches!)

One important drawback of unsupervised algorithms is thus that the nature of the dimensions produced requires intensive validation before they can be applied across different sets of texts and contexts (Grimmer and Stewart, 2013)