WZB

Wissenschaftszentrum Berlin
für Sozialforschung

What web-scraping and text-as-data approaches reveal about EU politics

Summer Institute in
Computational Social Sciences (SICCS) Berlin
July 7, 2023

Christian Rauh
WZB Berlin Social Science Center
Universität Potsdam

Welcome!

- Who am I (and how did I end up here)?
  - Studying politics beyond the nation state; facing data scarcity
  - Politics mostly happens in and through text (often available online)
  - ➢ Therefore, I became a kind of CSS autodidact ...

- What's in it for you?
  - Three exemplary and hopefully somewhat interesting applications (political science, but simplified ;))
  - ➢ Rather *simple* text-as-data measures applied to web-scraped data offer insights to pretty *big* substantial claims
  - ➢ Some methodological and practical insights on the way ...

- Please, do raise your questions!

Example I

To what extent does the European Commission shape European laws?

Motivation

- **The big claim:**
  European Commission is a powerful *legislative agenda–setter*
  - Formal monopoly of initiative
  - Anticipation of what is acceptable to co-legislators

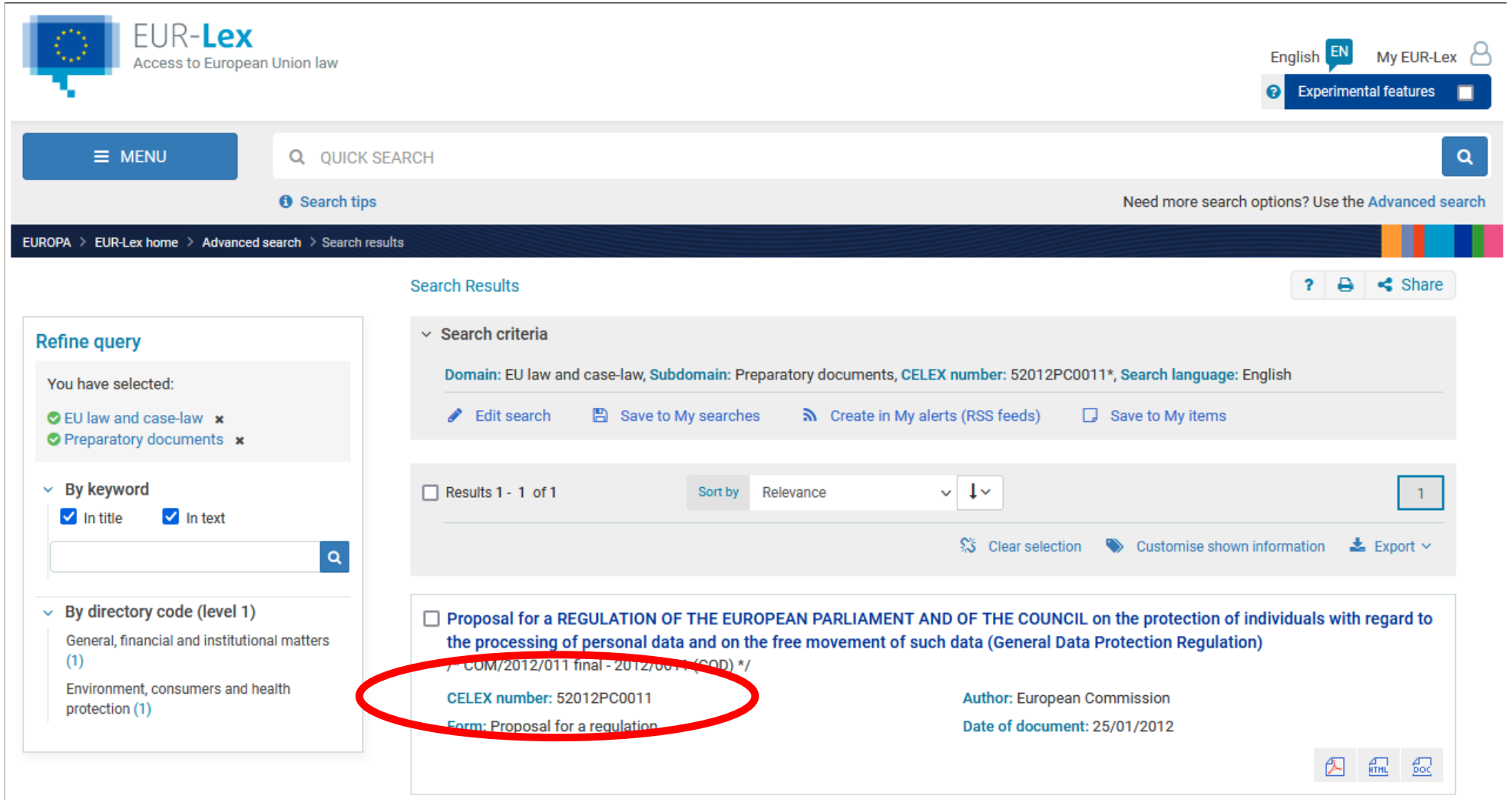
- **The debate:** *Extent* and *constraints* of this influence
  - Diverging preferences in the Council of Ministers
  - Degree to which the European Parliament is involved
  - Internal resources of Commission departments and coordination

➢**How to systematically assess and compare the Commission's ability to shape the contents of European law?**

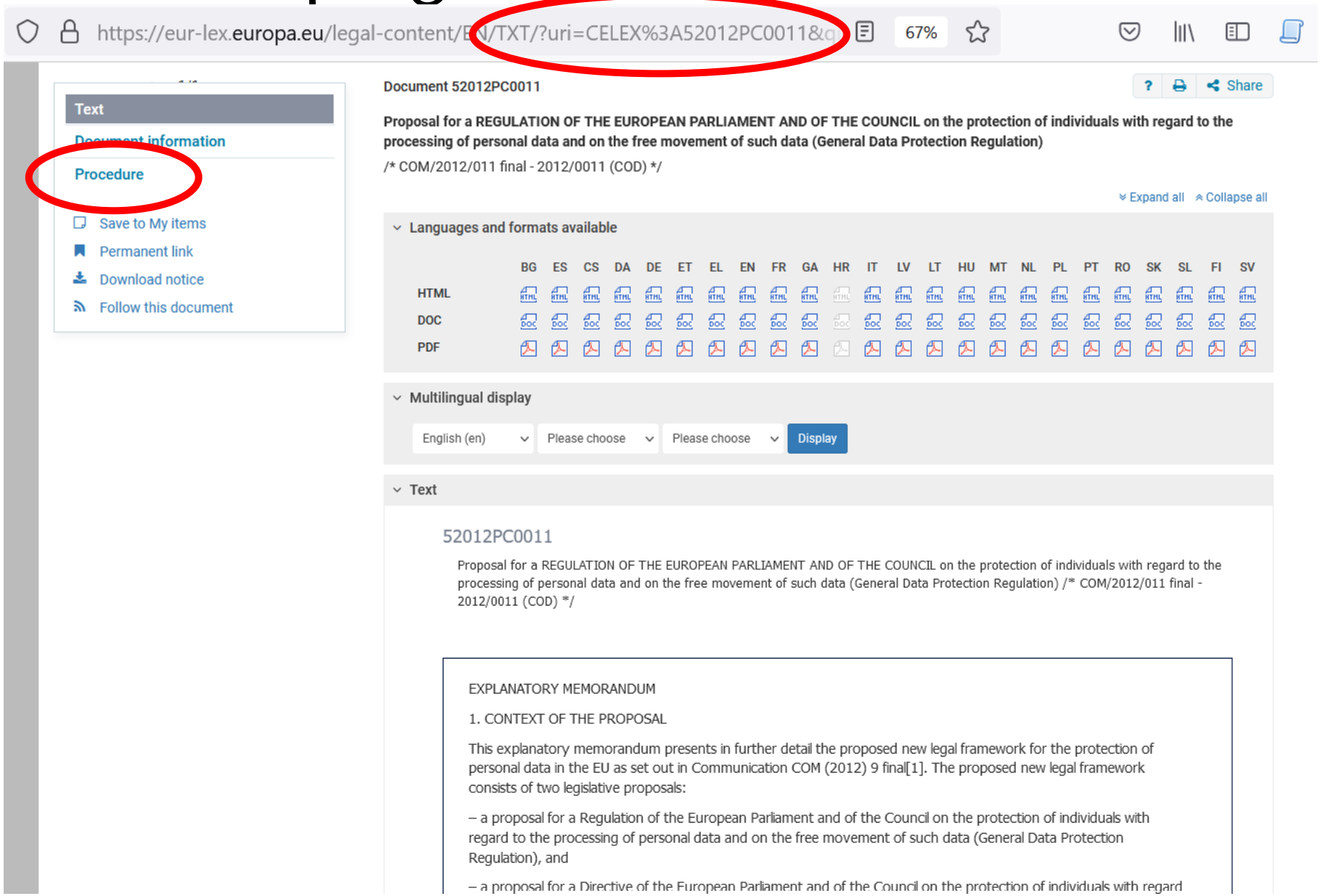
The (simple?) measurement idea

- *Textual change from legislative proposals to the finally adopted laws* reveals quality of Commission's anticipation of negotiations
- How to measure this textual change?
  - "*The economy is more important than the environment.*"
  - "*The environment is more important than the economy.*"
- Inverted Levenshtein minium edit distance
  - Matrix of ordered *words* in proposal and law
  - Count *insertion*, *deletions*, *substitutions* or *adjacent transposition* of words
  - Normalize to text length and subtract from one
    ~ Probability that an individual word from the proposal remains unchanged
    ~ Proportion of proposal text that remains unchanged
- Now we 'just' need full-text proposal/law pairs and metadata on Commission DGs, EP involvement, etc ...

Web scraping Eur-Lex

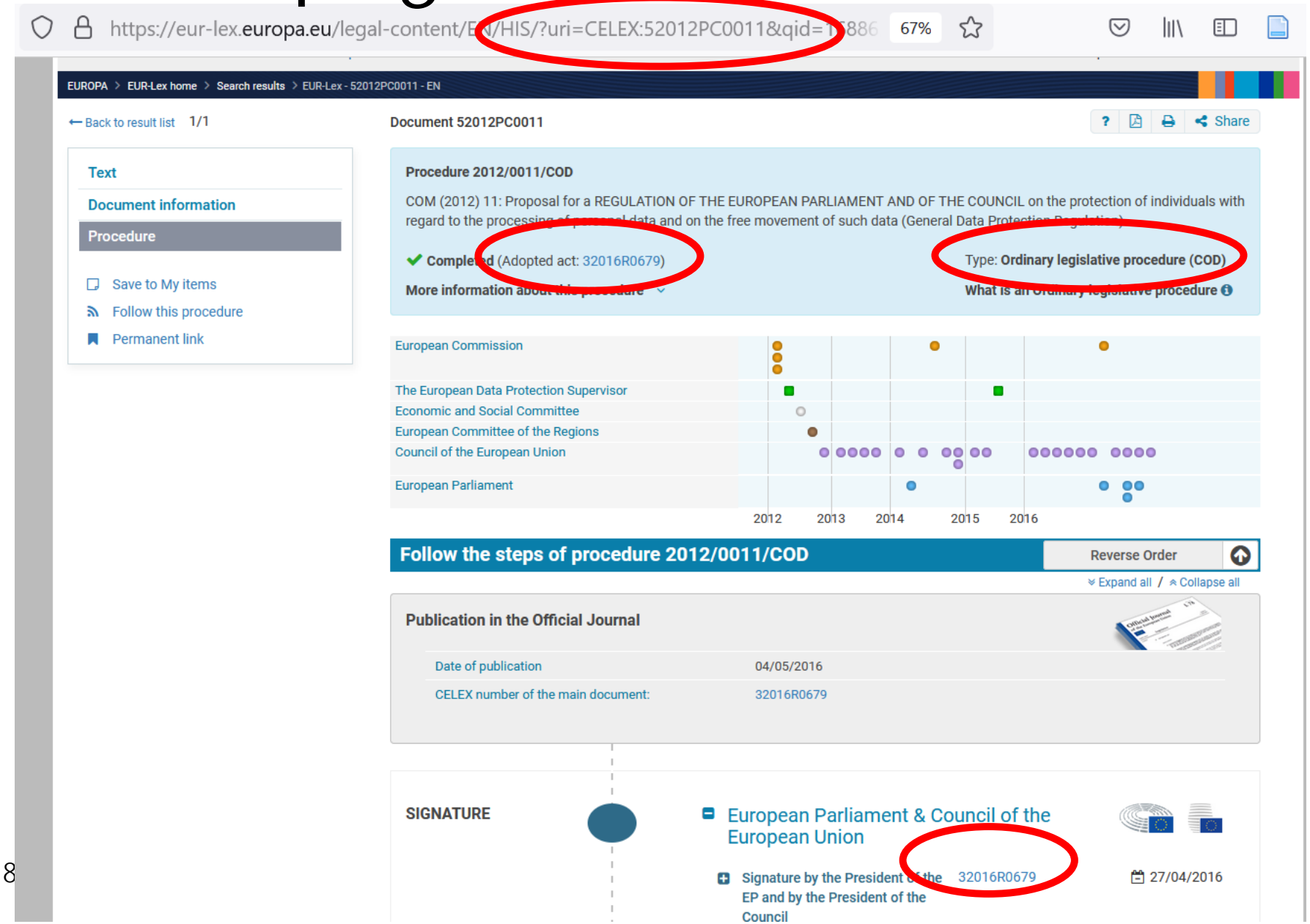


Web scraping Eur-Lex



WZB

Web scraping Eur-Lex



WZB

Step	Task	Details & Examples
1	Identify universe of document numbers	• CELEX identifiers: [SECTOR][YEAR][TYPE][DOCNUM] • Com proposals for binding secondary EU law: CELEX sector 5 (preparatory documents) of type 'PC' (COM – legislative proposals) • Manual search shows that the highest observed DOCNUM in this domain between 1985 and 2016 is 2282 • Proposals in legislative packages might share a CELEX number separated by bracketed numbers - the scraper allows for up to five of such sub-proposals; the highest number retrieved in random manual searches • This defines the possible numerical range: *5*1985*PC*0001(01) to *5*2016*PC*2282(05)
2	Check whether documents exist	• If EUR-Lex contains a document with these identifiers, it can be accessed via: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:[IDENTIFIER] • The scraper accordingly pings the respective URLs and stores the http response • '404' codes indicate that no such document exists in the database and corresponding identifiers are disregarded in all further steps
3	Download bibliographical information	• For existing identifiers, the scraper stores the html content of the landing URL • These files contain bibliographical information such as the proposal title, the adoption dates and – partially – the full legal text of the proposal • Execution: July 31 2017
4	Download procedural information	• EUR-Lex provides procedural information via a dedicated URL with the structure: http://eur-lex.europa.eu/legal-content/EN/HIS/?uri=CELEX:[IDENTIFIER] • The html content of these URLs is separately stored • These files contain further document information and various details on the inter-institutional process and its outcome (such as Council and EP readings) • Execution: July 31 2017
5	Parse html and extract data	• All downloaded files are then parsed to extract relevant information and store it in a combined data frame centred on the original Commission proposal • Besides procedural information, this also includes references to the finally adopted law, if applicable, which is also uniquely identified by a CELEX number in sector 3 (legislation) of type L (directives), R (regulations), or D (decisions)
6	Remove irrelevant cases	• Some documents have been wrongly classified as original Commission proposals by the database maintainers (as indicated in the document title) • First, I remove documents that are recommendations or reports only • Second, documents that do not initiate a new procedure but present inter-inst. steps only: 'opinions' and 'modified|revised|reexamined proposals' are removed
6	Download and parse adopted law	• Similar to steps 3 and 5, the scraper then exploits the CELEX numbers of the adopted laws to download and store bibliographical information and the full legal text (where available) • Execution: August 3 2017
7	Data cleaning	• Infer missing procedures from legal basis, or EP involvement and Treaty in force • Harmonize actor and procedure names that have varied over time • Check and correct false entries by exploiting overlapping data points

Table 1: Logic of the custom EUR-Lex scraper

Step	Task	Details & Examples
1	Identify universe of document numbers	• CELEX identifiers: [SECTOR][YEAR][TYPE][DOCNUM] • Com proposals for binding secondary EU law: CELEX sector 5 (preparatory documents) of type 'PC' (COM – legislative proposals) • Manual search shows that the highest observed DOCNUM in this domain between 1985 and 2016 is 2282 • Proposals in legislative packages r bracketed numbers - the scraper all highest number retrieved in random • This defines the possible numerical
2	Check whether documents exist	• If EUR-Lex contains a document http://eur-lex.europa.eu/legal-conte • The scraper accordingly pings the re • '404' codes indicate that no su corresponding identifiers are disrega
3	Download bibliographical information	• For existing identifiers, the scraper s • These files contain bibliographical adoption dates and – partially – the • Execution: July 31 2017
4	Download procedural information	• EUR-Lex provides procedural infor http://eur-lex.europa.eu/legal-conte • The html content of these URLs is s • These files contain further document information and various details on the inter-institutional process and its outcome (such as Council and EP readings) • Execution: July 31 2017
5	Parse html and extract data	• All downloaded files are then parsed to extract relevant information and store it in a combined data frame centred on the original Commission proposal • Besides procedural information, this also includes references to the finally adopted law, if applicable, which is also uniquely identified by a CELEX number in sector 3 (legislation) of type L (directives), R (regulations), or D (decisions)
6	Remove irrelevant cases	• Some documents have been wrongly classified as original Commission proposals by the database maintainers (as indicated in the document title) • First, I remove documents that are recommendations or reports only • Second, documents that do not initiate a new procedure but present inter-inst. steps only: 'opinions' and 'modified|revised|reexamined proposals' are removed
6	Download and parse adopted law	• Similar to steps 3 and 5, the scraper then exploits the CELEX numbers of the adopted laws to download and store bibliographical information and the full legal text (where available) • Execution: August 3 2017
7	Data cleaning	• Infer missing procedures from legal basis, or EP involvement and Treaty in force • Harmonize actor and procedure names that have varied over time • Check and correct false entries by exploiting overlapping data points

Table 1: Logic of the custom EUR-Lex scraper

```
for (i in 1:nrow(proposals)){

  # Construct filenames and paths
  dest.prop <- paste("./PROPOSALS/", proposals$celex[i], "_prop.html", sep = "")
  dest.proc <- paste("./PROCEDURES/", proposals$celex[i], "_proc.html", sep = "")

  # Download and save
  download.file(url = proposals$link.text[i], destfile = dest.prop, quiet = T)
  download.file(url = proposals$link.procedure[i], destfile = dest.proc, quiet = T)

  # Random pause between 0 and 1 seconds so as not overburden the server
  Sys.sleep(sample(0:1, 1))

  # Show progress
  print(round(i/nrow(proposals)*100, digits = 2))

}
```

Step	Task	Details & Examples
1	Identify universe of document numbers	• CELEX identifiers: [SECTOR][YEAR][TYPE][DOCNUM] • Com proposals for binding secondary EU law: CELEX sector 5 (preparatory documents) of type 'PC' (COM – legislative proposals) • Manual search shows that the highest observed DOCNUM in this domain between 1985 and 2016 is 2282 • Proposals in legislative packages might share a CELEX number separated by bracketed numbers - the scraper allows for up to five of such sub-proposals; the highest number retrieved in random manual searches • This defines the possible numerical range: *5*1985*PC*0001(01) to *5*2016*PC*2282(05)
2	Check whether documents exist	• If EUR-Lex contains a document with these identifiers, it can be accessed via: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:[IDENTIFIER]

```r
# Also in these cases, jump to next iteration
if (proposals$no.box[i] == TRUE){
  next
}

# Get information from proposal header (stored in a blue box)
title.box <- html_nodes(doc, ".blueBox")
proposals$proc.num[i] <- html_text(html_nodes(title.box, "strong"))

proposals$title[i] <- html_text(html_nodes(title.box, "p:nth-child(2)")) # Second paragraph - contains com number and title
proposals$com.num[i] <- sub(":.*$", "", proposals$title[i], fixed = F) # Extract COM Number from title
proposals$com.num[i] <- gsub(" ", "", proposals$com.num[i], fixed = T) # Clean

proposals$title[i] <- sub("^.*?:", "", proposals$title[i], fixed = F) # Clean title
proposals$title[i] <- gsub("[\r\n\t]", "", proposals$title[i], fixed = F) # Clean title
proposals$title[i] <- str_trim(proposals$title[i], side = "both") # Clean title
proposals$title[i] <- gsub("\\s+", " ", proposals$title[i], fixed = F)

proposals$celex.fin[i] <- html_text(html_nodes(title.box, "p:nth-child(3)")) # Third paragraph in blue box
proposals$celex.fin[i] <- sub("^.*?:", "", proposals$celex.fin[i], fixed = F) # Clean
proposals$celex.fin[i] <- gsub("[\r\n\t]", "", proposals$celex.fin[i], fixed = F) # Clean
proposals$celex.fin[i] <- gsub("\\s", "", proposals$celex.fin[i], fixed = F)# Clean

# Extract all tables from html file
tables <- html_table(doc, header = FALSE, trim = TRUE)

# First table contains general information on proposal
# Wrapping in ifelse necessary because loops breaks otherwise if row is not existing
info <- as.data.frame(tables[1])
proposals$legal.basis[i] <- ifelse(length(info$X2[which(info$X1 == "Legal basis:")]) > 0, info$X2[which(info$X1 == "Legal basis:")]
```

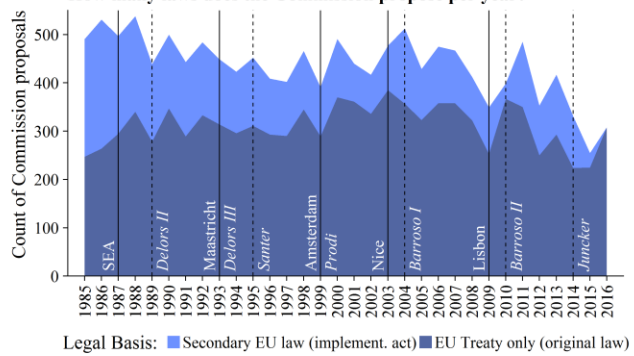
• Check and correct false entries by exploiting overlapping data points

Table 1: Logic of the custom EUR-Lex scraper

Legislative proposals of the European Commission 1985-2016
An aggregate perspective on Eur-Lex information

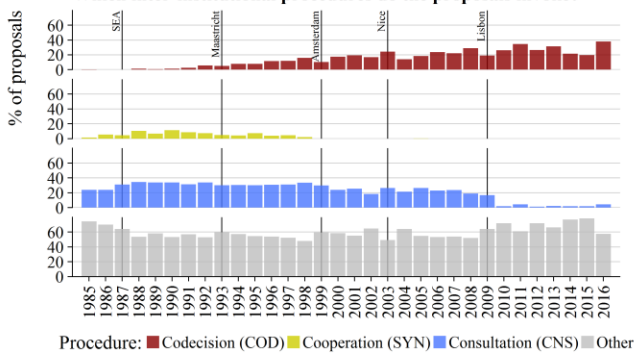
The proposals

How many laws does the Commission propose per year?



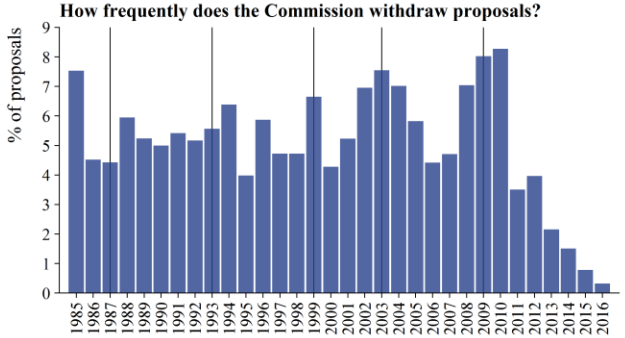
The inter-institutional process

Which inter-institutional procedures do the proposals invoke?

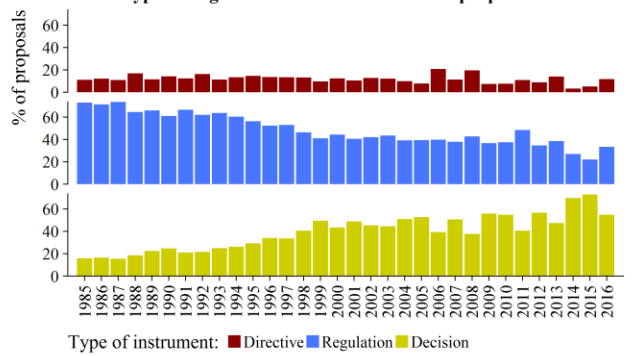


The results

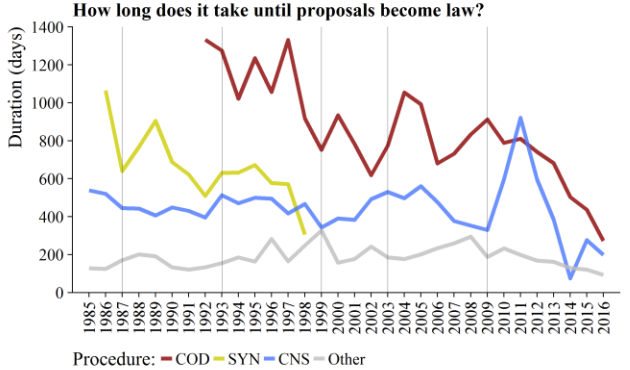
How frequently does the Commission withdraw proposals?



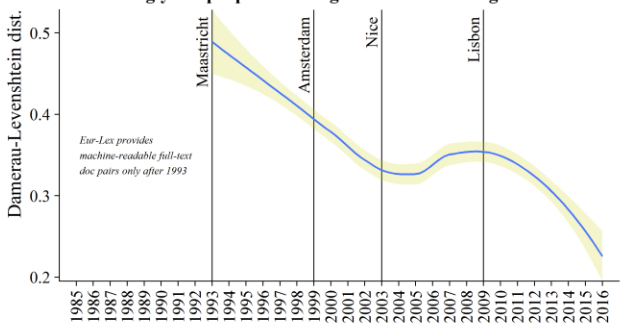
Which types of legislation does the Commission propose?



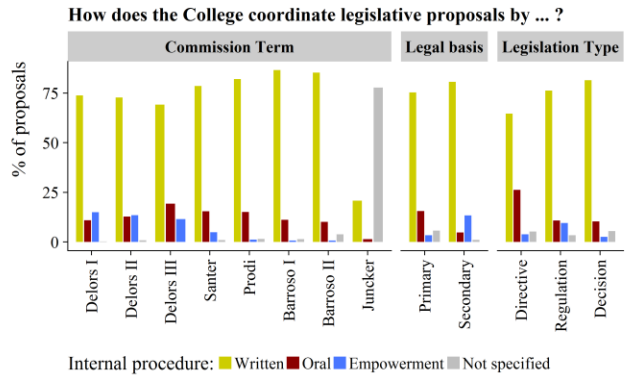
How long does it take until proposals become law?



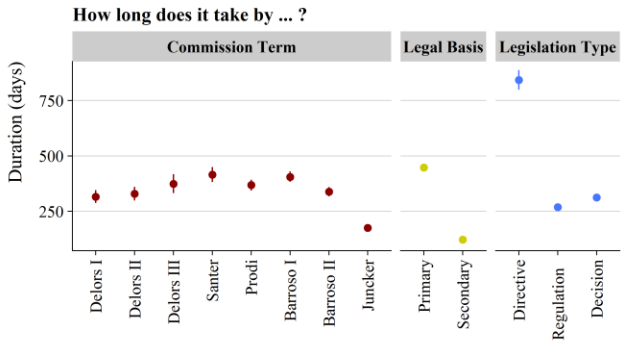
How strongly are proposals changed before becoming law?



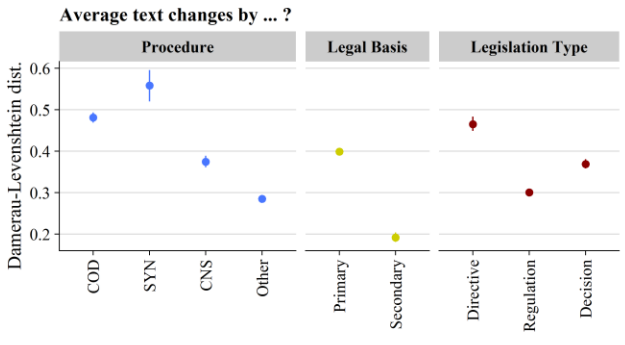
How does the College coordinate legislative proposals by ... ?



How long does it take by ... ?

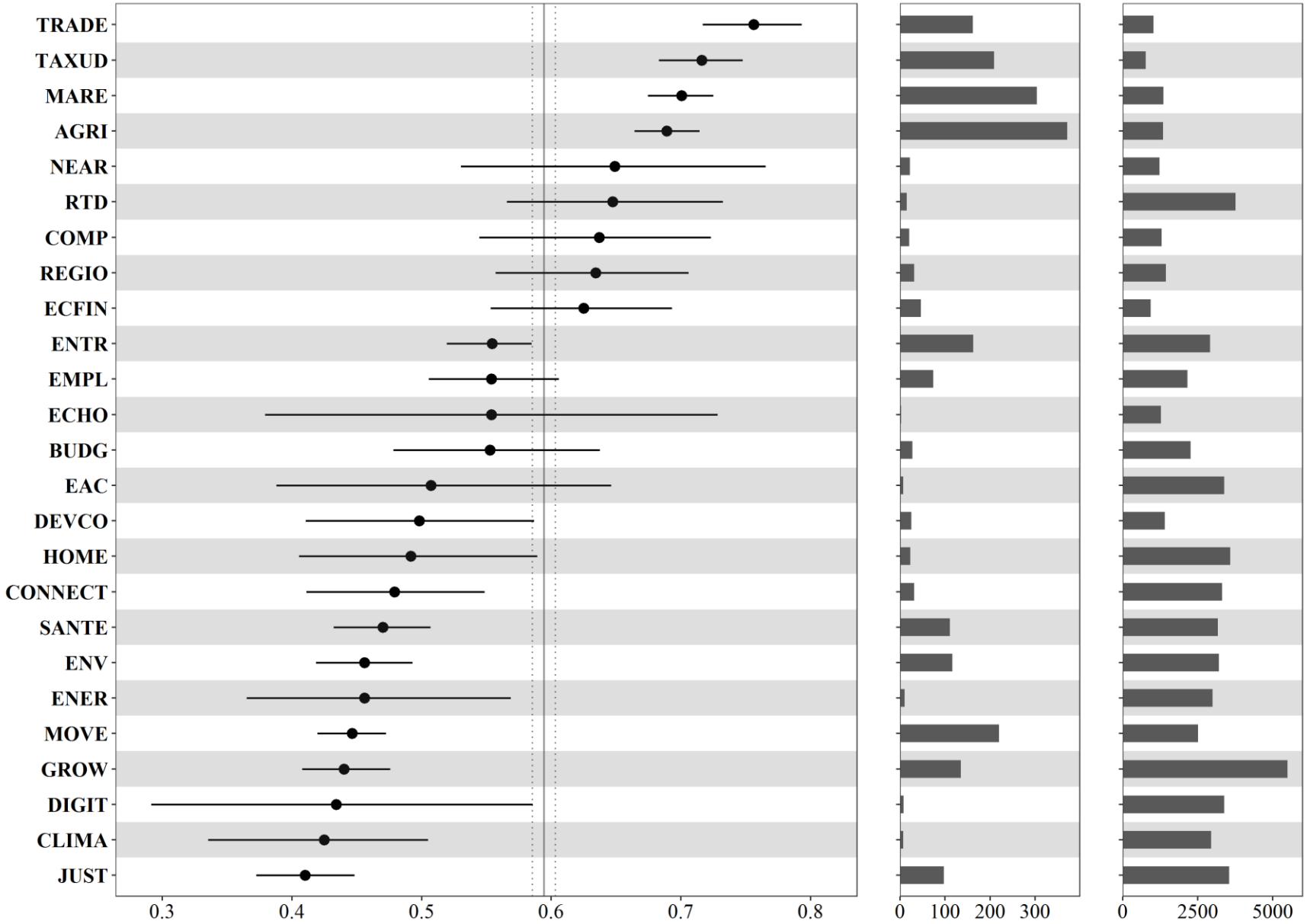


Average text changes by ... ?

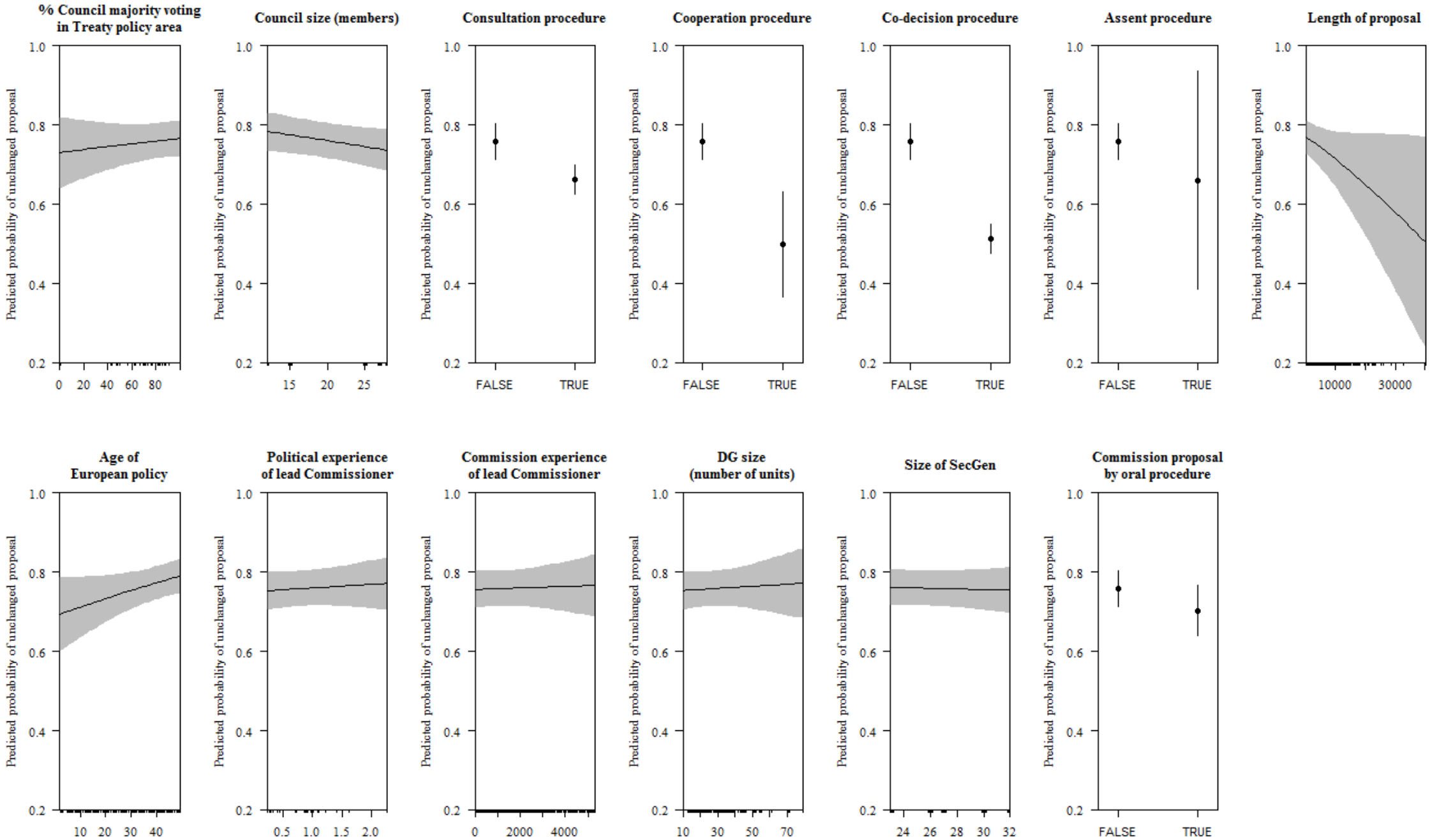


Based on 14,068 original Commission proposals for binding EU law identifiable in Eur-Lex for the period 1985-2016
Data scraped, cleaned, and visualized in January 2018 by @Chrauh (www.christian-rauh.eu)

WZB


TRADE
TAXUD
MARE
AGRI
NEAR
RTD
COMP
REGIO
ECFIN
ENTR
EMPL
ECHO
BUDG
EAC
DEVCO
HOME
CONNECT
SANTE
ENV
ENER
MOVE
GROW
DIGIT
CLIMA
JUST
0.3
0.4
0.5
0.6
0.7
0.8
Inverted Levenshtein distance between Commission proposal and adopted law
Means with bootstrapped 95% confidence bands
0
100
200
300
Number of proposals
0
2500
5000
Average words per proposal

WZB


% Council majority voting in Treaty policy area
Council size (members)
Consultation procedure
Cooperation procedure
Co-decision procedure
Assent procedure
Length of proposal
Age of European policy
Political experience of lead Commissioner
Commission experience of lead Commissioner
DG size (number of units)
Size of SecGen
Commission proposal by oral procedure
Predicted probability of unchanged proposal
FALSE
TRUE

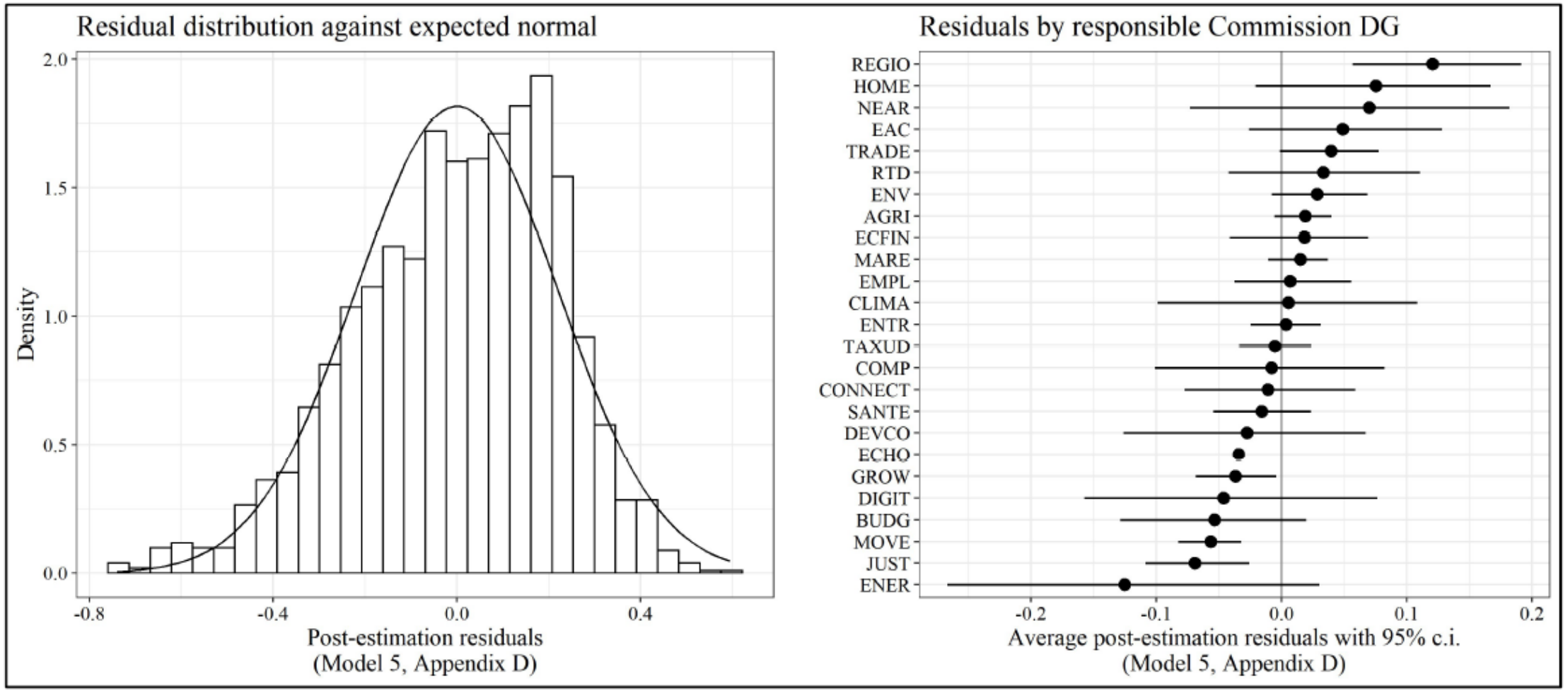


Figure 3: *Analysis of post-estimation residuals*

What to learn from this example

- **The *European Commission* is indeed a rather successful agenda setter:**
  40-80% of the legal text it proposes is adopted into binding European Law
- **But:** - The European Parliament constrains this heavily
  - There is significant (unexplained) variation across Commission DGs
- …

- **The (textual) information hidden in clunky official web archives can be *systematically extracted* to *shed light* on the distribution *of political power***

Example II

Does the European Commission communicate to European citizens?

Motivation

- **The big claim:**
  The *European Commission* is a *detached, technocratic actor*

- **The debate:** Increasing politicization indicates growing demand for justification of EU decisions, but unclear whether Commission provides it
  - Strategic self-legitimation
  - 'Technocratic mindset'
  - Strategic caution and de-politicization

➢**Does the European Commission try to speak to the 'ordinary' citizen? How did this change along the increasing EU politicisation over time?**

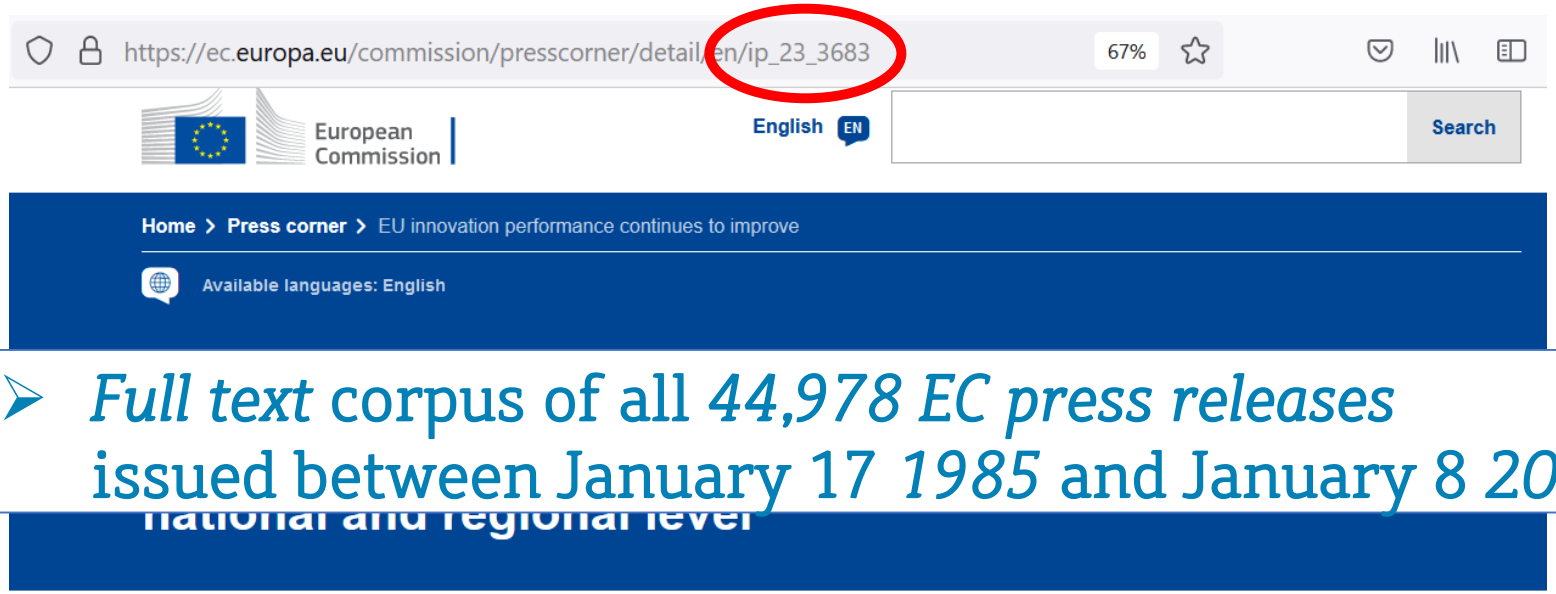
Basic measurement idea

- **The language by which the Commission communicates indicates which audiences it wants and can address**

  For the 'Brussels bubble' *highly specialized expert language* is ok,
  but the wider public requires *more accessible communication*

- **Commission press releases should be informative in this regard**
  - *Conscious efforts to convey messages to the general public*
  - *Preferred message before mediatization and framing by others*
  - *Most classical and consistently available public communication channel*

Web scraping the “EC press corner”



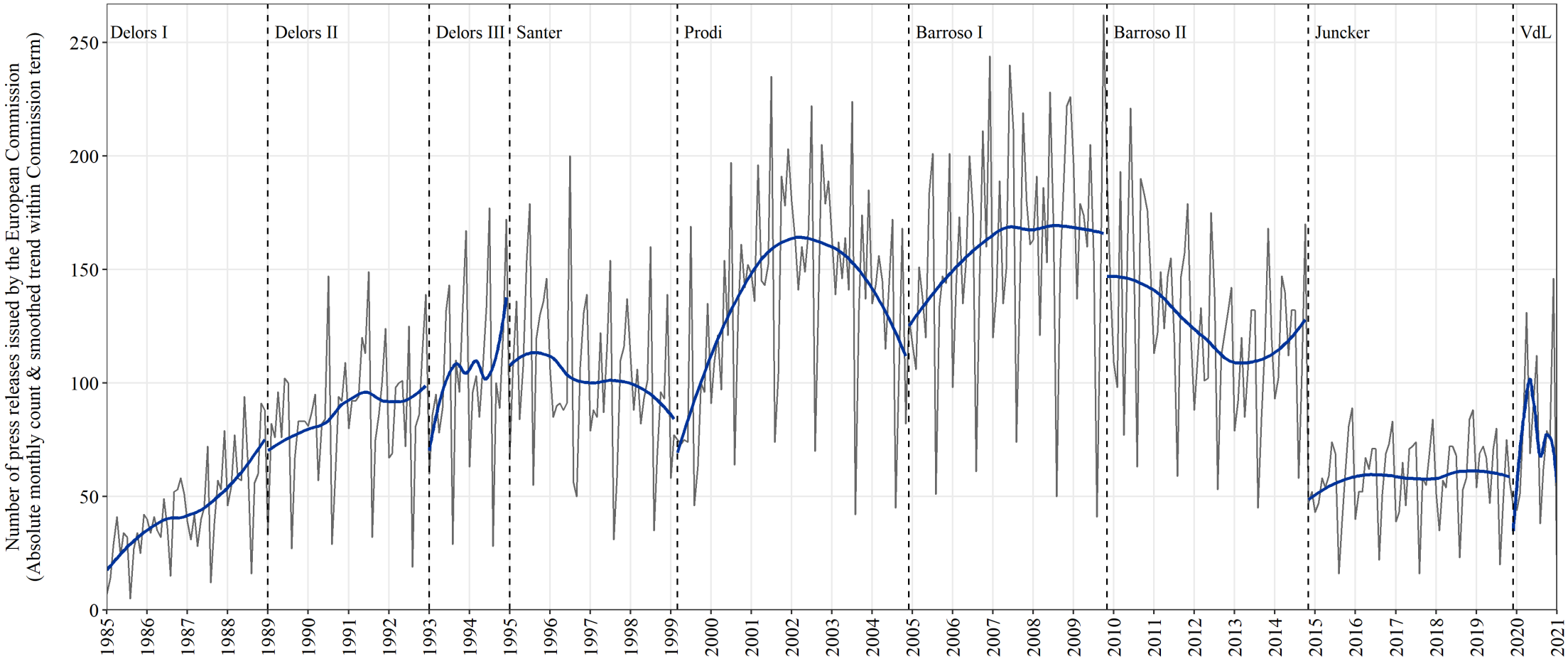
- *Full text* corpus of all *44,978 EC press releases* issued between January 17 *1985* and January 8 *2021*

Page contents

Top
Quote(s)
Print friendly pdf
Contacts for media

The 2023 edition of the European Innovation Scoreboard and the bi-yearly edition of the Regional Innovation Scoreboard published today showcase that, despite the recent crises, EU Member States and their regions keep improving their innovation performance.

The **European Innovation Scoreboard 2023** highlights a substantial improvement in the innovation performance of approximately 8.5% since 2016, confirming the EU's commitment to fostering a culture of innovation. The innovation performance of 25 countries improved during this period, although at a slower pace in the more recent years, and that 20 Member States experienced a significant rise in their innovation capabilities over the past year, while only seven observed a decline. Nevertheless, countries with less strong innovation systems tend to improve less rapidly than the EU average.


Delors I
Delors II
Delors III
Santer
Prodi
Barroso I
Barroso II
Juncker
VdL
Number of press releases issued by the European Commission
(Absolute monthly count & smoothed trend within Commission term)
0
50
100
150
200
250
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021

Technocratic vs. **accessible** language

Dimension	*Technocratic language*	*Accessible language*	Quantitative indicator
Grammar & Syntax	*Complex sentences and terminology*	*Short and concise text structure*	Flesch/Kincaid Reading Ease Score
Vocabulary	*Specialized jargon for experts*	*Familiar and commonly used words*	Average Google Books frequency
Action orientation	*Nominal style (abstract objects, states, processes)*	*Verbal style (personalized actions, clear temp. structure)*	Verb-to-noun Ratio

Benchmark data

- Expert-oriented political language

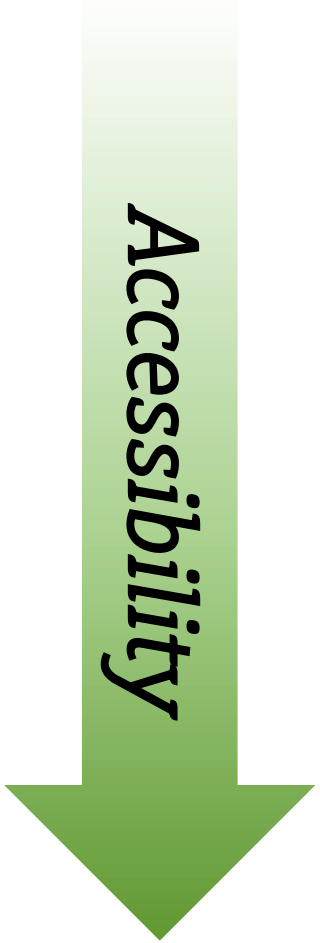
  2,332 political science abstracts scraped from the websites of five major academic journals

- **Political communication by national executives**

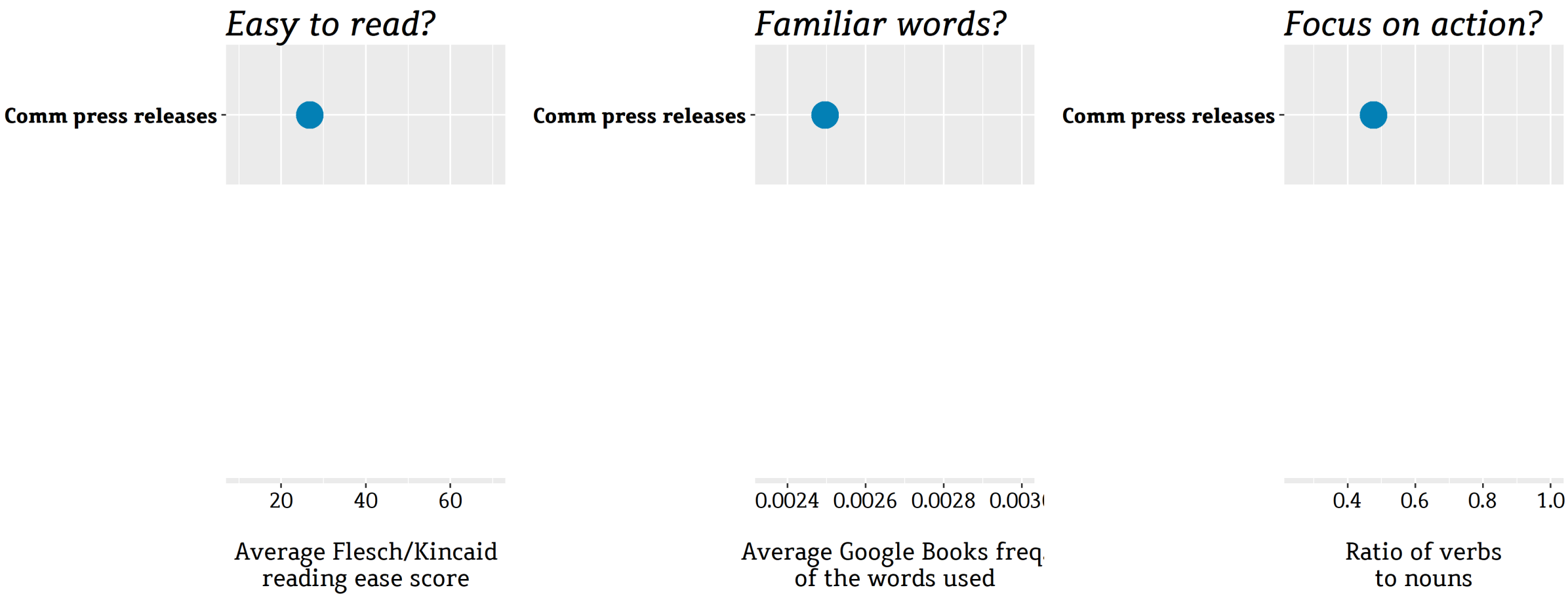
  92,070 press releases from different ministries, departments, and agencies of the UK and IRE governments (scraped from gov.uk / gov.ie)

- Public political discourse

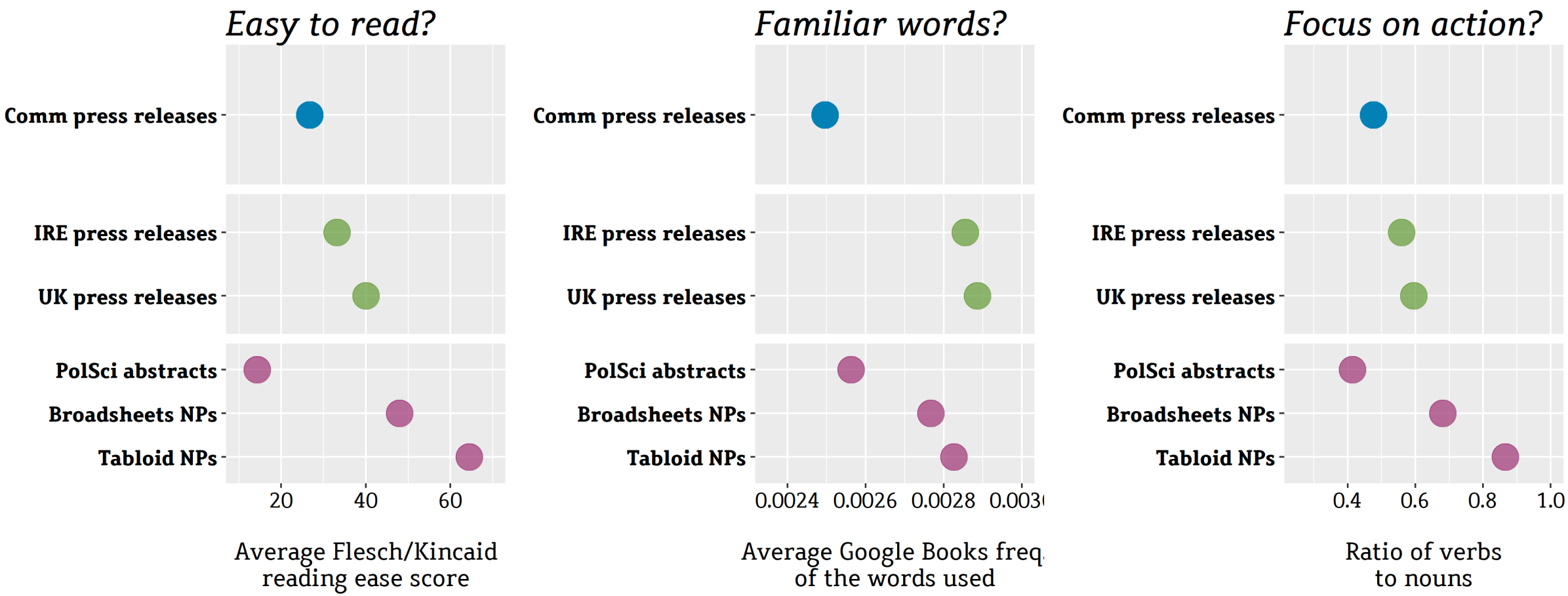
  Random text samples from the political sections of broadsheet and tabloid newspapers (57,765 and 22,160 paragraphs from the British National Corpus)



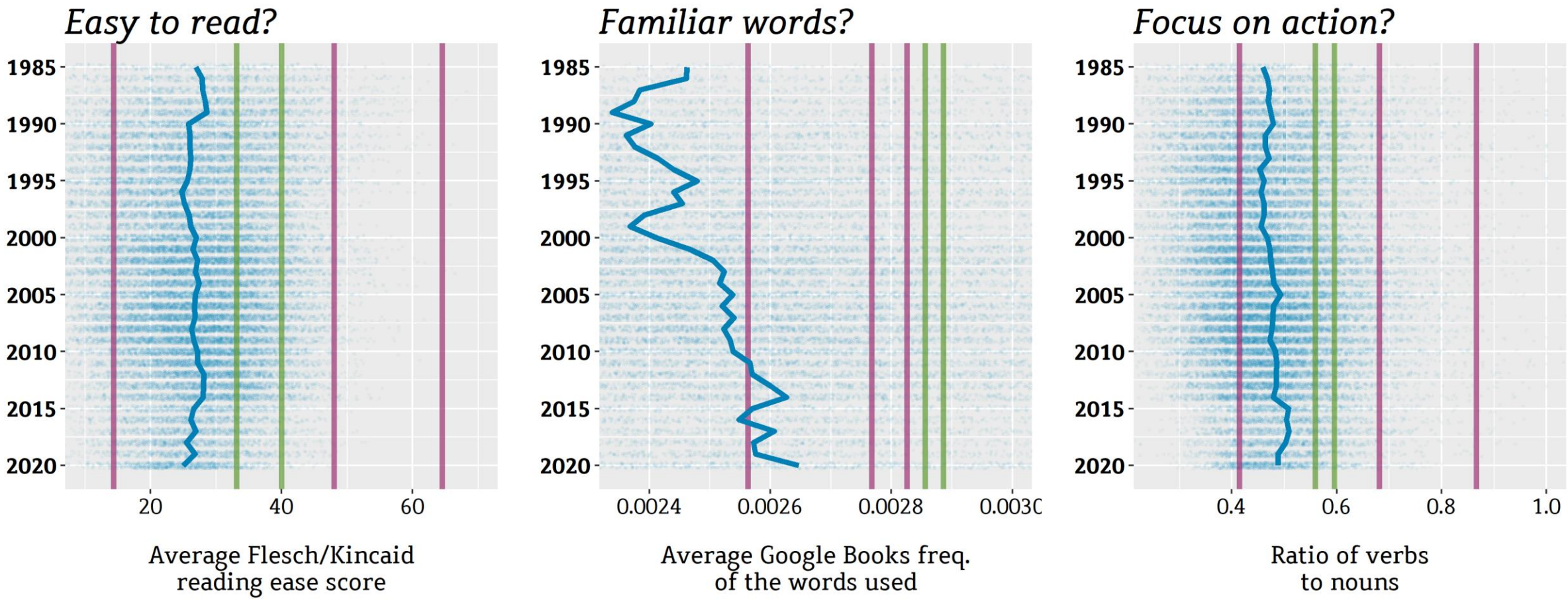
Rather technocratic communication



Rather technocratic communication



Little improvement over 35 years

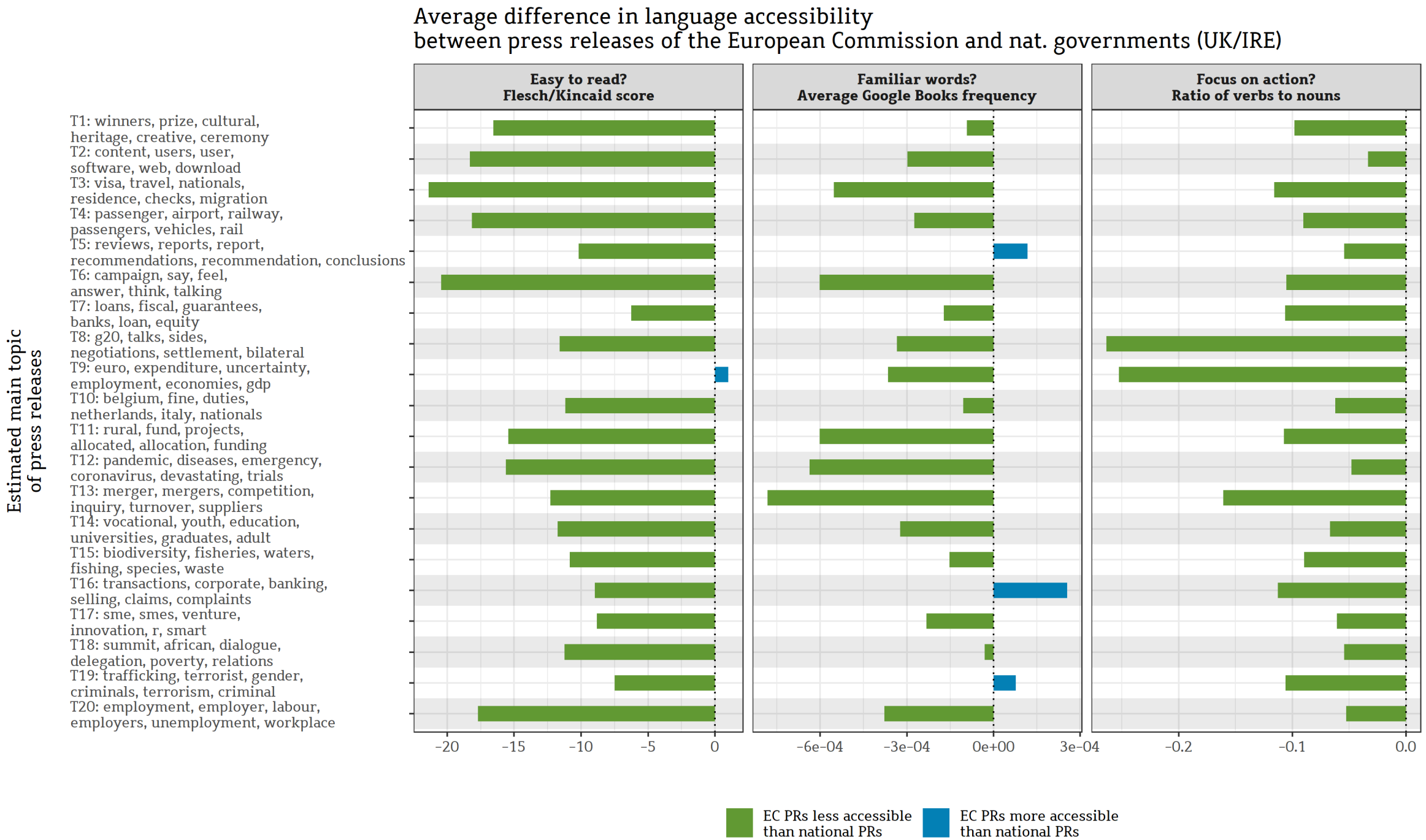


Does the Comm. just speak about different things?

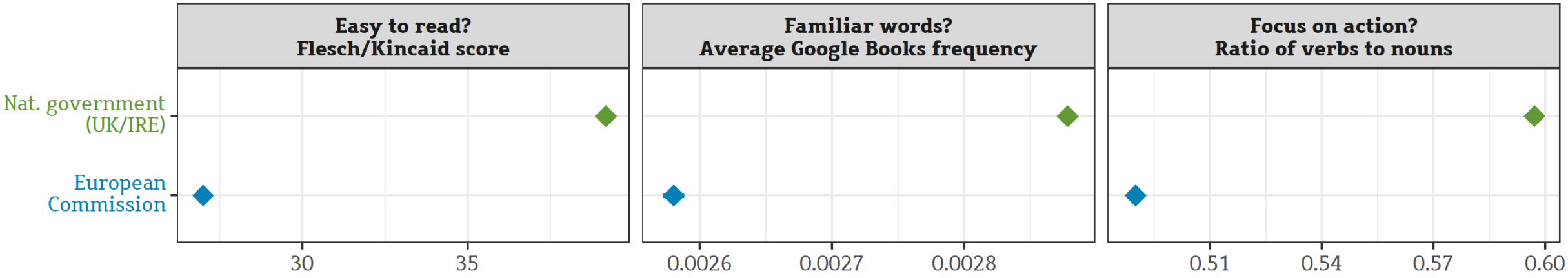
- The Commission maybe has to communicate about somehow inherently more complex policies !?
- *(Can complex policies not be clearly communicated to citizens?)*
- Test for or topic confounding
  - *Combine corpora of EC and national press releases (2010–2020)*
  - *Estimate k=20 structural topic model on this joint corpus (with content covariate for structural differences in word choice)*

  ➢ Compare language indicators by estimated main topic of texts

  ➢ Compare language indicators in a topic-matched sample (coarsened exact matching on topic distributions within texts)


Average difference in language accessibility
between press releases of the European Commission and nat. governments (UK/IRE)
Easy to read?
Flesch/Kincaid score
Familiar words?
Average Google Books frequency
Focus on action?
Ratio of verbs to nouns
Estimated main topic
of press releases
T1: winners, prize, cultural, heritage, creative, ceremony
T2: content, users, user, software, web, download
T3: visa, travel, nationals, residence, checks, migration
T4: passenger, airport, railway, passengers, vehicles, rail
T5: reviews, reports, report, recommendations, recommendation, conclusions
T6: campaign, say, feel, answer, think, talking
T7: loans, fiscal, guarantees, banks, loan, equity
T8: g20, talks, sides, negotiations, settlement, bilateral
T9: euro, expenditure, uncertainty, employment, economies, gdp
T10: belgium, fine, duties, netherlands, italy, nationals
T11: rural, fund, projects, allocated, allocation, funding
T12: pandemic, diseases, emergency, coronavirus, devastating, trials
T13: merger, mergers, competition, inquiry, turnover, suppliers
T14: vocational, youth, education, universities, graduates, adult
T15: biodiversity, fisheries, waters, fishing, species, waste
T16: transactions, corporate, banking, selling, claims, complaints
T17: sme, smes, venture, innovation, r, smart
T18: summit, african, dialogue, delegation, poverty, relations
T19: trafficking, terrorist, gender, criminals, terrorism, criminal
T20: employment, employer, labour, employers, unemployment, workplace
−20
−15
−10
−5
0
−6e−04
−3e−04
0e+00
3e−04
−0.2
−0.1
0.0
EC PRs less accessible than national PRs
EC PRs more accessible than national PRs

Comparing language of Commission and national government press releases in a sample matched on topic distributions in full texts



Mean value of respective indicator and 95% c.i.s

WZB

What to learn from this example

- The European Commission still cultivates *a very technocratic style* of communicating to the wider European citizenry
- In a politicized context this is *risky, if not dangerous*
  - *Leaves misunderstandings/interpretation/framing to others*
  - *Supports the narrative of a detached technocratic elite in Brussels*
- And again: The systematic information hidden in textual data and online archives can help to make relevant things visible …

Example III

Do supranational politicians play up the state of emergency?

Motivation

- **The claim:** In order to protect or expand their competences in the face of politicization, supranational institutions exploit and play up crises – 'emergency politics' (Draghi: Whatever it takes …)

- **The debate:** Hard to distinguish objective crisis pressures from strategic incentives in individual cases

➢ **Do supranational institutions emphasise the state of emergency more strongly over time and more than other actors?**

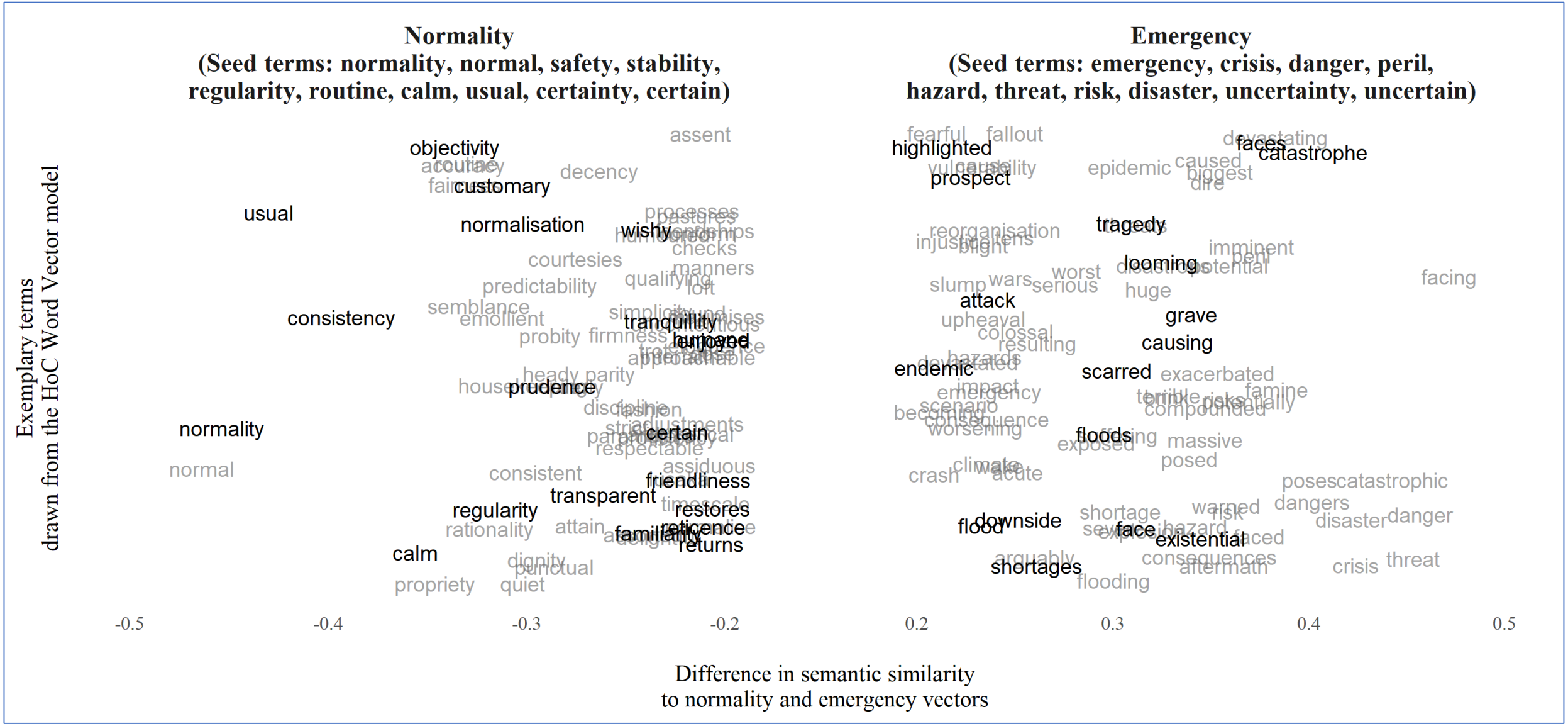
(Basic) measurement idea

- Capture whether words in speeches of supranational politicians (European Commission and ECB) imply an 'emergency'

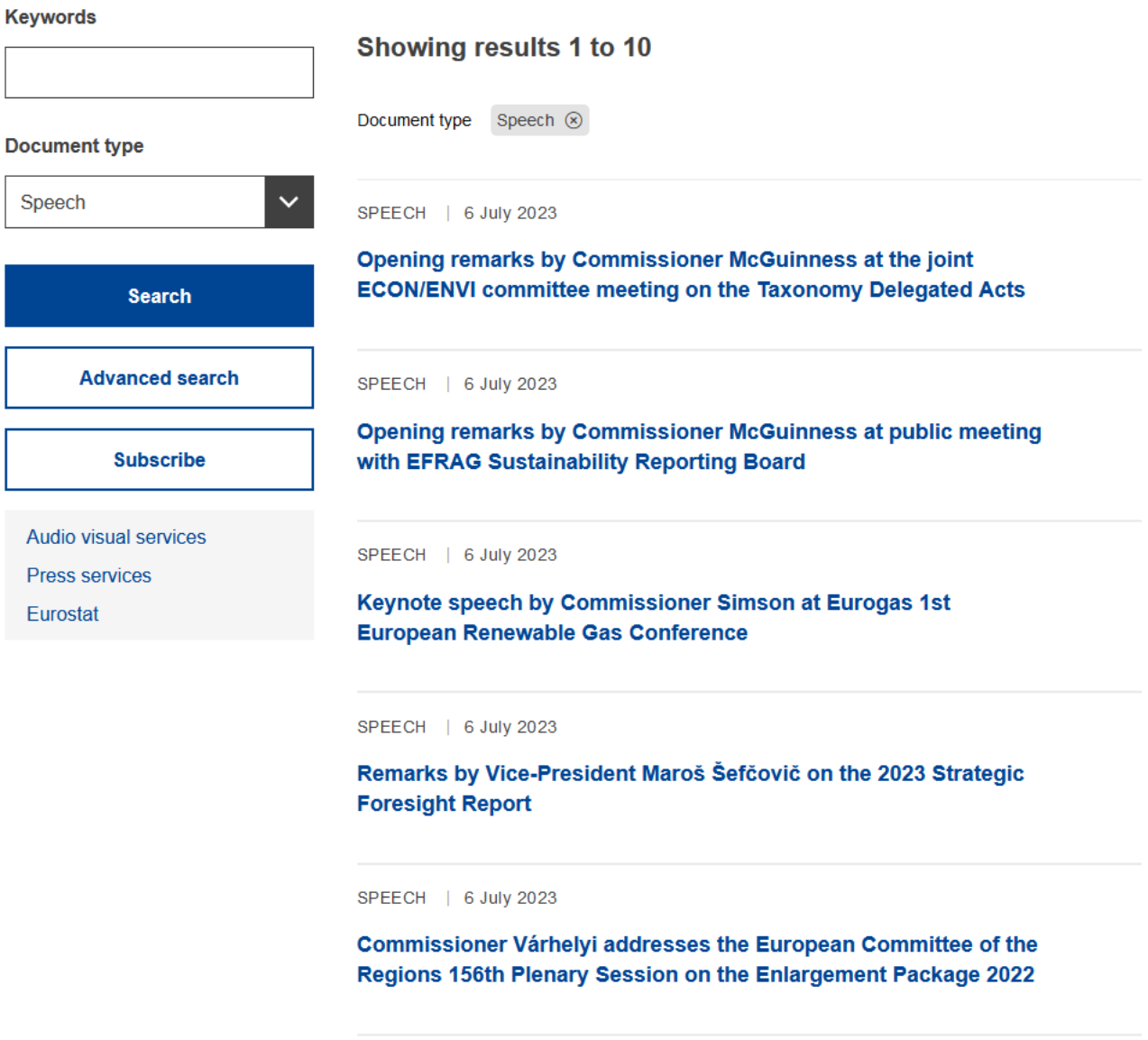
- Latent semantic scaling / semantic projection
  - Train a *word vector model* on a large corpus of political English (1.8 mio. House of Commons Speeches from ParlSpeech, Rauh & Schwalbach 2020)
  - Extract the average vector of *a few seed terms* related to *normality vs. emergency* from this model
  - Calculate cosine similarity of the vectors of all other words to these two seed vectors and subtract the two values from each other (for each word in the corpus)
    → normality|emergency scale
  - Weight all words in Commission and ECB speeches with the resulting values and average by speech

Word vector / word embedding intuition (distributional semantics)

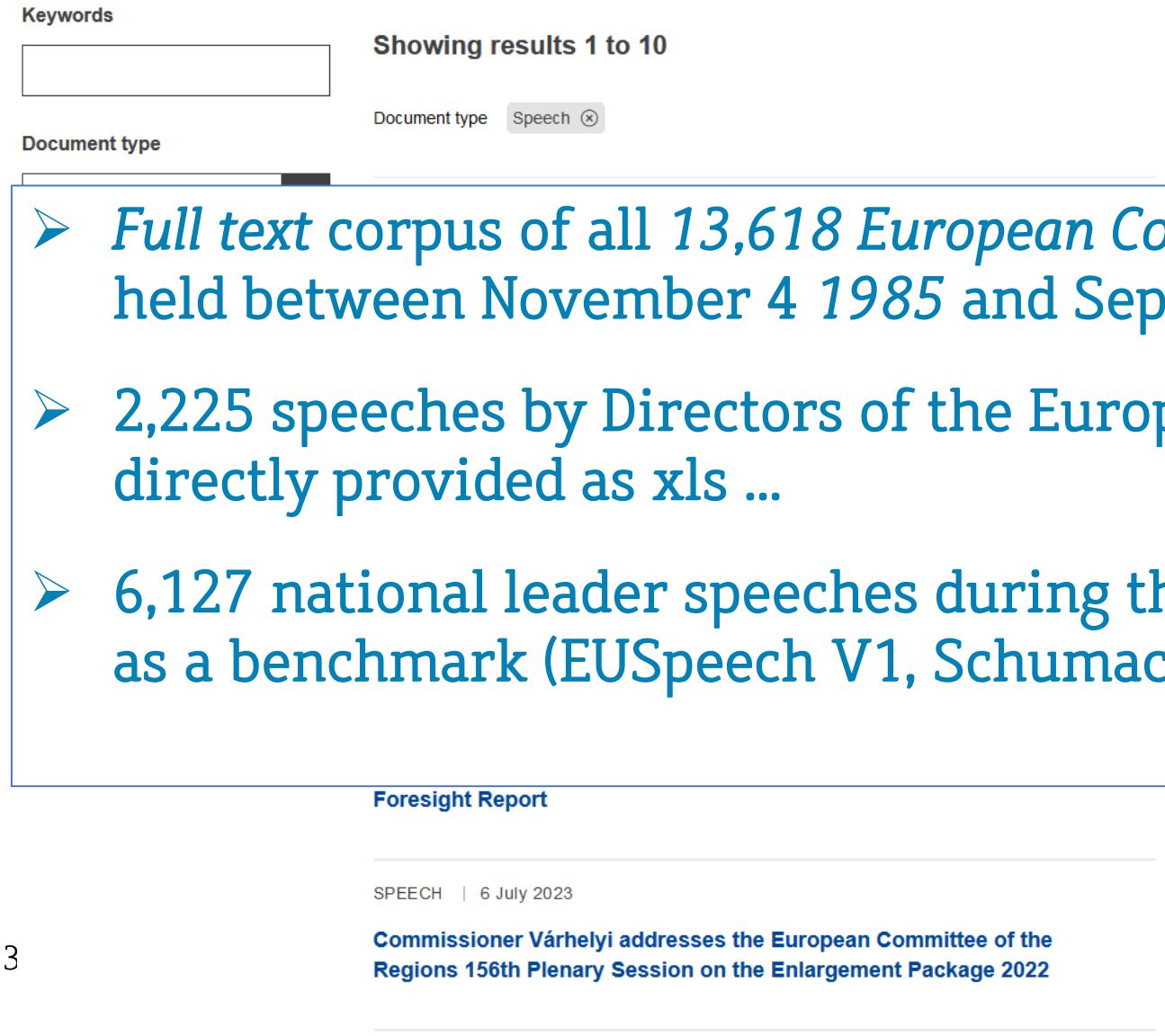
We are facing an emergency that threatens our safety.
We are facing a catastrophe that threatens our security.
…
We enjoy more safety.
We enjoy that everything is normal.
…
…
…
…


Normality
(Seed terms: normality, normal, safety, stability, regularity, routine, calm, usual, certainty, certain)
Emergency
(Seed terms: emergency, crisis, danger, peril, hazard, threat, risk, disaster, uncertainty, uncertain)
Exemplary terms drawn from the HoC Word Vector model
objectivity
customary
usual
normalisation
wishy
courtesies
checks
manners
predictability
qualifying
consistency
tranquility
prudence
normality
certain
normal
consistent
friendliness
transparent
regularity
restores
familiarity
returns
calm
highlighted
prospect
faces
catastrophe
tragedy
looming
attack
grave
causing
endemic
scarred
floods
downside
flood
face
existential
shortages
facing
-0.5
-0.4
-0.3
-0.2
0.2
0.3
0.4
0.5
Difference in semantic similarity to normality and emergency vectors

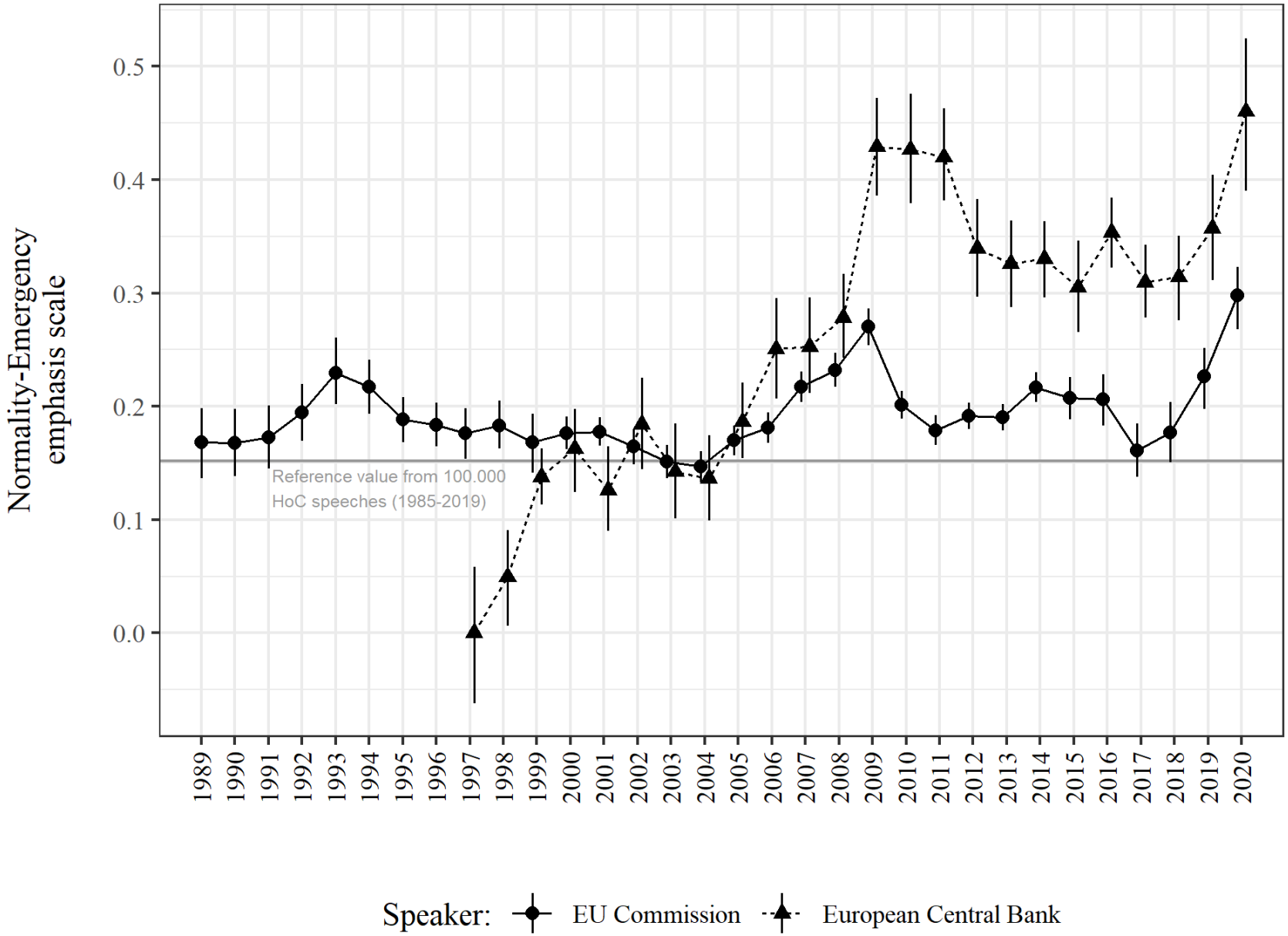
Web scraping the ‘EC press corner’ (again …)

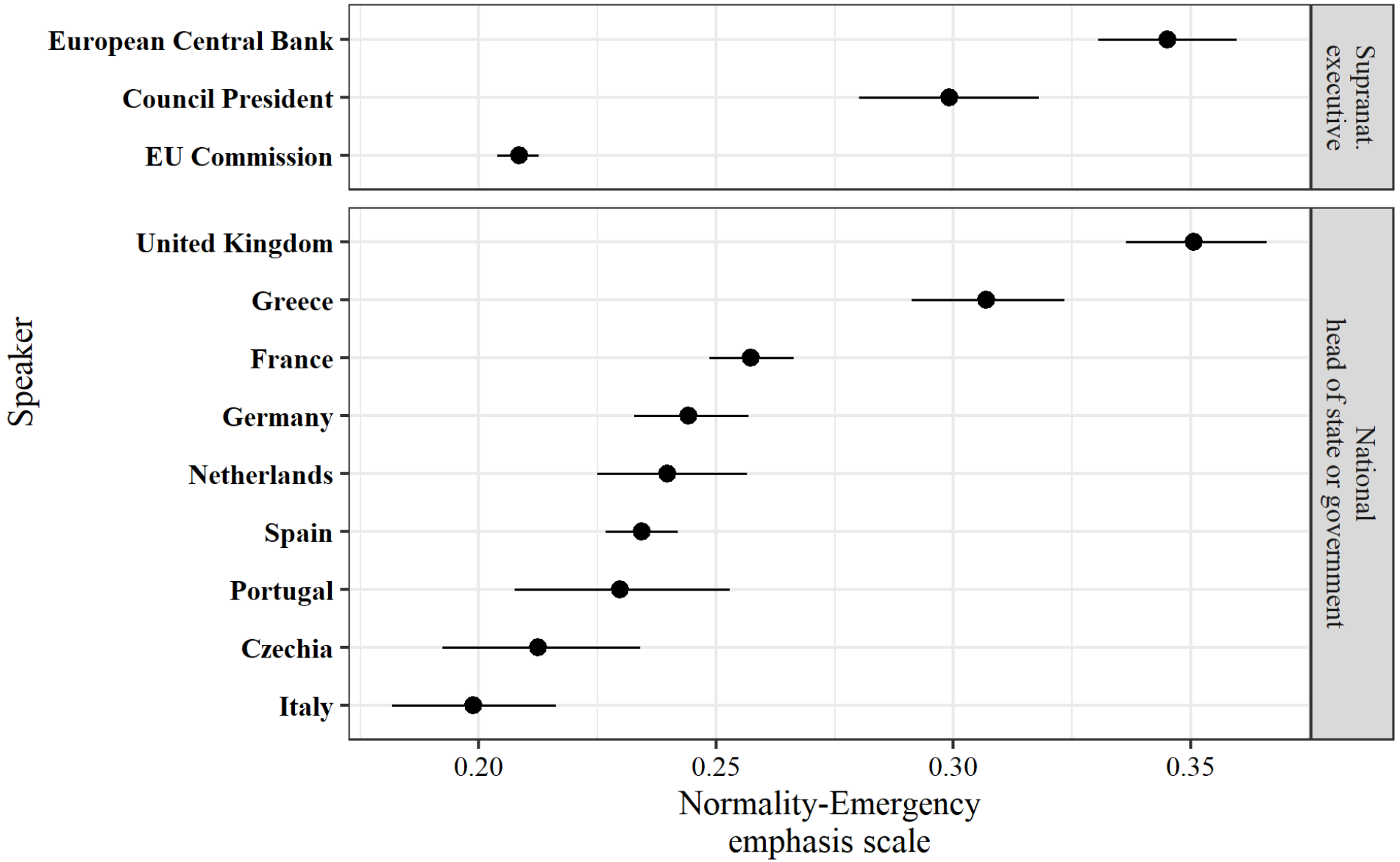


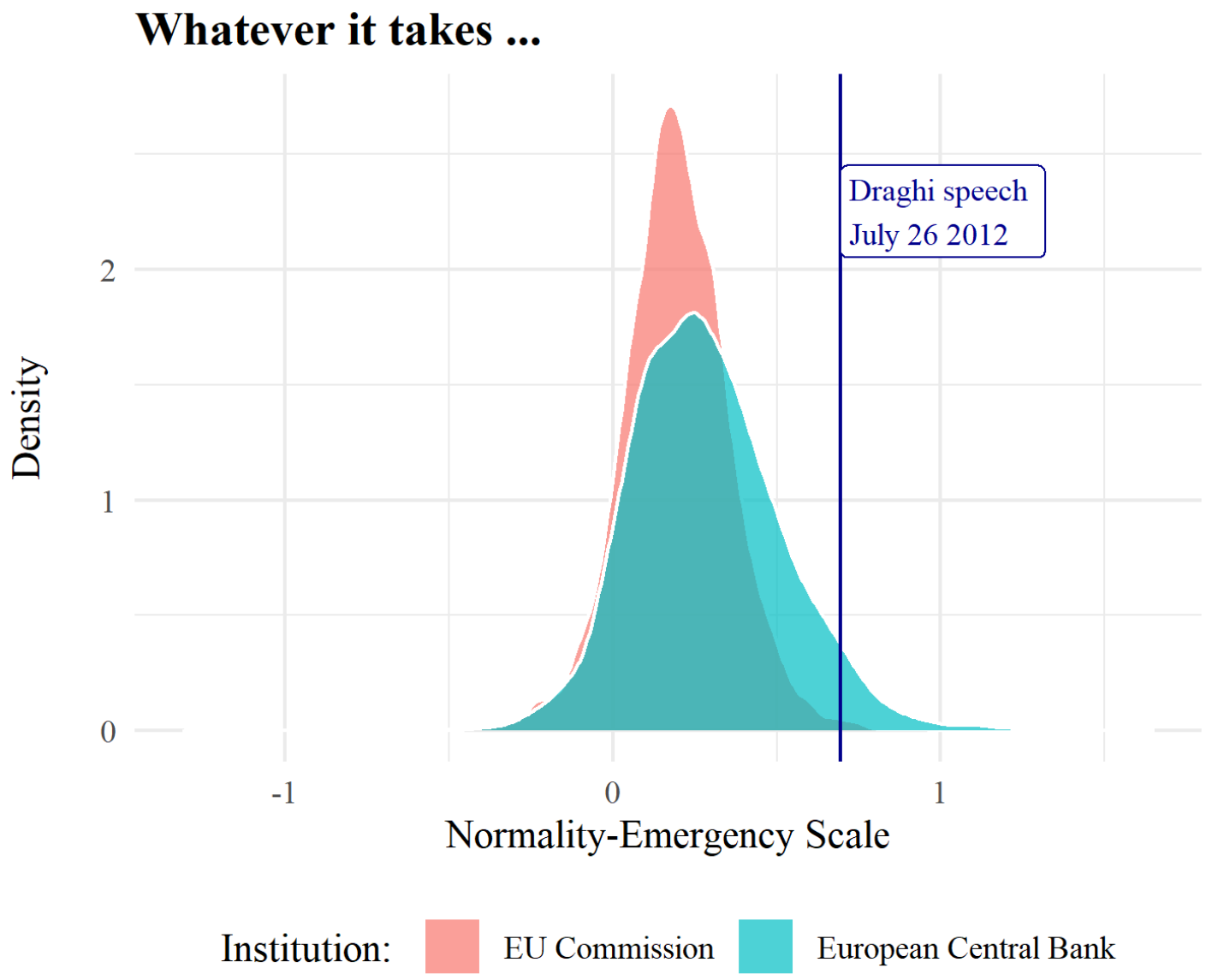
Web scraping the ‘EC press corner’ (again …)



- *Full text* corpus of all *13,618 European Commissioner speeches* held between November 4 *1985* and September 3 *2020*
- 2,225 speeches by Directors of the European Central Bank (ECB) directly provided as xls …
- 6,127 national leader speeches during the Eurocrisis 2009–2015 as a benchmark (EUSpeech V1, Schumacher et. al. 2019)


Normality-Emergency emphasis scale
0.5
0.4
0.3
0.2
0.1
0.0
Reference value from 100.000
HoC speeches (1985-2019)
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Speaker:
EU Commission
European Central Bank


European Central Bank
Council President
EU Commission
Supranat. executive
United Kingdom
Greece
France
Germany
Netherlands
Spain
Portugal
Czechia
Italy
National head of state or government
Speaker
0.20
0.25
0.30
0.35
Normality-Emergency emphasis scale



Takes-aways from this example

- Supranational actors sometimes play up the 'emergency' in their speeches
- But this is not a consistent phenomenon and can hardly be explained by contested competences alone ...

My main messages to you

- Rather big (political) claims can be addressed by rather simple text-as-data ideas! (KISS)
- Some learning curves show increasing returns!
- It's fun!
- Look around, transfer ideas, connect with neighbouring disciplines!
- Descriptives and benchmarks matter!
- Often 80% data collection & cleaning, 20% analyses
  - *Plan accordingly, start small and scale up*
  - *Be polite to the servers and store raw online data locally before parsing information*
  - *Share your data and scrapers (in a citable manner)!*

WZB

Wissenschaftszentrum Berlin
für Sozialforschung

Thanks for your attention and your questions!

Papers, data, scripts:
www.christian-rauh.eu

Contact:
christian.rauh@wzb.eu
@ChRauh