ParlaMint

The Impact of Parliamentary Datasets for Society and (Data) Science

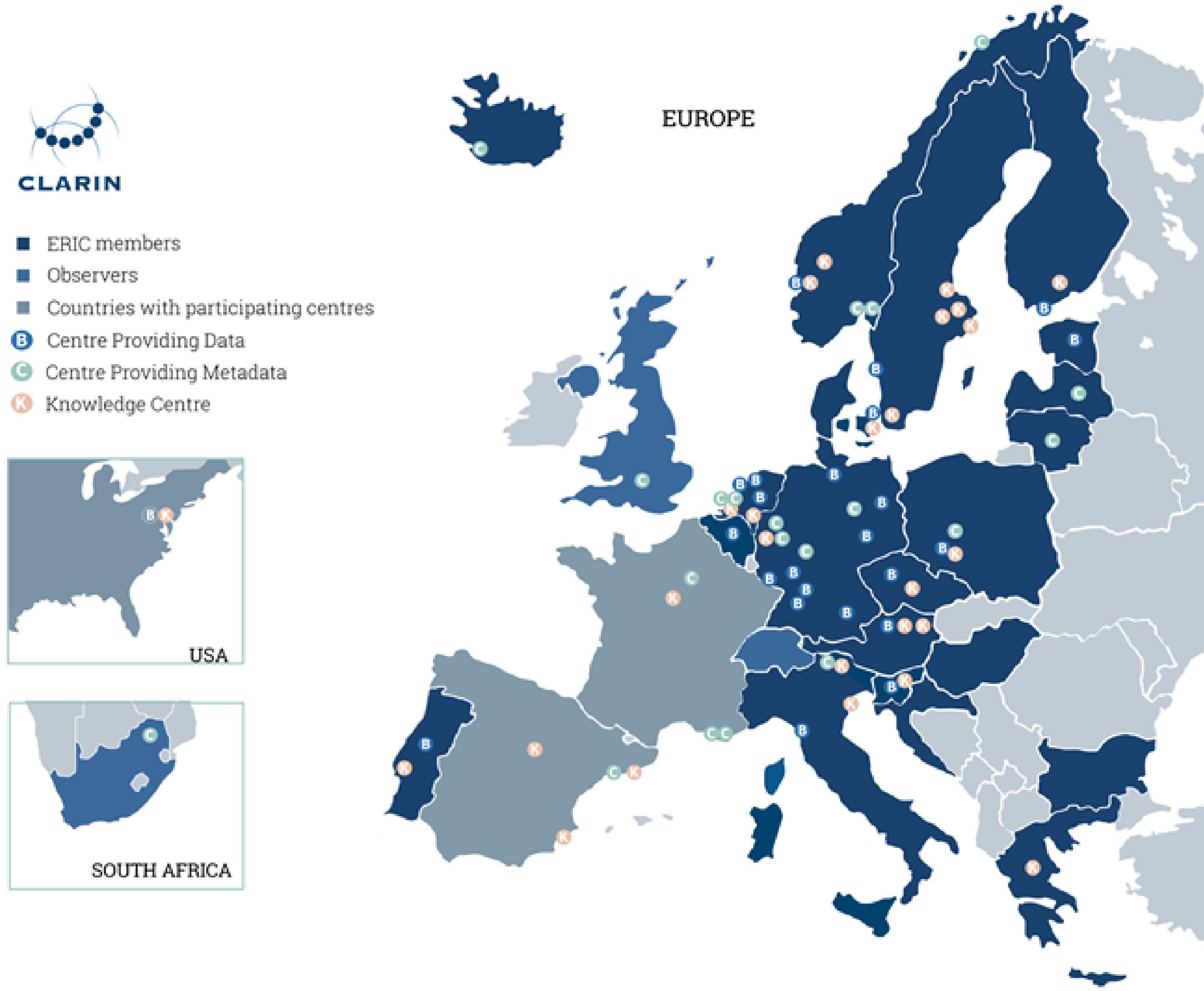
Maciej Ogrodniczuk and the ParlaMint Team

DH Forum | European Parliament's STOA Panel | Brussels, April 26, 2023

ABOUT CLARIN

Common Language Resources and Technology Infrastructure:

- a **distributed network** of more than 60 centres
- **ERIC** status since 2012
- provides easy and sustainable **access** for scholars in the **humanities and social sciences** and beyond
- to **digital language data** (in written, spoken, video or multimodal form)
- and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them
- through a **single sign-on** environment
- that serves as an **ecosystem for knowledge exchange**
- with many services **integrated in EOSC** (European Open Science Cloud)



THE MOTIVATION

Debates are:

- a verified communication channel between the elected political representatives and society
- a reflection of the interests of the whole national community

We need to be able to:

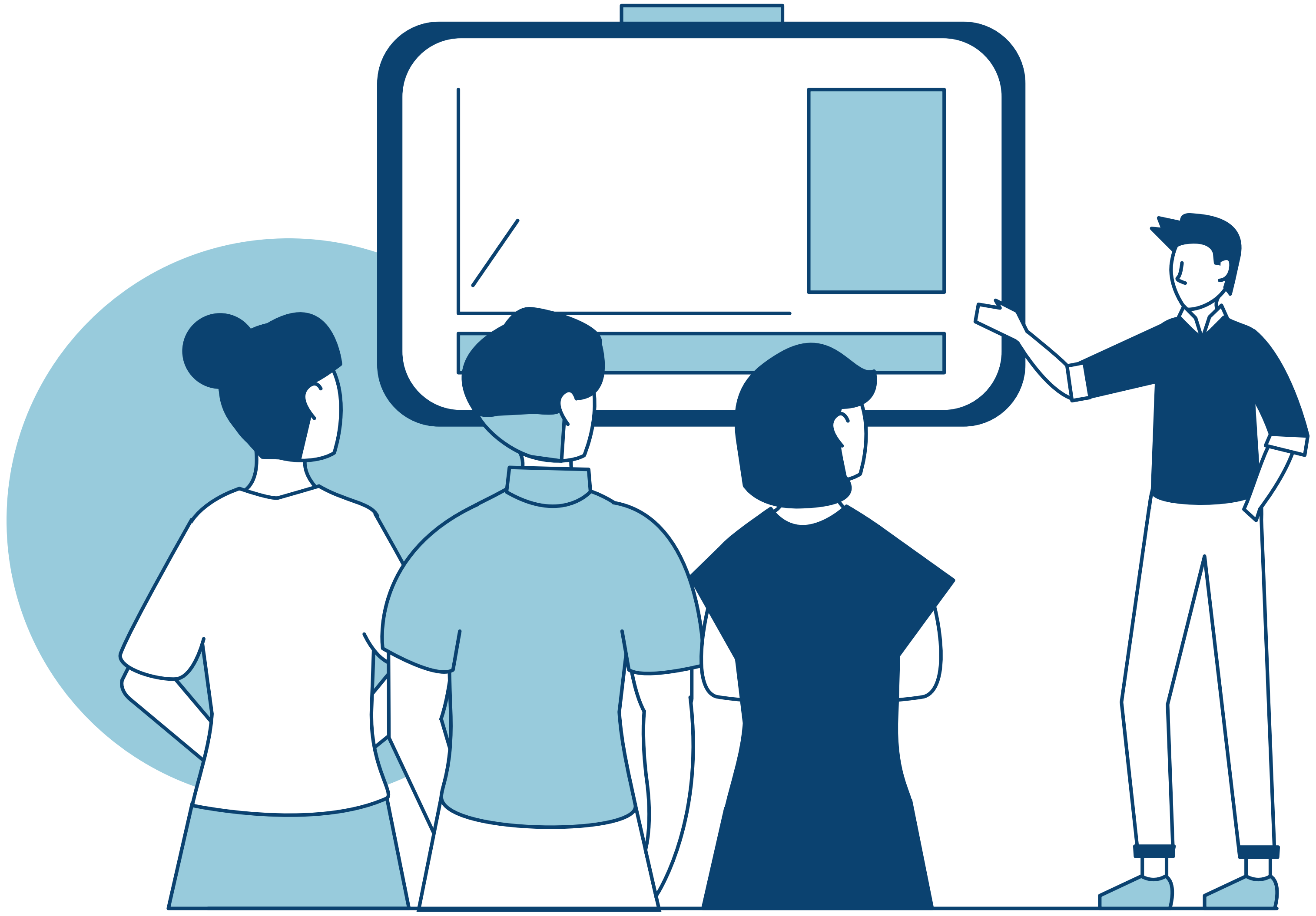
- **analyze** such data in multiple languages
- **compare** the data in a cross-lingual context
- **track** the pan-European discussion
- make this data **interpret**able and highly communicative with respect to society

THE PROBLEM

This data is really **valuable** and **timely released** by national parliaments.

But:

- each country has a **different parliamentary system**
- the data comes in **different formats**, with **different metadata**
- the debates have **varied structures**



THE SOLUTION: PARLAMINT

A CLARIN-ERIC project which managed to:

- propose a harmonised **representation format** for parliamentary debates
- compile a collection of **parliamentary datasets**
- process the compiled corpora **linguistically**
- make them available for **download and search**
- initiated several **use cases**

METHODOLOGY

Key properties:

- use of **virtual subcorpora** to enable the tracking of various phenomena in the data
- this methodology is **scalable** to current events: pandemics, economic crises, environmental issues, wars
- observe democratic processes through **parliamentary analytics**

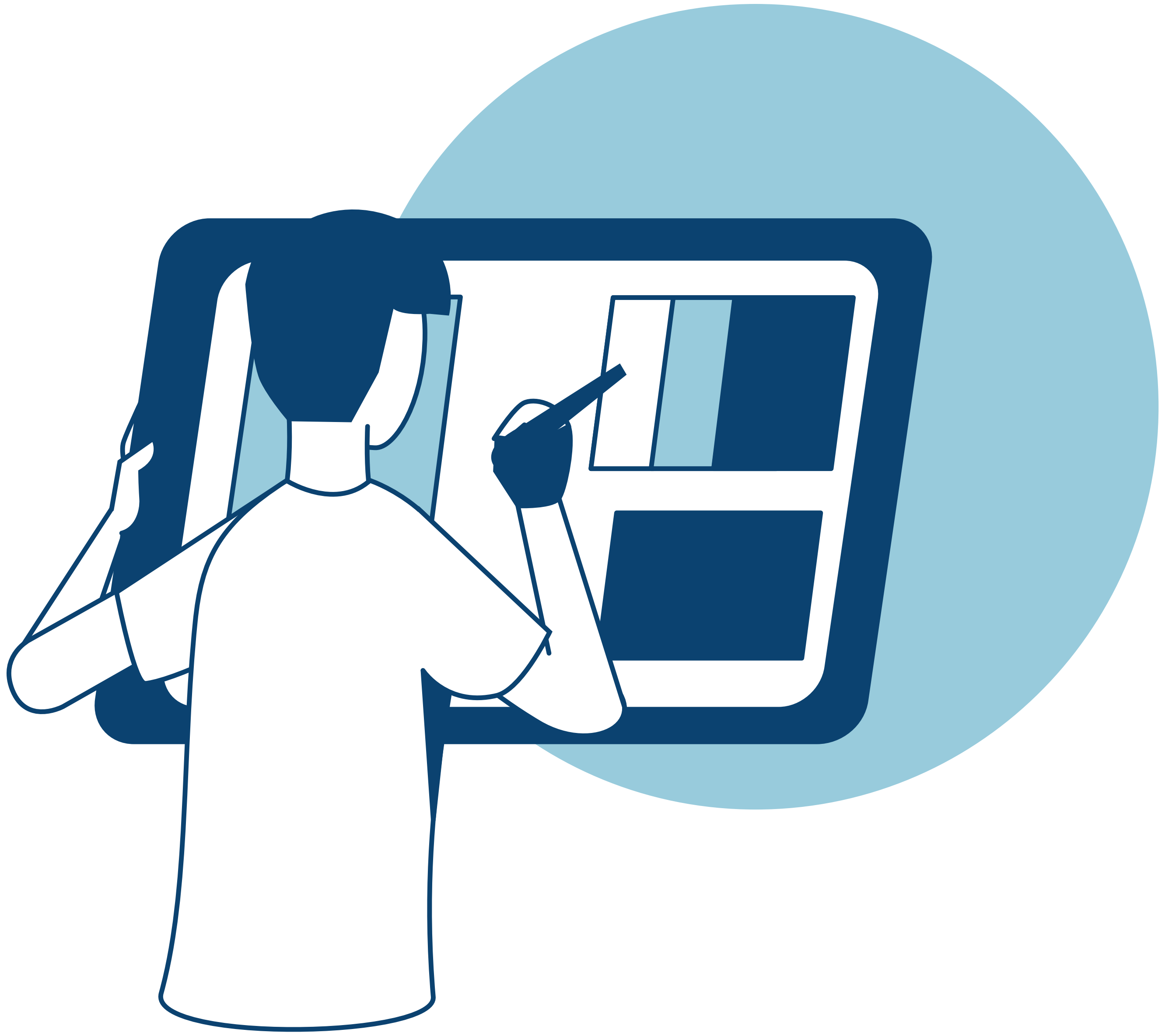
PARLAMINT INFRASTRUCTURE

Not just:

- **the dataset**
- its **joint representation model**
- common **metadata**, **document structure**, **linguistic annotation**

But also:

- good practice **guidelines**
- validation **procedures**
- **documentation** and **samples**



THE PARLAMINT TEAM

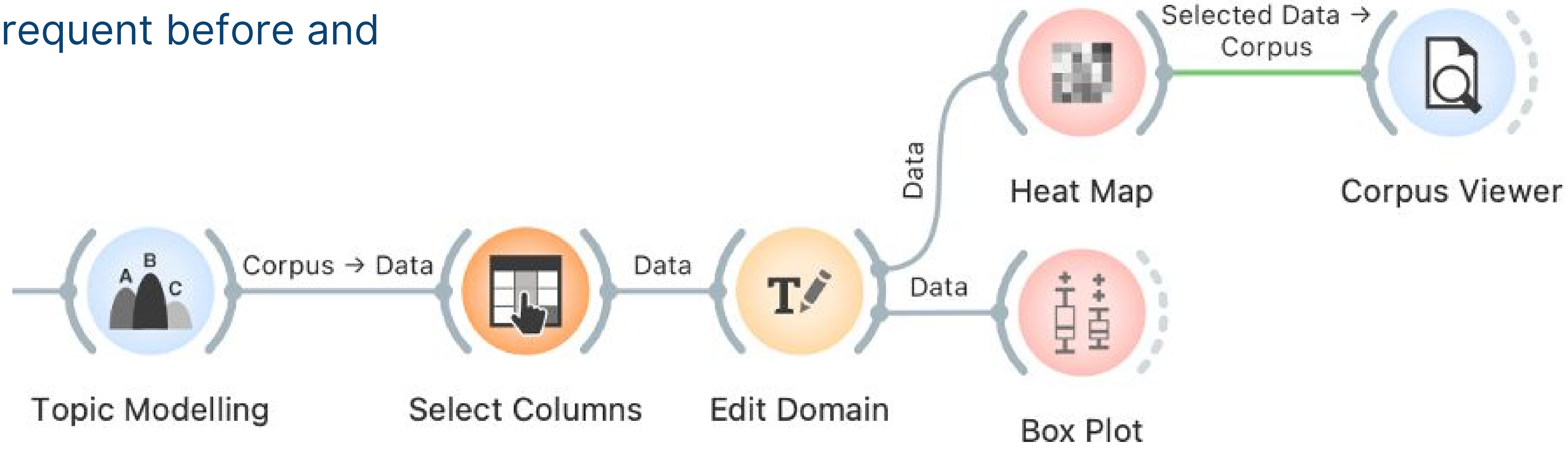
We are:

- a wide, **pan-European community**
- offering **linguistic expertise**
- developing **tutorials**
- creating **impact stories**
- organizing **hackathons**
- using this data for creating **language models** to be used in data science

SHOWCASE 1: What’s on the agenda?

Research questions:

- Which topics are characteristic of the corpus?
- Which topics did MPs debate the most?
- Which topics were more frequent before and during the pandemic?

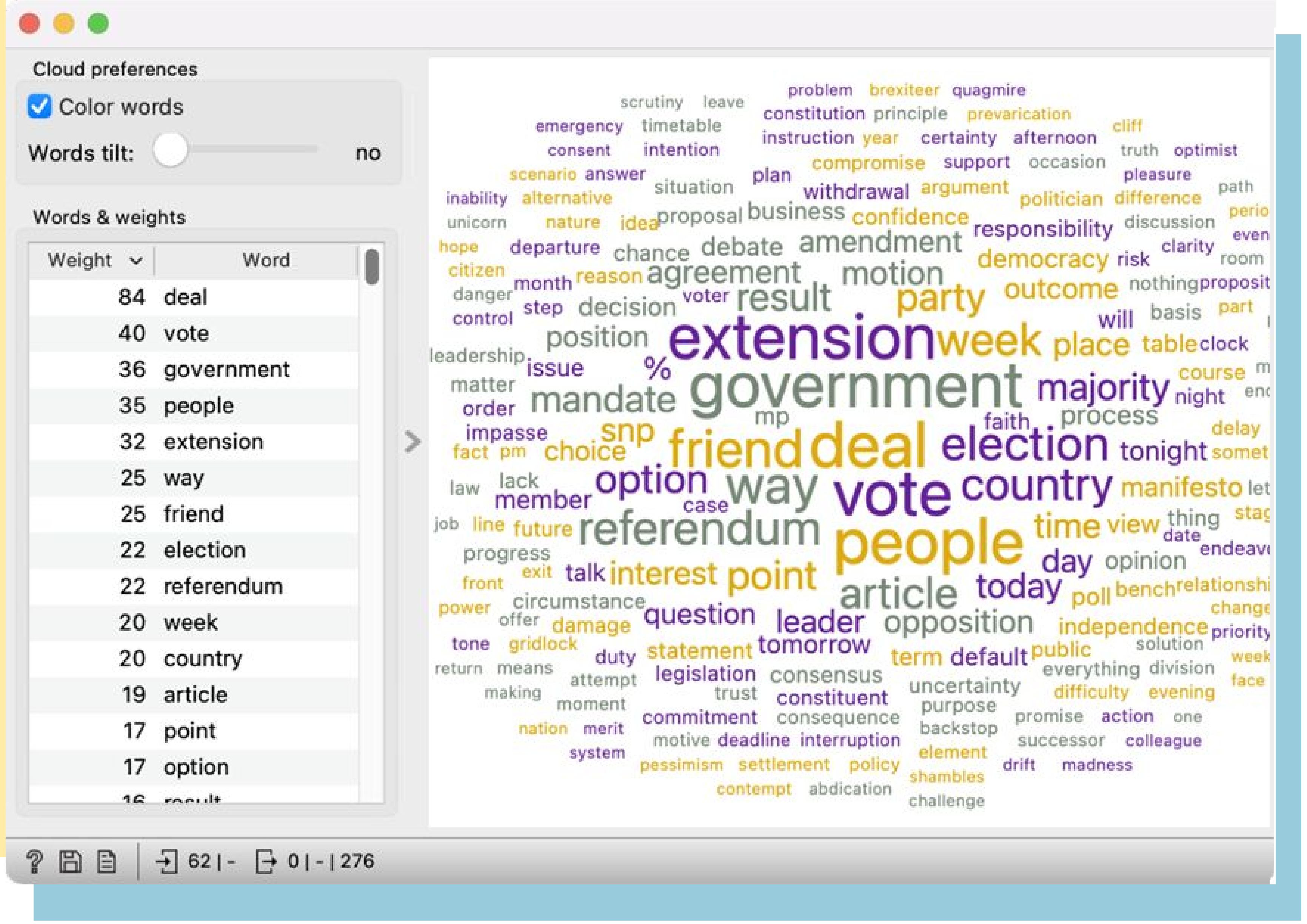


Pretnar Žagar A., Pahor de Maiti K., Fišer D. (2022). *What's on the agenda? Topic modelling parliamentary debates before and during the COVID-19 pandemic*. https://sidih.github.io/agenda/

SHOWCASE 1: What’s on the agenda?

A few findings from the British corpus:

- deals, voting, government, people, and extension were most frequently discussed
- speeches characterized by *referendum* probably refer to Brexit
- surprisingly, the word *virus* was quite infrequent



Cloud preferences

☑ Color words

Words tilt: no

Words & weights

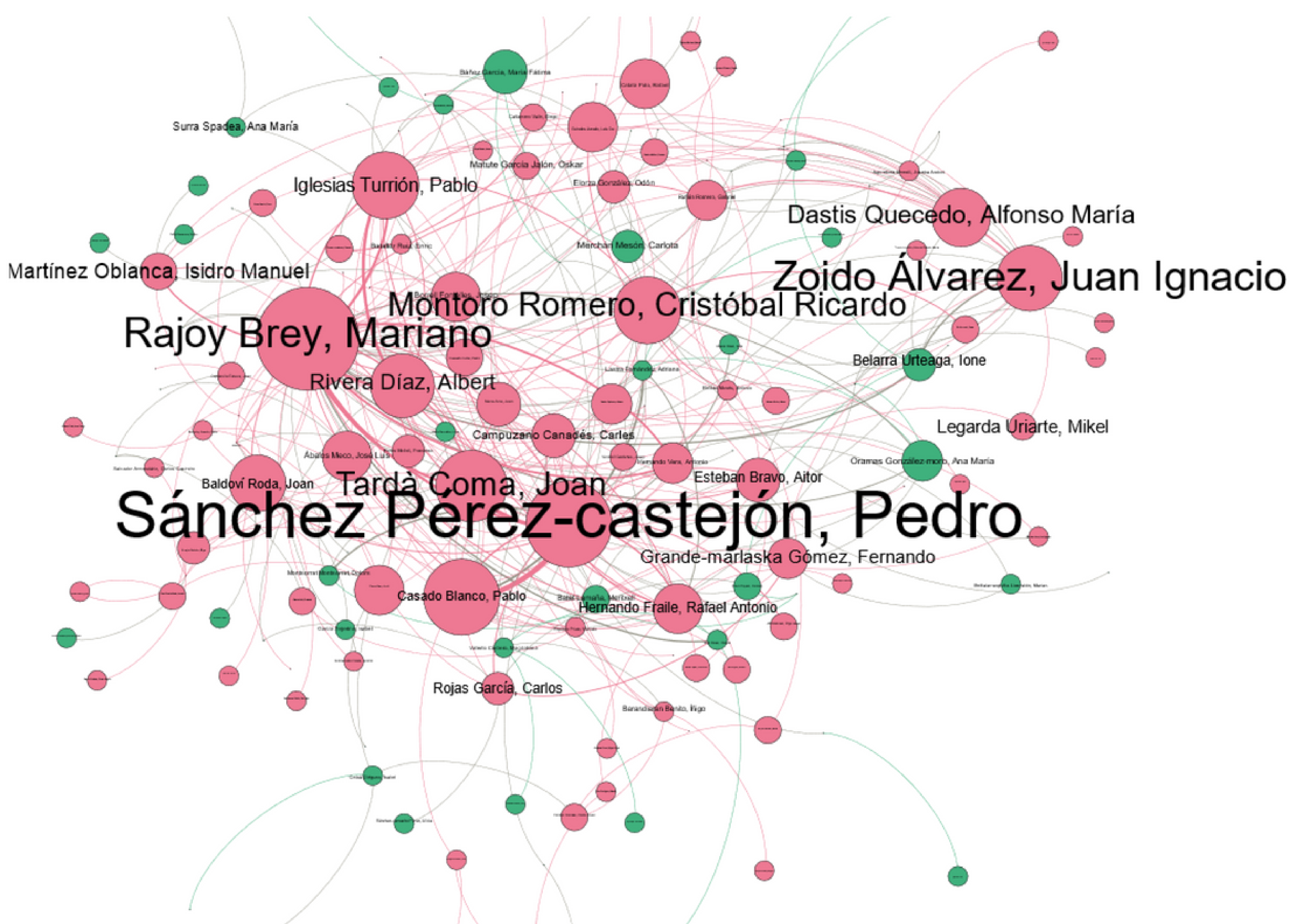
Weight	Word
84	deal
40	vote
36	government
35	people
32	extension
25	way
25	friend
22	election
22	referendum
20	week
20	country
19	article
17	point
17	option
16	result

62 | - 0 | - | 276

SHOWCASE 2: Networks of Power

Analyzing the networks that emerge from parliamentarians mentioning one another:

- **Argumentative Power**: How can speeches and mentions give insights into the power of MPs?
- **Structural Power**: How do the speech practices of female and male MPs relate to topic and power distribution?

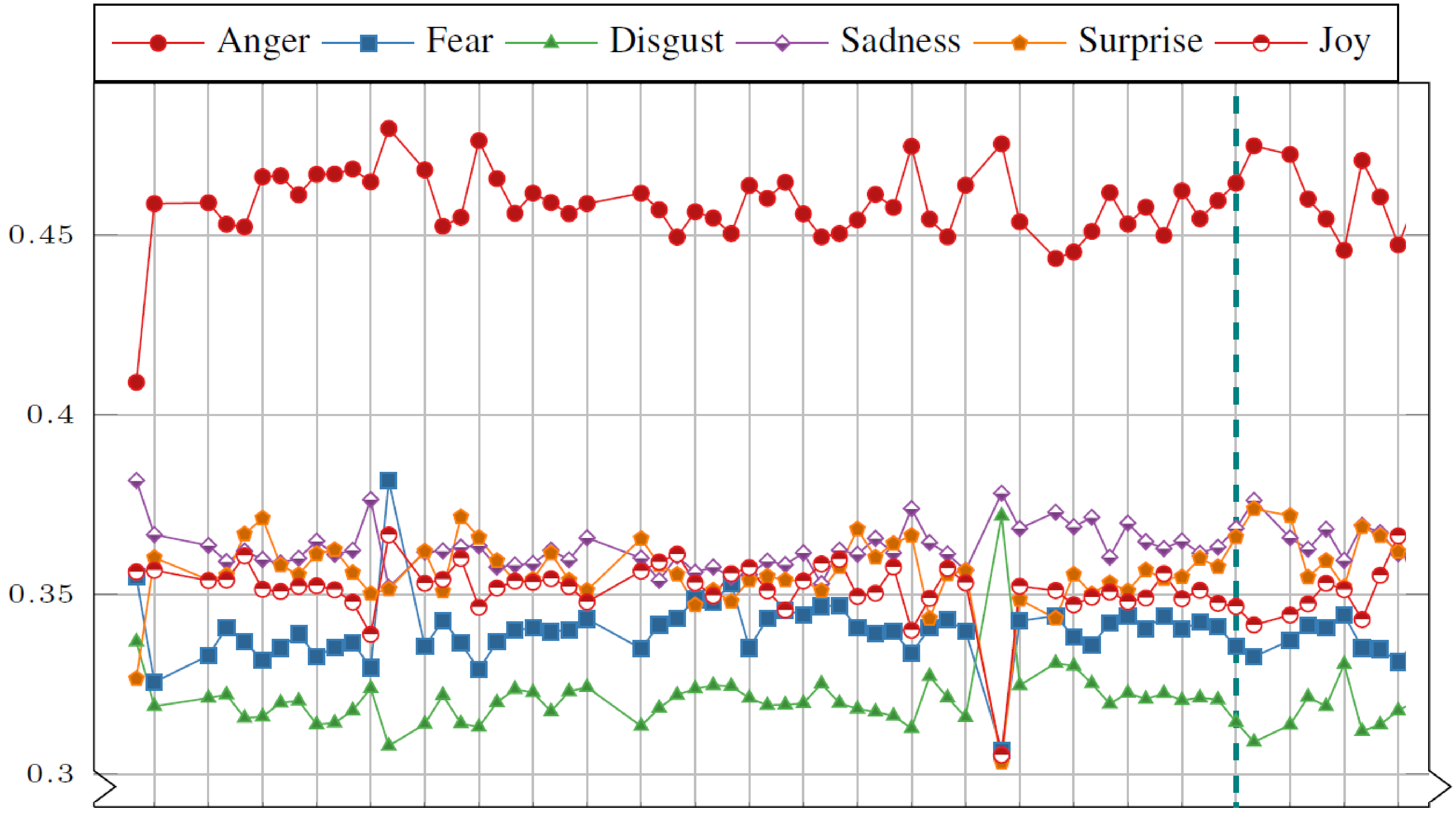


Surra Spadea, Ana María
Iglesias Turrión, Pablo
Dastis Quecedo, Alfonso María
Martínez Oblanca, Isidro Manuel
Zoido Álvarez, Juan Ignacio
Montoro Romero, Cristóbal Ricardo
Rajoy Brey, Mariano
Belarra Urteaga, Ione
Rivera Díaz, Albert
Legarda Uriarte, Mikel
Campuzano Canadés, Carles
Baldoví Roda, Joan
Tardà Coma, Joan
Esteban Bravo, Aitor
Sánchez Pérez-castejón, Pedro
Grande-marlaska Gómez, Fernando
Casado Blanco, Pablo
Hernando Fraile, Rafael Antonio
Rojas García, Carlos

Angermeier J., Bruncrona A., Evkoski B., Gu Z., Harlamov O., Islam J., Janicki M., Leiminger L., Marjanen J., Skubic J., Tamper M. (2022). *ParlaMint. Networks of Power.* https://www.helsinki.fi/assets/drupal/s3fs-public/from_d7/dhhparlamintdraftfinal3.pdf

SHOWCASE 3: Emotions Running High

- Investigating **polarization** of politics by assigning emotion scores to speeches
- **anger** being the dominant emotion
- the ruling party showing more stable emotions compared to the opposition

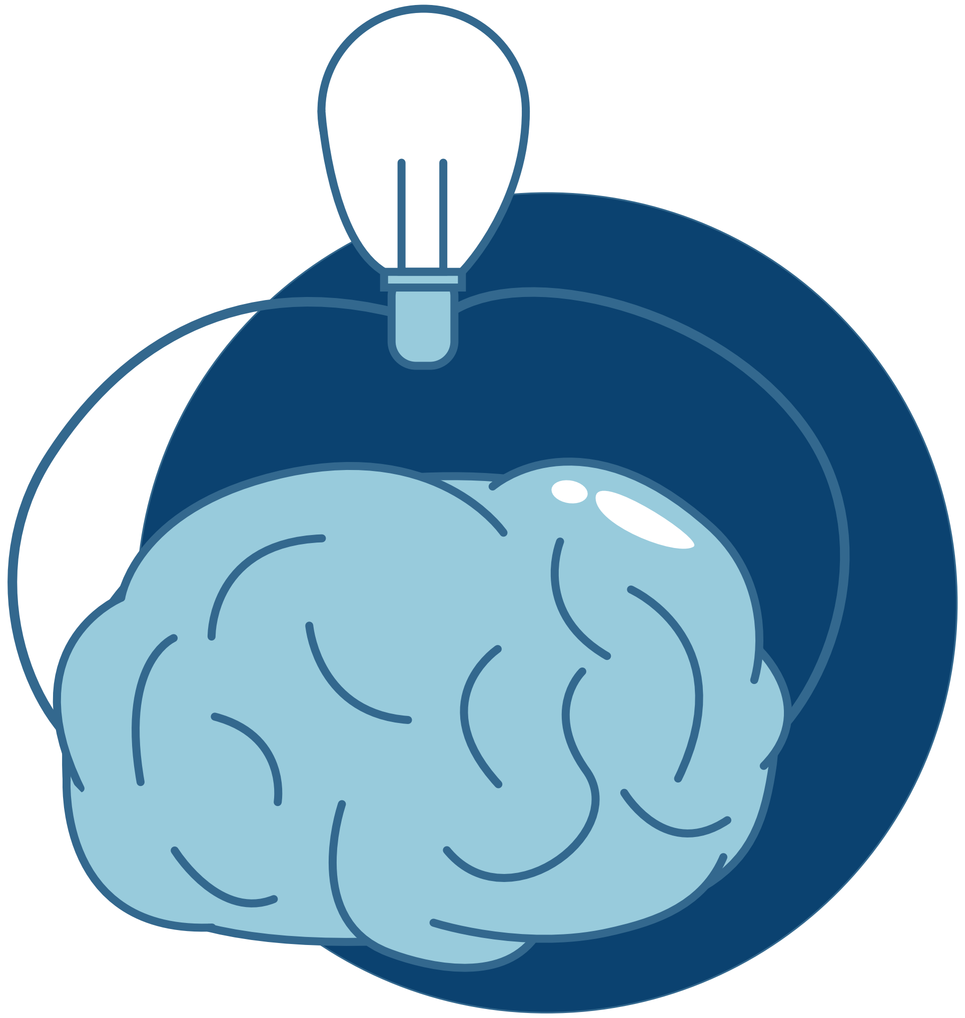


Kurtoğlu Eskişar G. M., Çöltekin Ç (2022). *Emotions Running High? A Synopsis of the state of Turkish Politics through the ParlaMint Corpus*. In Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference, pp. 61–70, ELRA. https://aclanthology.org/2022.parlaclarin-1.10.pdf

FROM SHOWCASES TO REAL APPLICATIONS

- The project will **boost research** in digital humanities, linguistics, political science, social science and other related fields.

- Different reference corpora could be produced with parliamentary records and different analyses could use **our methodology**.



CURRENT ACTIONS

ParlaMint is a constant effort!

- 10 more languages
- new subcorpus starting with the date of the Russian invasion on Ukraine
- translating all data to English to perform semantic tagging
- a multimodal experiment of aligning audio recordings with transcriptions

BEYOND PARLAMINT

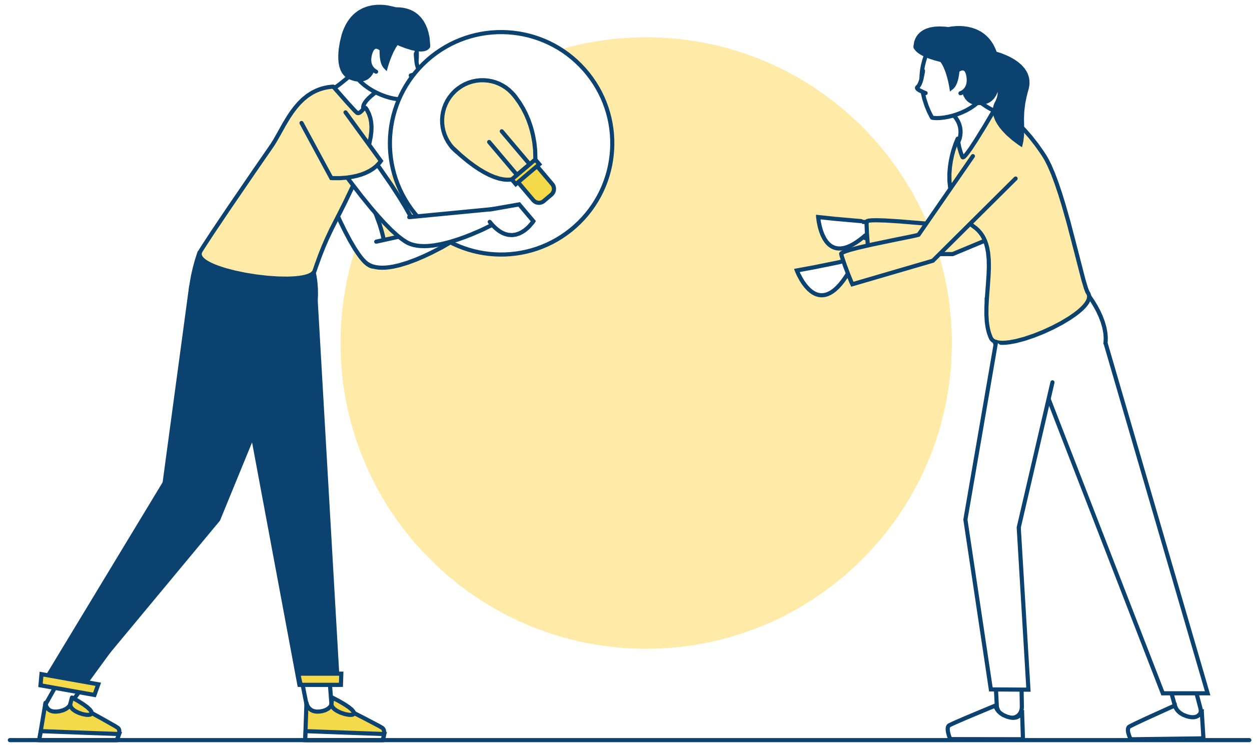
Into the future:

- data coming from the **European Parliament** and **regional parliaments**
- **voting results**
- documents related to the **law-making** process
- new and emerging technologies, concentrating on processing **multimodal data** or producing **live datasets**

KEY TAKEAWAYS

ParlaMint is a **solid data-intensive infrastructure** to make parliamentary debates across Europe more transparent and comparable.

It starts a long-term impact action of **bringing the accurate and trustworthy information** to all interested parties in a cross-lingual perspective never possible before!



Thank you for your attention!

https://www.clarin.eu/parlamint