

# ParlaMint



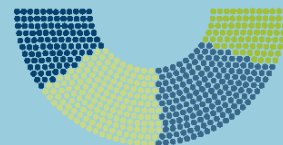
**Multilingual Comparable Corpora  
of Parliamentary Debates for Digital Humanities**

Maciej Ogrodniczuk and the ParlaMint Team

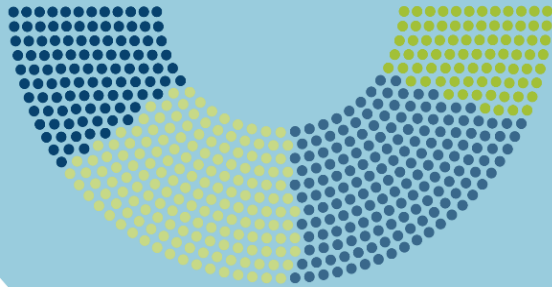
September 12, 2023 | CLARIN-LV conference



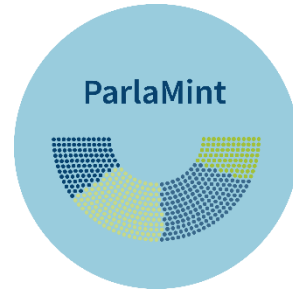
ParlaMint



# ParlaMint



- » The ParlaMint project
- The ParlaMint schema
- The ParlaMint process
- The ParlaMint data and analytics
- Usage examples (and inspirations)
- What's next?



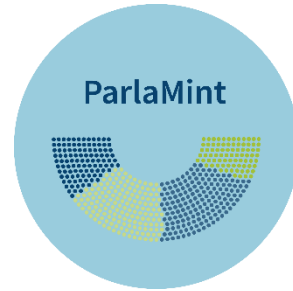
# Motivation and background

## Parliamentary data:

- shows real potential for reuse and re-purposing within many fields of study
- is considered a rich data type (in metadata but also extralinguistic clues)
- but: created under specific circumstances that need to be well understood before strong conclusions can be drawn

## Where did we start?

- many corpora existed, but were encoded in many different ways, limiting interchange
- there was a will of the community to converge:
  - parliamentary corpora one of the key [resource families in CLARIN](#)
  - [ParlaCLARIN](#) workshops at LREC 2018, 2020, 2022, [ParlaFormat](#) workshop (2019)
  - COST Action ParlaNT proposed (Parliamentary Data and Language Technology)



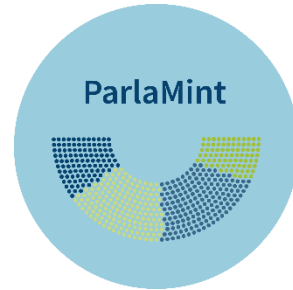
# What is ParlaMint?

A project financially supported by CLARIN-ERIC which contributed to the creation of multilingual corpora of parliamentary sessions which are:

- uniformly annotated
- comparable
- interpretable
- highly communicative to the society (researchers, journalists, NGOs, citizens, etc.)

But it's much more than data!

- the schema
- the process
- the community-building effort



# How was ParlaMint implemented?

## **Phase 1** (*July 2020 – Sep 2020*):

- 4 pilot languages: Bulgarian, Croatian, Polish and Slovene
- 2 subcorpora: COVID-19 (Nov 2019 – 2020) and reference (2015 – Nov 2019)

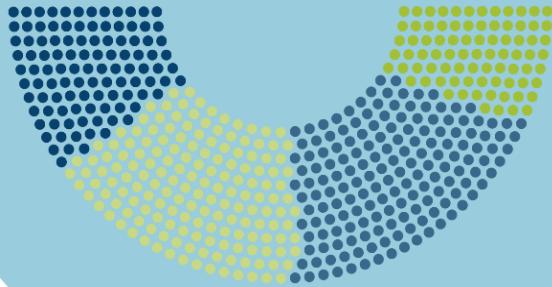
## **Phase 2** (*Dec 2020 – May 2021*):

- more languages: Czech, Danish, Dutch (also from Belgium), English, French (also Belgium), Hungarian, Icelandic, Italian, Latvian, Lithuanian, Spanish and Turkish

## **Phase 3** (*December 2021 – September 2023*):

- even more languages: Bosnian, Catalan, Croatian, Estonian, Galician, German (Austria), Greek, Norwegian, Portuguese, Russian, Serbian, Swedish and Ukrainian
- extending existing corpora with the most recent data
- MT of the data to English, semantic tagging, war subcorpus

# ParlaMint



The ParlaMint project

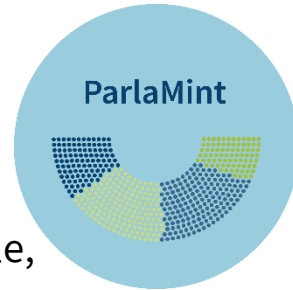
» The ParlaMint schema

The ParlaMint process

The ParlaMint data and analytics

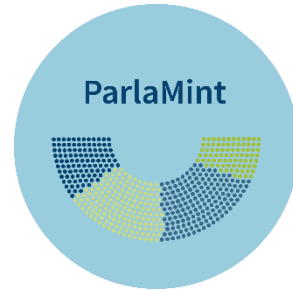
Usage examples (and inspirations)

What's next?



# ParlaMint encoding

- The corpora are encoded as uniformly as possible to make them interoperable, so that e.g. they can be validated, converted etc. automatically, however:
  - the debates have very different source encoding
  - they are differently structured, contain different information and reflect different parliamentary traditions
  - each corpus is produced by a separate partner
- ParlaMint schema is TEI-based (more accurately, based on Parla-CLARIN TEI Schema for Corpora of Parliamentary Proceedings)
- An important disclaimer: TEI Guidelines are large, complex and generic. Our schema concentrates only on aspects of the Guidelines most likely to be of use in encoding parliamentary proceedings.

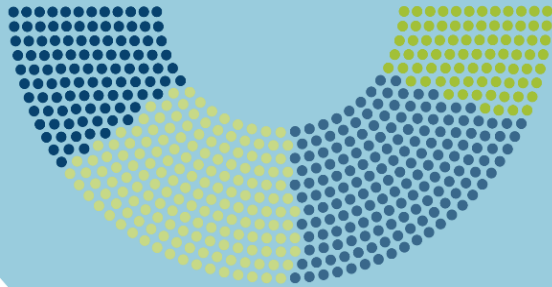


# Encoding aspects

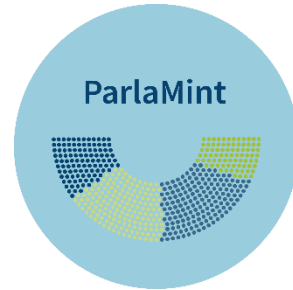
- **Structure:** legislative periods, sessions, topics, speeches, transcription variants
- **Metadata:** mandates, titles, parliamentary bodies, locations, dates and times
- **Speakers:** sex, date of birth, education, party membership, links to external resources
- **Political parties:** name(s), history, relations
- **Speeches:** speaker, text, comments, verbal interruptions, (non-)verbal incidents
- **Linguistic annotation:** PoS tagging, normalisation, syntax etc.
- **Multimedia:** audio and video, facsimile of original



# ParlaMint



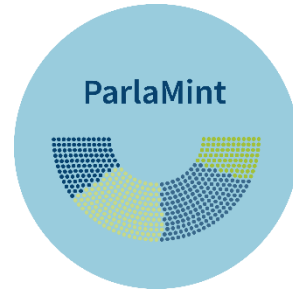
- The ParlaMint project
- The ParlaMint schema
  - » The ParlaMint process
- The ParlaMint data and analytics
- Usage examples (and inspirations)
- What's next?



# The ParlaMint process

1. Acquire the parliamentary data and metadata
2. Convert them into [the ParlaMint schema](#)
3. Validate (formally and qualitatively)
4. Annotate linguistically: UD morphosyntax and syntax + named entities
5. Machine-translate the corpus into English
6. Make the corpora available to download and search
7. Build use cases based on the corpus data

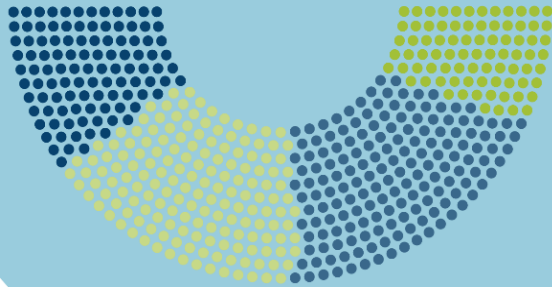




# Challenges at every step

- Different political and thus, parliamentary systems (unicameral, bicameral)
- Getting data, e.g.:
  - scraping it from the parliamentary websites
  - obtaining via parliamentary or third-party API
  - retrieving from an already maintained parliamentary corpus
- Converting data:
  - from HTML to basic TEI XML and then to the ParlaMint format
  - through XSLT stylesheets and Python, Perl and Bash scripts
  - writing own scripts in Perl/Java/Python with heuristics for difficult parts
- Linguistic processing:
  - UD-based annotation and NEs: PER, LOC, ORG, MISC
  - using own tools

# ParlaMint



The ParlaMint project

The ParlaMint schema

The ParlaMint process

- » The ParlaMint data and analytics
- Usage examples (and inspirations)
- What's next?

# Corpus download

Via CLARIN.SI repository:

- [the complete corpora](#)
- [the corpora with added linguistic annotations](#)
- [the corpora translated into English with linguistic annotations](#)

<b>Name</b>	ParlaMint-LV.tgz
<b>Size</b>	47.97 MB
<b>Format</b>	Unknown
<b>Description</b>	Latvian corpus

Download file

Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 3.0



Please use the following text to cite this item or export to a predefined format:

BIBTEX

CMDI

Erjavec, Tomaž; et al., 2023, *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 3.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1488>.



This resource is also integrated in following services:



Share:



noSketch

KonText

CLARIN.SI Data & Tools

Authors

Erjavec, Tomaž ; et al.

► show everyone

Item identifier

<http://hdl.handle.net/11356/1488>



Project URL

<https://www.clarin.eu/content/parlamint>

Demo URL

<https://github.com/clarin-eric/ParlaMint/>

Date issued

2023-07-04

Type

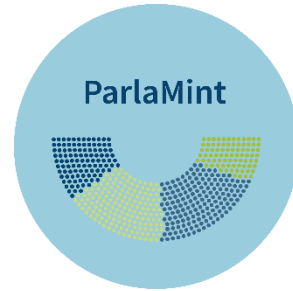
corpus, text

Size

7468674 utterances, 1147800265 words

Language(s)

Bosnian , Bulgarian , Catalan , Croatian , Czech , Danish , Dutch , English , Estonian , French , Galician , German , Hungarian , Icelandic , Italian , Latvian , Modern Greek (1453-) , Norwegian , Polish , Portuguese , Russian , Serbian , Slovenian , Spanish , Swedish , Turkish , Ukrainian



# Corpus access

Via concordancers:

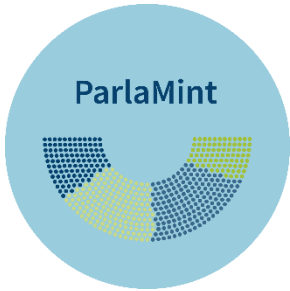
- Support for large corpora annotated with structural (e.g. speech, sentence, name) and positional (lemma, PoS, morphology, syntactic head) information
- Powerful corpus query language CQL
- Display of concordances, frequency lists, keyword lists (noSketch Engine only), collocations, saving and searching subcorpora, personalising the display, personal space (KonText only), REST interface
- **noSketch Engine:** <https://www.clarin.si/ske/>
- **Kontext:** <http://clarin.si/kontext>

# ParlaMint-LV 3.0 (Latvian parliament)

parlamint30\_lv

Latvian parliamentary corpus ParlaMint-LV, 2014-2022 v3.0 [see more](#)

- MANAGE SUBCORPORA
- TEXT TYPE ANALYSIS



## GENERAL INFO

Language: Latvian

CORPUS DESCRIPTION & BIBLIOGRAPHY

TAGSET

CORPUS WITHOUT WORD SKETCHES

CORPUS WITHOUT TERMS

## COUNTS ⓘ

Tokens	11,522,672
Words	8,962,269
Sentences	930,377
Paragraphs	162,782
Documents	162,782

## LEXICON SIZES ⓘ

word?	158,686
lc ⓘ	143,367
norm ⓘ	158,686
lemma	46,597
lemma_lc ⓘ	45,631
pos ⓘ	17
feats ⓘ	1,166
id ⓘ	251
dep ⓘ	37
dep_head_lemma ⓘ	35,203
dep_head_pos ⓘ	18
dep_head_feats ⓘ	1,062
dep_head_id ⓘ	250



# CONCORDANCE

ParlaMint-LV 3.0 (Latvian parliament)



simple **Rail Baltica** • 221

19.18 per million tokens • 0.0019%



KWIC ▾



☐ Details

Left context

KWIC

Right context

- |    |                          |  |  |  |
|----|--------------------------|--|--|--|
| 1  | <input type="checkbox"/> |  | MetsolaRoberta ... būvējot ātrgaitas infrastruktūras projektu <b>Rail Baltica</b> , kas šobrīd ir ļoti svarīgs arī militārajai m |  |
| 2  | <input type="checkbox"/> |  | Mūrniecelnāra •... - nav. Likums pieņemts. Likumprojekts “ <b>Rail Baltica</b> projekta likums”, trešais lasījums. Tautsa        |  |
| 3  | <input type="checkbox"/> |  | FeldmansKrišjān... ans. Nu tad izskatām arī likumprojektu “ <b>Rail Baltica</b> projekta likums” ( Nr. 1569/Lp13). Komis         |  |
| 4  | <input type="checkbox"/> |  | MedneLinda • 20... dnei. Kolēģi, šajā ģeopolitiskajā situācijā <b>Rail Baltica</b> projekta nozīmīgums ir ļoti augsts, tāpēc     |  |
| 5  | <input type="checkbox"/> |  | MedneLinda • 20... a, starptautiskā lidosta “ Rīga”, kur notiek <b>Rail Baltica</b> būvniecība... Taču likums ir svarīgs tieši p |  |
| 6  | <input type="checkbox"/> |  | MedneLinda • 20... is un būvniecības darbus, kas ir nozīmīgi <b>Rail Baltica</b> būvniecības fāzes uzsākšanai ārpus Rīga         |  |
| 7  | <input type="checkbox"/> |  | MedneLinda • 20... es uzsākšanai ārpus Rīgas. Lai sekmētu <b>Rail Baltica</b> projekta savlaicīgu pabeigšanu paredzē             |  |
| 8  | <input type="checkbox"/> |  | Mūrniecelnāra •... dzu zvanu! Balsosim par likumprojektu “ <b>Rail Baltica</b> projekta likums” trešajā lasījumā! Lūdzu l        |  |
| 9  | <input type="checkbox"/> |  | PūceJuris • 202... skaitā ļoti nozīmīgus likumus, piemēram, <b>Rail Baltica</b> projekta likumu mēs nupat pieņemām tre           |  |
| 10 | <input type="checkbox"/> |  | Mūrniecelnāra •... n - šā gada 11. oktobris. Likumprojekts “ <b>Rail Baltica</b> projekta likums”, otrais lasījums. Tautsair     |  |
| 11 | <input type="checkbox"/> |  | FeldmansKrišjān... katām otrajā lasījumā arī likumprojektu “ <b>Rail Baltica</b> projekta likums” ( Nr. 1569/Lp13). Likumj       |  |
| 12 | <input type="checkbox"/> |  | FeldmansKrišjān... skās lietošanas dzelzceļa infrastruktūras <b>Rail Baltica</b> - būvniecību un ieviešanu noteiktā termiņ       |  |



# Create subcorpus



Subcorpus name \*

COVID



Subcorpus from text types



Subcorpus from concordance

expand all collapse all

speech.subcorpus ^



COVID X



Reference

War

Parliamentary body v





# KEYWORDS

ParlaMint-LV 3.0 (Latvian parliament)



## ADVANCED

## ABOUT

Focus subcorpus ?

COVID



Reference corpus ?

ParlaMint-LV 3.0 (Latvian parliament)



Reference subcorpus ?

Reference



Focus on ?

rare

1

common

Minimum frequency ?

10

Maximum frequency ?

1000

Maximum items ?

1000



A = a ?



At least one alphanumeric ?



Only alphanumeric ?



Include nonwords ?



Exclude these words: ?



From list ?



Identify keywords



Identify terms



Identify n-grams

### Keywords settings

Attribute ?

lemma (lowercase)



### Terms settings

Matching regex ?

.\*

### N-grams settings

Attribute ?

word





# KEYWORDS

ParlaMint-LV 3.0 (Latvian parliament)



COVID ▾ ×

SINGLE-WORDS ✓



reference corpus: ParlaMint-LV 3.0 (Latvian parliament) subcorpus: Reference (items: 8,340)

Lemma (lowercase)			Lemma (lowercase)			Lemma (lowercase)			Lemma (lowercase)		
1	pandēmija	...	11	valstspilsēta	...	21	saulkrasti	...	31	bikše	...
2	kovids	...	12	reprezentācija	...	22	ločmele	...	32	ulbroka	...
3	dīkstāve	...	13	pārrāvums	...	23	pārslimošana	...	33	saslimstība	...
4	vakcinēties	...	14	ģirģena	...	24	antiviela	...	34	kovida	...
5	vakcinēt	...	15	respirators	...	25	flīģelis	...	35	vakcinēšanās	...
6	možvillo	...	16	sēlija	...	26	reindustrializācija	...	36	inficēt	...
7	vīruss	...	17	skaistumkopšana	...	27	sadarbspējīgs	...	37	pfizer	...
8	covid	...	18	baumane	...	28	vakcinēšana	...	38	inficēšanās	...
9	infekcija	...	19	pārslimot	...	29	zamurs	...	39	mārtuža	...
10	varakļāni	...	20	e-saeima	...	30	seplp	...	40	eitanāzija	...



# CONCORDANCE

ParlaMint-XX-en 3.0 (European parliaments, translation to English)



simple **Latvian border** • 111

0.08 per million tokens • 0.0000077%



KWIC



Details

Left context


KWIC

Right context


- | 1  | <input type="checkbox"/> |  | LV-en • 2022-10... | enced a pandemic, a hybrid attack at the <b>Latvian border</b> , a war in Ukraine, and we worked on a sp       |  |
|----|--------------------------|--|--------------------|--|--|
| 2  | <input type="checkbox"/> |  | LV-en • 2022-06... | the construction and maintenance of the <b>Latvian border</b> line infrastructure, the regulatory framew       |  |
| 3  | <input type="checkbox"/> |  | EE-en • 2022-04... | at they did not think that residents on the <b>Latvian border</b> could come to Estonia to buy petrol beca     |  |
| 4  | <input type="checkbox"/> |  | HU-en • 2021-12... | on the east, on the Polish, Lithuanian and <b>Latvian borders</b> . In the middle of this situation, the Europ |  |
| 5  | <input type="checkbox"/> |  | PL-en • 2021-10... | what is happening on the Lithuanian and <b>Latvian border</b> . Are we following this? Because some of         |  |
| 6  | <input type="checkbox"/> |  | PL-en • 2021-10... | military built here. The same is true on the <b>Latvian border</b> . There, too, you must know that there is a |  |
| 7  | <input type="checkbox"/> |  | EE-en • 2021-10... | isian-Lithuania border, on the Belarusian- <b>Latvian border</b> or on the Belarusian-Polish border also di    |  |
| 8  | <input type="checkbox"/> |  | PL-en • 2021-10... | lish border, not only on the Lithuanian or <b>Latvian border</b> , but also on the Mediterranean, and with     |  |
| 9  | <input type="checkbox"/> |  | EE-en • 2021-09... | y comes to Estonia primarily through the <b>Latvian border</b> . In Nord Pool's trading area, all electricity  |  |
| 10 | <input type="checkbox"/> |  | EE-en • 2021-09... | electricity from Russia is traded as at the <b>Latvian border</b> and mostly in the night hours. As I said, si |  |
| 11 | <input type="checkbox"/> |  | LV-en • 2021-09... | nd - so many Latvian citizens live outside <b>Latvian borders</b> . Live, study, work. Each has its own reasc  |  |
| 12 | <input type="checkbox"/> |  | EE-en • 2021-09... | at the police would monitor the Estonian- <b>Latvian border</b> . There was a police car in the Valga near     |  |



# PARALLEL CONCORDANCE





ParlaMint-XX 3.0 (European parliaments) 



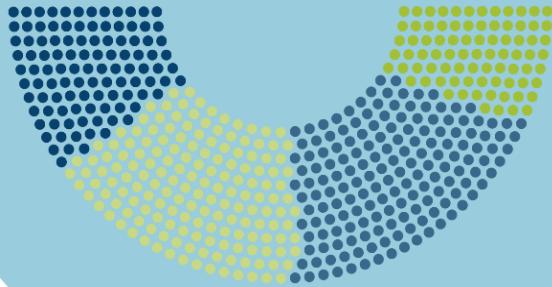
simple **karš** • 563  
0.43 per million tokens • 0.000043% 



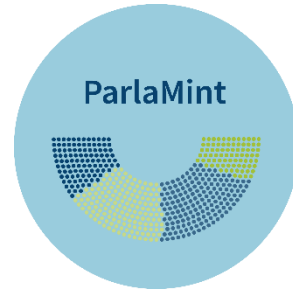
ParlaMint-XX en 3.0 (European parliaments, translation to English)

 LV • 2022-10-27	Mēs piedzīvojām pandēmiju, hibrīduzbrukumu pie Latvijas robežas, <b>karu</b> Ukrainā, un mēs strādājām īpašā režīmā gan klātienē, gan attālināti, un brīžam tas bija ārkārtīgi intensīvi.	<b>We experienced a pandemic, a hybrid attack at the Latvian border, a war in Ukraine, and we worked on a special regime both in person and remotely, and for a moment it was extremely intensive.</b>
 LV • 2022-10-27	pasaules enerģētikas aģentūra pateica, ka pirmoreiz pasaules vēsturē ir enerģētiskā krīze, ko izraisīja Krievijas <b>karš</b> pret Ukrainu.	<b>the World Energy Agency said that, for the first time in world history, there was an energy crisis caused by Russia's war against Ukraine.</b>
 LV • 2022-10-20	Tā kā de facto jūs esat, jūs varat pieņemt jebkādus lēmumus, varat pieteikt <b>karu</b> Lietuvai, ja gribat, bet de jure jūs neesat...	<b>As de facto you are, you can make any decisions, you can apply for war to Lithuania if you wish, but de jure you are not...</b>
 LV • 2022-09-29	Šis <b>karš</b> apliecina, ka būtiski ir augusi Krievijas vadības politiskā gatavība veikt pilna mēroga neslēptus militārus uzbrukumus kaimiņvalstīm, kuras Krievija uztver kā savu ietekmes sfēru.	<b>This war shows that the political readiness of the Russian leadership to carry out full-scale unhindered military attacks on neighbouring countries, which Russia sees as its sphere of influence, has increased significantly.</b>

# ParlaMint



- The ParlaMint project
- The ParlaMint schema
- The ParlaMint process
- The ParlaMint data and analytics
  - » Usage examples (and inspirations)
- What's next?



# Showcase 1: ‘Male’ vs. ‘female’ topics

## Top-ranking Topics by Term7-Female Freq.

Health	57
Labour, family and social affairs	22
Environment and spatial planning	4
Culture	3
Public administration	3
Other	3
Justice	2
Agriculture, forestry and food	2
Finance	1
Infrastructure	1
Interior	1
Multiple	1

## Top-ranking Topics by Term7-Male Freq.

Other	52
Infrastructure	11
Foreign affairs	10
Public administration	7
Economic development and technology	5
Defence	4
Agriculture, forestry and food	3
Interior	3
Justice	3
Finance	1
Multiple	1

Darja Fišer, Kristina  
Pahor de Maiti: [Voices  
of the Parliament](#)

## Showcase 2: 'COVID' keywords

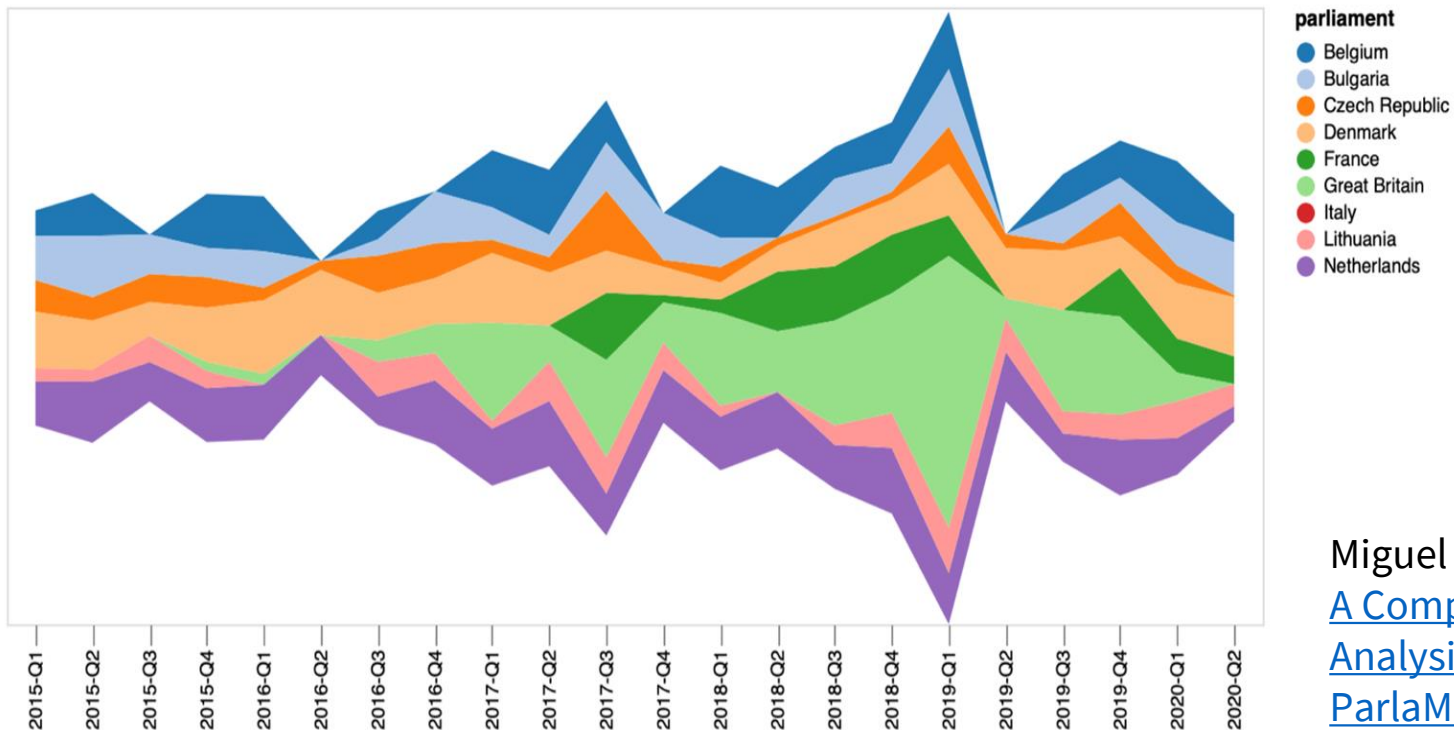
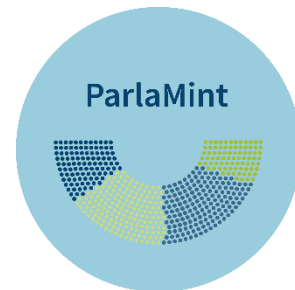
Italy	Poland	Slovenia	UK
pandemic	pandemic	epidemic	covid
covid	<i>bell</i>	ventilator	coronavirus
covid-19	coronavirus	coronavirus	<b>furlough</b>
coronavirus	covid-19	covid-19	lockdown
lockdown	mask	virus	pandemic
mask	epidemic	quarantine	distancing
<b>recovery</b>	quarantine	corona	<b>ppe</b>
European Stability Mechanism (ESM)	Kukiz15	mask	<i>inaudible</i>
distancing	the Left	pandemic	
fund	<b>shield</b>	<b>anti-corona (adj)</b>	
<b>serologic</b>	<b>anti-crisis</b>	/paramilitary group/	
virus	covid	/name/	

[Parliamentary Debates in the COVID Times](#)

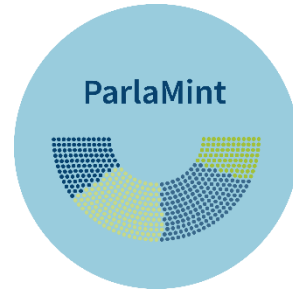
Helsinki DH Hackathon 21



## Showcase 3: Migration issues



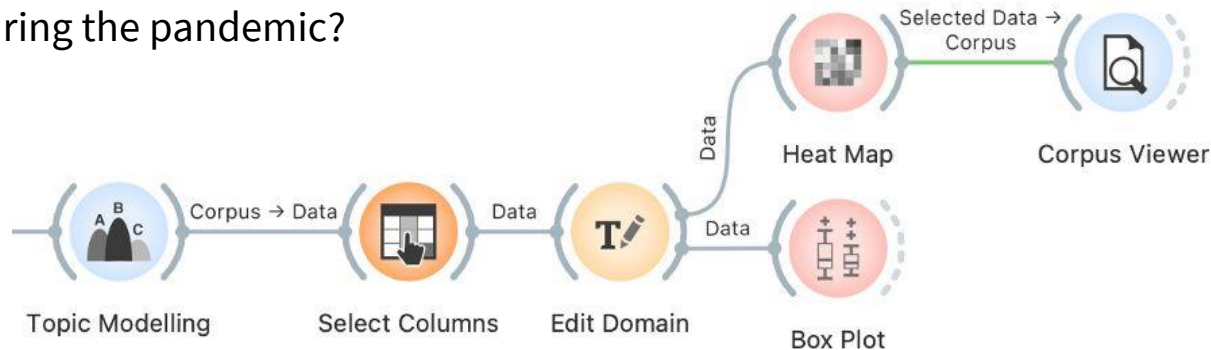
Miguel Pieters Thesis:  
[A Comparative  
Analysis on the  
ParlaMint Project](#)

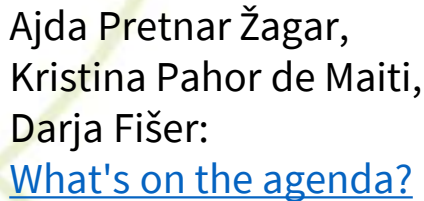


## Showcase 4: What's on the agenda?

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?



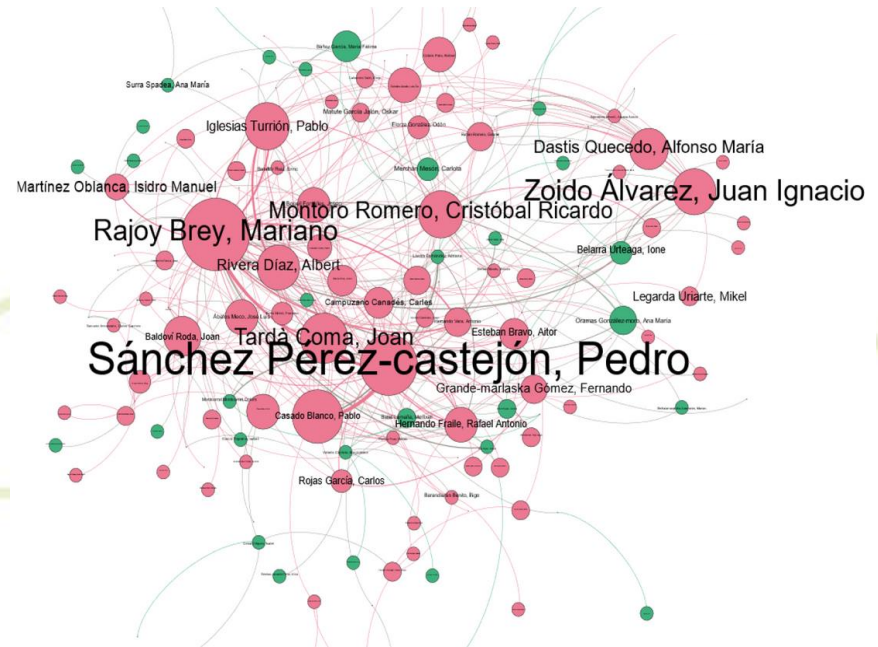


# Showcase 5: Networks of Power

Angermeier et al. (2022)  
[ParlaMint. Networks  
of Power](#)

Analyzing the networks that emerge from parliamentarians mentioning one another:

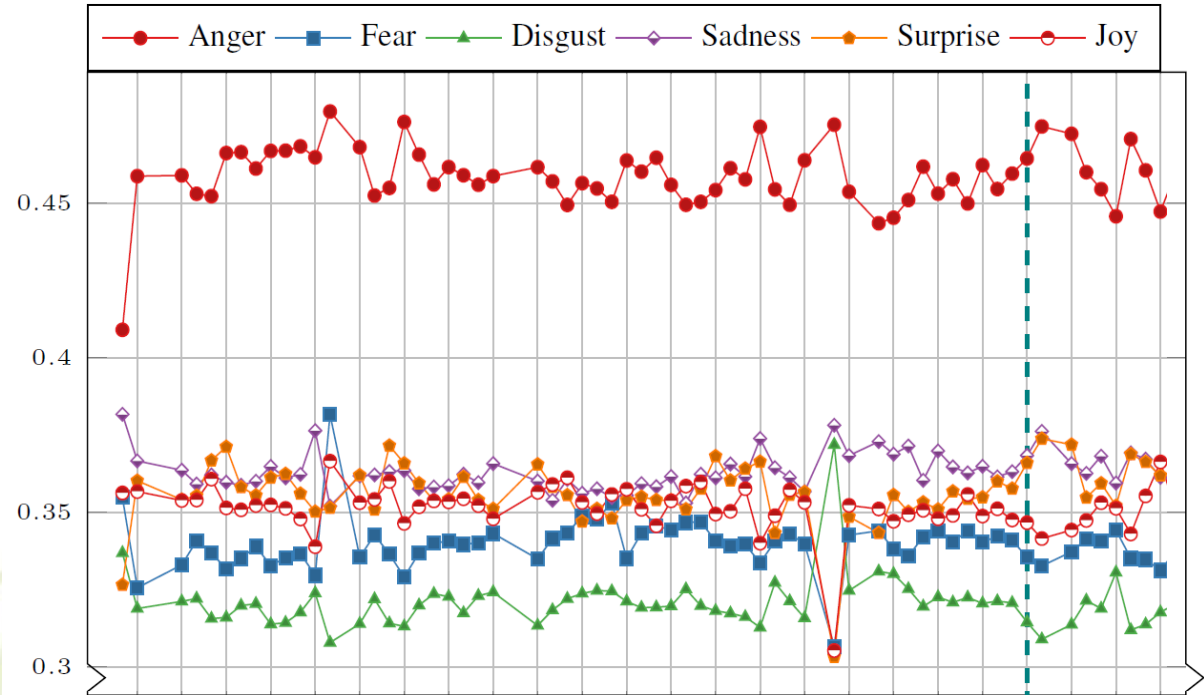
- **Argumentative Power:** How can speeches and mentions give insights into the power of MPs?
- **Structural Power:** How do the speech practices of female and male MPs relate to topic and power distribution?



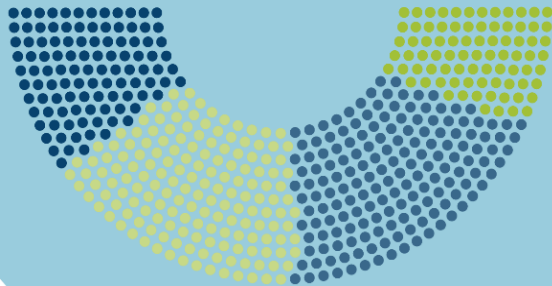
## Showcase 6: Emotions Running High

Kurtoğlu and Çöltekin (2022).  
[Emotions Running High?](#)

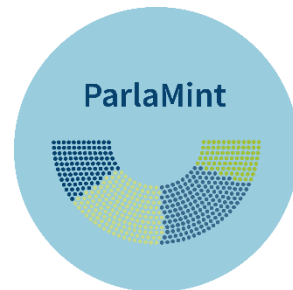
- Investigating polarization of politics by assigning emotion scores to speeches
- anger being the dominant emotion
- the ruling party showing more stable emotions compared to the opposition



# ParlaMint



- The ParlaMint project
- The ParlaMint schema
- The ParlaMint process
- The ParlaMint data and analytics
- Usage examples (and inspirations)
- » What's next?



# What's next in ParlaMint?

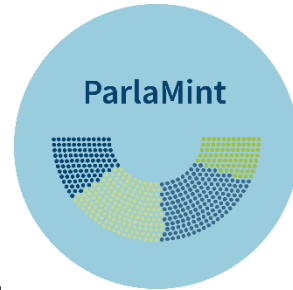
ParlaMint 3.1 to be released at the end of September 2023:

- main addition: semantic tagging
- and additional / better metadata: ministers and political orientations
- factorised common taxonomies with translations
- some corpora will extend the dates of speeches to mid-2023

Stay tuned for more events:

- a shared task at [CLEF 2024](#) on ideology and power identification
- hopefully another edition of ParlaCLARIN workshop at [LREC-COLING 2024](#)



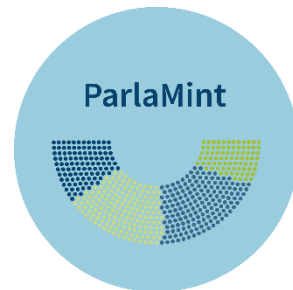


# What's beyond?

- Extending the coverage:
  - adding new national parliaments within and beyond Europe (e.g. Canada came to use the Parla-CLARIN schema for the encoding of their corpora!)
  - adding regional parliaments
  - incorporating EuroParl or other sources of European Parliament interventions
- Extending the data within individual corpora, e.g. adding audio/video recordings
- Enriching the existing data with more metadata and annotation of semantic content
- Linking the existing data to Wikipedia, DBpedia and other Linked Open Data
- Using the data for downstream tasks like text summarization, NE recognition etc.
- Applying data in real use cases based on focused research questions
- ...



# ParlaMint is a team effort!



## Our joint paper to cite:

Erjavec T., Ogrodniczuk M., Osenova P., Ljubešić N., Simov K., Pančur A., Rudolf M., Kopp M., Barkarson S., Steingrímsson S., Çöltekin Ç., de Does J., Depuydt K., Agnoloni T., Venturi G., Calzada Pérez M., de Macedo L. D., Navarretta C., Luxardo G., Coole M., Rayson P., Morkevičius V., Krilavičius T., Dargis T., Ring O., van Heusden R., Marx M., Fišer D. (2023). *The ParlaMint corpora of parliamentary proceedings*. Language Resources and Evaluation 57:415–448. <https://doi.org/10.1007/s10579-021-09574-0>

## Ready to play?

Use ParlaMint data in your research or create your own showcase!

The ParlaMint logo is a circular emblem composed of a grid of small dots. The dots are arranged in a circular pattern, with the top half being blue and the bottom half being yellow. The text "ParlaMint" is written in a dark blue, sans-serif font across the center of the logo.

ParlaMint

**Thank you – and see you at**

<https://www.clarin.eu/parlamint>