CLARIN Café: ParlaMint Unleashed

CLARIN ERIC

28 June 2021

CLARIN

Organisers

This edition of the CLARIN Café is organized by

The ParlaMint team

CLARIN host is
Darja Fišer, Head of the CLARIN User Involvement Committee

The event is recorded for further dissemination purposes.

Questions and comments? Put them in the chat box.

Plan

14:00 - 14:10 **CLARIN 101, CLARIN Resource Families** (Darja Fišer)
14:10 - 14:25 **ParlaMint: what was it all about?** (Maciej Ogrodniczuk & Petya Osenova)
14:25 - 14:45 **Creating comparable multilingual corpora of parliamentary debates** (Tomaž Erjavec)
14:45 - 15:00 **Parliamentary debates in COVID times** (Marta Kołczyńska)
15:00 - 15:15 **A comparative analysis on the ParlaMint project** (Miguel Pieters)
15:15 - 15:30 **ParlaMint and Parlameter: How standardized data formats empower end users** (Filip Dobranić)
15:30 - 15:50 **Discussion panel: Lessons learnt from Czech, Icelandic, Italian and UK groups** (Mini-grant recipients)
15:50 - 16:00 **QA and Closing**

CLARIN 101

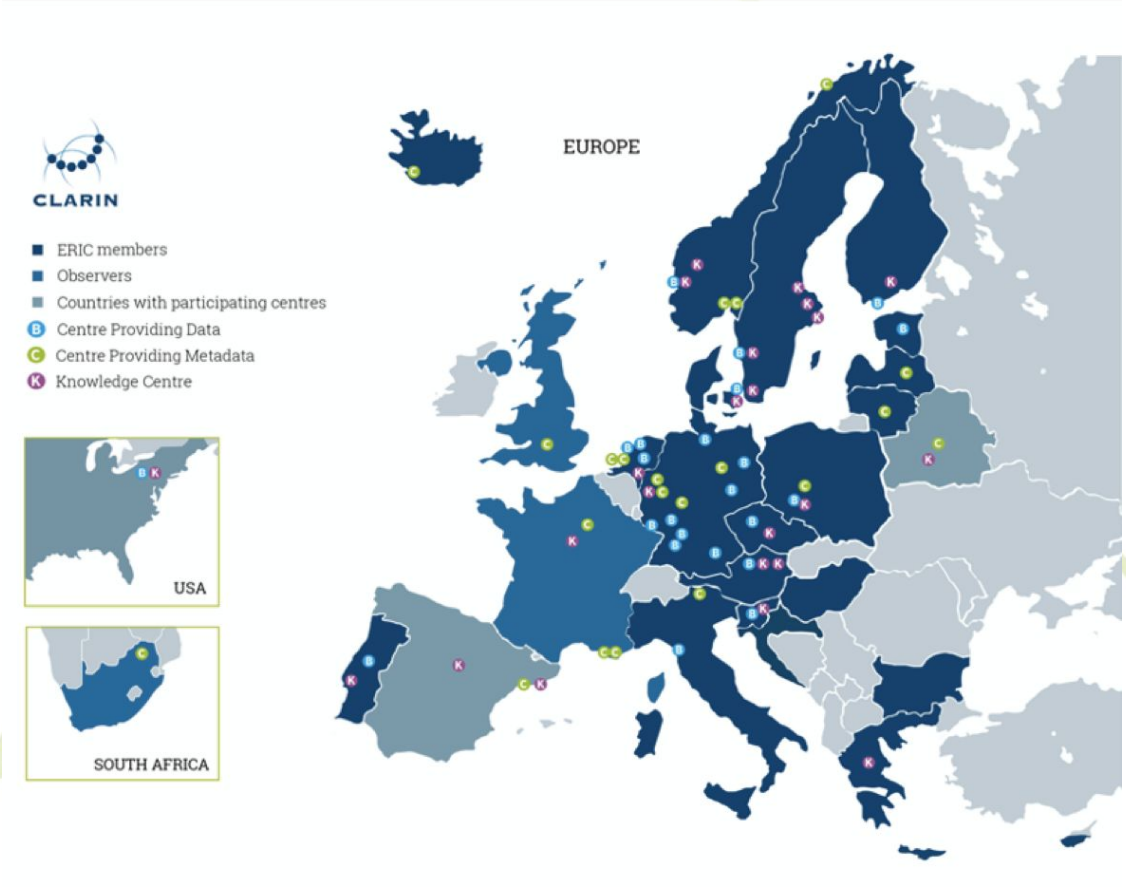
https://www.clarin.eu/content/clarin-in-a-nutshell

CLARIN …

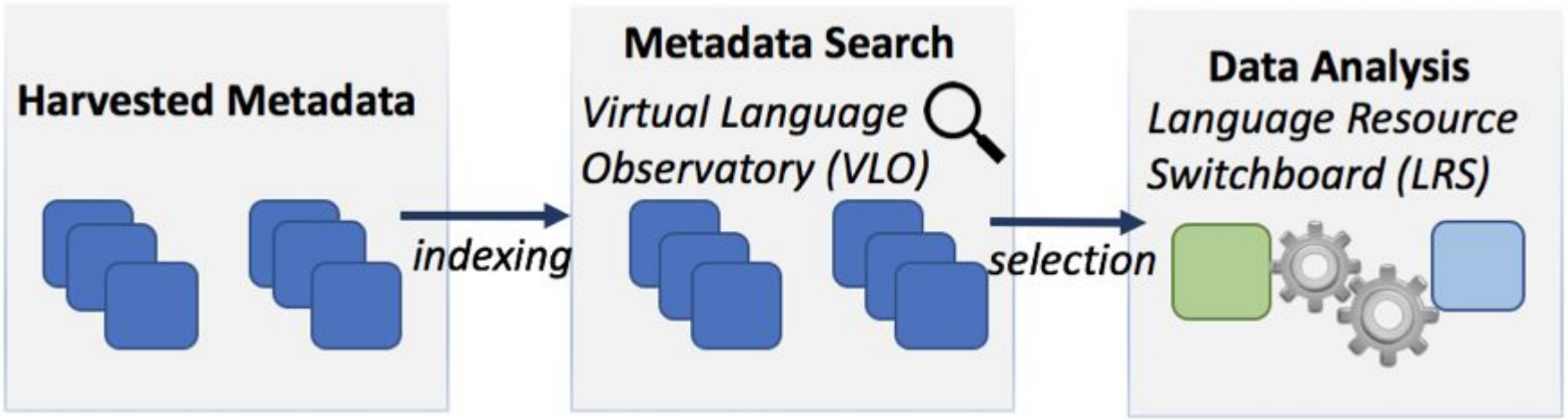
- is the *Common Language Resources and Technology Infrastructure*
- has the **ESFRI** ERIC status since 2012, Landmark since 2016
- provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
  - to **digital language data** (in written, spoken or multimodal form)
  - and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
  - through a **single sign-on** environment
- serves as an ecosystem for **knowledge sharing**
- is an integral part of **the European Open Science Cloud**
  - See **clarin.eu/eosc**

CLARIN today

- **68 centres**
- **21 members**: (AT, BG, CY, CZ, DE, DK, EE, FI, GR, HR, HU,IS, IT, LT, LV, NL, NO, PL, PT, SE, SI)
- 3 observers: FR, UK, ZA



The Technical Infrastructure



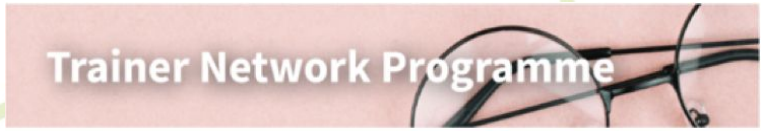
Findable Accessible Interoperable Reusable

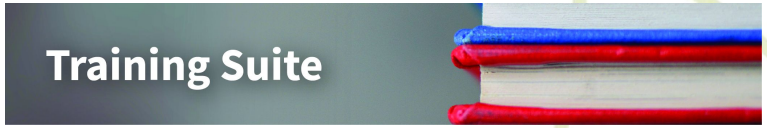
clarin.eu/fair

vlo.clarin.eu

switchboard.clarin.eu

The Knowledge Infrastructure





https://www.clarin.eu/content/clarin-for-researchers
https://www.clarin.eu/content/knowledge-sharing

CLARIN Resource Families

- What is CRF?
  - user-friendly overviews per data type of the available language resources in CLARIN
  - https://www.clarin.eu/resource-families
  - 12 corpora families, 5 families of lexical resources, and 4 tool families
  - the most important metadata and brief descriptions as well as links to download pages and concordancers
- Who is CRF for?
  - aimed at the needs of researchers from digital humanities, social sciences and human language technologies
  - facilitates comparative research
- How can you get involved?
  - report missing resources
  - deposit your resources
  - apply for funding for small projects that can help extending the scope of the initiative: https://www.clarin.eu/content/clarin-resource-families-project-funding

The café

ParlaMint:

What was it all about?

Maciej Ogrodniczuk and Petya Osenova

CLARIN

Background

Parliamentary corpora are one of the key resource families in CLARIN (https://www.clarin.eu/content/parliamentary-corpora)

- Several CLARIN-supported activities:
  - CLARIN Traveling Campus 'Talk of Europe' (2014 and 2015)
  - CLARIN-PLUS workshop *Working with parliamentary records* (2017)
  - ParlaCLARIN workshop at LREC 2018
  - ParlaFormat workshop (2019)
  - ParlaCLARIN II workshop at LREC 2020
- Many corpora exist, but are encoded in many different ways, limiting interchange and comparability

What is ParlaMint?

A mini-project supported by CLARIN-ERIC during the pandemic.

Budget: 135,000 €
Duration: July 1, 2020 – May 30, 2021

Direct motivation: Parliamentary data directly corresponds to the most recent events with a global impact on *human health, social life* and *economics* such as the current COVID-19 pandemic.

Goal: Provide resources and tools for focused observations on trends, opinions, decisions on *lockdowns* and *restrictive measures* as well as on *the consequences* with respect to health, medical care systems, employment, etc. during pandemic times.

How was ParlaMint implemented?

Phase 1 (*July 2020 – September 2020*):

The pilot corpus of 4 parliaments – *Bulgarian, Croatian, Polish* and *Slovene* – was created and linguistically annotated with two parts marked:

- COVID-19 subcorpus (November 2019 – July 2020)
- reference subcorpus (2015 – October 2019).

Phase 2 (*December 2020 – May 2021*):

Corpora for 13 more parliaments were added according to the methodology established in *Phase 1*: Belgian, Czech, Danish, Dutch, English, French, Hungarian, Icelandic, Italian, Latvian, Lithuanian, Turkish. Spanish joined with their own funds.

Useful information

Website: https://www.clarin.eu/content/parlamint

Participation in various events:

- CLARIN Bazaar Poster at the Virtual CLARIN Annual Conference 2020: ParlaMint: Towards Comparable Parliamentary Corpora (October 7, 2020)
- CLARIN Café – Join Our Parliamentary-flavoured Coffee: ParlaMint (November 3, 2020)
- Helsinki Digital Humanities Hackathon 2021 (May 28, 2021)
- CLARIN Café: ParlaMint Unleashed (June 28, 2021)
- Accepted joint paper at CLARIN Annual Conference 2021

Highly multilingual workflow

1. Getting the parliamentary data and metadata
2. Converting them into the ParlaMint schema
3. Validation (formal and qualitative)
4. Linguistic annotation: Universal Dependencies morphosyntax and syntax + Named Entities
5. Making corpora available
   - through the CLARIN.SI repository
   - through *concordancers* (noSketch/KonText)
   - and Parlameter
6. Building use cases in Political Sciences and Digital Humanities based on the corpus data

Cross-parliament challenges

- Different countries have different political and thus, parliamentary systems.
- This fact inevitably reflects the incorporation of the data into the common standard.
- For example, there are
  - *unicameral* (Bulgaria, Denmark, Hungary, Iceland, Latvia, Lithuania, Turkey)
  - and *bicameral* parliaments (Belgium, France, Italy, Spain, the Netherlands, UK),

  each with its own specifics.

Getting data

- Scraping it from the parliamentary websites (Belgium, Bulgaria, Czech Republic, Hungary, Iceland, France, Latvia, Spain, Turkey)
- Obtaining via Parlameter API, which returned results in JSON (Croatia)
- Retrieving from an already maintained parliamentary corpus (Poland and Slovenia)
- Downloading from a server (Denmark, the Netherlands)
- Obtaining through parliamentary API (UK) or through a service center at the parliament (Italy)

Data conversion

Various strategies such as:

- Incremental and semi-automatic transformation from HTML to basic TEI XML and then to the ParlaMint format through XML constraints (Bulgarian) or
- Through XSLT stylesheets and Python, Perl and Bash scripts (Belgian, Dutch, French, Spanish)
- Automatic conversion through Perl scripts with heuristics only for difficult parts such as the transcriber comments (Croatian, Czech, Danish)

Data conversion

- Automatic conversion through Python scripts with possible corrections of data during the process (Hungarian, Icelandic, Latvian, Polish, Turkish)
- Transformation with XSLT, and some manual interventions upstream (Slovene)
- Adding necessary extensions to XSLT (English)
- Automatic conversion with JAVA code (Italian)

Linguistic processing

- Included the *UD-based morphosyntactic and syntactic annotation* and *NEs*: PER, *LOC, ORG,* MISC.
- This step was also approached differently by the groups depending on factors like:
  - the availability of the tools for the language
  - their quality and performance
  - their suitability to the parliamentary domain.

Summary

ParlaMint project establishes an innovative strategy for handling parliamentary data and processing it in times of any emergency period (*COVID-19 is just a showcase*).

The novelties relate to:

- the proper and unified handling of cross-lingual and across-parliament comparable data, and
- to the quick access of all interested parties to these data.

Thus, different reference corpora could be produced with parliamentary records from previous times with global crisis states, e.g. the great economic recession, periods of floods in Europe, the Ebola outbreak etc.

What's next?

Many open research horizons...

- Extending the data as parliament coverage:
  - adding new national parliaments within and beyond Europe
  - adding regional parliaments
  - incorporating EuroParl or other sources of EP interventions.
- Extending the data:
  - adding speech recordings of the debates
  - translation of all corpora into English for enabling better comparative research
- Enriching the existing data with *more metadata* and annotation of *semantic content*.

What's next?

Many open research horizons...

- Linking the existing data to Wikipedia, DBpedia and other LOD.
- Using the data for downstream tasks like text summarization, NER and NEL, semantic similarity etc.
- Applying data in real use cases based on focused research questions.
- Observations over democratic processes through e.g.:
  - speaker and parties statistics
  - topic modeling
  - time and context-bound social tendencies

Creating comparable multilingual corpora of parliamentary debates

Tomaž Erjavec

CLARIN

The importance of encoding

- The idea of ParlaMint was that the corpora are encoded as uniformly as possible
- This allows the corpora to be interoperable, so that e.g. they can be converted to other formats by the same scripts
- However:
  - the debates have very different source encoding
  - they are differently structured, contain different information,and reflect different parliamentary traditions
  - each corpus produced by a separate partner
- The definition of a rich but constrained format and the possibility to validate the corpora is crucial!

Parla-CLARIN recommendations for encoding of parliamentary corpora

”CLARIN Parla Format” workshop in 2019:

- Introduced a ”standard” format for parliamentary corpora called ”Parla-CLARIN” (Erjavec and Pančur, 2019)
- A simple customisation of the TEI Guidelines: https://github.com/clarin-eric/parla-clarin
- However, we did write extensive annotation guidelines: https://clarin-eric.github.io/parla-clarin
- First Parla-CLARIN encoded corpus:
  Pančur, Andrej; Erjavec, Tomaž; Ojsteršek, Mihael;Šorn, Mojca and Blaj Hribar, Neja, 2020,Slovenian parliamentary corpus (1990-2018) siParl2.0, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1300.

ParlaMint encoding validation

- Start of ParlaMint project: encode BG,HR,PL,SI corpora in Parla-CLARIN
- Successive modifications of the encoding (unification, but also corpus-specific information), almost till the end of the project
- XML schemas made just for the ParlaMint corpus (i.e. not a TEI schema)
- XSLT scripts to validate the corpus further
- Functional validation through conversion scripts, and their use:
  - CoNLL-U files: checked with the UD validation script
  - vertical files: conversion and indexing logs, analyzing at the corpora through noSketch Engne (hackathon)
- Of course, bugs still remain...

Corpus title statement

```
<?xml version="1.0" encoding="UTF-8"?>
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0" xml:id="ParlaMint-CZ" xml:lang="cs">
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title type="main" xml:lang="cs">Český parlamentní korpus ParlaMint-CZ [ParlaMint SAMPLE]</title>
            <title type="main" xml:lang="en">Czech parliamentary corpus ParlaMint-CZ [ParlaMint SAMPLE]</title>
            <title type="sub" xml:lang="cs">Parlament České republiky, Poslanecká sněmovna</title>
            <title type="sub" xml:lang="en">Parliament of the Czech Republic, Chamber of Deputies</title>
            <meeting ana="#parla.term #parla.lower #parliament.PSP7" n="ps2013">ps2013</meeting>
            <meeting ana="#parla.term #parla.lower #parliament.PSP8" n="ps2017">ps2017</meeting>
            <respStmt>
               <persName ref="https://orcid.org/0000-0001-7953-8783">Matyáš Kopp</persName>
               <resp xml:lang="en">Data retrieval</resp>
               <resp xml:lang="en">TEI XML corpus encoding</resp>
            </respStmt>
            <funder>
               <orgName xml:lang="en">CLARIN research infrastructure</orgName>
               <orgName xml:lang="cs">Výzkumná infrastruktura CLARIN</orgName>
            </funder>
            <funder>
               <orgName xml:lang="cs">LINDAT/CLARIAH-CZ: Digitální výzkumná infrastruktura pro jazykové technol
               <orgName xml:lang="en">LINDAT/CLARIAH-CZ: Digital Research Infrastructure for Language Technolog
            </funder>
         </titleStmt>
```

Speakers

```
<listPerson>
   <head>List of speakers</head>
   <person xml:id="SayeedaWarsi">
      <persName>
         <forename>Sayeeda</forename>
         <surname>Warsi</surname>
      </persName>
      <sex value="F">Female</sex>
      <affiliation from="2007-10-11" ref="#parla.lower" role="MP"/>
      <affiliation from="2007-10-11" ref="#party.CON" role="member"/>
      <affiliation from="2010-05-12" ref="#PoGB" role="minister" to="2012-09-06"/>
      <affiliation from="2012-09-06" ref="#PoGB" role="minister" to="2014-08-05"/>
      <idno subtype="contact" type="URI">https://members.parliament.uk/member/3839/contact</idno>
      <figure>
         <graphic url="https://api.parliament.uk/photo/Paa3j0vS.jpg?crop=CU_1:1"/>
      </figure>
   </person>
```

```
<category xml:id="parla.lower">
   <catDesc>
      <term>Lower house</term>
   </catDesc>
</category>
```

```
<org role="politicalParty" xml:id="party.CON">
   <orgName full="yes">Conservative</orgName>
   <orgName full="init">CON</orgName>
</org>
```

Transcription

```
<div type="debateSection">
   <head>Sul 35° anniversario del rapimento di Aldo Moro e dell'uccisione dei componenti d
   <note type="role">PRESIDENTE</note>
   <u ana="#chair"
      who="#ColomboEmilio"
      xml:id="ParlaMint-IT_2013-03-16-LEG17-Sed-2.u2">
      <incident type="action">
         <desc>Si leva in piedi e con lui tutta l'Assemblea.</desc>
      </incident>
      <seg xml:id="ParlaMint-IT_2013-03-16-LEG17-Sed-2.seg2">Prima di iniziare la seduta c
      <seg xml:id="ParlaMint-IT_2013-03-16-LEG17-Sed-2.seg3">Il loro sacrificio, la barbar
      <seg xml:id="ParlaMint-IT_2013-03-16-LEG17-Sed-2.seg4">Come una volta fu ricordato i
      <kinesic type="applause">
         <desc>Segni di commozione del Presidente. Applausi.</desc>
      </kinesic>
      <seg xml:id="ParlaMint-IT_2013-03-16-LEG17-Sed-2.seg5">..la gioia più alta che ciasc
      <seg xml:id="ParlaMint-IT_2013-03-16-LEG17-Sed-2.seg6">Vi invito ad un minuto di rac
```

Linguistic annotation

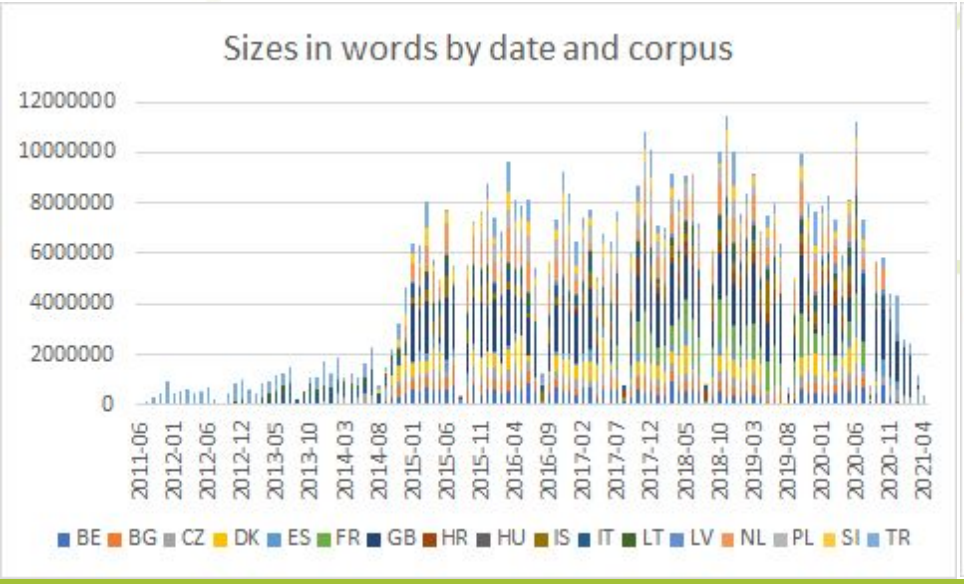
```
<u ana="#chair" who="#Sadık.Yakut" xml:id="tbmm-2014-07-19s11p047">
   <seg xml:id="tbmm-2014-07-19s11p047.seg0">
      <s xml:id="tbmm-2014-07-19s11p047-000010">
         <w msd="UPosTag=NOUN|Case=Nom|Number=Sing"
            lemma="kabul"
            xml:id="tbmm-2014-07-19s11p047-000010.t1">Kabul</w>
         <w msd="UPosTag=VERB|Case=Nom|Number=Plur|Tense=Pres|VerbForm=Part"
            lemma="et"
            xml:id="tbmm-2014-07-19s11p047-000010.t2"
            join="right">edenler</w>
         <pc msd="UPosTag=PUNCT" xml:id="tbmm-2014-07-19s11p047-000010.t3">...</pc>
         <linkGrp targFunc="head argument" type="UD-SYN">
            <link ana="ud-syn:obj"
                  target="#tbmm-2014-07-19s11p047-000010.t2 #tbmm-2014-07-19s11p047-000010.t1"/>
            <link ana="ud-syn:root"
                  target="#tbmm-2014-07-19s11p047-000010 #tbmm-2014-07-19s11p047-000010.t2"/>
            <link ana="ud-syn:punct"
                  target="#tbmm-2014-07-19s11p047-000010.t2 #tbmm-2014-07-19s11p047-000010.t3"/>
         </linkGrp>
      </s>
```

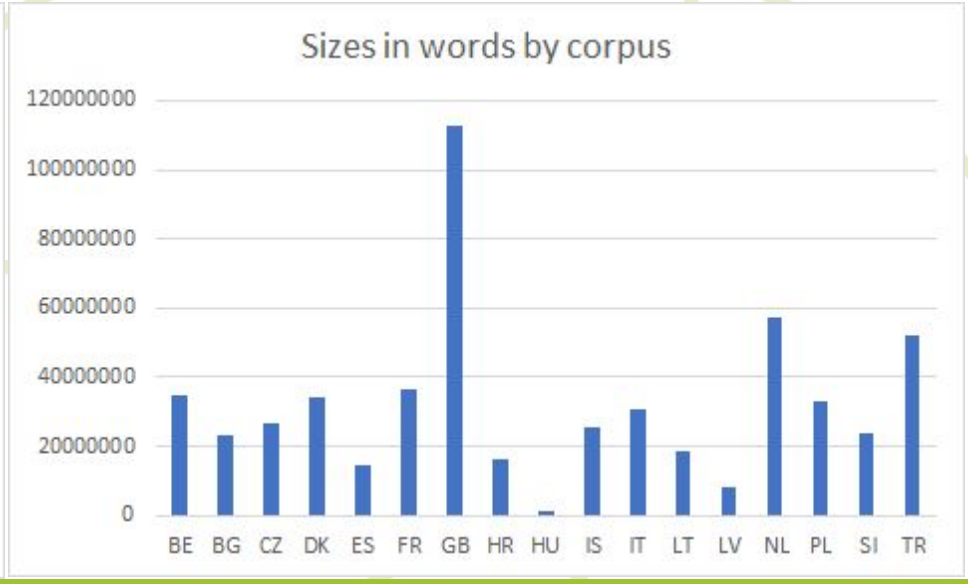
Corpus compilation

- Each partner produced their corpus
- Initial corpora each had their own mistakes and problems
- Communication and problem solving:
  - GitHub (70 issues, 500 commits)
  - email (200 emails)
- “Central services”
- A “polish” script for releases
- XML files converted with XSLT to other immediately useful formats (e.g. vertical files for concordancers)

ParlaMint in numbers

- 17 corpora:
  BE, BG, CZ, DK, ES, FR, GB, HR, HU, IS, IT, LT, LV, NL, PL,SI
- 16 languages:
  fr+nl, bg, cs, dk, es, fr, en, hr, hu, is, it, lt, lv, nl, pl, sl
- 22 thousand files, 5 mil. speeches, 500 mil. words
- 3,600 speakers, 1,680 “organisations”
- from 2011-06 / 2017-07 to 2020-06 / 2021-04

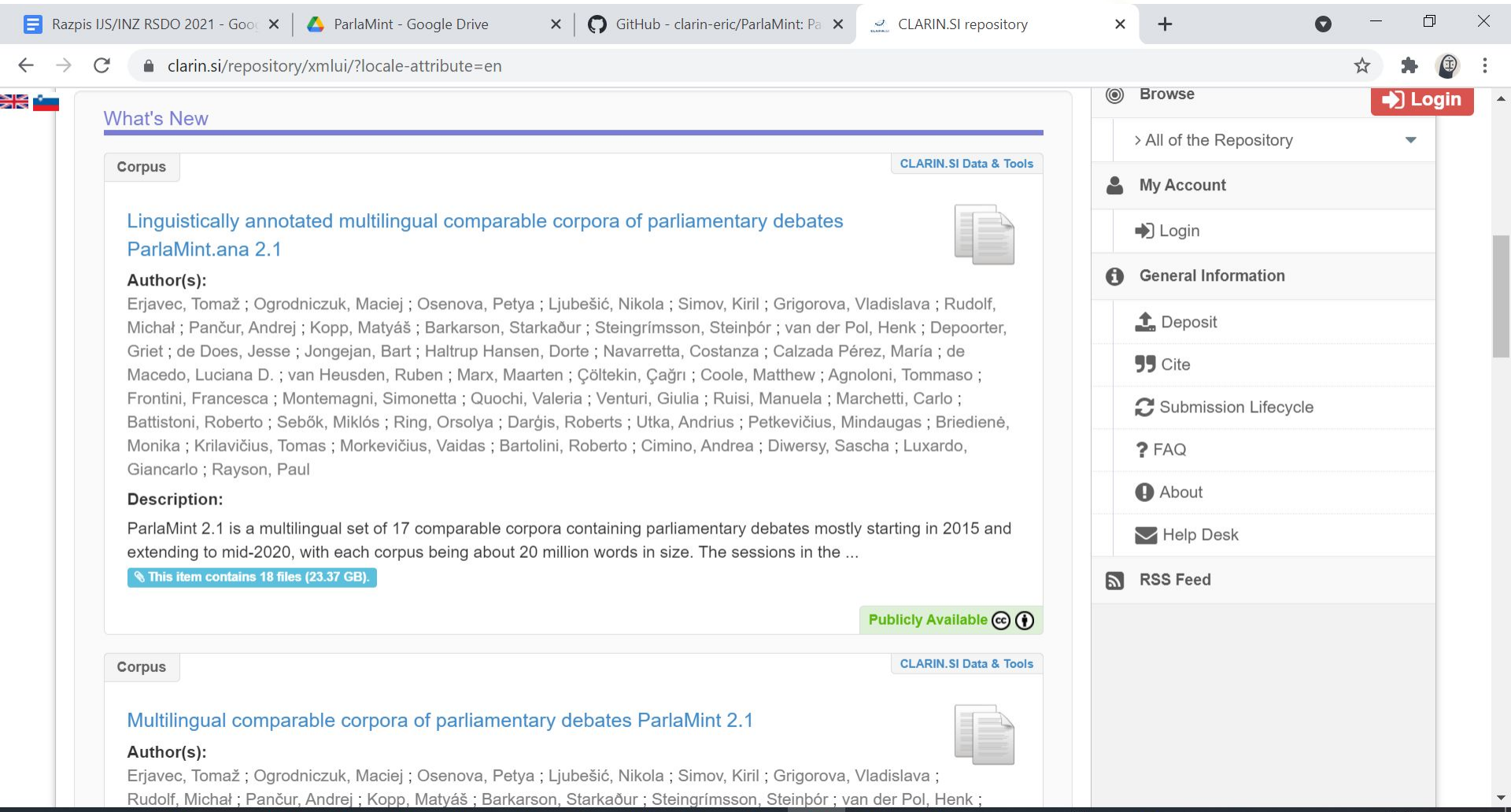




ParlaMint 2.1

- All scripts, schemas and sample files available on https://github.com/clarin-eric/ParlaMint
- Complete corpora
  - http://hdl.handle.net/11356/1432
  - http://hdl.handle.net/11356/1431
- Together 2 x 17 .tgz bitstreams with 2 + 22,000 x 5 files:
  - ParlaMint / Parla-CLARIN XML +
  - TSV metadata,
  - plain text of speeches (with speech IDs),
  - CoNLL-U,
  - vertical with registry

Team effort


Razpis IJS/INZ RSDO 2021 - Goo
ParlaMint - Google Drive
GitHub - clarin-eric/ParlaMint: Pa
CLARIN.SI repository
clarin.si/repository/xmlui/?locale-attribute=en
What's New
Corpus
CLARIN.SI Data & Tools
Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1
Author(s):
Erjavec, Tomaž ; Ogrodniczuk, Maciej ; Osenova, Petya ; Ljubešić, Nikola ; Simov, Kiril ; Grigorova, Vladislava ; Rudolf, Michał ; Pančur, Andrej ; Kopp, Matyáš ; Barkarson, Starkaður ; Steingrímsson, Steinþór ; van der Pol, Henk ; Depoorter, Griet ; de Does, Jesse ; Jongejan, Bart ; Haltrup Hansen, Dorte ; Navarretta, Costanza ; Calzada Pérez, María ; de Macedo, Luciana D. ; van Heusden, Ruben ; Marx, Maarten ; Çöltekin, Çağrı ; Coole, Matthew ; Agnoloni, Tommaso ; Frontini, Francesca ; Montemagni, Simonetta ; Quochi, Valeria ; Venturi, Giulia ; Ruisi, Manuela ; Marchetti, Carlo ; Battistoni, Roberto ; Sebők, Miklós ; Ring, Orsolya ; Darģis, Roberts ; Utka, Andrius ; Petkevičius, Mindaugas ; Briedienė, Monika ; Krilavičius, Tomas ; Morkevičius, Vaidas ; Bartolini, Roberto ; Cimino, Andrea ; Diwersy, Sascha ; Luxardo, Giancarlo ; Rayson, Paul
Description:
ParlaMint 2.1 is a multilingual set of 17 comparable corpora containing parliamentary debates mostly starting in 2015 and extending to mid-2020, with each corpus being about 20 million words in size. The sessions in the ...
This item contains 18 files (23.37 GB).
Publicly Available
Corpus
CLARIN.SI Data & Tools
Multilingual comparable corpora of parliamentary debates ParlaMint 2.1
Author(s):
Erjavec, Tomaž ; Ogrodniczuk, Maciej ; Osenova, Petya ; Ljubešić, Nikola ; Simov, Kiril ; Grigorova, Vladislava ; Rudolf, Michał ; Pančur, Andrej ; Kopp, Matyáš ; Barkarson, Starkaður ; Steingrímsson, Steinþór ; van der Pol, Henk ;
Browse
Login
All of the Repository
My Account
Login
General Information
Deposit
Cite
Submission Lifecycle
FAQ
About
Help Desk
RSS Feed
SL
17°C
01:11
27. 06. 2021

Parliamentary debates in COVID times

Experience from the 2021 DHH Hackathon

Marta Kołczyńska

CLARIN

ParlaMint team at #DHH21

Helsinki Digital Humanities Hackathon #DHH21

19-28.05.2021

ParlaMint team:

Isabella Calabretta, Courtney Dalton, Richard Griscom, Marta Kołczyńska, Kristina Pahor de Maiti, Ruben Ros

Ajda Pretnar, Matej Klemen, Darja Fišer

Blog post:
https://dhhackathon.wordpress.com/2021/05/28/parliamentary-debates-in-the-covid-times/

Questions

How was the COVID-19 pandemic reflected in parliamentary speeches?

- How do speeches on COVID differ from regular debates?
- Which topics arise in COVID debates? Which topics are shared between the countries and which are country-specific?
- Do the debates highlight any major shifts in topics or priorities over time?
- What is the frequency of COVID-related debates over time, and is there any connection between debates and COVID cases reported?

Data

Multilingual comparable corpora of parliamentary debates ParlaMint 2.0
https://www.clarin.si/repository/xmlui/handle/11356/1388

Dataset	N words	N speakers	Time span	House
Italy	26,571,966	716	15.03.2013-18.11.2020	Upper
Poland	26,882,964	1,121	16.11.2015-14.08.2020	Lower & upper
Slovenia	19,933,836	353	01.08.2014-16.07.2020	Lower
UK	100,967,492	1,895	01.05.2015-01.03.2021	Lower & upper

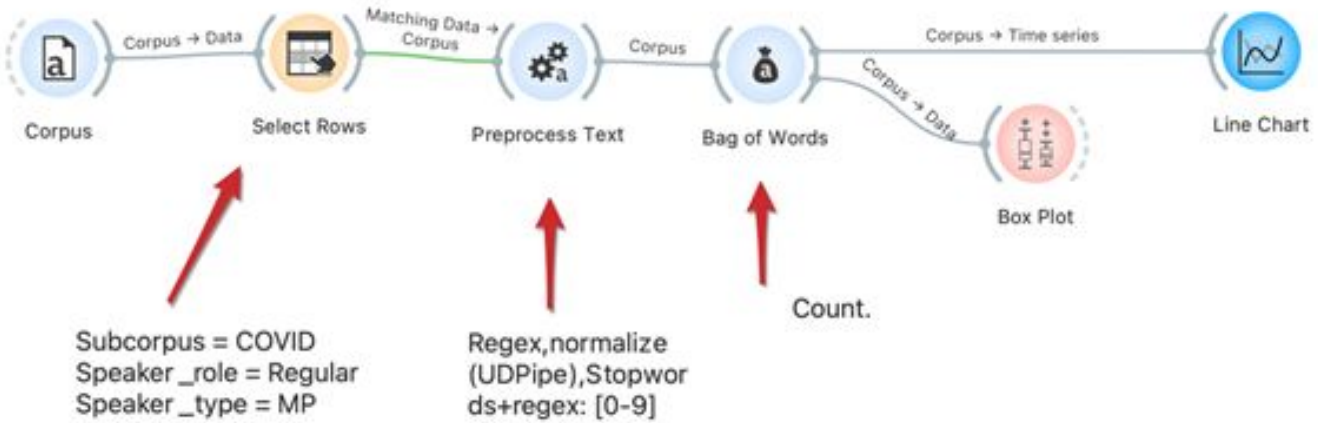
COVID subcorpus: starting 2019-11-01
Reference subcorpus: up to 2019-10-31
Subset of regular MPs (excluding chairpersons and guests)

COVID infections data: Johns Hopkins University COVID-19 Data Repository
https://github.com/CSSEGISandData/COVID-19

Tools

https://www.clarin.si/noske/index-en.html

Orange data mining
orangedatamining.com

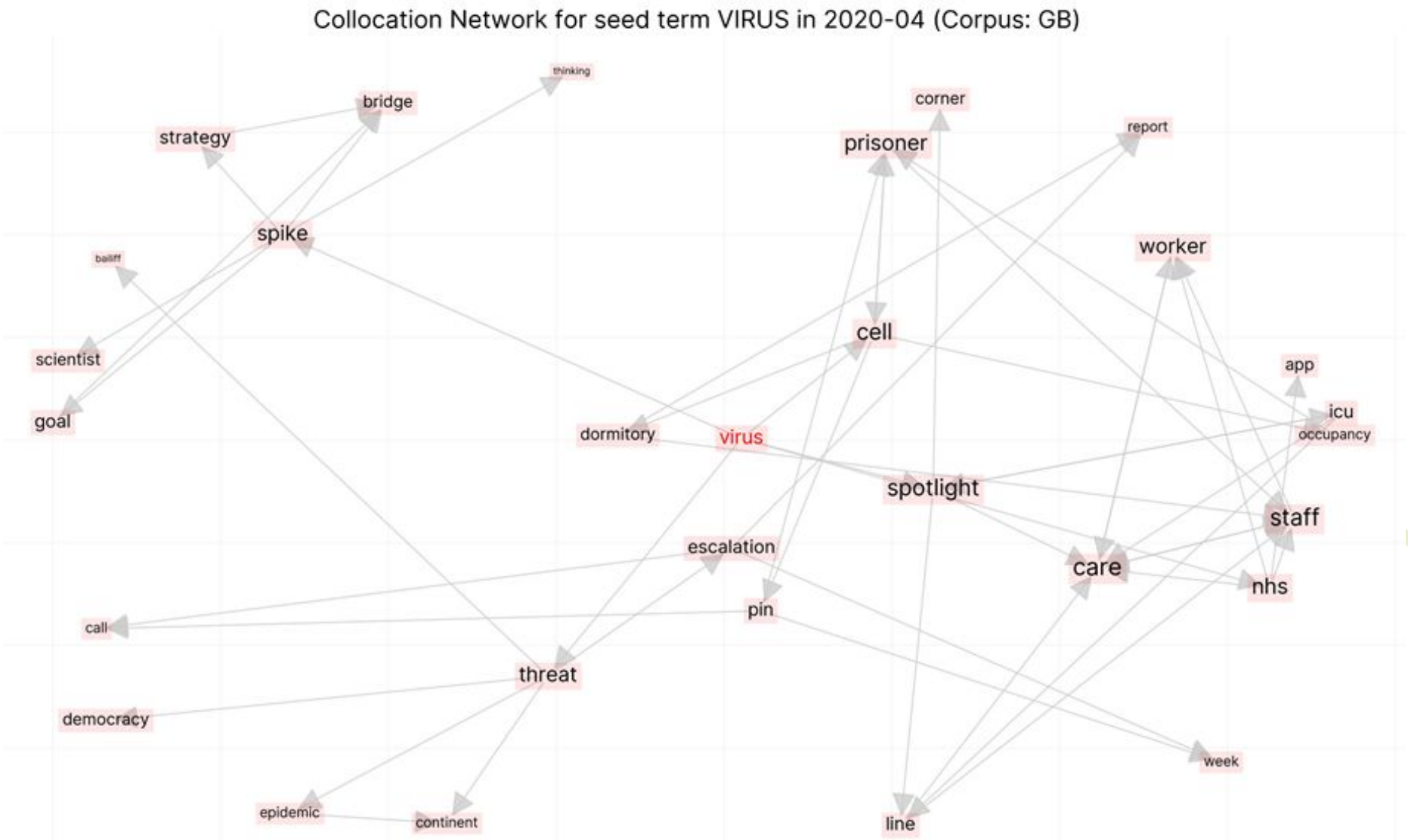


Python, R

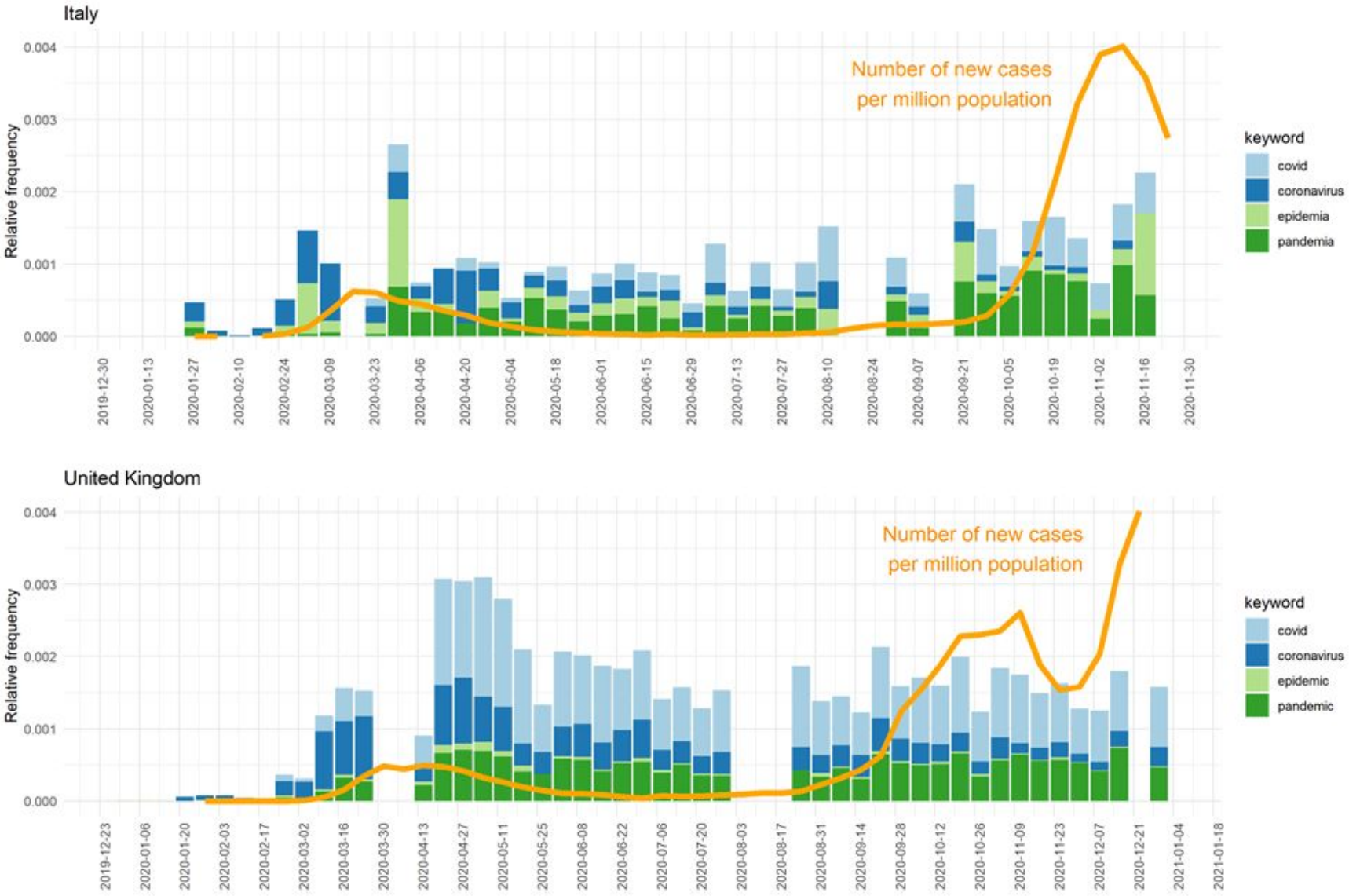
Keywords

Italy	Poland	Slovenia	UK
pandemic	pandemic	epidemic	covid
covid	*bell*	ventilator	coronavirus
covid-19	coronavirus	coronavirus	**furlough**
coronavirus	covid-19	covid-19	lockdown
lockdown	mask	virus	pandemic
mask	epidemic	quarantine	distancing
recovery	quarantine	corona	**ppe**
European Stability Mechanism (ESM)	Kukiz15	mask	*inaudible*
distancing	the Left	pandemic	*unmute*
fund	**shield**	**anti-corona (adj)**	oim
serologic	**anti-crisis**	/paramilitary group/	biden
virus	covid	/name/	kickstart
pandemic	**epidemic (adj)**	covid	ima
generation	ventilator	**anti-corona (noun)**	fcdo
infection	**remote**	infection	chis
next	**Szumowski**	**respirator**	quarantine
/name/	individual pension account	/name/	**asymptomatic**
fibp (protein)	small bottle of liquor	/name/	**post-covid**
Ministerial Decree	**sars-cov-2**	conscription	polygraph
headache	*applause*	**respirator**	virus

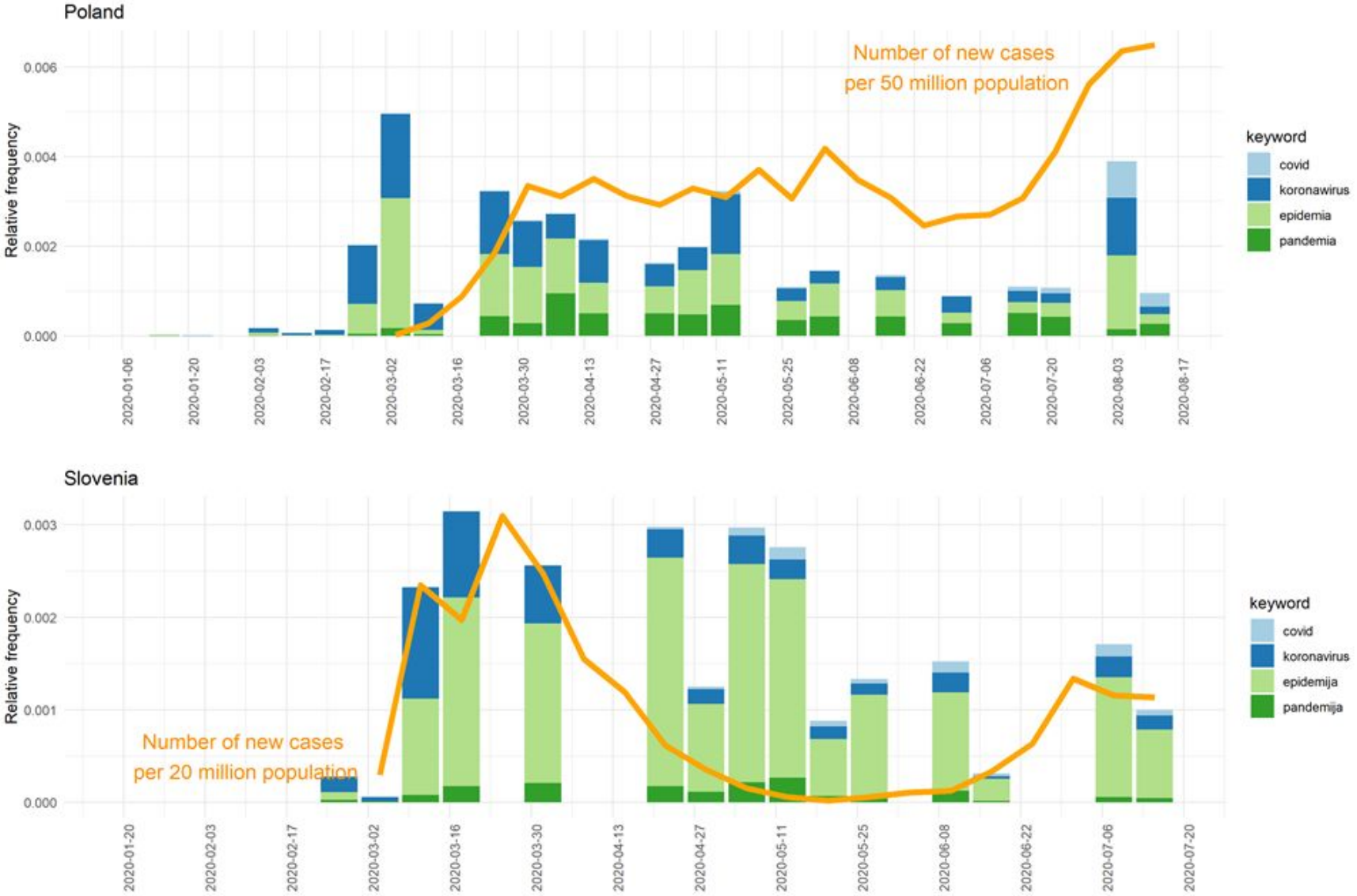
Collocation networks



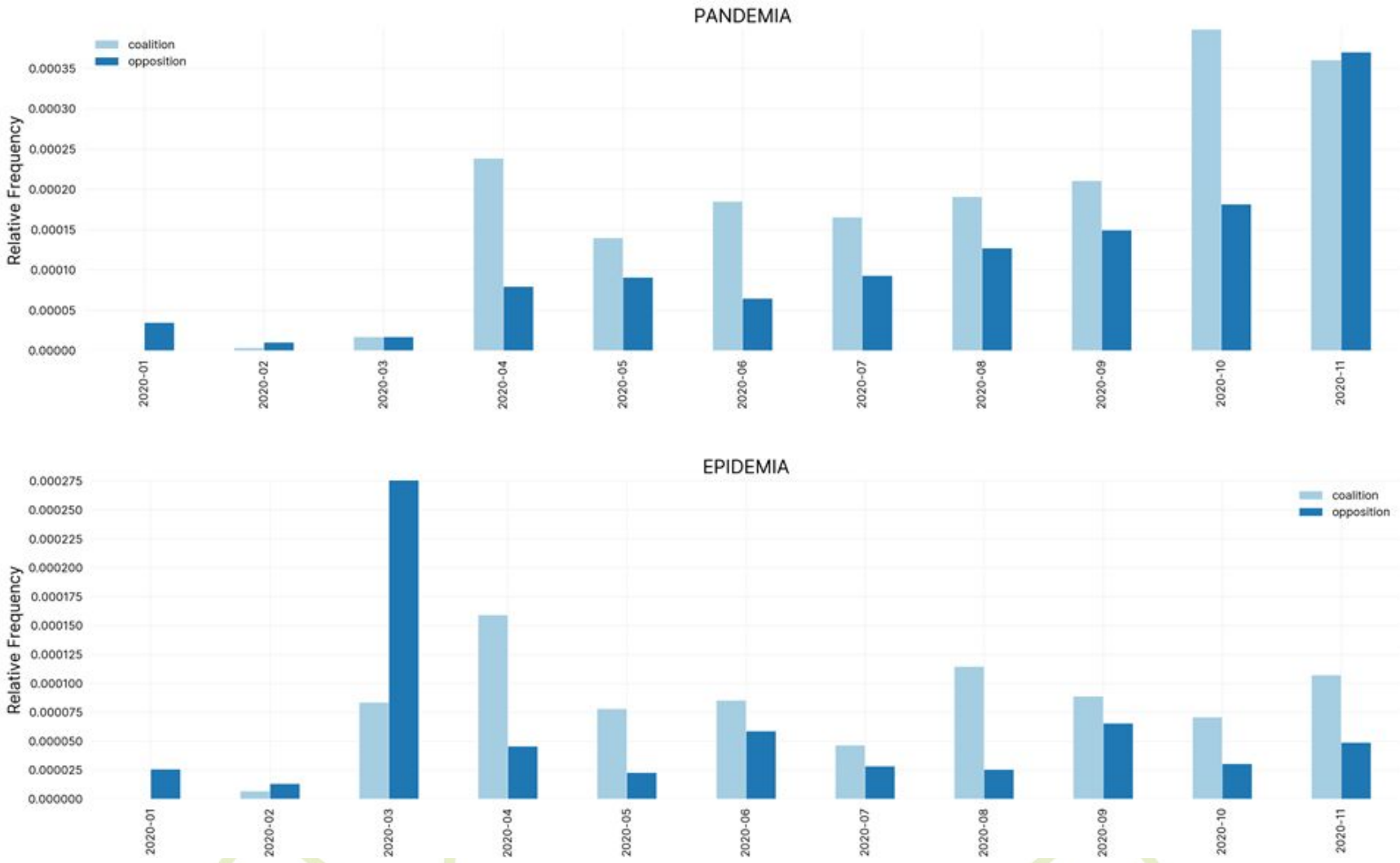
Timelines


Italy
Relative frequency
0.004
0.003
0.002
0.001
0.000
Number of new cases
per million population
keyword
covid
coronavirus
epidemia
pandemia
2019-12-30
2020-01-13
2020-01-27
2020-02-10
2020-02-24
2020-03-09
2020-03-23
2020-04-06
2020-04-20
2020-05-04
2020-05-18
2020-06-01
2020-06-15
2020-06-29
2020-07-13
2020-07-27
2020-08-10
2020-08-24
2020-09-07
2020-09-21
2020-10-05
2020-10-19
2020-11-02
2020-11-16
2020-11-30
United Kingdom
Relative frequency
0.004
0.003
0.002
0.001
0.000
Number of new cases
per million population
keyword
covid
coronavirus
epidemic
pandemic
2019-12-23
2020-01-06
2020-01-20
2020-02-03
2020-02-17
2020-03-02
2020-03-16
2020-03-30
2020-04-13
2020-04-27
2020-05-11
2020-05-25
2020-06-08
2020-06-22
2020-07-06
2020-07-20
2020-08-03
2020-08-17
2020-08-31
2020-09-14
2020-09-28
2020-10-12
2020-10-26
2020-11-09
2020-11-23
2020-12-07
2020-12-21
2021-01-04
2021-01-18

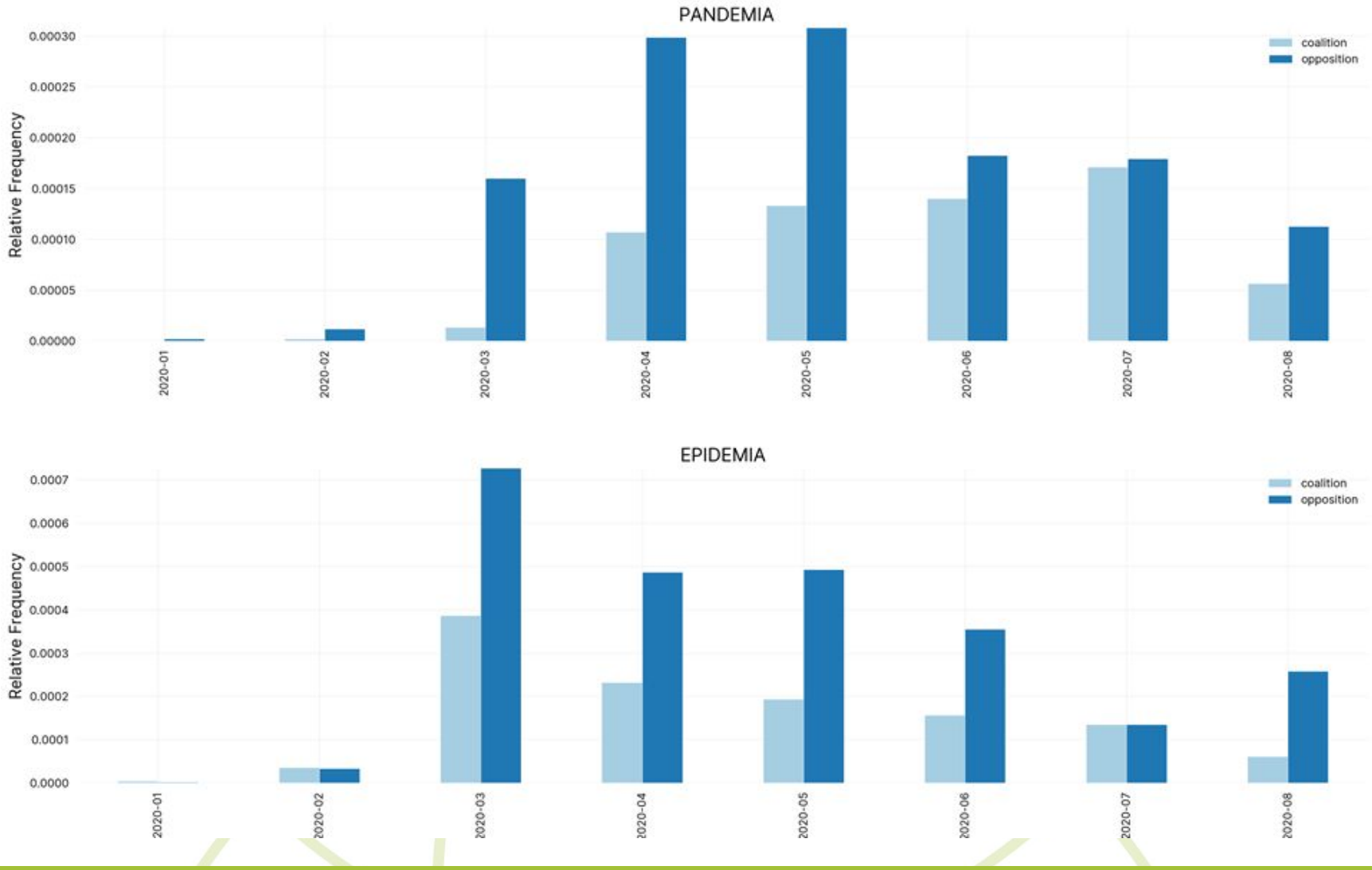
Timelines


Poland
Relative frequency
0.006
0.004
0.002
0.000
Number of new cases
per 50 million population
keyword
covid
koronawirus
epidemia
pandemia
2020-01-06
2020-01-20
2020-02-03
2020-02-17
2020-03-02
2020-03-16
2020-03-30
2020-04-13
2020-04-27
2020-05-11
2020-05-25
2020-06-08
2020-06-22
2020-07-06
2020-07-20
2020-08-03
2020-08-17
Slovenia
Relative frequency
0.003
0.002
0.001
0.000
Number of new cases
per 20 million population
keyword
covid
koronavirus
epidemija
pandemija
2020-01-20
2020-02-03
2020-02-17
2020-03-02
2020-03-16
2020-03-30
2020-04-13
2020-04-27
2020-05-11
2020-05-25
2020-06-08
2020-06-22
2020-07-06
2020-07-20

Coalition vs. opposition: Italy



Coalition vs. opposition: Poland



Challenges and ideas

- identifying specific COVID-related measures and tracing them over time
- examining differences between political parties (and possible commonalities among parties with similar ideologies in different countries)
- analyzing mentions of named entities

More:

https://github.com/rubenros1795/ParlaMintCase

https://github.com/mkolczynska/parlamint

A comparative analysis of the ParlaMint corpus

Master thesis Data Science, University of Amsterdam

Miguel Pieters

CLARIN

ParlaMint project

Mission: Creating comparable multilingual corpora of parliamentary debates

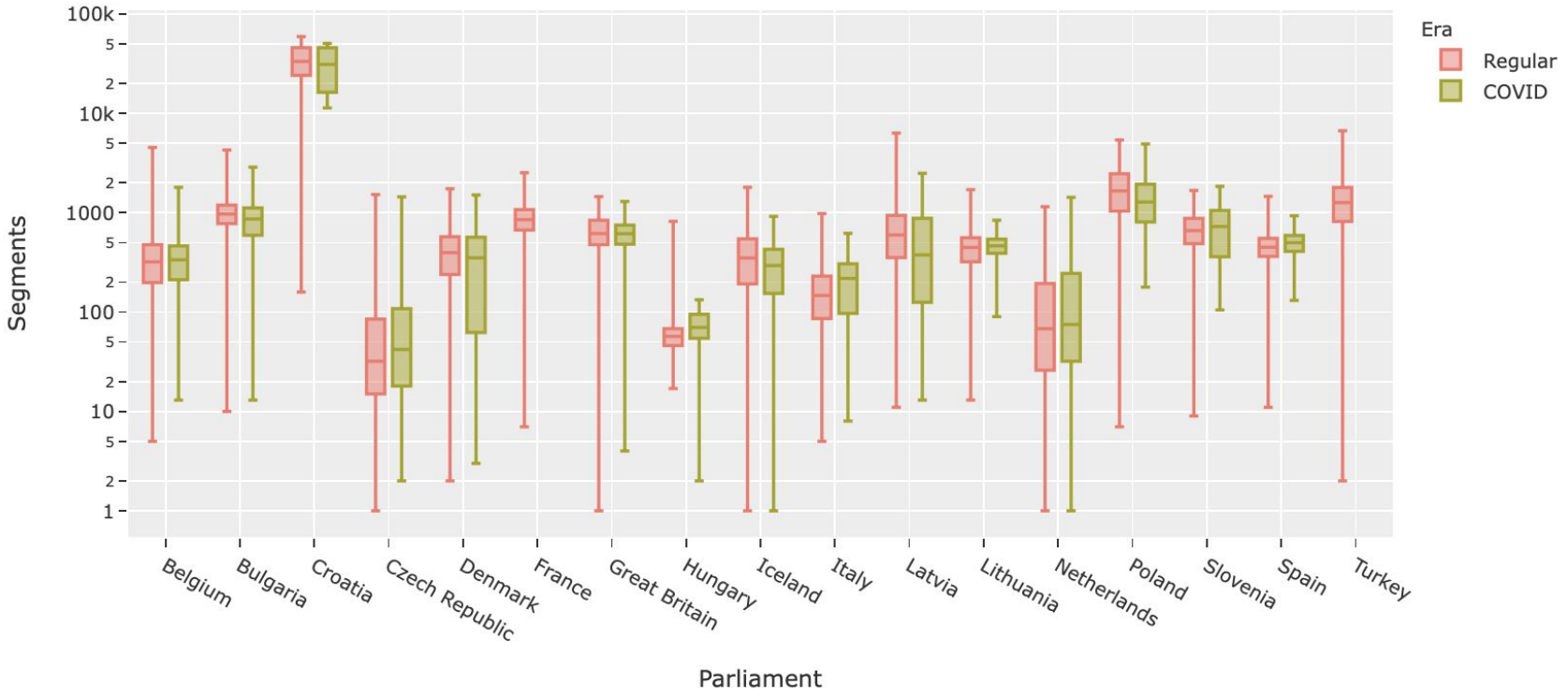
Interesting to evaluate the comparability of the corpus as a thesis study!

Study analysis

1. Create a high-level overview of the corpus
2. Compare the Part-of-Speech tag distribution of parliamentary debates against a background corpora such as Wikipedia dumps
3. Evaluate the role of women within parliaments
4. Comparing how the different parliaments handle migration, climate and corona topics

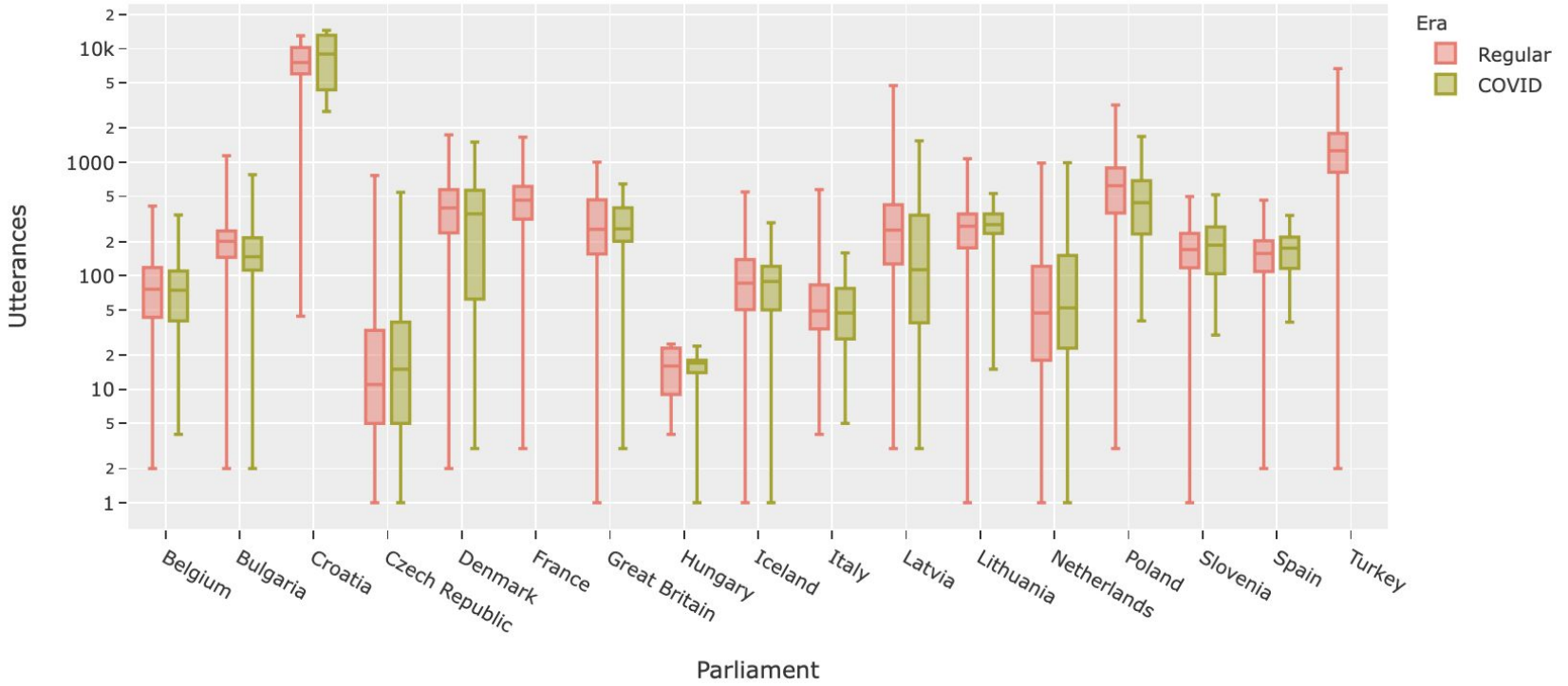
Corpus overview

Segments per debate per parliament



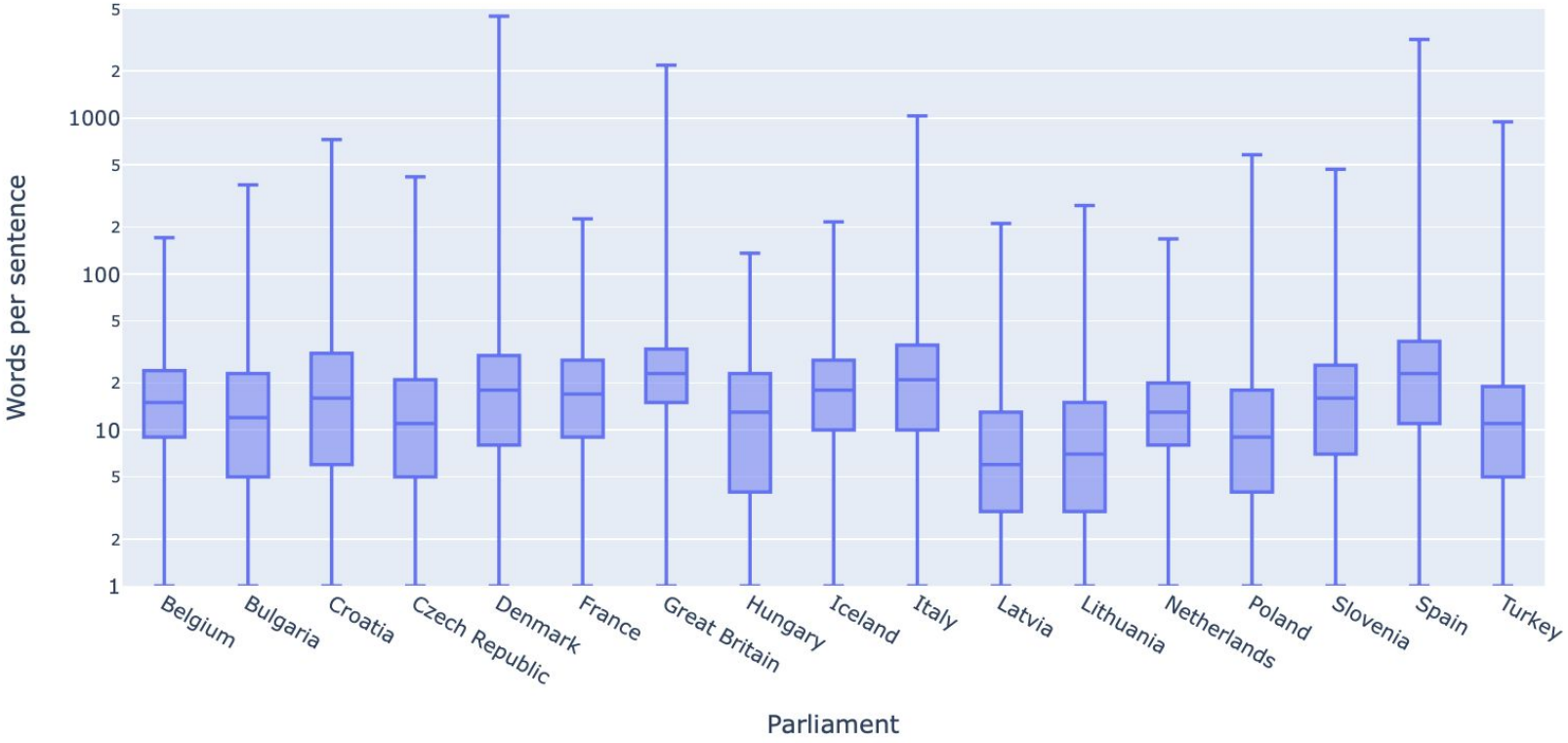
Corpus overview

Utterances per debate per parliament



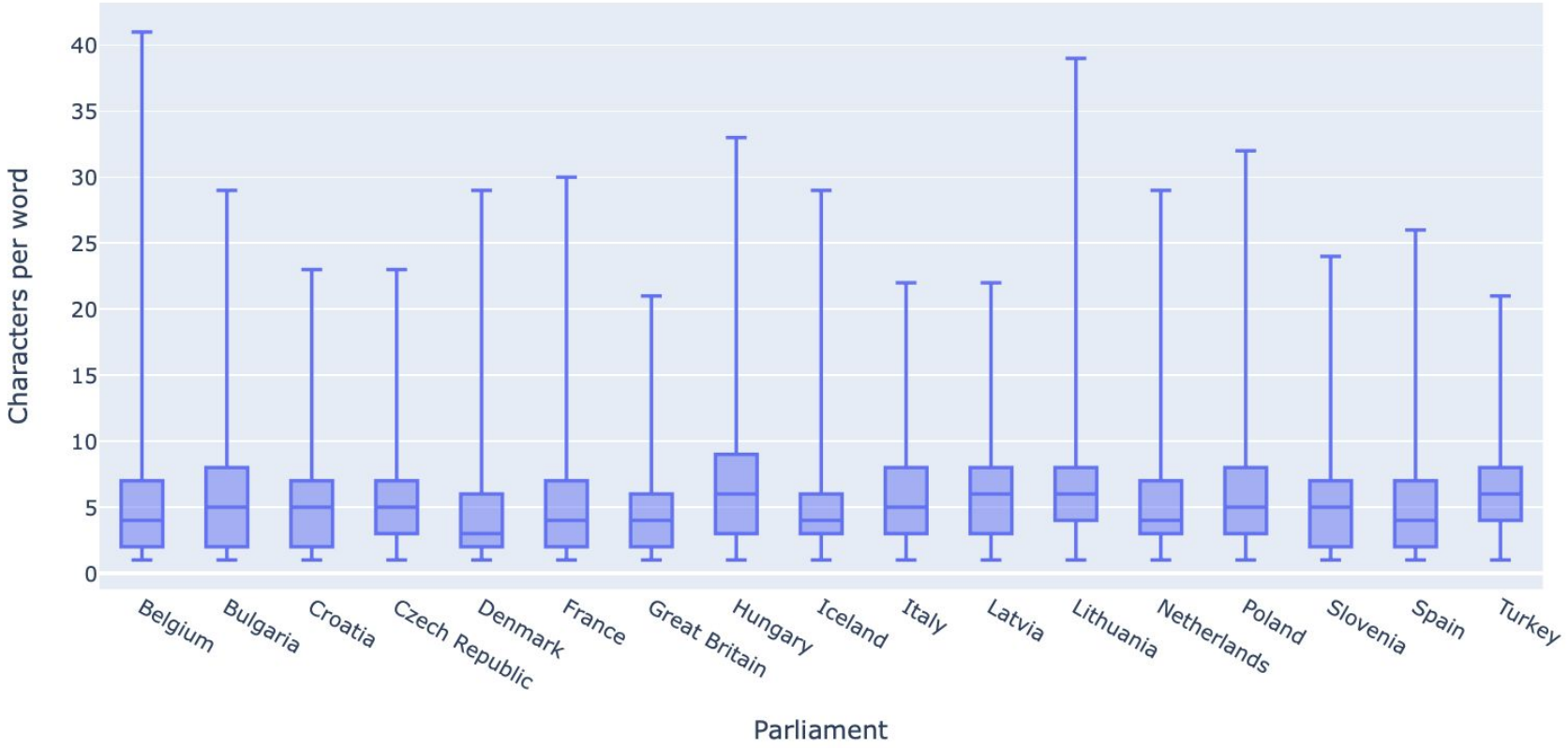
Corpus overview

Words per sentence per parliament



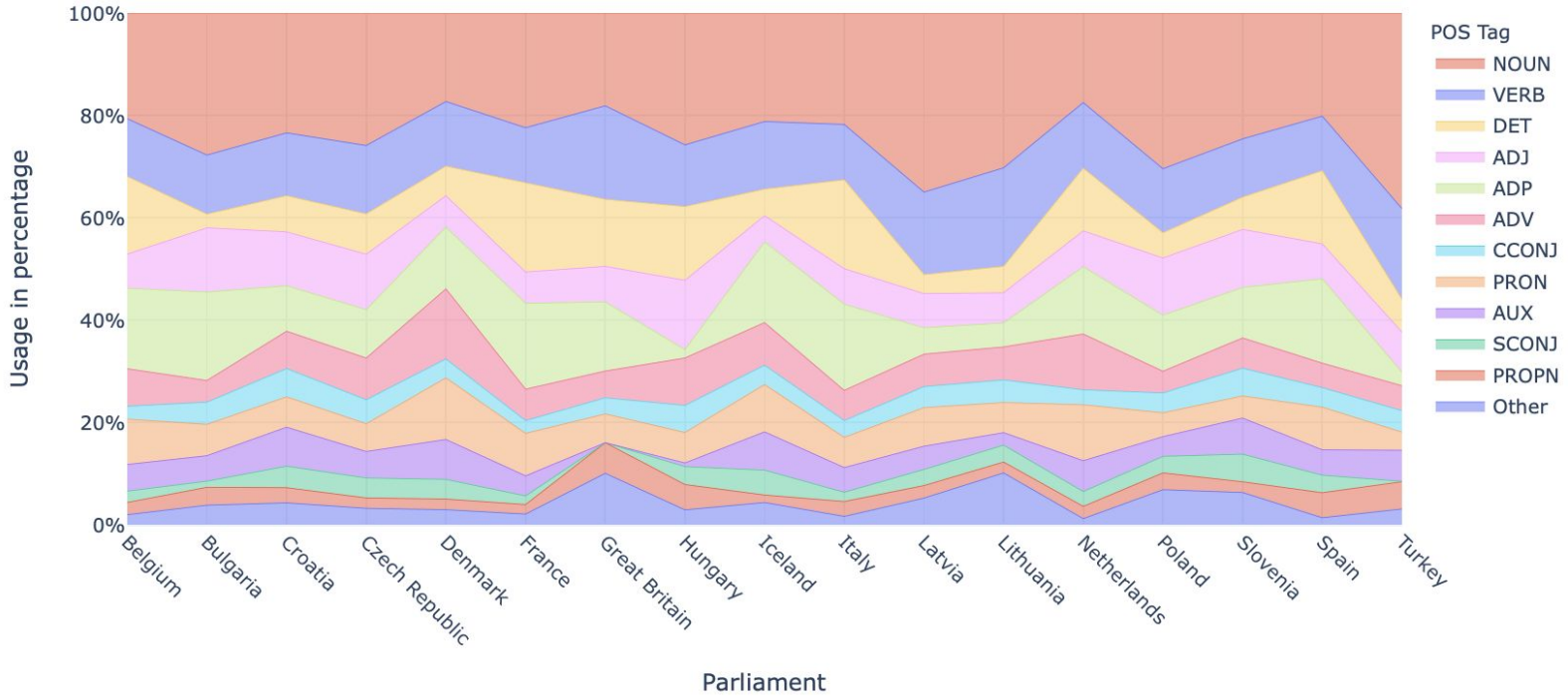
Corpus overview

Characters per word per parliament



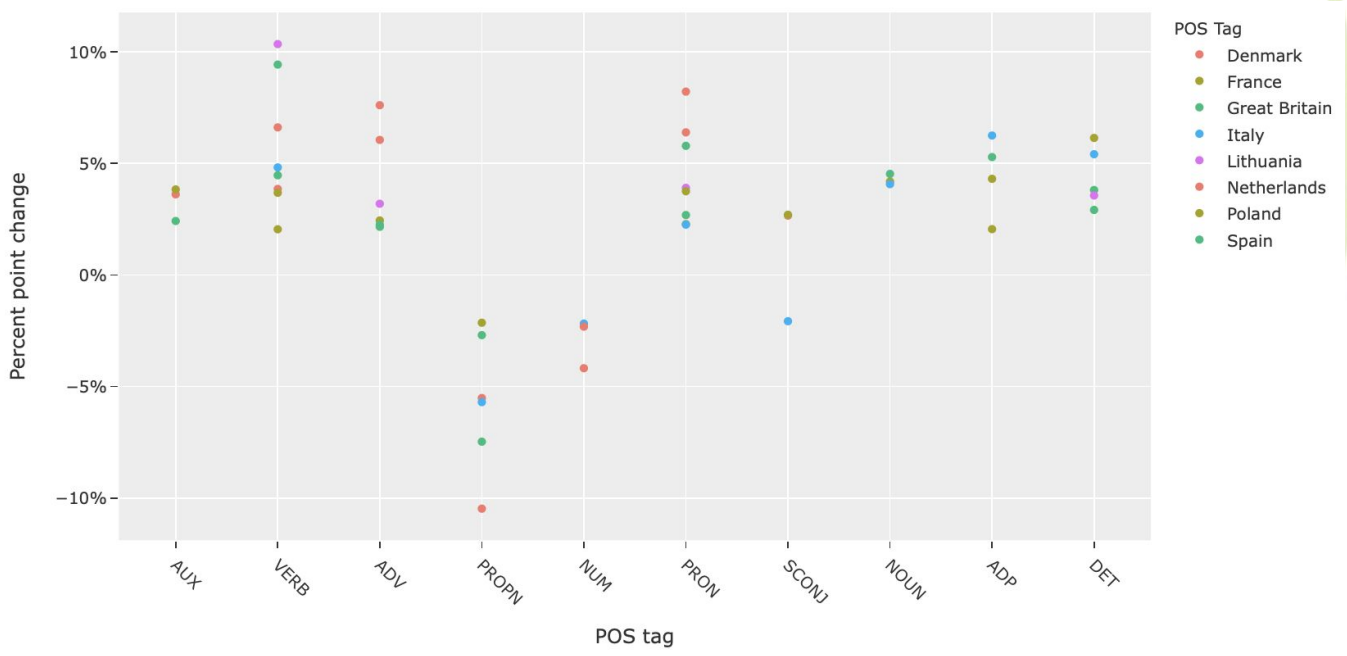
Part Of Speech tag analysis

POS tag distribution per parliament provided in the annotated dataset



POS tags: ParlaMint vs. Wikipedia

POS tags which differed at least 20% relatively and 2% absolutely. A positive change means a more frequent occurrence in the ParlaMint corpus.



A Feminized Democracy?

- Women in parliament are important. Female MPs represented a hitherto marginalized ‘women’s interest’, placed ‘women’s issues’ on the agenda, and added a feminine perspective to existing discussion
  - Study: A Feminized Language of Democracy? The Representation of Women at Westminster since 1945.
- By measuring the attention given to female party leaders during the Dutch’s 2021 elections, researchers from the Vrije Universiteit Amsterdam concluded they received less attention than male party leaders.
  - Study: VU Election Research 2021 Dutch Parliamentary Elections
- Can we measure if women receive equal attention as their male colleagues in parliament using the ParlaMint corpus?

A Feminized Democracy?

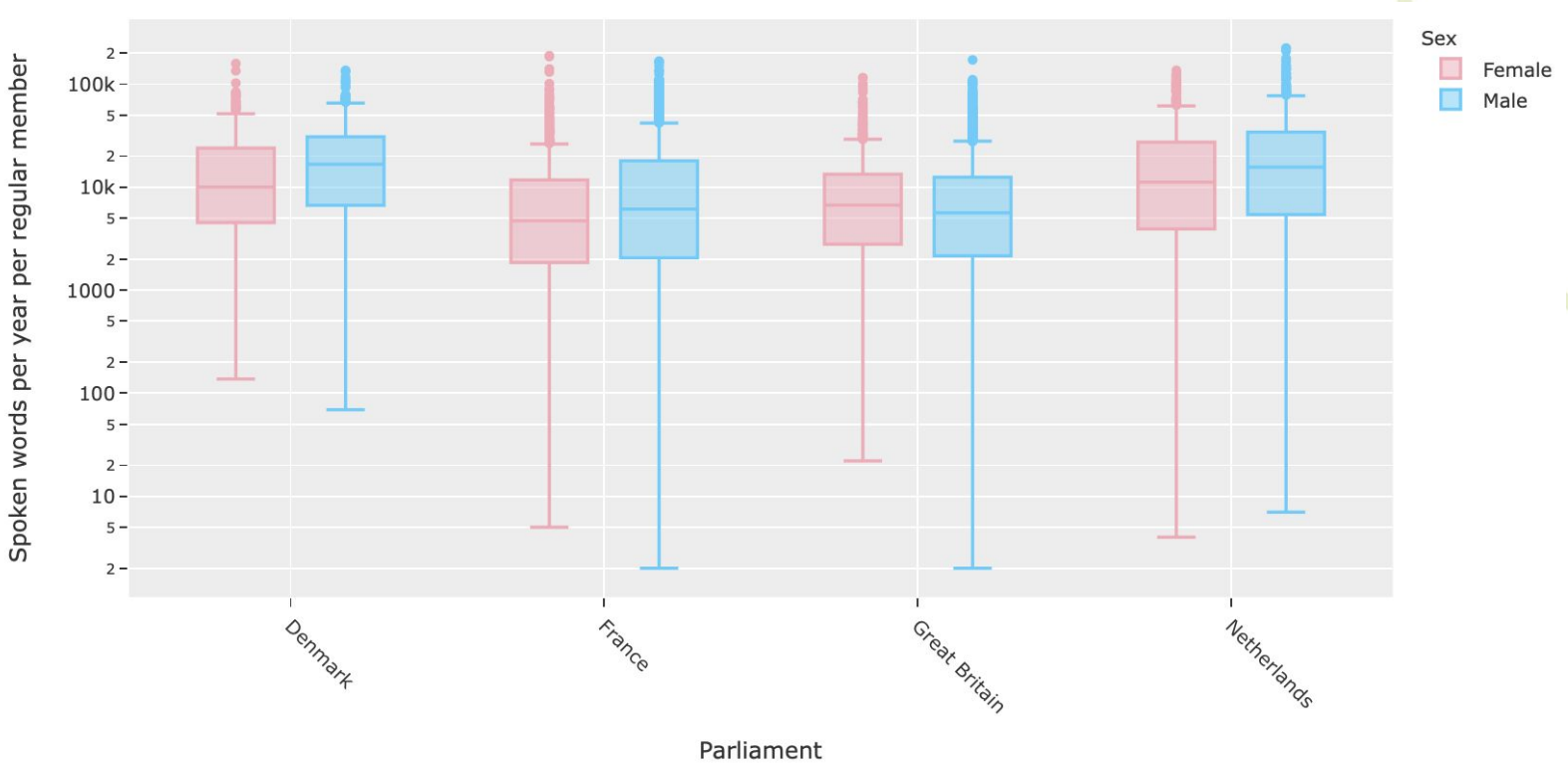
We must discriminate regular members of parliament from ministers, because ministers naturally speak substantially more than regular members of parliament.

Table 7. Population sizes research groups

Parliament	Gender	Minister	Count	Mean tokens
Denmark	Female	Yes	155	23.864,57
		No	393	17.382,93
	Male	Yes	252	28.162,63
		No	650	22.048,16
France	Female	Yes	1	2.665,00
		No	897	10.707,70
	Male	Yes	2	26.261,00
		No	1.346	14.818,64
Great Britain	Female	Yes	289	33.391,99
		No	2.507	9.963,38
	Male	Yes	710	31.150,66
		No	5.760	9.353,64
Netherlands	Female	Yes	45	77.388,58
		No	634	19.835,56
	Male	Yes	74	140.897,35
		No	1.152	24.294,35

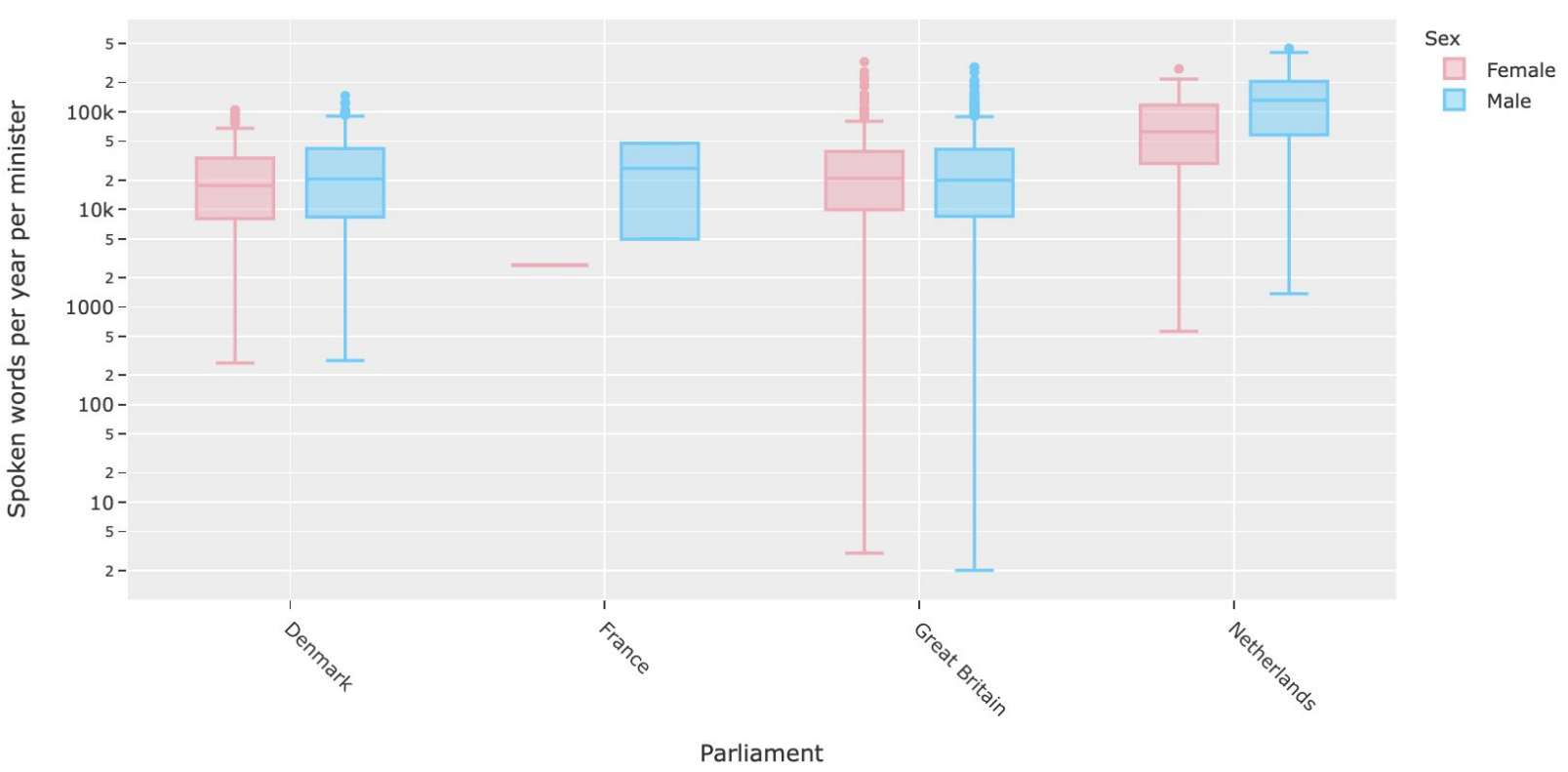
A Feminized Democracy?

- Words per year per gender for <u>regular members of parliament</u>
  - Differences appear smaller than they are, the Y-axis on a log-scale!



A Feminized Democracy?

- Words per year per gender for ministers
  - Differences appear smaller than they are, the Y-axis on a log-scale!



A Feminized Democracy?

- Spoken words per gender hypothesis:
  - $H0: \mu_{male} - \mu_{female} = 0$
  - $HA: \mu_{male} - \mu_{female} > 0$

Table 4. Gender t-test p-values

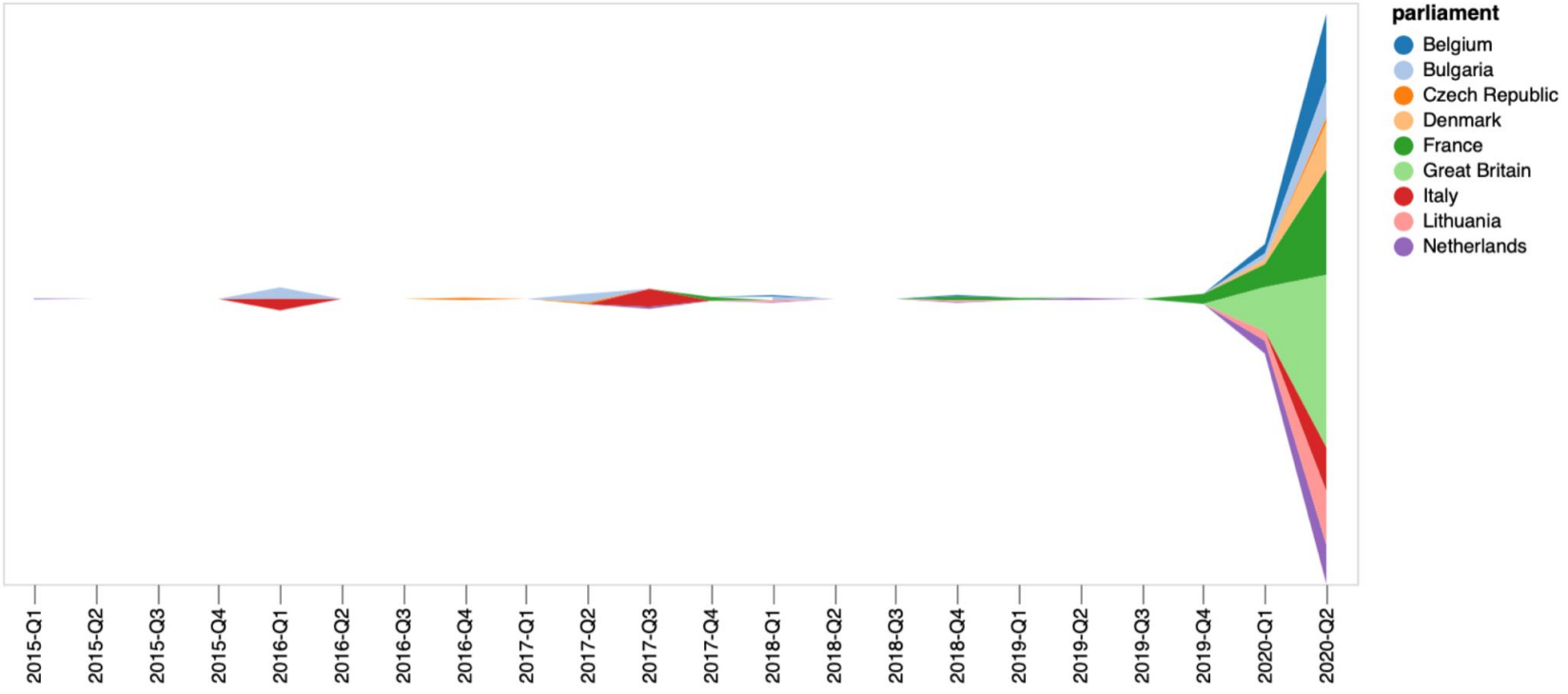
#	Parliament	ministers	non-ministers
1	France	N/A	<0.01
2	Netherlands	<0.01	<0.01
3	Denmark	0.07	<0.01
4	Great Britain	0.92	>0.99

Hurdle of a multilingual corpora

- ParlaMint corpus
  - 17 parliaments
  - 16 languages
  - 4 alphabets
- How can we compare textual semantics and refer to the same real-world entity?
  - Google Translate? Neural Machine Translations?
    - It depends on your budget…
  - Topic modelling on the translated corpus

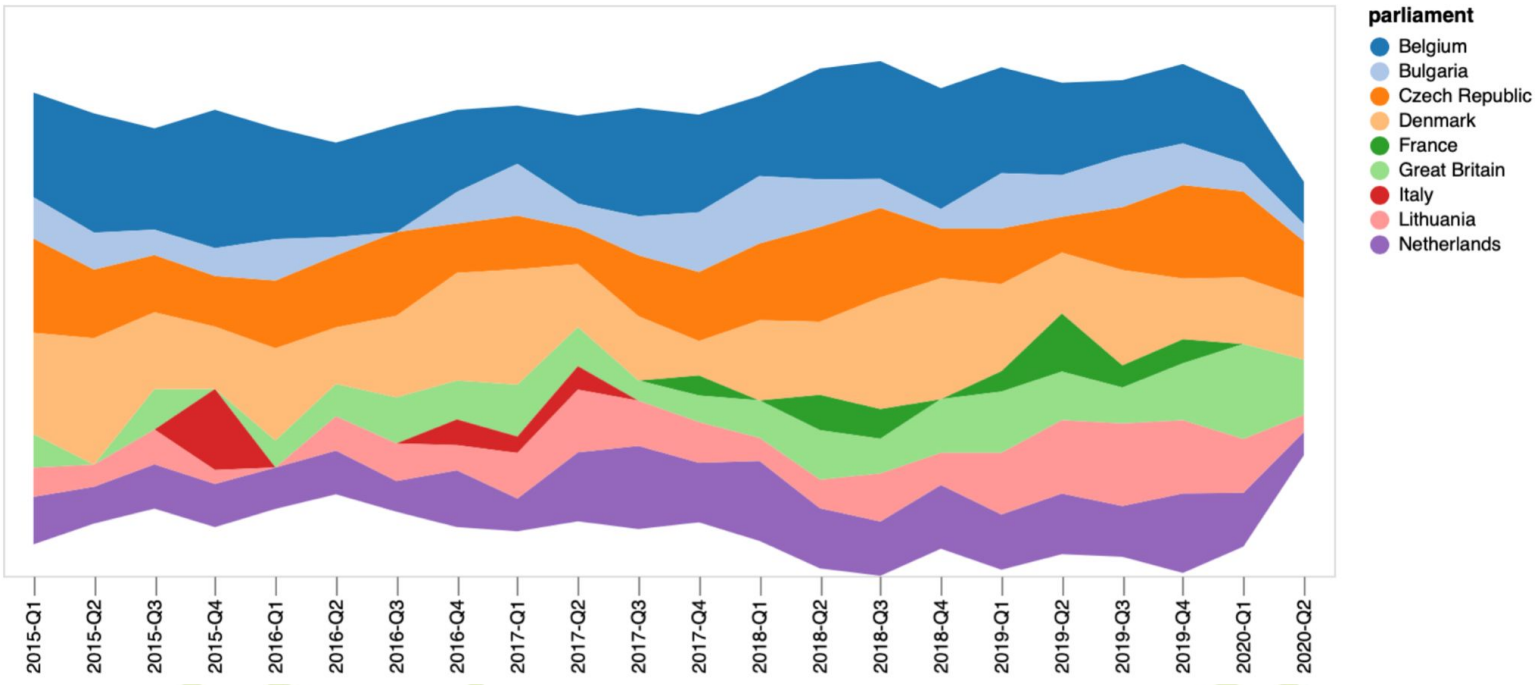
Corona, climate and migration

Corona



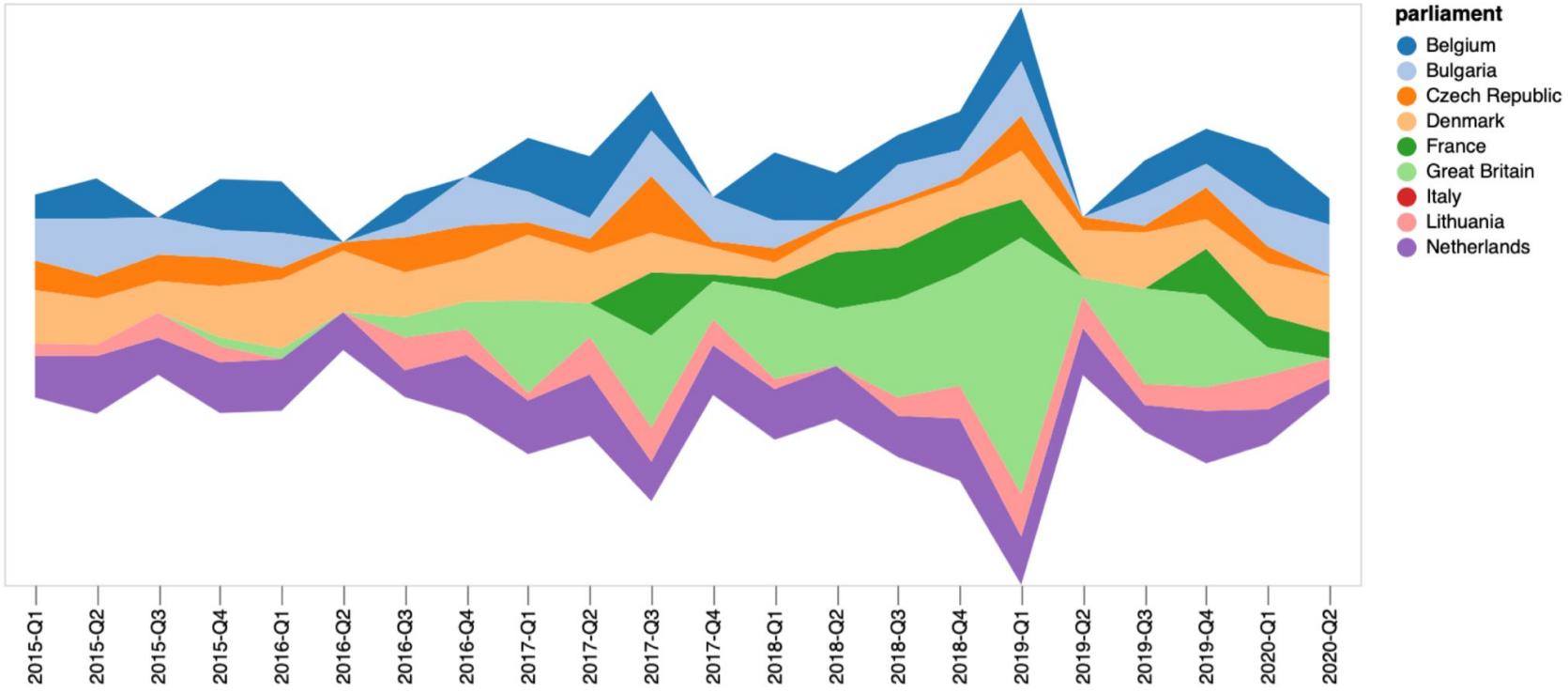
Corona, climate and migration

Climate



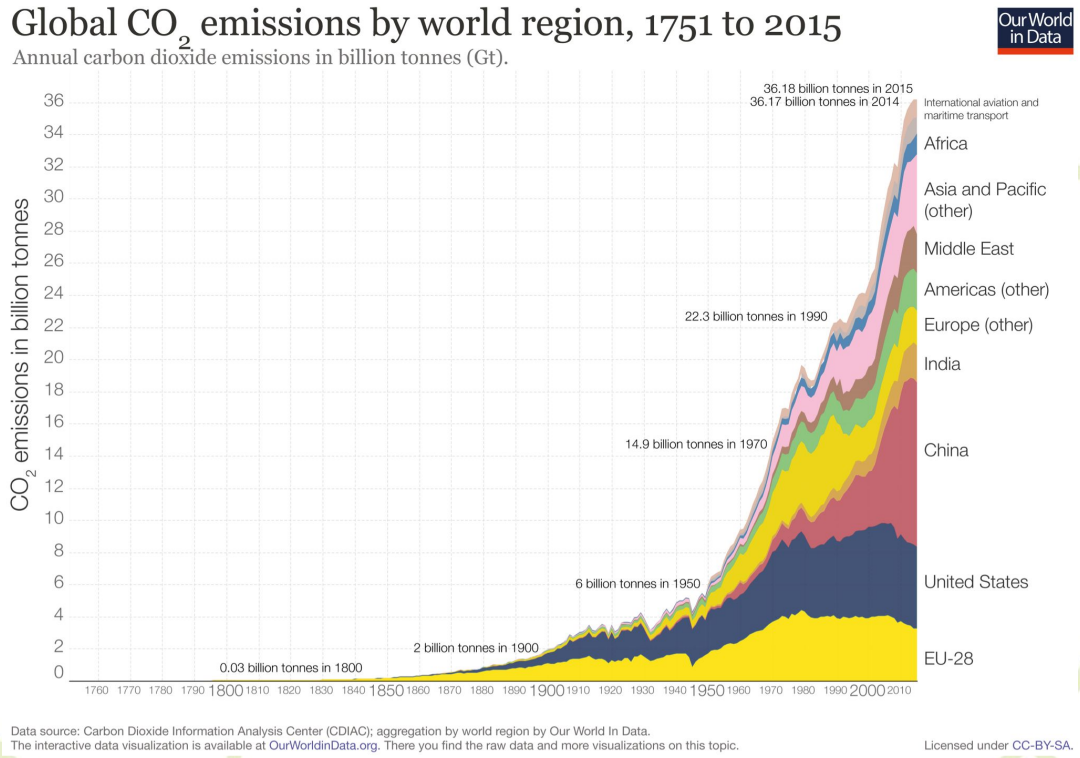
Corona, climate and migration

Migration



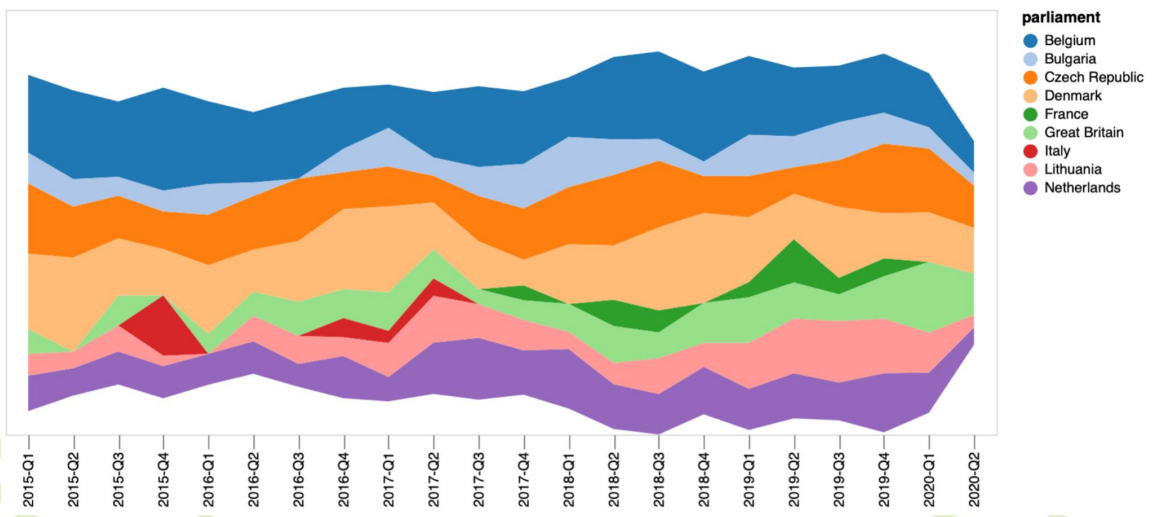
A question to think about

- In the past 30 years we have emitted 50% of all CO-2 emissions since 1751. We have been climate-aware for 32 years.



A question to think about

- Dutch court: Climate change affects human rights, including the right to live. The court rules that companies such as Shell must respect these rights, and therefore is obligated to reduce its emissions.
- *Is it a good thing that law starts to interfere with climate change instead of solely relying on politics to get the job done?*



Parlamint and Parlameter: How standardized data formats empower end users

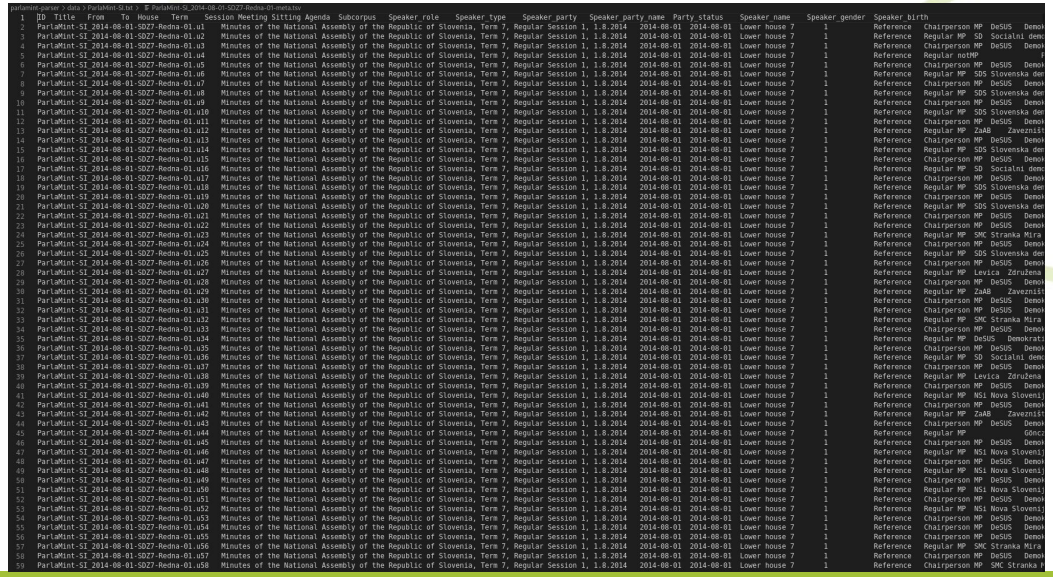
Filip Dobranić
Today is a new day, Institute for other studies

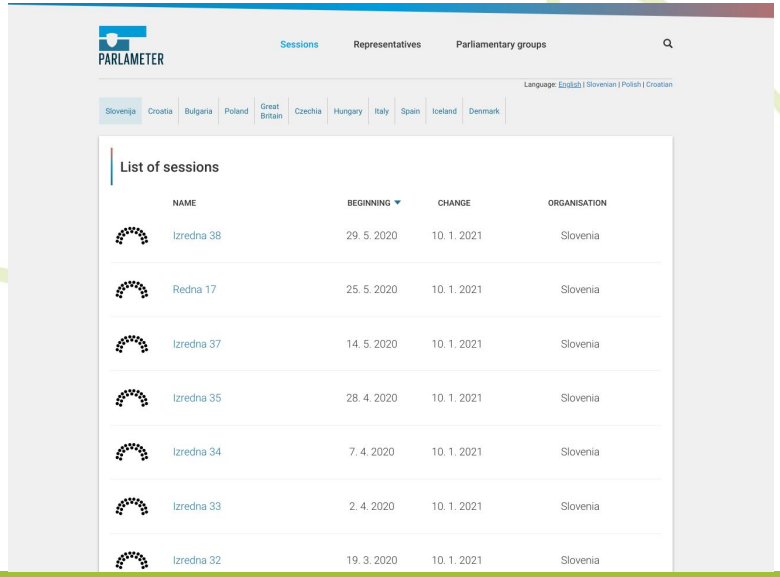
CLARIN

Communication is everything

If no one is there to see the data, does it really exists?

- Datasets are representations - who are they for?
  - Computers
  - Researchers
  - The Interested Public
- Data literacy is unevenly distributed, and that's OK.





Parlameter

Parlameter is available to open any
parliament
to the public.

Speed up research and journalistic work

If unstructured, data is practically inaccessible. Help the media and research institutions explore your datasets, daily serving them analyses and visualizations ready for immediate publishing.

Give a boost to developers

Show developers some love without troubling your IT people. With Parlameter, your data will be exportable, machine-readable, and ready for action!

Indulge your audience

People will benefit twice - through the work of journalists and developers, as well as by sifting through data by themselves.

Lessons learned

Data stewardship is not an easy job.

- Centralized silos of data representations are hard to maintain.
  - Linguistic challenges
  - Scale requires optimization
  - Cost (h/$)
  - Agility and updates
- Tools vs. Services
- Nobody cares as much about the data as the people who created it.

People should be empowered, not deskilled.

Links

https://github.com/danesjenovdan/parlamint

https://github.com/danesjenovdan/parlamint-parser

https://github.com/danesjenovdan/parladata

https://github.com/danesjenovdan/parlanode

https://github.com/danesjenovdan/parlasite

https://github.com/danesjenovdan/parlassets

https://parlamint.parlameter.org/

Discussion panel

Lessons learnt from Czech, Icelandic, Italian and UK groups

Mini-grant recipients

Parliament of the Czech Republic, Chamber of Deputies:

Matyáš Kopp and Barbora Hladká
Institute of Formal and Applied Linguistics (ÚFAL), Charles University, Prague

- The motivation:
  - no corpus-based project on Czech parliamentary data
  - ParlaMint's vision of uniform encoding
  - ParlaMint I participants that guarantee the quality of the output
- The main challenges:
  - Data harvesting
  - Merging multiple speaker profiles of a single person
  - Detecting and patching bugs in annotating tools
- Lessons learnt:
  - Running more checks in the pipeline

Corpus of Italian Senate:

T. Agnoloni (Institute of Legal information and Judicial systems, IGSG-CNR); F. Frontini, M. Monachini, S. Montemagni, V. Quochi, G. Venturi (Institute for Computational Linguistics, ILC-CNR)

- The motivation:
  - continue our long standing collaboration with national institutions in
    - developing/adapting linguistic annotation pipeline to domain-specific language varieties,
    - managing, processing and publishing legal corpora and in developing metadata models

- The main challenges:
  - conversion into Parlamint XML-TEI from the source HTML format (not always consistent) required several transformation rules
  - manual retrieval of speakers metadata not always contained in the Senate open data portal (only members of Senate)
  - lack of a NER component included in a UD-based linguistic annotation pipeline (rules to align two different schema)

- Lessons learnt:
  - it would be preferable to start from a standard source format, e.g. Akomantoso (available since 2018)
  - need for an evaluation campaign of to test the performance of the linguistic annotation components (on-going activity)

UK Parliament (Hansard)

Matthew Coole, Paul Rayson, UCREL, Lancaster University

- Motivations
  - SAMUELS (Semantic Annotation and Mark-Up for Enhancing Lexical Searches) project (https://www.gla.ac.uk/samuels/) 2014-5, 1803-2005 Hansard and EEBO
  - Matt Coole's PhD "Novel database design for extreme scale corpus analysis" (2021) including a UK Hansard case study (https://doi.org/10.17635/lancaster/thesis/1236)
  - Re-engage in CLARIN
  - Collaboratively create very large scale comparable corpora

- Main challenges
  - Two parallel toolchains reduced to one and a new pipeline tool created (kjede) using Stanford CoreNLP mapped to UD
  - Combining Hansard API with Parliamentary Open Data API for metadata extraction for standard name, party, photo and social media information

- Lessons learnt
  - Metadata extraction, conversion and encoding is 'fun' 
  - Existing expertise is crucial for the short timescale

Q&A and Closing

More resources

- Showcases:
  - *A Return of Science? Mapping attitudes towards science and expertise in COVID-19 parliamentary debates* by Ruben Ros
  - GitHub repository with code and research report
  - *A Comparative Analysis on the ParlaMint Project* by Miguel Pieters
  - *ParlaMint and ParlaMeter: How standardised data formats empower end users* by Filip Dobranić
- Tutorial:
  - Voices of the Parliament by Darja Fišer and Kristina Pahor de Maiti

Getting involved in CLARIN

- Join our NewsFlash
  - https://www.clarin.eu/content/newsflash
- Check out our events
  - https://www.clarin.eu/events
- Open calls
  - https://www.clarin.eu/content/funding-opportunities

- Follow us on Twitter @CLARINERIC
- And stay tuned for the next cafés
  - https://www.clarin.eu/content/clarin-cafe
  - **#clarincafe**

See you at the next café

CLARIN Café is coming back in September

Stay tuned: https://www.clarin.eu/content/

Share your #SummerCafé moments with @CLARINERIC