What can you do with the CLARIN research infrastructure? The example of ParlaMint

Corpus Linguistics 2023
Pre-conference Workshop

Dario Del Fante
CLARIN-IT | University of Ferrara

2nd July 2023

CLARIN

CLARIN-IT

FERRARIAE UNIVERSITAS · 13 91 · EX LABORE FRUCTUS

Università degli Studi di Ferrara

ParlaMint:

What was it all about?

Maciej Ogrodniczuk and Petya Osenova

CLARIN

Creating comparable multilingual corpora of parliamentary debates

Tomaž Erjavec

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

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.

The Project

The project is being conducted in two stages:

- **ParlaMint I** (July 2020 – May 2021)
- **ParlaMint II** (December 2021 – May 2023)

What is ParlaMint I?

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 I 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*):

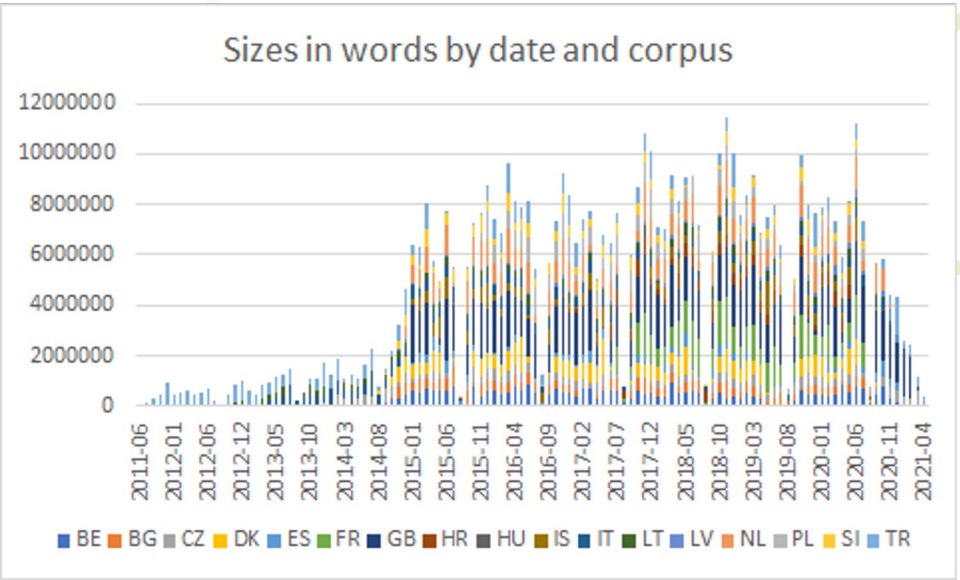
Corpora for 13 more parliaments were added according to the methodology established in *Phase 1*: Belgian, Czech, Danish, Dutch, English, French, Hungarian, Icelandic, Italian, Latvian, Lithuanian, Turkish. Spanish joined with their own funds.

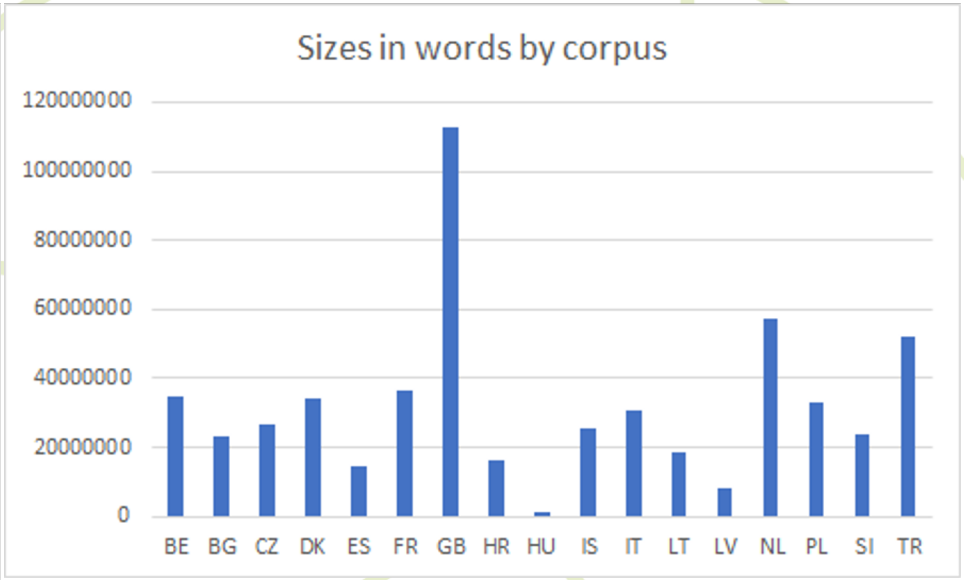
ParlaMint 2.1

- 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

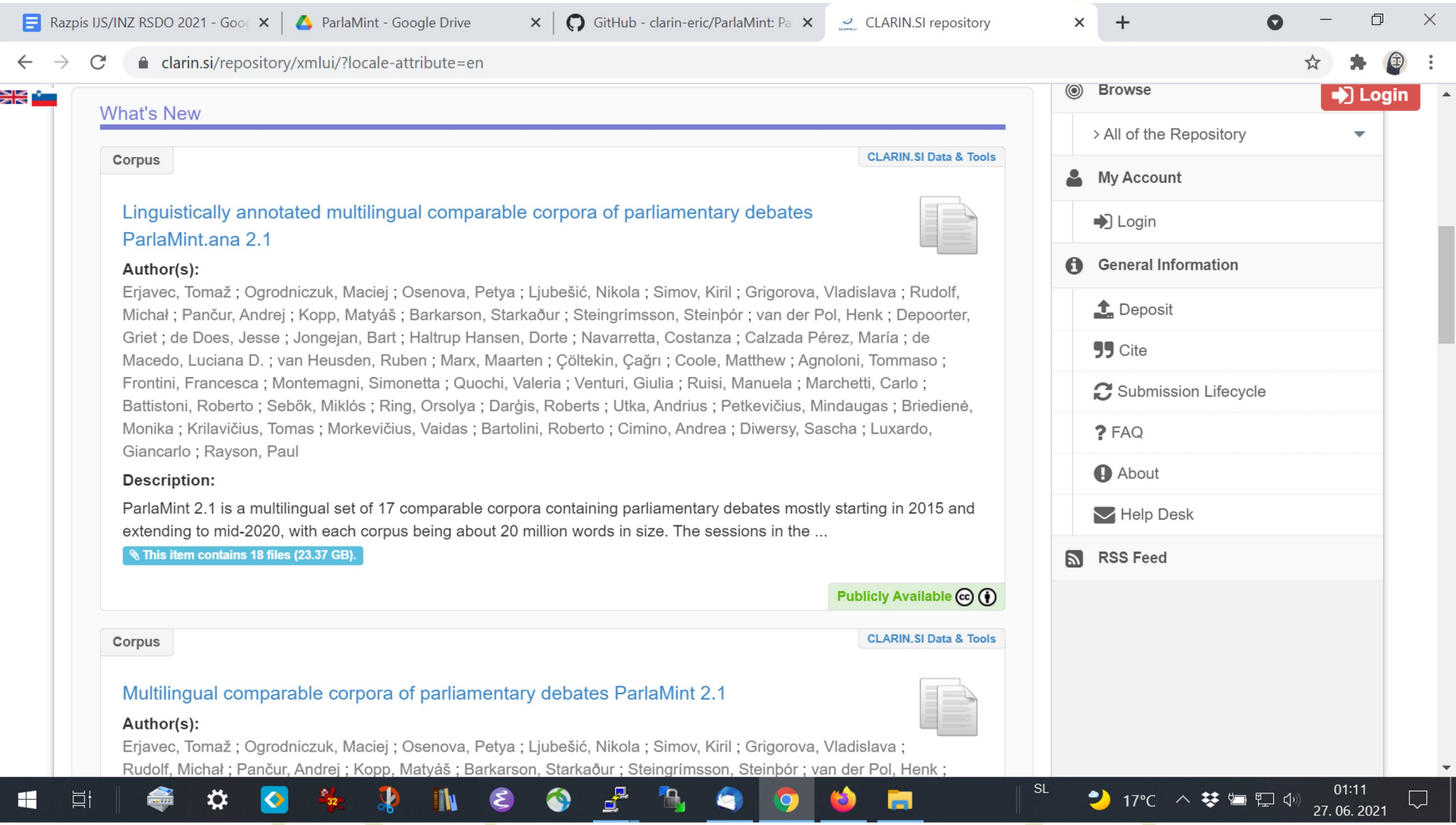
ParlaMint in numbers

- 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





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

ParlaMint II

ParlaMint II will upgrade the XML schema and validation, extend the existing corpora to cover data at least to July 2022, add corpora for new languages, further enhance the corpora with additional metadata; and improve the usability of the corpora.

Corpus Download

- The ParlaMint corpora can be downloaded from the Core Trust Seal and CLARIN certified CLARIN.SI repository in two formats. One corpus file unpacks into the source ParlaMint XML files:
  - Per-speech full metadata (19 columns) TSV files;
  - Plain text files, each line marked with speech ID;
  - CoNLL-U files, which also include NE annotations in IOB format;
  - Vertical files as used by the concordancers including the registry files, so they can be indexed and mounted on any other noSketch Engine installation, on the
  - commercial Sketch Engine, which supports more advanced features for corpus exploration, or (with some small changes) on any CWB-type (Evert and Hardie, 2011) concordancer.

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

Status of ParlaMint 3.0

ParlaMint 3.0 is ready:

- Multilingual comparable corpora of parliamentary debates ParlaMint 3.0 http://hdl.handle.net/11356/1486
- Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 3.0 http://hdl.handle.net/11356/1488

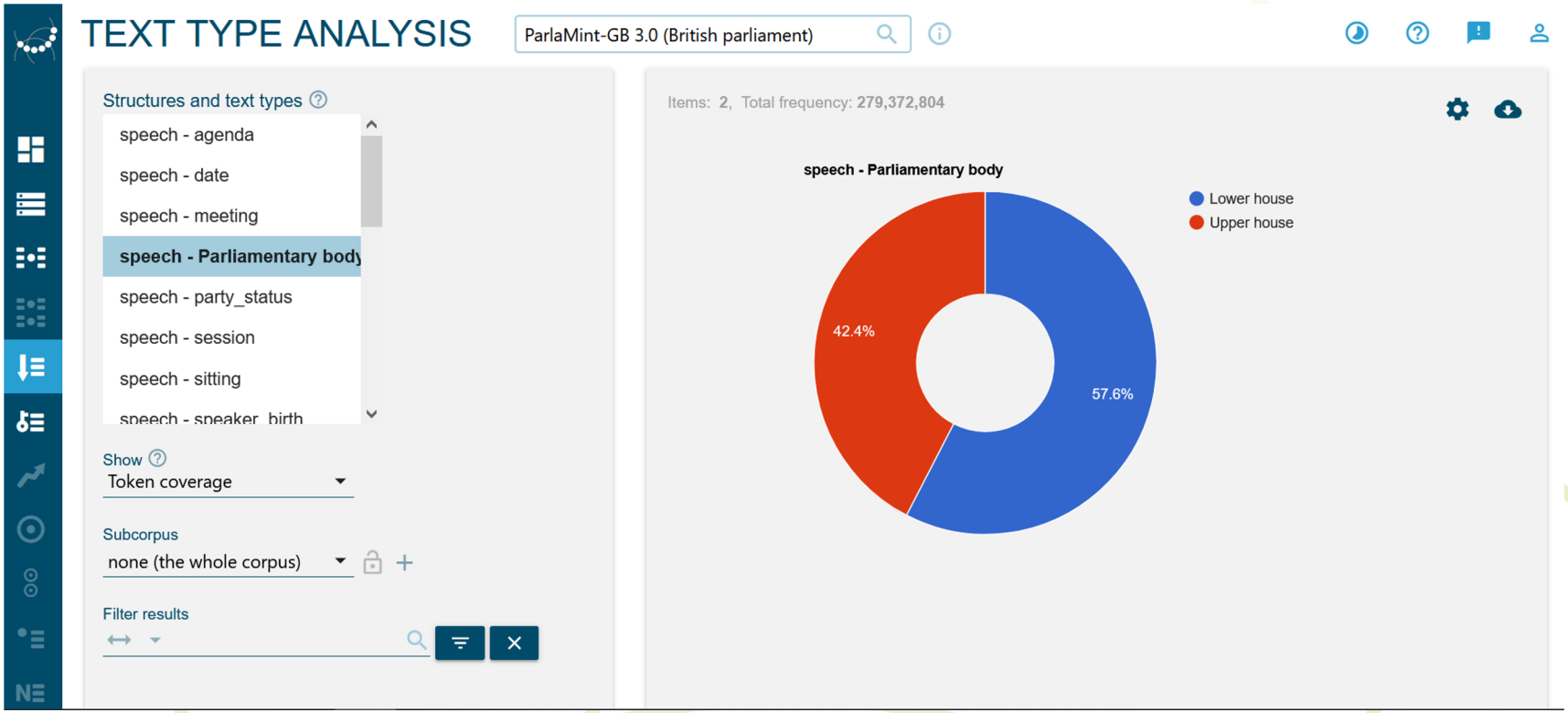
Authorship:

1. Tomaž Erjavec, Matyáš Kopp, Maciej Ogrodniczuk, Petya Osenova
2. respStmt names sorted by country

Acknowledgements of supporting projects:

1. Any *new* local project that supported the development of ParlaMint 3.0

Example screen



ParlaMint 3.1

Version 3.1 will be released in September 2023

- missing corpora: ES, ES-PV, FI, LT, RO
- correction of found bugs, cf. Milestone 3.1
- date extension to mid-2023 (so far CZ, UA)
- semantic tagging
- additional metadata for ministers and political orientations
- common taxonomies with translations
- better encoding of particular phenomena
- maybe a new MT release

Hands-on practical with Parlamint GB

- A corpus-assisted gender analysis
- What?

  The differences/similarities between the way women and men politicians discursively depict the concept of crisis in UK and Italy.

Parlamint GB

- The UK parliamentary corpus data from 2015 to March 2021 was gathered using the UK Parliament's Hansard API.
- Access to speeches from the House Commons and Lords in XML format and metadata on speakers and parties

Corpus-assisted Discourse analysis

- Interest in non-obvious meanings across discourse
- Data from Corpora
- Computer-mediated analysis
- Mixed qualitative-quantitative methodology, where quantitative techniques, characterized by their statistical reliability are assisted by the qualitative, close, analysis of linguistic data
  (e.g. Baker *et al.* 2008; 2013; Partington, Duguid & Taylor 2013; Marchi & Taylor 2018)

A corpus-assisted gender analysis

Does gender represent an influence on verbal behaviour in political and other public or institutional settings ?

- Gender and the Language of Illness (Chateris-Black & Seale 2010)
- Gender and Language in Political Institutions (Shaw 2020)
- Gender, Power and Political Speech (Cameron & Shaw 2016)
- Corpora & Gender (Baker 2014)

I know, we should move beyond the binary male VS female !

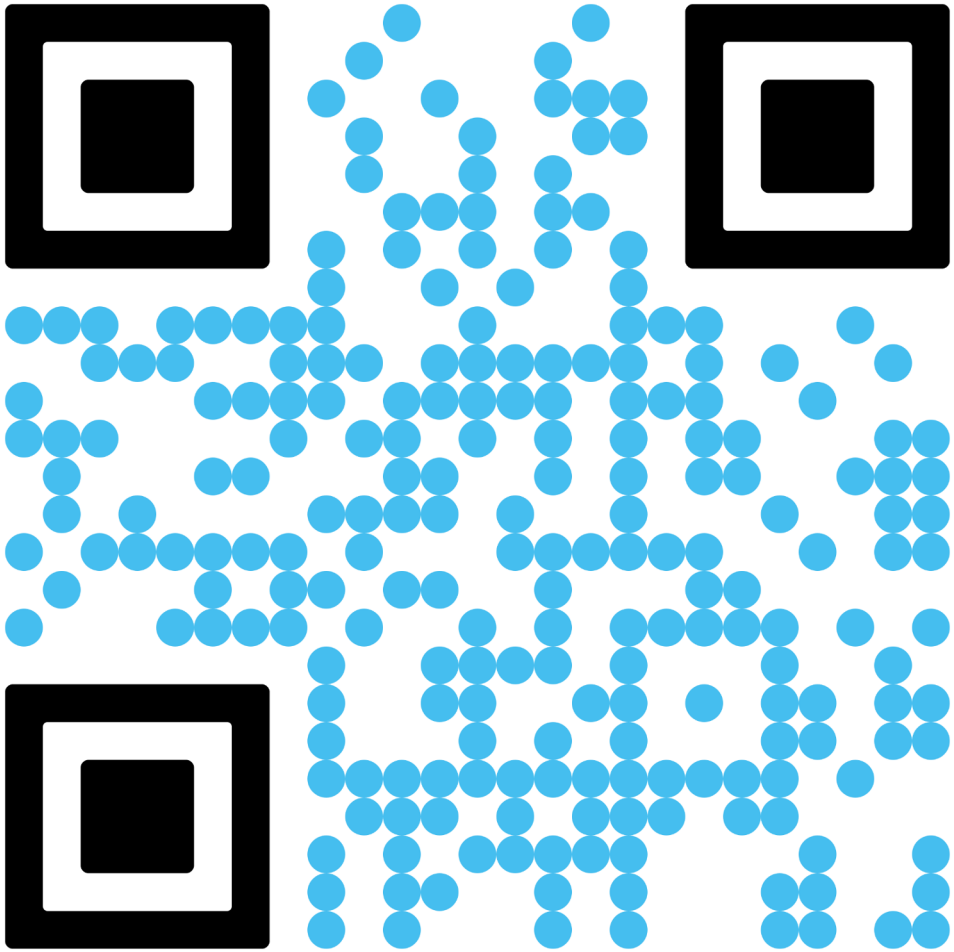
The notion of crisis - in few words

ENGLISH

- crisis
- emergency
- change
- urgent
- normality
- disaster
- distress
- dilemma
- turning point
- access
- opportunity

Select Corpus and Create a Sub-corpus

- Go to http://www.clarin.si/noske/
- Select the corpus
- Create a subcorpus
- Concordance Search
- Keyword Analysis
- Explore the corpus



CRISIS - CQL

[word="crisis" & pos="N.*"][lemma="emergency" & pos="N.*"][lemma="change" & pos="N.*"][lemma="urgency" & pos="N.*"][lemma="disaster" & pos="N.*"]

Q&A and Closing

References

Baker, P. (2014). *Using corpora to analyze gender*. A&C Black.

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, *19*(3), 273–306. https://doi.org/10.1177/0957926508088962

Baker, P., Gabrielatos, C., & McEnery, T. (2013). *Discourse Analysis and Media Attitudes: The Representation of Islam in the British Press*. Cambridge University Press. https://doi.org/10.1017/CBO9780511920103

Cameron, D., & Shaw, S. (2016). *Gender, Power and Political Speech*. Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-58752-7

Charteris-Black, J., & Seale, C. (2010). *Gender and the Language of Illness*. Palgrave Macmillan UK. https://doi.org/10.1057/9780230281660

References

Del Fante, Dario. (2022). *ParlaMint – IT – Il corpus del Senato Italiano. Una guida pratica per l'interrogazione del corpus ParlaMint-IT con NoSketch Engine, a supporto dell'analisi del discorso politico.* https://doi.org/10.5281/ZENODO.6526914

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., Pérez, M. C., De Macedo, L. D., Navarretta, C., Luxardo, G., … Fišer, D. (2023). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, *57*(1), 415–448. https://doi.org/10.1007/s10579-021-09574-0

Evert, S., & Hardie, A. (n.d.). *Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium*.

Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and Meanings in Discourse: Theory and practice in corpus-assisted discourse studies (CADS)* (Vol. 55). John Benjamins Publishing Company. https://doi.org/10.1075/scl.55

Shaw, S. (Ed.). (2020). Gender and Language in Political Institutions. In *Women, Language and Politics* (pp. 24–46). Cambridge University Press. https://doi.org/10.1017/9781139946636.002

Taylor, C., & Marchi, A. (Eds.). (2018). *Corpus approaches to discourse: A critical review*. Routledge.

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**