ParlaMint

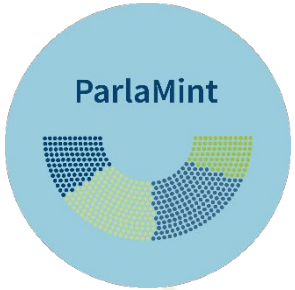
Mini-grant kick-off meeting

December 1, 2020

Agenda

- **Administrative issues** (Quirijn → Maciej)
- **Scheduling our work** (Maciej)
- **ParlaMint encoding** (Tomaž)
- **‘New’ languages and how to tame them** (grant winners)
- **Showcase idea** (Ruben)
- **Questions, discussion, AOB** (all)

Administrative issues (by Quirijn Backx from the CLARIN office)

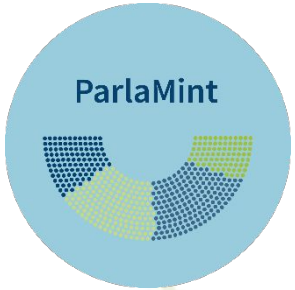


- **Eligibility**: Personnel costs, including the relevant indirect and administrative costs, are eligible for funding.
- **Financial procedure**: Each participant is responsible for assembling all relevant cost claims. The sum will be paid in one installment (Mar/Apr 2021).
- **Max amount**: € 5000
- **Invoicing**: After getting the confirmation that the corpora have been delivered the invoice should be sent to invoice@clarin.eu and addressed to:

  CLARIN ERIC
  Drift 10
  3512 BS Utrecht
  The Netherlands
  Reference: ParlaMint external grant

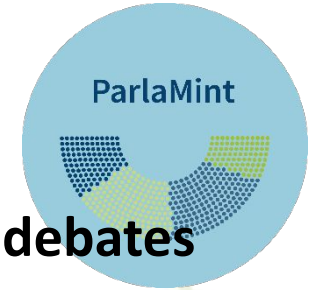
- **Questions?** Please mail quirijn@clarin.eu

Scheduling our work



- **Main aims**:
  - coverage of the same periods
  - the same depth of standard schema and linguistic processing
- **Final delivery date**: 31 March 2021
- **Progress monitoring**: meetings every last Wednesday of the month, 13:00 CEST
- **Subtask completion dates**:
  - Jan 27: gathering data and encoding it in Parla-CLARIN format
  - Feb 24: linguistic processing
  - Mar 31: documentation and delivery
- **Key contact person for corpus encoding**: Tomaž Erjavec (tomaz.erjavec@ijs.si)

ParlaMint Encoding

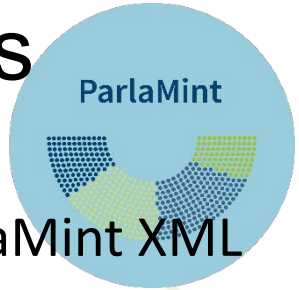


- **Multilingual comparable corpora of parliamentary debates ParlaMint 1.0.** http://hdl.handle.net/11356/1345
- The encoding of these corpora follows the Parla-CLARIN recommendations, which are, however, very general and have not changed much since before the start of ParlaMint
- In the process of the development of ParlaMint 1.0 corpora we developed a much tighter unification of encoding practices
- These should be followed, as much as possible, in version 2

Structure of the ParlaMint corpora

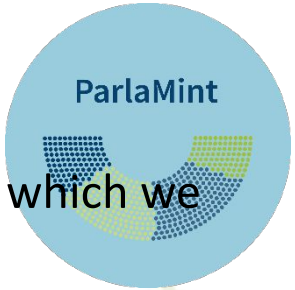
- Each of 4 corpora has a linguistically annotated variant, e.g. ParlaMint-SI.zip has ParlaMint-SI.ana.zip
- The 8 <teiCorpus> documents have similar (but not identical!) encoding
- Each corpus variant consists of a corpus root + corpus components (say, a sitting)
- One more ZIP with :
  - **Parla-CLARIN documentation** and schema (https://github.com/clarin-eric/parla-clarin)
  - **ParlaMint schemas**, which constrain the encoding (interoperability)
  - **XSLT scripts** to convert ParlaMint TEI into other formats: per-speech tabular meta-data, plain text; CoNLL-U, vertical files

How to validate your ParlaMint files



- Your corpora should validate according to the 4 ParlaMint XML schemas:
  (corpus root + corpus component) × (unannotated + annotated)
  - how to run a validator:
    https://github.com/clarin-eric/parla-clarin/wiki/Validating-your-data
- Take (probably) the Slovene (i.e. ParlaMint-SI) as the exemplar for structuring your corpora:
  - semi-fixed information in the <taxonomy> elements
  - English + localisation into corpus language
    (common ontology / vocabulary)
  - if there are reasons for them to be different, pls. discuss first

What will hopefully also be done

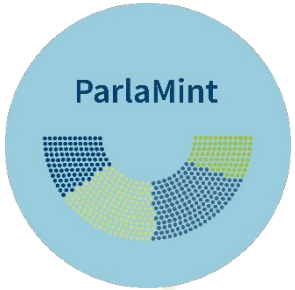


Not everything was unified completely and sensibly for version 1, which we could correct for version:

- the exact contents of some <taxonomy> elements, e.g. legislature (SI has source by Andrej Pančur)
- better linguistic annotation documentation and validation (UD / MULTEXT-East)
- better typology of transcriber notes, i.e. the values of the @type attribute on <note>, <gap>, <incident>, <kinesic>, <vocal>
- incorporate lessons and examples from ParlaMint into Parla-CLARIN
- put ParlaMint into Git
  - communication about the corpora would be better via issues than email
  - a part of Parla-CLARIN with example files
  - any Git gurus around? I‘m looking for help in managing the GitHub project!

Corpus of the Saeima (Latvian Parliament)

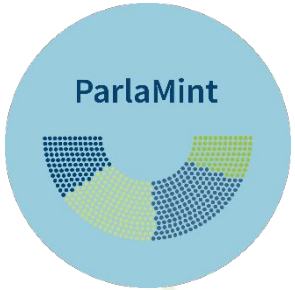
Roberts Darģis, IMCS, University of Latvia



- **Data characteristics**:
  - Everything since Latvian regain independence in 1993
  - 24M tokens until May 2020
  - Metadata: Speaker name, gender, role, age and date
  - Licence: public domain
- **Linguistic processing** – available NLP tools:
  - The current version of morphological tagger achieves 95.1% accuracy for full morphological tag and lemmas and 97.1%
  - NER – 82.6% F1-score for nine entity categories
  - UD – 89.9% Labeled Attachment Score
  - ?: Experimental PropBank/FrameNet parser
  - ?: Experimental keyword-based topic parser

Danish Parliament Corpus

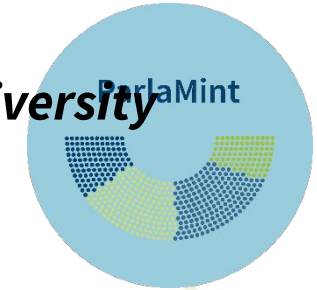
Hansen, Jongejan, Navarretta – University of Copenhagen



- **Data characteristics**:
  - 2015–2020 (Hansards), size in tokens: not known yet
  - Availability: 2015-17 already available with some metadata, 2018-2020 is being downloaded
  - Licence: public domain
  - The data is in XML, but must be converted to ParlaMint format.
- **Linguistic processing**:
  - Text Tonsorium tools under CLARIN-DK
  - PoS, lemma, syntax (NE?)
- **The method:** Add missing metadata, apply NLP tools to Hansards, mark 2019/2020 data with COVID-19 terms, convert all data to the ParlaMint format
- **Known risks:** parsing and NER not tested on larger data sets, and their performance not evaluated.

Language: Czech

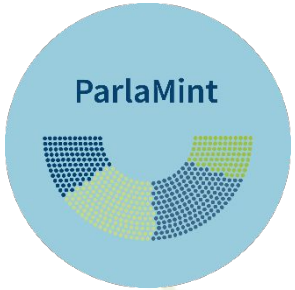
Barbora Hladká, Matyáš Kopp, Pavel Straňák, Charles University



- **Data characteristics**: https://ufal.mff.cuni.cz/parczech
  - timespan Nov 2013-Sep 2019 (pre-Covid), Oct 2019-Jan 2021 (Covid)
  - size in tokens - stenographic protocols of the Lower House of the PCR, 25M
  - availability - data already scraped, re-running the scraping in Jan 2021
    - metadata available but not collected yet
  - licence - Public Domain Dedication (CC Zero)
  - anything specific about the data and the conversion process
    - divide the speeches according to the topics
    - keep the urls to the original sources (pages, speeches) and audio (to do alignment in the future)
    - keep the urls inside speeches (voting, parliament prints)
    - convert ParlaMint format into the TEI-like format used by TEITOK
- **Linguistic processing**:
  - UDPipe 2, NameTag 2, i.e. all layers available; NameTag 2 tagset richer than the one used in ParlaMint
- **The method** 1. scraping, 2. converting, 3. annotating, 4. including the metadata

Language: Romanian

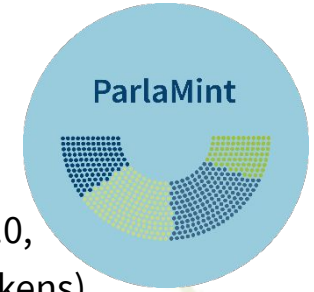
Petru Rebeja, Alexandru Ioan Cuza University of Iași
Mădălina Chitez, West University of Timișoara



- **Data characteristics**:
  - timespan: 2015-2020
  - size in tokens: unknown, stats will be updated on project page
  - availability: still to be crawled; licence: public domain
  - the data from the senate is not interlinked (PDF format that only mentions the speaker name)
- **Linguistic processing**:
  - framework to be used: UDPipe
  - all layers available: yes
- **The method:** get data & metadata, UD processing, Parla-CLARIN encoding
- **Known risks:** linking the speaker from the PDF with the personal page; needs additional crawling.

Language: Lithuanian

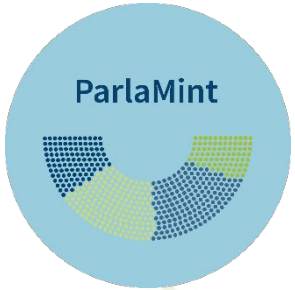
Tomas Krilavičius, Vytautas Magnus University



- **Data characteristics**:
  - timespan: 2 terms (11/2012 - 11/2020) of the Seimas; licence: CC BY 4.0,
  - size in tokens: ~1,000 of sittings during the period (each 11–35,000 tokens),
  - availability: data for 2020 still have to be collected; metadata will have to be collected,
  - data has to be parsed in order to identify speeches/speakers; large part of metadata is available in XML (some will be keyed-in), will have to be merged with speech data.
- **Linguistic processing**:
  - framework to be used: UDPipe, TBD.
  - all layers available: lemma, POS, syntax (?)
- **The method**: download and pre-process speech data, download metadata and add missing parts, apply linguistic processing, convert to the ParlaMint format.
- **Known risks**: (1) extracting speaker data and linking it to speech (record) data, (2) converting data/metadata to the ParlaMint format, (3) adding a syntax layer could be problematic due to Lithuanian parser initial maturity level.

Language: Dutch

Ruben van Heusden, Jaap Kamps, Maarten Marx
University of Amsterdam & INT

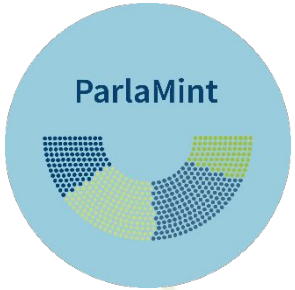


- **Data characteristics**:
  - **timespan?** *2015–2020 (if time permits 1995–2015 will also be included)*
  - **size in tokens?** *for 2015–2020 roughly (40M tokens)*
  - **availability?** *Publicly available, scraping tools ready* **licence?** *Public domain*
  - **specifics?** *Mostly available in XML / some in PDF*
- **Linguistic processing**:
  - Alpino parser (or something like UDPipe) for Dependencies
  - Experiments with coreference resolution model for Dutch
  - Manually annotating data for quality control
- **The method:** Scraping data & checking it is correct and complete, convert to TEI, run linguistic processing (INT)
- **Known risks:** Ling. processing can be time consuming, possibly inconsistent XML formats

Language: Dutch/French

Belgian federal parliament

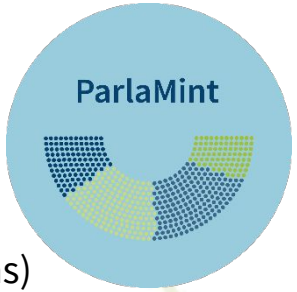
Katrien Depuydt, Jesse de Does, Griet Depoorter, Henk van der Pol, Vincent Vandeghinste; INT (Dutch Language Institute)



- **Data characteristics**:
  - **Period**: 2015-2020; time permitting also include extra reference material from 2007-2014
  - **Size**: total (including extra reference) ~39M tokens; Language: mixed French and Dutch
  - **Data harvesting**: data already collected; metadata have to be enhanced
  - **Availability**: public domain
  - **Conversion**: HTML exported from Microsoft Word → XHTML → (XSLT) ParlaMint TEI
- **Linguistic processing (in cooperation with UvA team)**:
  - Python + Flair + Spacy (backend implementation tbd by targeted evaluation: probably one of udpipe, udify or other implementation of Dozat/Manning biaffine dependency parser with more Dutch training data); fallback to Alpino also possible
  - Layers: PoS, lemma, dependency, NER
- **The method:**
  - Write XSLT's for conversion to target format; Build relevant entity database for speakers
  - Integrate dependency parsing and NER in INT processing pipeline
- **Known risks:**
  - Input data not formally structured; may have quirks → fine-tune conversion; some manual correction; accept some degree of noise
  - Metadata enhancement might take considerable effort → too bad, just do the work
  - NER may require domain adaptation/retraining → work with UvA team on this

Corpus of Hungarian National Assembly

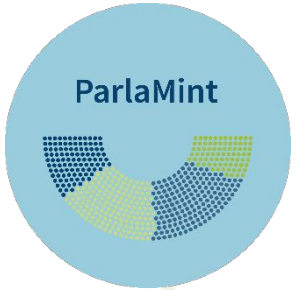
Miklós Sebők, Centre for Social Sciences, Budapest



- **Data characteristics**:
  - 6. May 2014 – 1. December 2020. (Interpellations and Urgent Questions)
  - Size in tokens: not known yet
  - 6. May 2014 – 19. Febr 2018 available with all metadata. 2018–2020 is being compiled
  - Licence: public domain
  - The data are in txt, html and csv, must be converted to ParlaMint xml format
- **Linguistic processing**:
  - Python + spacy-hungarian-models - Hungarian multi-task CNN trained on Universal Dependencies data. Assigns context-specific token vectors, Brown cluster IDs, word probabilities, POS tags, dependency parse, named entity tags and lemmata.
    (F1: tokenizer 99.89, sentencizer 96.97, NER 93.95, Acc: lemmatizer 95.51, tagger 94.81)
- **The method**: add missing texts and metadata, convert data to the ParlaMint format, apply NLP tools, quality control

Corpus of Italian Senate:

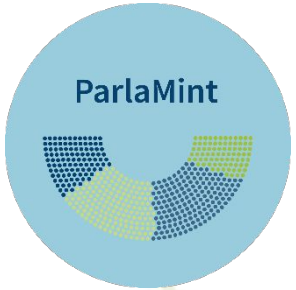
T. Agnoloni (IGSG-CNR); F. Frontini, M. Monachini, S. Montemagni, V. Quochi, G. Venturi (ILC-CNR); M. Palmirani (University of Bologna)



- **Data characteristics**:
  - Covid corpus: Oct 2019 - Dec 2020; Reference: 23 March 2018 - Sept 2019
  - Size: approx. 12 M tokens
  - Already collected availability; metadata will be collected
  - Licence: public domain
  - All the available data for the considered time span are in AKN format. Previous data are only available in HTML while future data will be in AKN
- **Linguistic processing**:
  - We plan to run tests using both UDPipe and Stanza, in order to choose the best pipeline in terms of performance for the Italian language
  - Concerning NER: we will use spaCy with possibile adaptations specific to the domain of texts
- **The method**:
  - XML Conversion from AKN to Parlamint format
  - Creation of a pipeline fine-tuned to deal with language and domain peculiarities
- **Known risks:**
  - difficulties in converting AKN variants into Parlamint format → collaboration with UNIBO who developed the AKN format and the other Parlamint teams working with AKN; difficulties in metadata retrieving → collaboration with Senate personnel who will provide missing information; performances of linguistic processing results lower than expected → domain adaptation strategies

Turkish Parliament (TBMM)

Çağrı Çöltekin, University of Tübingen

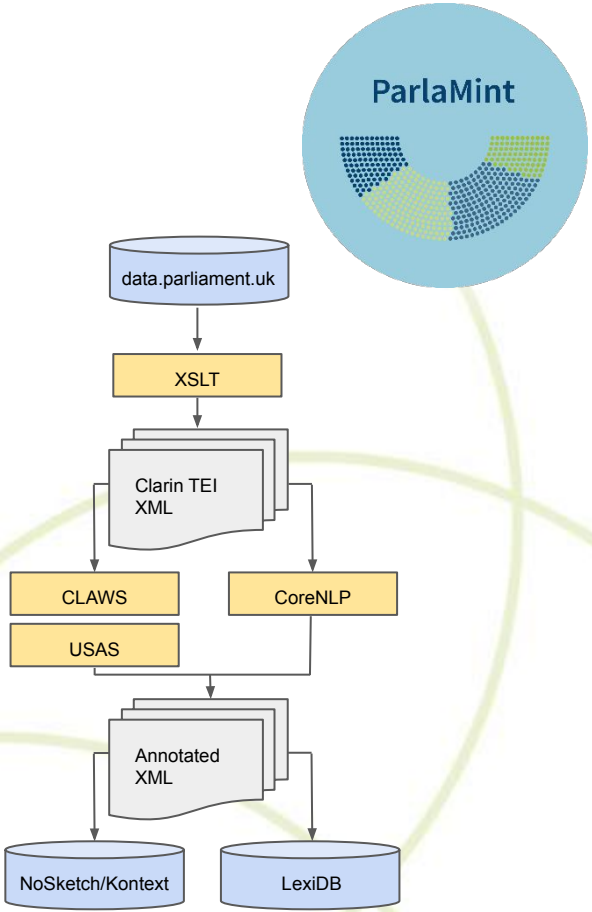


- **Data characteristics**:
  - Reference: 2015 – 2020 (50M tokens, intention to extend it back to 1920)
  - COVID-19: Nov 2019 – Mar 2021
  - Scripts for retrieving published/additional data available
  - Licence: Public domain
  - PDF/HTML → text → CoNLL-U → XML (Clarin TEI)
- **Linguistic processing**:
  - Mixture of various tools (most are not determined yet)
    - Morphology: TRMorph, Syntax: UDPipe, NER: the morphological analyzer has limited support
- **Output**
  - Currently experimenting with converting to CoNLL-U, eventually XML (Clarin TEI)
- **Potential risks:**
  - Not-so-we-structured source format.
  - Scarcity of students/help with right skill set.

UK Parliament (Hansard)

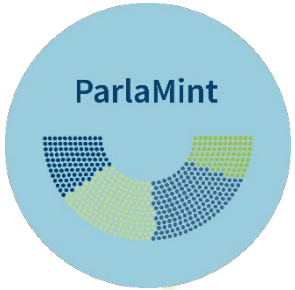
Paul Rayson, Matthew Coole, Lancaster University

- **Data characteristics**:
  - Reference: 2015 – Oct 2019 (120M tokens)
  - COVID-19: Oct 2019 – Mar 2021 (38M tokens)
  - Data gathered up to Dec 2019 [1]
  - Scripts for retrieving additional data available [2]
  - Licence: Open Parliament Licence v3.0 [3]
  - XML → XSLT → XML (Clarin TEI)
- **Linguistic processing**:
  - Lancaster Toolchain and Stanford CoreNLP [4]
    - POS: CLAWS [5]
    - Semantic tags: USAS [6]
    - UD & NER: CoreNLP
- **Output**
  - LexiDB [7], NoSketch, Kontext formats (VRT, CoNLL-U, TEI)
  - Toolchain for corpus extension



The National Assembly (France)

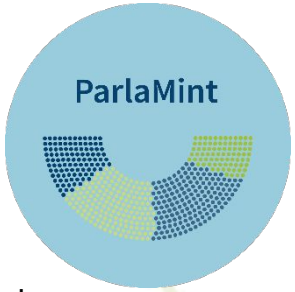
Sascha Diwersy, Giancarlo Luxardo
(Praxiling - Université Paul-Valéry Montpellier, CNRS)



- **Data characteristics**:
  - *Comptes rendus des débats en séance publique de l'Assemblée nationale*
  - Covid corpus: October 2019 - July 2020 (2 sessions)
  - Reference corpus : January 2014 - September 2019
  - not collected yet, estimated size of the reference corpus: > 50M tokens (~10M Covid corpus)
  - all data is XML-based (different versions to be converted to ParlaCLarin TEI)
  - licence: public domain
- **Linguistic processing**:
  - Implemented by a pipeline including tokenisation, sentence segmentation, lemmatisation, annotation with UD, part-of-speech and morphological features, UD dependency relations and named entity markup (PER, ORG, LOC, MISC). The processing will be carried out by means of a Python script combining an XML parser module with the Stanza NLP package.
- **The method**:
  - The XML conversion will be done with the previously TAPS demonstrated procedure developed in Python, including new features and with an XSLT post-processing.
- **Known risks:**
  - The source format has been updated since 2014 and involves different processings.
  - Planning internships, need of NLP skills (while in a linguistic lab).

Alþingi: The Icelandic Parliament

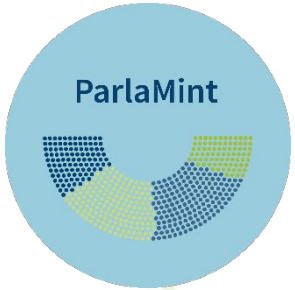
Steinþór Steingrímsson, Starkaður Barkarson, Einar Freyr Sigurðsson
The Árni Magnússon Institute for Icelandic Studies



- **Data characteristics**:
  - Covid corpus: October 2019 - July 2020; possibly extended to December 2020
  - Reference corpus : 1938 – September 2019
  - Licence: CC BY 4.0, availability: data for 2020 and some metadata is being collected
  - size in tokens: ~220 million
  - we already built a parliamentary corpus, published earlier this year, which we are now extending and adding text and metadata to
- **Linguistic processing**:
  - Sentence segmentation, tokenisation, morphosyntactic POS-tagging, lemmatisation, annotation with UD and named entity markup.
  - Various tools used, Icelandic rule-based tokenizer for sentence segmentation and tokenisation. BiLSTM tagger for POS-tagging. Lemmatiser using a morphological lexicon for lemmatisation. UDPipe and/or a BiLSTM tagger for UD. A recent Icelandic model for NER.
- **The method**: amend our scripts for converting the Althingi data to conform with the ParlaMint format. Add missing metadata.
- **Known risks**: (1) We haven't annotated with UD before. (2) We haven't used the new NER model before.

Case Study

Ruben Ros, Luxembourg Centre for Contemporary & Digital History



- **Topic:**
  - Contested Expertise in COVID-19 Debates
  - Expertise between technocratic adoption and conspiratory rejection
- **Hypothesis:**
  - Relatively high overall levels of scientific knowledge in debates
  - The introduction of ‘counter’-knowledge by opposition parties
  - Initial ‘rally around the flag’-effects, followed by gradual politicization
- **Methods:**
  - Detection of references to scientific knowledge (reports, institutions) → NER (ORG)
  - Detection of references to other countries as policy models → NER (LOC)
  - Trend analysis of specific terms (‘expertise’, ‘science’)
  - Polarity analysis of text surrounding keywords
  - …