Minutes of the ParlaMint meeting

20 January 2021, 13.00–14.00 CEST (online)

Participants: Filip Dobranić, Tomaž Erjavec, Nikola Ljubesić, Maciej Ogrodniczuk, Petya Osenova, Kiril Simov

1. Status of 'new' languages

- ➔ Maciej will send everyone information about the plenary meeting planned for the next week. We will ask everyone to inform us about the status of their work.
- ➔ We will try to maintain the status table to keep track of progress for individual languages.
- ➔ Several people already contacted Tomaž:
  - ➔ Czech: put some content on GitHub
  - ➔ Dutch, Hungarian: got in touch with Tomaž, sent him a preliminary data
  - ➔ Icelandic: also contacted Tomaž

2. ParlaMeter

The website is looking better and better. Still, there are several issues to discuss:

- ➔ country-level issues (e.g. sorting, translations etc.)
- ➔ general issues (do we plan to have the same information about MPs for all countries or we want to keep it separate? can we? how do we fit another 14 languages in the interface etc.)
- ➔ where do we stop in our interface changes?

We decided that:

- ➔ Petya will send Filip the list of all parties from opposition and coalition + photos of MPs.
- ➔ Filip will try to fix the remaining known interface issues by Monday (Jan 25).
- ➔ Filip will set up a Trello project for ParlaMint in Parlameter and register older issues.
- ➔ Everyone will test and register new issues.
- ➔ We will all think about interface changes / how the interface should function and exchange mails about that or have a separate meeting about ParlaMeter.

3. Showcases

- ➔ DH showcase:

- ➔ Most participants already filled in Ruben's spreadsheet, only Turkey, Latvia and Lithuania still missing. Ruben will send them the reminder this week.
- ➔ Ruben created a GitHub repository for the use case. After data preparation he envisages 3 analytical steps; he will put them in the project and create deadlines for himself.

➔ Voices of Parliament showcase:

- ➔ We discussed it with Darja last year that we intended to start working on selected sections of Voices of Parliament in order to contribute to this guide from a multilingual aspect and that we would like to start at some point with such contributions for Bulgarian, Polish, Croatian and Slovene.
- ➔ Petya will contact Darja about this issue.

4. ParlaSpeech

➔ Elizabeth Nielsen, a PhD student at the University of Edinburgh asked us about audio recordings of various parliamentary corpora.

➔ Several people replied: María Calzada Pérez (Spain), Roberts Darģis (Latvia), Matyáš Kopp (Czech Republic), Nikola, Kiril, Costanza Navarretta (Denmark), Sebők Miklós (Hungary).

➔ Nikola came up with the idea of ParlaSpeech, a ParlaMint follow-up project that would add at least the audio layer to the corpora we produce.

➔ Several issues still remain to discuss:

- ➔ Can we use CLARIN funds for that?
- ➔ Which level of alignment we target?
- ➔ How do we tackle this issue?

➔ Here is a preliminary document listing our ideas.

5. The Swedish seminar about mining parliamentary data

➔ We have been approached by Fredrik Norén, a researcher at Humlab – the DH center at Umeå University in Sweden who, together with Jo Guldi (Southern Methodist University) and Pelle Snickars (Umeå University) is arranging a online seminar series on the topic of text mining parliamentary data, open to the international scholarly community.

➔ So far they have had three meetings and with up to 30 attending participants, from different disciplines and countries. Among our guest speakers they have had Luke Blaxill (University of Cambridge), Kaspar von Beelen's (The Alan Turing Institute), Eero Hyvönen's (University of Helsinki) and Paul Nulty (University College Dublin).

➔ During spring 2021 they will continue on with the series, focusing on a couple of thematic oriented panel discussions. The first will be on practices of marking up parliamentary data, sometime in March, related to different things as curation and metadata. The idea is to have a moderated conversation about this theme with some selected representatives from institutions (responsible for digitizing parliamentary data), researchers (studying the data), and technicians (curating and marking up the data).

- ➔ They invited Maciej to participate in the panel about processing historical Polish parliamentary data on March 18.
- ➔ ParlaMint topics will be presented later.

6. Next telco and other issues

- ➔ Next (plenary) meeting: Jan 27, 13:00 CET.
- ➔ Main points of the agenda:
  - ➔ Status of new languages.
  - ➔ Demonstration of ParlaMint in Parlameter (Filip).
  - ➔ Notes on the showcase (Ruben).

Minutes of the ParlaMint plenary meeting

27 January 2021, 13.00–14.00 CEST (online)

1. Preparation of language corpora

- ➔ Everyone informed us about the status of their work and potential problems.
- ➔ We will maintain the status table to keep track of the progress for individual languages and get a snapshot of it at the end of each remaining month.

Language	Reporting person	Status
Belgian Dutch/ French	Jesse de Does	Corpus downloaded, conversion in progress, mostly done, metadata needs some cleaning up. Will finish soon and start working on linguistic annotation. Data is already TEI P5-valid but not yet Parla-CLARIN-valid. Next version should be ready on February 1.
Czech	Barbora Hladká	Corpus downloaded and converted, several files submitted to GitHub, meta-anotations almost done (there are some anomalies in persons), Annotation pipeline done (TODO: NER 46 to 4 cats transformation, UD syntax – fixing `ana` attributes, validation issues).
Danish	Costanza Navarretta	Corpus downloaded, comments received from Tomaž, meta annotations done. Workflow for processing linguistic annotations prepared (TODO: running the linguistic annotations).
Dutch	Ruben van Heusden	Corpus downloaded, plain text version corpus converted, started annotating and fixing bugs.
English	Paul Rayson, Matthew Coole	Corpus downloaded (http://ucrel-hansard-l.lancs.ac.uk/hansard/parlamint/), annotation pipeline built, annotation underway.

French	Giancarlo Luxardo	Corpus and metadata downloaded (assemblee-nationale.fr), TEI encoding for utterances and incidents in progress. Samples sent to Tomaž but some files are not validating due to excessive attributes. Incidents marked with italics so they will not be easy to convert. Metadata not included yet, linguistic annotation pipeline to be implemented.
Hungarian	Orsolya Ring	Corpus and metadata downloaded, TEI encoding is almost done, linguistic annotation pipeline built.
Icelandic	Starkaður Barkarson	Corpus downloaded, meta annotations in progress, samples sent to Tomaž.
Italian	Tommaso Agnoloni	Corpus and metadata downloaded (sources: senato.it and dati.senato.it), TEI encoding in advanced progress (TODO: finalize corpus encoding, integrate linguistic annotation pipeline). Samples not sent yet – planned for the beginning of February.
Latvian	Roberts Dargis	Data downloaded, samples to be sent in the beginning of February.
Lithuanian	Vaidas Morkevičius	Data collected, conversion started, samples to be sent in the beginning of February. Linguistic annotation attempted.
Romanian	Madalina Chitez	Corpus downloaded, annotations in progress. Raw data for the proceedings was obtained and the team is working on creating pilot annotations and automation of the process.
Spanish	María Calzada Pérez	Corpus data (COVID part only) downloaded in Spanish format, meta annotations in progress.
Turkish	Çağrı Çöltekin	Çağrı was not present but sent an e-mail with his status saying that he is able to extract almost all linguistic data from the transcripts since 2008. He still needs to map speakers to parties but data processing is almost complete. On the annotation side, tokenization/sentence segmentation is done. Initial experiments with UD parsing show LAS barely above 50% on a small test set so if he is unable to improve it, he will accept it and continue. He is currently working with CoNLL-U-like files

		with meta information stored in the sentence comments. He will attempt to convert it to XML at the end of February. He did not send samples to Tomaž yet but can send it very soon.

2. Q&A session

- ➔ Jesse: How should we include different taxonomies in the header?
- ➔ Tomaž: We should adapt the Slovenian taxonomy now to avoid complicating things. Eventually we would like to have a nice European view of the parliamentary system.

- ➔ Jesse: Is there a fix for XML validation issues?
- ➔ Tomaž: Indeed, the RelexNG validation does not check that IDs that are referred to actually exist, but there is the check-links.xsl script that does that. It has an explanation how to use it at the start of the script. I hope to also write another XSLT to check for more things. Another thing I might do is to run my script for converting data to the vertical format which will also validate the metadata.

- ➔ Giancarlo: Do we all have access to GitHub?
- ➔ Tomaž: Not everyone has access yet but if you feel confident enough in using GitHub, write to Tomaž and he will grant access. But you can also send me samples via email. Note also that the GitHub repo is meant for samples only - once you have the complete corpus, pls. put it on Dropbox or somewhere else on the Web, and I will download it.

- ➔ Giancarlo: What do you do when you have attributes on utterances that are not allowed by ParlaMint / TEI?
- ➔ Tomaž: I can add new attributes or elements, but only those that are valid in TEI. We would have to see on a case-by-case basis what information you have, and how it could be encoded in ParlaMint.

- ➔ Giancarlo: How to process incidents when they are just presented in italic in the source?
- ➔ Tomaž: The minimum is to annotate such pieces of text that represent transcriber comments as `<note>` elements. To use `<incident>`, `<vocal>`, `<kinesic>` you need to write a heuristic that classifies them according to their content (that is what we did for V1 corpora).

3. Demonstration of ParlaMint in Parlameter

- ➔ Filip showed a live demonstration of ParlaMint Parlameter (https://parlamint.parlameter.org/).
- ➔ When more data is uploaded we would like to ask everyone to test it.
- ➔ Interface localisations are not obligatory but will be welcome. It can be done online:
  - ➔ the page: https://crowdin.com/project/parlasite

➔ individual cards: https://crowdin.com/project/parlanode

4. Notes on the showcase

➔ Ruben thanked everyone for filling in the spreadsheet.
➔ Ruben created a repository with the showcase overview: https://github.com/rubenros1795/ParlaMintCase

5. Next telco and other issues

➔ Next plenary meeting: Feb 24, 13:00 CET.

Minutes of the ParlaMint plenary meeting

24 February 2021, 13.00–14.00 CEST (online)

1. Dissemination activities

- ➔ Feb 17: Maciej presented ParlaMint at the Dan Cristea's course (Natural Language Engineering Techniques)
- ➔ Feb 23: Petya, Maria and Maciej presented ParlaMint at INTELE Webinar
- ➔ ParlaMint slides from various activities are free to take from the Events folder. Please add your slides there if you are talking about ParlaMint.

2. Status of language corpora preparation

- ➔ Everyone informed us about the status of their work and potential problems.
- ➔ The status table now has several columns showing the progress for individual tasks.
- ➔ Documentation template will be sent to all ParlaMinters but you can already see what we prepared for Phase 1.
- ➔ A few comments from Tomaž:
  - ➔ you should create your account in GitHub – even if you do not plan to add issues, you will still be able to track what is happening
  - ➔ all TEI headers will be available to let us compare our metadata etc.
  - ➔ when you have your corpus ready, you can send it in advance to check whether the data can be converted to the format of concordancers for additional validation
  - ➔ Tomaž already started working on the TEI header for the complete corpus; it will be also put in GitHub for everyone to comment.

3. Next telco and other issues

- ➔ Next plenary meeting: Mar 24, 13:00 CET.

Minutes of the ParlaMint meeting

10 March 2021, 13.00–14.00 CEST (online)

Participants: Tomaž Erjavec, Nikola Ljubesić, Maciej Ogrodniczuk, Petya Osenova

1. Status of language corpora preparation

- ➔ Czech and Icelandic corpora are finished. Turkish seems the least advanced – I will mail Cagri and Tomaz will mail the others.
- ➔ The status table now has several columns showing the progress for individual tasks.
- ➔ María Calzada Pérez managed to produce a corpus from the 2016-2020 period (2015 is a different term and the HTML is slightly different). We hope UD and NER annotation is ready until the end of March.
- ➔ We prepared a documentation template based on what we prepared for Phase 1. We will put our descriptions until the end of the week. Petya will do the first round and we will check.

2. Articles/presentation/dissemination

- ➔ Feb 17: presented ParlaMint at the Dan Cristea's course (Natural Language Engineering Techniques).
- ➔ Feb 23: Petya, Maria and Maciej presented ParlaMint at INTELE Webinar.
- ➔ Mar 18: Maciej will be taking part in the Text Mining Parliamentary Data Seminar dedicated to a panel discussion on "Practices of parliament", related to experiences and challenges in working with marking up parliamentary data, such as curation and metadata: https://www.umu.se/en/events/practices-of-parliament_10147805/.
- ➔ May 19–28: Slovenian ParlaMint team (Tomaž and Nikola) supported by Ajda Pretnar from Orange will be offering a group at the Helsinki Digital Humanities Hackathon (online): https://www.helsinki.fi/en/helsinki-centre-for-digital-humanities/helsinki-digital-humanities-hackathon-2021-dhh21. The registration has just opened.
- ➔ Jun 2–4: Ruben sent an abstract for DH Benelux – he is thinking about presenting a case there.
- ➔ Let's have a short paper/position paper/poster about ParlaMint at CLARIN Annual Conference 2021 (Sep 27–29). Abstract to be delivered by 14 April 2021. Read the full 14 April 2021. Authors: everyone who participate
- ➔ We are thinking about writing a LRE paper about ParlaMint. One author per language probably.

3. ParlaMeter

- ➔ Todo board: https://pad.muki.fyi/kanban/#/2/kanban/edit/+RDJ1WwqEqCO2kiUu9Nbyyu2/ was set up by Filip. We should add feature requests and bugs into the *To do* column. Once we enter a name we can click on the pencil to add more information about what is happening.

4. Showcases

- ➔ DH showcase:
  - ➔ Ruben developed several scripts for frequency analysis, PMI analysis, training of word2vec; should be finished this week.
  - ➔ He will also take pre-trained word embeddings.
  - ➔ A short preliminary report will be put in the GitHub repository.
  - ➔ The pipeline on sentiment analysis on different units using Polyglot is already finished.
- ➔ Voices of Parliament showcase:
  - ➔ Petya prepared an abstract for the special issue in Gender and Language in Eastern and Central Europe extending Darja's tutorial to our 4 languages
  - ➔ We would try to do the following: representation of women in the Slovenian Parliament, issues addressed by women, topics related to women.

5. ParlaSpeech

- ➔ Before we can proceed with ParlaSpeech, we need a proof of concept.
- ➔ There has been one development in February regarding the potential of aligning ParlaMint data with speech, which Nikola's collaboration with some Serbian colleagues on a Serbian spoken corpus (very low audio quality), for which we used, inter alia, the approach by Plüss et al. who use a commercial TTS and perform alignment over the two texts. Roughly 50% of the segments were well aligned, which is a very good result given the bad audio quality, but also spontaneous, informal speech. Nikola has hired Michel Plüss to perform his method over their data and obtained the necessary code base from him as well.
- ➔ Now Nikola is looking for someone who would perform a proof-of-concept with the available code base over a sample of Croatian data.

6. Next telco and other issues

- ➔ Next meeting: Mar 17, 13:00 CET?

Minutes of the ParlaMint plenary meeting

24 March 2021, 13.00–14.00 CEST (online)

1. Status of language corpora preparation

- ➔ Everyone informed us about the status of their work and potential problems.
- ➔ We are tracking the status of corpus preparation (please clear the comments so that they reflect the current situation) and gathering documentation descriptions.
- ➔ Tomaž will put all the data somewhere to let us download it even before it gets official.
- ➔ Ruben's plans for the showcase are getting more and more concrete, he will get in touch with us soon.

2. Plans for the next months, next telco and other issues

- ➔ The call text requires the corpora to be delivered until the end of March but our contract sets the final deadline at the end of May.
- ➔ The consequence is that the corpora delivered until the end of March will become part of Ruben's use case while the others will continue with the corpora delivery and tuning.
- ➔ We plan to have at least three more plenary meetings: at the end of March, April and May.
- ➔ We will consider publication at the CLARIN Conference and also discuss it next week.
- ➔ Before we have a formal publication we will have the repository handle to cite.
- ➔ Francesca Frontini invited us to present our showcase at CLARIN Cafe in June.
- ➔ Next plenary meeting: Mar 31, 13:00 CET.

Minutes of the ParlaMint plenary meeting

31 March 2021, 13.00–14.00 CEST (online)

1. Status of language corpora preparation

- ➔ BE, CZ, DK, IS and NL are completed.
- ➔ EN: some minor things found by Tomaž, should be fixed after Easter.
- ➔ FR: still a lot of work to do, little progress from last week. Still a few weeks to go.
- ➔ HU: Tomaž received a bunch of versions, several errors corrected, level 2 CONLL validation fails (roots and syntactic heads). We'll see whether we can fix it – if not – we will keep it.
- ➔ IT: several examples sent to Tomaž, linguistic annotation is ready and confirmed. The corpus will be ready at the end of the week or just after Easter.
- ➔ LV: the corpus is almost ready,
- ➔ LT: nothing changed in the last days.
- ➔ RO: no root file yet, status unclear.
- ➔ ES: annotation complete, still to be validated. Tomaž will take a look.
- ➔ TR: another round of sample files sent to Tomaž just before the meeting; some version will be produced today, there might still be problems with identification of speakers because of spelling errors in the official transcript.

2. Acceptance and payment

- ➔ Tomaz has the final say about the acceptance.
- ➔ As for the payments, we have the pre-financing, interim payment (if requested) and the final payment. Quirijn told us that some partners received the first two, some only the pre-financing. After May CLARIN will settle the balance and everyone gets the final payment.

3. Publication plans

- ➔ CLARIN Annual Conference 2021:
  - ➔ We plan to submit an abstract about "ParlaMint: Comparable Corpora of European Parliamentary Data" (until April 14). Petya volunteered to help with the first draft. Please think about co-authors.
  - ➔ We also plan to have a demo of our achievements.
  - ➔ We will probably skip the Bazaar since it did not bring much attention last year.
- ➔ Language Resources and Evaluation:
  - ➔ We also plan to publish a more extensive article (most likely project report type).

- ➔ We will wait until summer to write it; should be done by August at the latest.
- ➔ Costanza: we would probably need an evaluation section. Tomaž: please think about ideas for evaluation; we can e.g. take samples that we have in GitHub and manually evaluate them manually.

4. Dissemination activities

- ➔ Feb 17: Maciej presented ParlaMint at the Dan Cristea's course (Natural Language Engineering Techniques)
- ➔ Feb 23: Petya, Maria and Maciej presented ParlaMint at INTELE Webinar
- ➔ ParlaMint slides from various activities are free to take from the Events folder. Please add your slides there if you are talking about ParlaMint.

5. Q&A session

- ➔ Roberts: are samples consistent? Tomaž: the idea is to get a small but varied sample. But there is no randomness in the process.
- ➔ Costanza: how about storing extra information from our original corpora, e.g. subjects assigned to speeches? Tomaž: before ParlaMint we had Parla-CLARIN recommendation where you could write anything that's TEI-compatible. But now we have specialized schemata, with very reduced set of information and converters that work with what we have. I'm against making the schema more permissive. But we have ana attributes which work as pointers and we can store links there.
- ➔ Maria: do we plan to extend ParlaMint? Tomaž: it would be silly not to extend it – but it can be extended in many directions, e.g. by putting speech into ParlaMint, automatically translate all the corpora into English to make it more comparative, extend it with more languages or extend the existing corpora with more data.
- ➔ Cagri: even if we don't manage to get the extension, do we plan to continue updating the corpora. Tomaž: it's not very difficult to validate the new updated version but releasing a new version is tricky because if we update one language, the whole corpus gets a new release. We need to discuss it.

6. Next telco and other issues

- ➔ Next plenary meeting: Apr 28, 13:00 CET.

Minutes of the ParlaMint plenary meeting

28 April 2021, 13.00–14.00 CEST (online)

1. Status of language corpora preparation

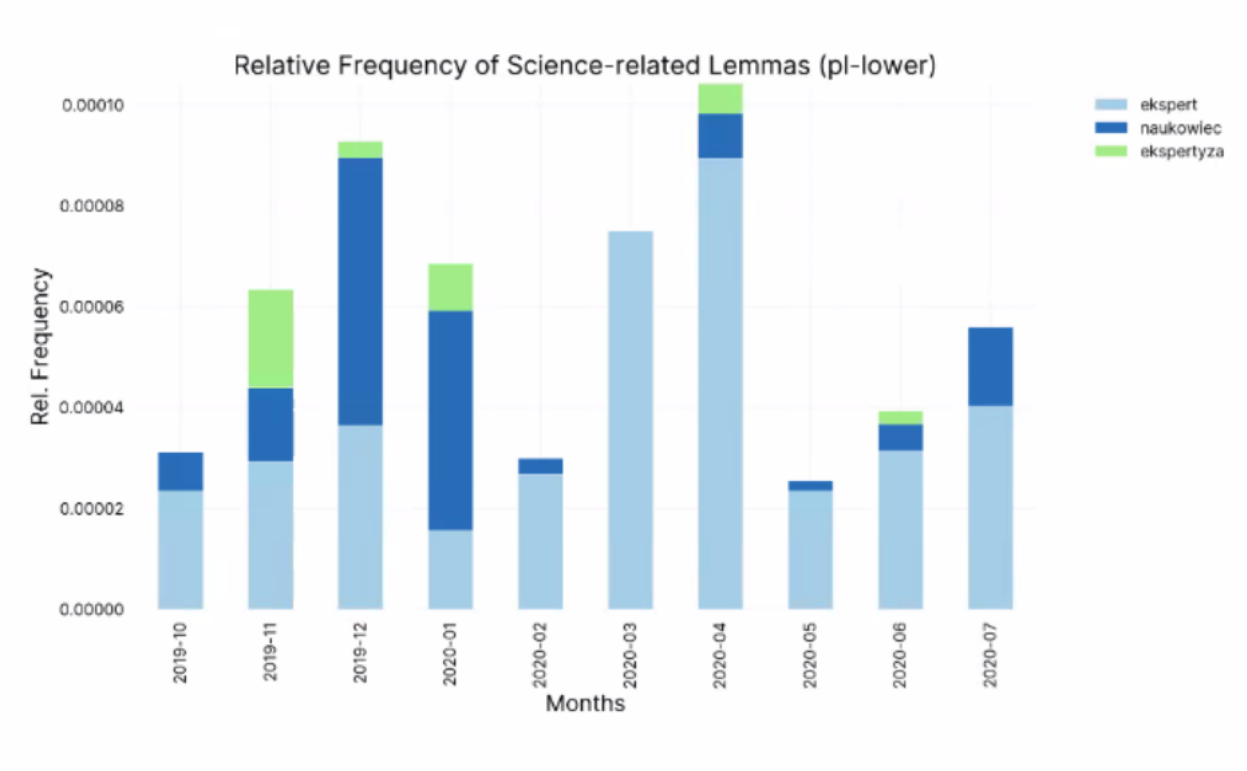
- ➔ All languages are more or less complete except for Romanian – Maciej will contact Petru.
- ➔ You can invoice CLARIN ERIC independently on taking part in the demonstration.

2. CLARIN Annual Conference paper on ParlaMint

- ➔ 4-page paper was submitted today.
- ➔ Here is the link to Overleaf.
- ➔ If it's accepted, we will be given another page of content and we will be able to decide whether we publish there or go for a journal paper.

3. The case study

- ➔ We are in the process of adding coalition/opposition metadata to Ruben's spreadsheet, with new rows for every coalition/opposition, split by their start and end dates. Tomaž has made another file with all the party names. They have also prepared a short guide for the encoding.
- ➔ The case study will concentrate on the large role for expertise in corona policy measures by investigating changing language use around terms such as "experts", "expertise", "science" etc. looking at metadata, relative frequency of specific terms, language diversity (normalised type-token ratio), polarity measurements using different lexicons, e.g.



- ➔ To do: TF-IDF vectors for detecting concept drift of expertise-related terms
- ➔ Challenges: data scarcity in specific months, multilingual resources (sentiment/lexicon/embedding models) and coalition–opposition metadata.

4. Next telco and other issues

- ➔ You are welcome to take part in the online Conference Day on 'Parliament & Media: A continuously evolving love-hate affair?' on May 20th, which is part of this year's Entangled Media Histories (EMHIS) Forum. Zoom links will be sent to those who, register their interest by April 30th by emailing betto.van_waarden@kom.lu.se.
- ➔ We will contact everyone about the CLARIN Cafe planned for July – apart from showing the use case we would like to let everyone show what they managed to do in ParlaMint.
- ➔ Last ParlaMint plenary meeting: May 26, 13:00 CET. This is also the date when the decision on the COST Action Parlant results comes to us. So bring your own cup to celebrate either the success of both the COST Action and ParlaMint or just ParlaMint.

Minutes of the ParlaMint plenary meeting

26 May 2021, 13.00–14.00 CEST (online)

1. Status of language corpora preparation

- ➔ ParlaMint 2.0 is out there! There will be another official release on May 31 (or a few days later).
- ➔ Most corpora have this coalition/opposition information (except for the Hungarian).
- ➔ French: plain text version is ready, there was some delay related to linguistic annotation. If Tomaz gets the corpus until June 4, we will include the corpus.
- ➔ Romanian: the first version will be ready this week except for NE annotation (RRT model of the UDPipe does not have a NER module). Petru would like to submit the corpus anyway because Romanian is one of the under-resourced languages.

2. The showcases

- ➔ DH showcase: Ruben is still improving it and working on the interpretation. CLARIN Cafe sets a useful deadline for the showcase.
- ➔ DH Hackathon: Ruben is also active in the Parlamint task (https://www2.helsinki.fi/en/helsinki-centre-for-digital-humanities/dhh21-hackathon/dhh21-themes#section-10475). The presentation summing up the Hackathon will happen on May 28, Ruben will send the link.
- ➔ Parlameter showcase: Czech and English data are being indexed, others to follow.
- ➔ Miguel Pieters analysis: we watched his pre-recorded presentation (https://www.youtube.com/watch?v=QCrGl8wcFmw) and discussed his findings.

3. Other issues

- ➔ CLARIN conference paper: Lithuanian and Belgian Dutch/French groups: please send the co-authors. Miguel's graphs will be used in the paper.
- ➔ Another ParlaMint paper: first version to be drafted by Tomaz in the summer period.
- ➔ CLARIN Cafe: we plan to discuss it with CLARIN on May 31.
- ➔ COST Parlant: unfortunately "our proposal was not retained for funding". We received 41/50 points (first time: 41/50, second time: 45/50 with more or less the same text). I will send everyone the evaluation results.
- ➔ ParlaMint follow up: stay tuned for CLARIN ERIC-related activities and please think about using our parliamentary network in your new proposals.

This was the last ParlaMint plenary meeting! Thank you for taking part in this successful project. Cheers to all of you – and see you around in the near future!