A new ParlaMint corpus for Hungarian
30m tokens of annotated parliamentary data

Noémi Ligeti-Nagy[1], Réka Dodé[1], Kinga Jelencsik-Mátyus[1], Zsófia Varga[1,2],
Enikő Héja[1], Tamás Váradi[1]

[1]Hungarian Research Centre for Linguistics
H-1068 Budapest, Benczúr u. 33.
{vezeteknev.keresztnev}@nytud.hu
[2]University College London
2 Wakefield St, London, WC1N 1PF, UK
zsofia.varga.20@ucl.ac.uk

Abstract. Parliamentary data constitute a rich source for research for academic fields in the social sciences and humanities (SSH). To facilitate such research, comparable, high-quality parliamentary corpora are needed. The ParlaMint project, funded by CLARIN-ERIC, aims to create such corpora for languages spoken in European parliaments in a shared framework consisting of uniform encoding schemas, metadata structure, and Universal Dependencies-type linguistic annotation. The newly built Hungarian corpus of ParlaMint II focuses on the minutes of the Hungarian National Assembly between May 2014 and June 2022 and can be considered a major improvement from the Hungarian corpus of ParlaMint I. It has a wider time frame, more extensive metadata on speakers and their affiliations, and more sophisticated linguistic analysis than what was available in ParlaMint I. The Hungarian ParlaMint II corpus is openly available, just as all the ParlaMint corpora for other languages. Some potential applications of ParlaMint corpora in SSH research are also discussed.

Keywords: language data, parliamentary data, corpus, linguistic annotation, metadata, TEI

1 Introduction

In the fields of SSH there is an ever-growing need for language data as a source of research. Parliamentary debates as a special kind of spoken language data are especially valuable as they provide opportunities for numerous branches of SSH (e.g., political science, sociology, history, discourse analysis or sociolinguistics) to study various aspects of the transcripts of the verbal exchanges. At the same time, parliamentary language data are also valuable as they are not subject to copyright, and thus can be easily used in building datasets (Erjavec et al., 2022).

In 2019, CLARIN (clarin.eu) – the pan-European research infrastructure funded to support researchers in SSH – launched ParlaMint, a large-scale project with the aim of building corpora of parliamentary texts for the participating languages. In the first phase of the project, ParlaMint I, corpora for parliamentary debates of 17 national parliaments were created, spanning almost 500 m words in total. The project's main objective was to produce comparable, highly useful language resources of contemporary parliamentary data for SSH research and education. The language data in each corpus of ParlaMint I were analysed using the Universal Dependencies framework. As such, the 17 corpora of ParlaMint I provide comparable datasets, following the same encoding schema, with uniform linguistic annotation and rich metadata. These corpora are freely available at CLARIN.SI repository[1] to download, and at the NoSketch Engine[2] and KonText[3] platforms for analysis (Erjavec et al., 2021).

In December 2021 CLARIN ERIC decided to create ParlaMint II. As part of this second phase, new languages (and regional varieties) were added to the project, and a decision was made to extend participating languages' corpora up to July 2022 or further. In this new phase, the schema is updated and the validation processes are informed by the experiences of the first phase. ParlaMint II aims to include parliamentary texts in 31 languages. The project also incorporates data from regional parliaments such as Basque country or Galicia.

In the present paper, we introduce the Hungarian corpus of ParlaMint II created by the Language Technology Research Group of the Hungarian Research Centre for Linguistics. First, we review ParlaMint I and its Hungarian corpus. Then, we discuss the objectives of ParlaMint II and the details of the completed corpus, focusing on (meta)data collection and encoding and the linguistic annotation process. Lastly, we overview some potential applications of the new corpus.

2 ParlaMint I

Although there are several datasets of parliamentary data in numerous languages, they follow different encoding schemas and contain different kinds of information, making it difficult to conduct comparative studies. Following many preparatory events organised by CLARIN (see Erjavec et al., 2022), the ParlaMint project started in 2019 with the aim of creating a framework for parliamentary corpora for the four participating countries (Bulgaria, Croatia, Poland, Slovenia). Shortly later, 13 additional members joined the project.

2.1 Encoding the corpus

All 17 corpora in ParlaMint I have the same metadata and data structure, were created using the same encoding schema and follow the same type and degree of linguistic annotation. This framework was initially laid out in the Parla-CLARIN

[1] https://www.clarin.si/repository/xmlui/handle/11356/1432
[2] http://www.clarin.si/noske/
[3] https://www.clarin.si/kontext/corpora/corplist

recommendations (Erjavec and Pančur, 2019) and is based on the Text Encoding Initiative Guidelines (TEI, 2017). To make the corpora of the project interoperable, RelaxNG schemas were created and XSLT scripts were added to help the validation of the corpora. (Erjavec et al., 2022) Prior to the launch of the ParlaMint project, the 200-million-word siParl corpus was encoded using this framework, and all the lessons learned were later built into the newer version of Parla-CLARIN.

As a result, ParlaMint I contains transcriptions of speeches made in 17 European national parliaments running to half a billion words. The final corpora contain metadata of about 11 thousand speakers and are linguistically annotated following the Universal Dependencies formalism.[4]

2.2 The Hungarian corpus in ParlaMint I

The Hungarian parliamentary corpus in ParlaMint I was assembled by the Centre for Social Sciences (CSS), Hungary (Üveges and Ring, 2022). It contained two types of speeches (interpellations and urgent questions) from the plenary sessions of the National Assembly of Hungary, terms 7 and 8 (May 2014 – December 2020), amounting to 870,000 tokens in total. The authors used three separate tools for linguistic annotation and a fourth Java program to merge the output of the three tools. For morphosyntactic annotation, an old version of the magyarlanc linguistic toolkit was used (Zsibrita et al., 2013). Syntactic analysis was performed by UDPipe (Straka and Straková, 2017). Named entity recognition was done by a tool created by the MTA-SZTE Research Group on Artificial Intelligence (Szarvas et al., 2006). This approach has several drawbacks as the output of each module must be adjusted to provide a suitable input for the next module. This of course also means that additional bugs could appear.

For metadata, the political party, gender and date of birth were gathered for the 194 speakers of the corpus. Given the type of speeches collected, all speeches have an MP as speaker, and none of them are given by the chair of the session.[5]

3 ParlaMint II

ParlaMint II, while also upgrading the XML schema and validation, aims to feature new languages and extend the existing corpora to cover data to at least mid-2022. Originally, 30 languages were planned to be featured in this second phase of ParlaMint II, with Ukrainian being added to the project later on. Ukrainian parliamentary speeches are being processed at the moment.

[4] Samples of the corpora and conversion scripts are available from the project's GitHub repository (https://github.com/clarin-eric/ParlaMint), and the complete set of corpora is openly available at the platforms mentioned in the Intorduction of the present paper.

[5] The proportion of speeches given by MPs in the corpora of ParlaMint I are almost over 90%. The proportion of technical speeches varies widely between countries in PMI, although most are about half of all the speeches.

As the large number of partners led to a wide variety of parliamentary systems being featured, with different kinds of structures, positions, and speakers, encoding was very much like an iterative process both in ParlaMint I and II. In the second phase of the ParlaMint project, our institution took on the task to expand the Hungarian ParlaMint corpus both in terms of the time period covered and the range of text types covered. As a result of our work, this second ParlaMint-HU corpus now contains the minutes of the National Assembly from 2014 to 2022, comprising terms 7, 8, and the spring session of term 9 of the Third Republic of Hungary. The documents of the corpus thus range from 6 May 2014 to 14 June 2022 and contain the official textual transcriptions of 514 sitting days in this period.

The National Assembly is the unicameral legislative body of Hungary, currently consisting of 199 Members and advocates for nationalities that could not reach the threshold to elect any MPs (between 1990 and 2014 the number of MPs was 386). Currently, there are 12 advocates. Members and advocates are elected to 4-year terms, and a chair is chosen from MPs at the beginning of each parliamentary term. The National Assembly has 25 standing committees dedicated to areas of parliamentary activity that oversee ministers' activities and discuss and report on introduced bills.

While the first version of ParlaMint-HU was confined to interpellations, the PMII corpus includes all types of speeches and verbal exchanges made in the National Assembly. These equal 104,115 speeches, comprising 1,540,325 sentences and 32,353,437 tokens (27,533,236 words and 4,820,201 punctuation marks).

We detail our contribution to ParlaMint II in the following sections.

3.1 Data source and acquisition

The parliamentary texts were downloaded from the official website of the Hungarian National Assembly (https://www.parlament.hu/).

Metadata were gathered manually from official sources, primarily the website of the National Assembly. Other sources include *Magyar Közlöny* (the official journal of Hungary that publishes new laws, appointments, etc.), and various newspapers (e.g., for dates of resignations).

In total, we collected metadata for all 426 speakers appearing in the Hungarian National Assembly between May 2014 and June 2022, representing 91 organisations. Speakers include Members of Parliament, advocates of recognised minorities living in Hungary, secretaries of state, Hungarian members of the European Parliament, ombudsmans, visiting politicians from other countries, and (deputy) heads of offices and organisations such as the Hungarian National Bank, the State Audit Office, etc.

Among the 91 organisations, there are 17 parties, 21 parliamentary groups (the same party has a separate parliamentary group in each term), 17 ministries, and 36 other organisations. The last category includes parliamentary committees, relevant institutions of the European Union, etc. We decided not to indicate speakers who are commissioners of the state and/or members of parliamentary subcommittees.

Normalisation Before processing, the texts were manually normalised and – where applicable, – typos and other small errors were corrected. To avoid having false duplicates due to name changes or forms that are official for the persons but are not publicly used, speakers' names were unified and are included in their current form used for public appearances. (For example, Katalin Novák has some utterances as *Veresné Novák Katalin*, or Szilárd Németh appears as *Németh Szilárd István* in some cases. These were uniformly changed to *Novák Katalin* and *Németh Szilárd*, respectively. 14 names were changed in the corpus.)

Non-breaking spaces, non-breaking hyphens and soft hyphens were replaced by their regular counterpart. Quotation marks were left unmarked in texts. Occurrences of double opening or closing parentheses were uniformly changed to single parentheses.

Language usage Some speeches between 2014-2022 were held in languages other than Hungarian. These are the languages used by the representative of the German nationality in Hungary, and by advocates of other nationalities (Slovakian, Slovenian, Serbian, Bulgarian, Ukrainian, etc.). We used langdetect[6] to identify these segments. After the manual correction of langdetect's output, we had 213 segments detected as non-Hungarian. These are distributed as follows: de (85), pl (27), sk (14), sr (11), bg (10), sl (9), uk (5), cs (3), ro (3), el (2), hr(1), rom (1), ru (1). 42 segments were identified as multilingual, i.e. they contain chunks in two different languages.

3.2 Data encoding process in brief

Data processing and encoding were carried out using Python scripts. Each source txt already covered one day of a sitting, so we kept this structure. For extraordinary sittings, a separate file was created. We split the text into speeches and the speeches into segments and notes. For segments, the paragraphs of the original texts on the official National Assembly website were used. Speakers and transcription notes were detected, and the latter were categorised into TEI note types using regular expressions primarily. Segments were given a unique id, and id-segment pairs served as input for HuSpacy (Orosz et al., 2022), the linguistic analyser (more details on the linguistic analysis of the corpus in section 3.4). HuSpacy's output was converted into the required XML structure using another Python script.[7]

As the code creating the XML format is built to work on the quite rigid original format of the source files, input files had to be formatted accordingly.

[6] https://pypi.org/project/langdetect/

[7] All the scripts used to create the corpus are available at https://github.com/nytud/HuParlaMintII along with the metadata we collected. However, it is important to note that our primary aim was to create a corpus of good quality – hence the manual corrections –, so a reusable pipeline is just a side product of our project and by no means the main goal.

The presiding chair's technical remarks have *ELNÖK* ('president') as the speaker in the original transcription. The paraticular person acting in this role could not always be identified from the text alone, therefore, we needed to collect the chairs for all sessions manually. To do this, we used the data available on the official website of the National Assembly. Parallel to this, we manually collected all metadata.

3.3 Metadata

We collected metadata for the 426 speakers appearing in the Hungarian National Assembly between May 2014 and June 2022, and the 91 organisations they represent. Every organisation and person received a unique id, and we always included start and end dates for specific roles (indicating the length of the 'tenure' of a person in a role) and the duration of operation for organisations. For organisations, the start and end dates should be understood as the founding date and the date of disassembly/renaming.

If an organisation was renamed, the date of the renaming is featured as the end date of its entry on its original name, and a separate entry was created with the new name. People with multiple names in the corpus were only included under their current name, to allow for more traceability.

For each sitting day in each separate file, we also provided links to the video and the text of that sitting day available on the National Assembly website. We also annotated transcriber notes according to the ParlaMint TEI XML guidelines.[8]

Whilst some of the roles and categories we used in the metadata for organisations and speakers are not necessarily intuitive, these were dictated by the rigid metadata structure of ParlaMint.

Metadata of organisations We collected metadata on organisations according to the following categorisation:

1. *political parties*;
2. *parliamentary groups*;
3. *ministries*;
4. *other organisations*;
5. *opposition and coalition.*

We collected 17 parties, 9 parliamentary groups, 17 ministries, and 41 other organisations. The last category includes 24 committees, as well as the republic, some NGOs, and several European institutions (e.g., the European Parliament, the European Commission, and 'europeanInstitution' for additional ones).

Following a requirement of the XML structure, we marked each party as coalition or opposition in each term, depending on whether they were governing or not. Below, the entry for the Párbeszéd parliamentary group is shown as an example of the final XML structure.

[8] https://clarin-eric.github.io/ParlaMint/#chp-intro

```
<org xml:id="parliamentaryGroup.Parbeszed"
role="parliamentaryGroup">
   <orgName full="yes" xml:lang="hu">A Párbeszéd Magyarországért
   parlamenti frakciója</orgName>
   <orgName full="yes" xml:lang="en">Parliamentary group of the
   Dialogue for Hungary</orgName>
   <orgName full="abb">Párbeszéd-frakció</orgName>
   <listEvent>
      <event xml:id="parliamentaryGroup.Parbeszed.8"
            from="2018-05-08"
            to="2022-05-01">
         <label xml:lang="hu">A Párbeszéd Magyarországért
         parlamenti frakciója a 8. parlamenti ciklusban
         (2018-05-08 - 2022-05-01)</label>
         <label xml:lang="en">Parliamentary group of the Dialogue
         for Hungary in the 8th parliamentary term (2018-05-08 -
         2022-05-01)</label>
      </event>
      <event xml:id="parliamentaryGroup.Parbeszed.9"
      from="2022-05-02">
         <label xml:lang="hu">A Párbeszéd Magyarországért
         parlamenti frakciója a 9. parlamenti ciklusban
         (2022-05-02 - )</label>
         <label xml:lang="en">Parliamentary group of the Dialogue
         for Hungary in the 9th parliamentary term (2022-05-02 -
         )</label>
      </event>
   </listEvent>
</org>
```

XML snippet 1. A snippet of the XML file showing the encoded metadata of the Párbeszéd parliamentary group in the Hungarian corpus of ParliaMint II

Metadata of persons We collected the metadata on speakers appearing in the corpus as follows:

1. *basic personal data*: the name of the person (with any additional prefixes, such as *dr.*), their gender, and date of birth[9];
2. *membership of political parties and parliamentary groups:* start and end date of speakers' membership were included, and the specific role the person had in that entity[10];

[9] Out of the 426 speakers in the corpus, the speaker's date of birth could not be established in 37 instances

[10] We treated head and deputy head as roles taken on by some speakers in addition to being a member of a party. Being party-independent was treated as not having any roles in any party rather than as an additional label. MPs can be members of parliamentary groups of parties not their own. This happens, for example, if they are members of a party without parliamentary seats but they choose to join the parliamentary group of a different party.

3. *parliamentary membership*: the start and end date of the person's mandate in each parliamentary term;
4. *other roles*: any other important political role held by a person. For example, being in parliamentary committees, serving as a secretary of state, etc. We decided not to indicate speakers who are commissioners of the state and members of parliamentary subcommittees.

We collected each MP's committee positions for the period covered by the corpus, and 'party joining dates' in case of ongoing party affiliations that started before 2014. For judges and individuals operating irrespective of parliamentary terms, we added the start dates of their relevant appointments. For secretaries of state, whenever possible, we included a detailed version of their role names in English and Hungarian, as these are highly specific and variable across parliamentary terms and even under the same ministry.

The ParlaMint TEI XML schema[11] is quite rigid in terms of the possible roles a person or an organisation may have. The @role attribute has only a set number of valid values, so sometimes, we had to follow this schema instead of our intuition. This means that presidents of the republic (János Áder and Katalin Novák in the time period our corpus covers) are, for example, annotated as heads of the republic. A minister is encoded with four roles in the corpus: they are heads *and* members of a ministry (if not without portfolio), and ministers *and* members of the government. As the ParlaMint TEI XML does not have `notary` as a valid value for the @role attribute, notaries of the Hungarian Assembly are annotated as secretaries, and their exact role (*notary*, *junior notary* or *senior notary*) are listed as a natural language description of their role. A similar solution was used for the senior chair ('korelnök') which was merged into the role of chairperson with an additional note. Thus, the 426 speakers in the corpus have 3273 affiliations altogether. Figure 1 shows the number of roles assigned to the speakers. There are five speakers without any affiliation (but with basic personal data); they are representatives of foreign countries (e.g. Stanislaw Tillich), or presidential candidates giving a speech in the Assembly only once (László Majtényi and Péter Zsolt Róna). The ParlaMint TEI XML schema does not support the encoding of this kind of person metadata. Their role is encoded in the corpus as a `note` at the beginning of their speech. XML snippet 2 shows the corpus entry for MP Ágh Péter with all his encoded affiliations between 2014 and 2022.

[11] https://github.com/clarin-eric/ParlaMint/blob/main/Schema/README.md

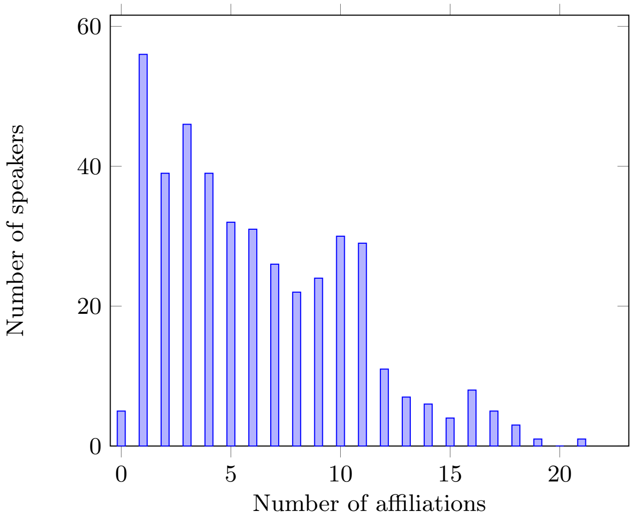

Fig. 1: Bar plot showing the number of affiliations the speakers in our corpus have. As can be seen, most speakers are affiliated with at least one role, and the majority of them have 2-10 roles. The recorder is Csaba Hende with 21 roles. Speakers without any role are guest speakers, representatives of other countries.

```
<person xml:id="AghPeter">
     <persName>
        <surname>Ágh</surname>
        <forename>Péter</forename>
     </persName>
     <sex value="M"/>
     <birth when="1982-01-30"/>
     <affiliation role="member"
                 ref="#OGY"
                 from="2014-06-06"
                 to="2018-05-07"
                 ana="#OGY.7"/>
     <affiliation role="member"
                 ref="#OGY"
                 from="2018-05-08"
                 to="2022-05-01"
                 ana="#OGY.8"/>
     <affiliation role="member" ref="#OGY" from="2022-05-02"
     ana="#OGY.9"/>
     <affiliation role="member" ref="#party.Fidesz" from="1999"/>
     <affiliation role="deputyHead"
                 ref="#parliamentaryGroup.Fidesz"
                 ana="#parliamentaryGroup.Fidesz.7"
                 from="2014-07-04"
                 to="2015-11-01">
```

```
      <roleName xml:lang="en">Deputy Head</roleName>
   </affiliation>
   <affiliation role="member"
               ref="#parliamentaryGroup.Fidesz"
               ana="#parliamentaryGroup.Fidesz.7"
               from="2014-05-06"
               to="2018-05-07"/>
   <affiliation role="member"
               ref="#parliamentaryGroup.Fidesz"
               ana="#parliamentaryGroup.Fidesz.8"
               from="2018-05-08"
               to="2022-05-01"/>
   <affiliation role="member"
               ref="#parliamentaryGroup.Fidesz"
               ana="#parliamentaryGroup.Fidesz.9"
               from="2022-05-02"/>
   <affiliation role="deputyHead"
               from="2018-10-15"
               to="2020-05-05"
               ref="#org.HRB">
   </affiliation>
   <affiliation role="member" from="2014-06-06" to="2018-05-07"
   ref="#org.HRB"/>
   <affiliation role="member" from="2018-05-18" to="2020-05-05"
   ref="#org.HRB"/>
   <affiliation role="member" from="2014-06-06" to="2015-09-22"
   ref="#org.TB"/>
   <affiliation role="member" from="2020-05-05" to="2022-05-01"
   ref="#org.TB"/>
   <affiliation role="member" from="2022-05-02" ref="#org.TB"/>
</person>
```

XML snippet 2. A snippet of the XML file showing the rich metadata of Ágh Péter, a member of parliament featured in the Hungarian ParlaMint II corpus

Other metadata In addition to metadata about speakers and organisations, we have further data on each session day. These include the type of session (regular, special or ceremonial), links to the text and the video recording of the session, the number of speeches made, and the relative place of a sitting day situated in a sitting that is a specific, often several-day-long session of that parliamentary term.

Transcriber comments Transcriber comments are also noted in the corpus, containing information about the speaker, time, voting results, or, more importantly, incidents in the assembly hall. Such an element is encoded as a <note>, where *incidents* can be further specified as vocal, kinesic and incident.

Vocals are vocalised – but not necessarily lexical – happenings, with further subtypes of laughter, speaking, interruption, shouting or murmuring. Sometimes, there are multiple speakers in an interruption. The further processing of such notes is part of our future work. Kinesic describes communicative – though not necessarily vocalised – events, with subtypes such as applause, ringing, snapping or any gestures.

Incidents can be very versatile, describing not necessarily communicative events. Subtypes of this category are break, pause, entering or leaving.

Coding the transcriber comments was not always easy as the categories were not very distinct. Transcriber comments represent very rich metadata in our corpus – more than 37 thousand occurrences in total. The processing of these data was partly automatic but lots of additional manual analysis was needed due to the detailed framework.

3.4 Linguistic annotation

ParlaMint II requires all corpora to be linguistically annotated as follows:

- tokenisation with preserving information on inter-token spaces while separately tagging words and punctuation marks
- sentence splitting
- lemmatisation
- part-of-speech and morphological features using the UD tagset, while possibly also providing part-of-speech tags of a different (local) tagset
- named entity tagging, using the four standard named entity classes: `PER`, `LOC`, `ORG`, `MISC`
- UD dependency syntactic parse of the sentences

Only the text content of the speech segments was linguistically annotated. Transcriber notes and speeches held in languages other than Hungarian were excluded from the linguistic analysis.

Hungarian is in the fortunate position of having several processing pipelines to choose from for carrying out the linguistic analysis required by ParlaMint II. e-magyar (Váradi et al., 2018) and HuSpaCy (Orosz et al., 2022) both provide the necessary morphological and syntactic analysis in addition to named entity recognition. emDep (the dependency parser of e-magyar) does not have the UD format as an available output, and no converter is available to create that either. The output of HuSpaCy fitted perfectly well with ParlaMint II's requirements, so we chose this pipeline over e-magyar.[12] For the linguistic annotation of the corpus, we used HuSpaCy and its hu_core_news_lg model. The lower-level analyses, such as sentence splitting, tokenisation, and lemmatisation resulted in an accurate and easily validatable annotation. The named entity recognition, on the other hand, generated erroneous labels in some cases that popped up during the automatic central validation of the XML files, causing the break of

[12] We are aware that the Stanza and UDPipe parsers are available as alternative dependency parsers in e-magyar – as emStanza and emUDPipe –, but to avoid conversion problems and to stick with the more compact option, we chose HuSpaCy.

the script generating the CoNLL-U format from the XML, among others. We manually corrected the named entity tags of these tokens.[13]

4 Potential applications

Numerous fields of SSH and especially digital humanities may benefit from well-structured, linguistically annotated parliamentary data with rich metadata. There is extensive research on political debates in Hungary as well as worldwide, for example in political science (Kiss, 1998; van Dijk, 2010), discourse analysis (Schirm, 2021), translation studies (Kovács, 2012) or history (Pančur and Šorn, 2016). By increasing the level of processing, more in-depth research is made possible on an ever larger scale of resources. For example, researchers can look at a longer time span, include more speeches, or better identify the ideal targets of their analysis. The importance of research into parliamentary data is further borne out by the fact that there were Parla-CLARIN workshops on creating corpora of parliamentary data at each of the 2018, 2020, and 2022 International Conferences on Language Resources and Evaluation (LREC). The keynote speaker of the 2022 workshop[14] was Luke Blaxill, a historian of British politics and monarchy from 1750 to present, who talked about the possible ways parliamentary data could be used in history and political science. Some specific use cases of the Danish and Basque ParlaMint corpora are demonstrated by Navarretta and Haltrup Hansen (2022) and Escribano et al. (2022), respectively. We welcome enquiries from researchers working in the social sciences and humanities regarding the use of the Hungarian ParlaMint II corpus.

5 Summary

We presented the newly built Hungarian corpus of the second phase of the ParlaMint project, funded by CLARIN-ERIC. The Hungarian corpus of ParlaMint II – alongside all other corpora of European parliamentary data in the participating languages – is a comparable, high-quality parliamentary corpus, which was built with encoding schemas, metadata structure, and Universal Dependencies-type linguistic annotation that are uniform across the corpora of ParlaMint II. Our corpus focuses on the minutes of the Hungarian National Assembly between May 2014 and June 2022. It has a wider time frame, more extensive metadata on speakers and their affiliations, and higher quality linguistic analysis than what was available before for Hungarian. The corpus will be released and be openly available online alongside with the rest of the ParlaMint II corpora at the end of the CLARIN ParlaMint II project (March 2023). Some potential applications of ParlaMint corpora in SSH research are also discussed.

[13] The – detected – false named entity tagging was mainly in connection with punctuation marks. For example, *… képviselő úr*[, *Jobbik*]$_{ORG}$. or *„Bal kéz*[, *jobb kéz”*]$_{MISC}$ *címmel.* both created errors.

[14] For more information on the workshop see https://www.clarin.eu/ParlaCLARIN-III

References

TEI P5: Guidelines for electronic text encoding and interchange. (2017), http://www.tei-c.org/Guidelines/P5/

van Dijk, T.A.: Political identities in parliamentary debates. In: Ilie, C. (ed.) European Parliaments under Scrutiny, chap. 1, pp. 29–56. John Benjamins Publishing Company, Amsterdam (2010)

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, c., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M.C., de Macedo, L.D., Navarretta, C., Luxardo, G., Coole, M., Rayson, P., Morkevičius, V., Krilavičius, T., Darġis, R., Ring, O., van Heusden, R., Marx, M., Fišer, D.: The ParlaMint Corpora of Parliamentary Proceedings. Language Resources and Evaluation (2022), https://link.springer.com/article/10.1007/s10579-021-09574-0

Erjavec, T., Ogrodniczuk, M., Osenova, P., Pančur, A., Ljubešić, N., Agnoloni, T., Barkarson, S., Pérez, M.C., Çağrı Çöltekin, Coole, M., Darġis, R., de Macedo, L., de Does, J., Depuydt, K., Diwersy, S., Hansen, D.H., Kopp, M., Krilavičius, T., Luxardo, G., Marx, M., Morkevičius, V., Navarretta, C., Rayson, P., Ring, O., Rudolf, M., Simov, K., Steingrímsson, S., Üveges, I., van Heusden, R., Venturi, G.: ParlaMint: Comparable Corpora of European Parliamentary Data. In: CLARIN Annual Conference Proceedings 2021. pp. 20–25. CLARIN ERIC, Utrecht, The Netherlands (2021)

Erjavec, T., Pančur, A.: Parla-CLARIN: TEI guidelines for corpora of parliamentary proceeding (2019), https://doi.org/10.5281/zenodo.3446164

Escribano, N., Gonzalez, J.A., Orbegozo-Terradillos, J., Larrondo-Ureta, A., Peña-Fernández, S., Perez-de Viñaspre, O., Agerri, R.: BasqueParl: A Bilingual Corpus of Basque Parliamentary Transcriptions. In: Proceedings of the Language Resources and Evaluation Conference. pp. 3382–3390. European Language Resources Association, Marseille, France (June 2022), https://aclanthology.org/2022.lrec-1.361

Kiss, J.: A pártok szimbolikus arculata és érvelési sajátosságai a parlamenti vitákban. Politikatudományi Szemle 7/2, 27–60 (1998)

Kovács, M.: Az európai frazeológiai univerzálék konceptualizációja és fordítási megfeleltetései. Fordítástudomány XIV/1., 48–68 (2012)

Navarretta, C., Haltrup Hansen, D.: The Subject Annotations of the Danish Parliament Corpus (2009-2017) - Evaluated with Automatic Multi-label Classification. In: Proceedings of the Language Resources and Evaluation Conference. pp. 1428–1436. European Language Resources Association, Marseille, France (June 2022), https://aclanthology.org/2022.lrec-1.153

Orosz, Gy., Szántó, Zs., Berkecz, P., Szabó, G., Farkas, R.: HuSpaCy: an industrial-strength Hungarian natural language processing toolkit. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia (2022)

Pančur, A., Šorn, M.: Smart big data: use of Slovenian parliamentary papers in digital history 56(3), 130–146 (2016)

Schirm, A.: Diskurzusjelölők szövegeken innen és túl. Loisir Könyvkiadó, Budapest (2021)

Straka, M., Straková, J.: Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 88–99. Association for Computational Linguistics, Vancouver, Canada (Aug 2017), https://aclanthology.org/K17-3009

Szarvas, G., Farkas, R., Kocsor, A.: A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In: Todorovski, L., Lavrač, N., Jantke, K.P. (eds.) Discovery Science. pp. 267–278. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)

Váradi, T., Simon, E., Sass, B., Mittelholcz, I., Novák, A., Indig, B., Farkas, R., Vincze, V.: E-magyar – A Digital Language Processing System. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), https://aclanthology.org/L18-1208

Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP. pp. 763–771 (2013)

Üveges, I., Ring, O.: A CLARIN ParlaMint magyar korpusza. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia (2022)