One Year of Continuous and Automatic Data Gathering from Parliaments of European Union Member States

Ota Mikušek
Lexical Computing, Brno, Czech Republic
ota.mikusek@sketchengine.eu

Abstract

This paper provides insight into automatic parliamentary corpora development. One year ago, I created a simple set of tools designed to continuously and automatically download, process, and create corpora from speeches in the parliaments of European Union member states. Despite the existence of numerous corpora providing speeches from European Union parliaments, the tools are more focused on collecting and building such corpora with minimal human interaction. These tools have been operating continuously for over a year, gathering parliamentary data and extending corpora, which together have more than one billion words. However, the process of maintaining these tools has brought unforeseen challenges, including issues such as being blocked by some parliaments due to overloading the parliament with requests, the inability to access the most recent data of a parliament, and effectively managing interrupted connections. Additionally, potential problems that may arise in the future are provided, along with possible solutions. These include problems with data loss prevention and adaptation to changes in the sources from which speeches are downloaded.

Keywords: parliamentary protocols, continuous downloading, corpus processing, automatic tools, corpus development, maintenance of tools

1. European Parliamentary Corpora

Between July 2020 and May 2021, the ParlaMint I (Erjavec et al., 2022) project aimed to create corpora of transcriptions from the sessions of 17 European Union parliaments from 2015 to October 2019. ParlaMint I was the largest project of its kind for European parliamentary corpora at the time. Each parliamentary corpus had a dedicated lead developer.

In December 2021, the ParlaMint II (Erjavec et al., 2021) project extended the work of ParlaMint I by including parliamentary transcriptions up to July 2022. This project also involved updates to the schema, validation, and enhancement of corpora with additional metadata.

In July 2023 ParlaMint 3.0 (Erjavec et al., 2023b) and in October 2023 ParlaMint 4.0 (Erjavec et al., 2023a) follows ParlaMint II and extend it. Currently, ParlaMint 4.0 provides 29 corpora, namely for Bulgarian, Croatian, Polish, Slovenian, Czech, Icelandic, Belgian, Danish, Spanish, Dutch, Turkish, Italian, Hungarian, Latvian, French, Bosnian, Catalonian, Galician, Greek, Norwegian, Serbian, Swedish, Ukrainian, Finnish, Estonian, Basque, United Kingdom, Portuguese and Austrian parliament.

For all corpora, ParlaMint 4.0 provides unified metadata, including timestamps, speaker details, transcriber notes, and source URLs for documents. Expanding coverage to include other parliaments is a future objective for the ParlaMint project.

In addition, there are other initiatives to create parliamentary corpora, such as the Polish Parliamentary Corpus (Ogrodniczuk, 2018), which covers debates from 1919 to the present, and the German Parliamentary Corpus (GerParCor) (Abrami et al., 2022), which includes transcripts from Germany, Liechtenstein, Austria, and Switzerland up to 2021, with plans for continuous development. The Czech Parliamentary Corpus (CzechParl) (Jakubíček and Kovář, 2010) is based on Czech parliament stenographic protocols from the 1990s. The Dutch Parliamentary Corpus (DutchParl) (Marx et al., 2010) aims to collect Dutch-language parliamentary documents and has different sized corpora for Belgium, Flanders, and the Netherlands, with ongoing development efforts.

2. Automatic Tools

A year ago, I created a toolset written in Python language providing continuous automatic development of corpora from transcriptions of parliamentary chambers from selected members of the EU. From suitable sources of parliamentary protocols on chamber websites, created scripts are gathering protocols in different formats and unifying their format as preverticals[1]. The prevertical format is a file format containing plain text and structures. The structures enclose the text and provide metadata about the text. An example of a document in prevertical format, created by the tools, is shown in Figure 1.

Created scripts are independent of each other and work autonomously, automatically, and atom-

[1] https://www.sketchengine.eu/my_keywords/prevertical/

ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora, pages 149–153

icly. Each script consists of three parts: shared code, a tool for discovering and downloading new protocols, and a tool for processing downloaded protocols into prevertical files. In case of any error, scripts are able to log this error, notify the script administrator, and roll back to the last consistent state.

2.1. Downloading of Data

Reliable sources of protocols were searched on parliamentary official websites. For a source to be considered reliable, it must come directly from the parliament, it has to provide an option to discover newly added protocols, and it must not rely on website-provided scripts (mainly javascript).

The reason why script execution to access or discover new protocols is unwanted is that user-side scripts can change over time, and these changes may cause errors during the automatic download process. Such dependency is unwanted because it increases maintenance difficulty.

Found sources provided data in plain text, HTML, JSON, CSV, XML, XLSX, and DOCX format. PDF file format was also available. However, PDF format introduced problems with the ordering of the paragraphs, and text extraction, when words were split at the end of the line by “-" character. In cases when the source was not found on the parliament website, the parliament was contacted via email.

Created scripts are downloading protocols from sources automatically and atomically. If the downloading of a protocol fails, this information is logged, and the download will be retried during the next script execution.

2.2. Processing of Protocols

A script that processes downloaded protocols called prevertbuilder was created for each chamber website. The prevertbuilder is responsible for metadata extraction and unifying downloaded protocols into prevertical format. Common metadata across all corpora are the speaker name, the date, the source URL, the URL access time, and the filename where prevertical is stored. More metadata, like notes of transcriber, are also provided for some corpora.

The prevertbuilder works like a pipe. It contains the initialization, writing, and finalization methods, which process downloaded protocols linearly and do not require the whole protocol to be loaded in memory. This capability is used, for example, in the Swedish parliament, where one downloaded document consists of protocols from a month period.

A protocol is marked as successfully processed only when prevertbuilder process the protocol without an error. Prevertbuilders are capable of detecting the presence of new information (for example, new tags or attributes) in processed protocols. By default, in these cases, protocols are processed without these new elements. However, their occurrence is logged as a warning in the script log.

```
<doc source_url="https://www.oireachtas.ie/en/debates/
debate/select_committee_on_justice/2022-06-28/"
url_access_time="2023-05-10 10:41:45 UTC"
filename="select_committee_on_justice_2022-06-28.prev-
ert" date="2022-06-28" date_day="28" date_month="6"
date_year="2022">
<note type="Other">
Tháinig an Roghchoiste le chéile ag 03:00 p.m.
</note>
<note type="Other">
The Select Committee met at 03:00 p.m.
</note>
<speaker name="Chairman">
<p>
I welcome the Minister of State, the departmental
officials and ...
</p>
</speaker>
<speaker name="Minister of State at the Department of
Justice (Deputy James Browne)">
<p>
I wish to mention something. As the Minister for
Justice, ...
</p>
</speaker>
</doc>
```

Figure 1: Example of prevertical format from the upper chamber of the Irish parliament (modified)

The final corpora is created using (No)Sketch Engine (Kilgarriff et al., 2014) infrastructure and are available on Sketch Engine[2] under the name Parliament debates.

3. Flaws of Current Design

The original toolset was the first of its kind. Some of the original goals, like zero human interaction and the ability to have the most recent data could be considered naive after running them for over one year. During the maintenance of these tools, several problems were encountered.

3.1. Speaker name attribute detection

In some cases, sources do not provide the name of a speaker but just their role. This can be seen in Figure 1, where in one case, only the speaker role "Chairman" is provided, without the actual name of the speaker. This information can be acquired elsewhere and could be resolved at a possible cost

[2] https://app.sketchengine.eu/

De Nederlandse landbouwsector kiest voor geïntegreerde gewasbescherming en is wat dat betreft koploper in Europa. Europese landen kunnen hier nog veel van leren. Pas in allerlaatste instantie worden chemische middelen ingezet. Daarbij is belangrijk dat de toelating van groene gewasbescherming wordt verbeterd. Ik heb daar eerder dit jaar Kamervragen over gesteld. Heeft de staatssecretaris al met de agrarische sector overlegd over de acute knelpunten voor de groene gewasbescherming?

De **voorzitter**:
Daarmee zijn we gekomen aan het eind van de eerste termijn van de Kamer.

De vergadering wordt enkele ogenblikken geschorst.

De **voorzitter**:
Ik geef de staatssecretaris het woord voor zijn beantwoording in eerste termijn.

Figure 2: Lower parliament chamber of the Netherlans

of more dependencies and, therefore, higher maintenance.

3.2. Notes of transcriber detection

Notes of transcriber are hard to detect in the lower parliament chamber of the Netherlands. In Figure 2 is a sample of discussion[3]. All sentences were spoken except the penultimate sentence, which is a note from the transcriber saying that the sitting is suspended for a few moments.

The format of the note is indistinguishable from the rest of the spoken text. Currently, these notes remain undetected and are added as spoken text to the current speaker.

3.3. Overloading of Parliaments

In the original release of the tools, none of the tools were using delays between requests to parliamentary source. The Parliament of Denmark started to require human verification to access its website two weeks after the first run of the original tools. The Parliament of the Netherlands banned the IP address of the server where the tools were originally running. This led to a quick fix by adding random time delays between requests to each source. No more problems that could be related to overloading the parliaments were encountered since the fix.

3.4. Delay in Data Source

During the selection of a suitable source for the Finnish Parliament, I was unable to find any reliable source for the Finnish Parliament website. I contacted the Finnish Parliament via email to ask for such a reliable source. The Finnish Parliament responded with webpage[4] where, according to the Finnish Parliament, new data should be available only twice a year.

However, it seems that data are not updated two times per year but only once a year. Unfortunately, in both cases, this is not ideal since one of the core ideas was to have up-to-date parliamentary transcripts with just a little delay from the time they are published on the source.

3.5. Connection Errors

Whenever tools encounter a problem, the problem is classified as a warning if the tool can continue or an error when the tool cannot continue. In both cases, an email is sent to the tool administrator to resolve the issue.

The most common type of error encountered during tools execution are connection errors when connection with sources is interrupted. The correct reaction to this error is waiting until another day when tools are automatically executed again. However, email is sent anyway, which leads to spamming the tool administrator's inbox with errors that require no action.

The collection of errors and warnings frequency is important. If connection errors become frequent in some tools, action may be required. I recommend solving this issue by creating an email filter that automatically archives this specific error. In cases when the connection errors would persist for a longer period, other tools safeguards will inform about the error, like checking if the tools data were recently compiled.

4. Future Work

As I present my tools, others also express their concerns about them. Currently, these concerns center on two main issues.

4.1. No Backups of Downloaded Sources

Downloaded sources are processed in memory. Only the output of the prevertbuilder is stored. This means that in the case when it would be found out that some part of the tools was working incorrectly, currently, the only way to reprocess incorrectly processed transcripts is to rely on their presence in parliamentary sources.

This situation may require redownloading a bigger portion of transcripts from the source, which could be a problem since, in the past, some parliaments were actively blocking the toolset because it was gathering too much data. Because of that, the current main focus is on fixing this issue.

[3] https://www.tweedekamer.nl/kamerstukken/plenaire_verslagen/detail/2016-2017/85

[4] https://avoindata.eduskunta.fi/#/fi/dataset-search

4.2. Major Change of the Source

One of the main ideas was the toolset's ability to adapt to changes in parliamentary transcript sources. Most of the time, only new elements or segments are added to the parliament source, which provides no more information to the gathered transcripts. In other cases, useful information is added to the overall structure of the transcripts, which does not interrupt the continuous and automatic download process. Fortunately, there was never a change that would require a complete rewrite of the downloading tool.

This means that the toolset currently could be run on sources that were available at the time of toolset creation and still work correctly. Problems may arise when parliamentary sources would undergo complete renewal. This would mean that the ability to go back to older transcripts would be lost.

5. Conclusion

Currently, tools are running for over one year and have collected over 1,200 million words, as can be seen in Table 1. Development and maintenance of the automatic parliamentary corpora toolset have revealed several flaws in its original design.

The most important flaws are attribute detection and connection errors. The connection errors show the importance of atomicity. Problems with attribute detection still remain to be solved.

Still, the tools are working and doing what was expected from them. Small human interaction is still required, but those interactions are not critical for tools correct function.

The source code of all the tools is licensed under GNU Lesser General Public License 3.0 and available in a GitLab repository.[5]

corpus name	words	words now
bg_deputies	5.40M	5.86M
cz_deputies	18.41M	20.79M
cz_senate	11.32M	11.58M
dk_deputies	79.00M	79.59M
nl_deputies	71.20M	80.25M
nl_senate	9.99M	11.01M
ir_deputies	40.70M	87.31M
ee_deputies	9.04M	10.49M
fi_deputies	21.09M	21.11M
be_deputies	54.94M	56.77M
be_senate	0.06M	0.69M
fr_deputies	21.09M	59.57M
fr_senate	169.08M	173.53M
at_deputies	6.94M	7.21M
at_senate	2.73M	2.88M
de_deputies	125.03M	125.53M
gr_deputies	58.31M	59.48M
hu_deputies	3.08M	3.93M
it_deputies	3.32M	5.16M
it_senate	13.31M	14.62M
pl_senate	20.08M	20.26M
pt_deputies	141.10M	154.37M
ro_deputies	14.02M	14.86M
ro_senate	26.36M	26.88M
sk_deputies	6.76M	8.74M
si_deputies	15.49M	23.70M
es_deputies	66.66M	68.73M
se_deputies	131.74M	131.75M
sum	1,146.25M	1,286.65M

Table 1: Comparison of processed data from May 2023 to April 2024 (word count from final corpora)

6. Bibliographical References

Giuseppe Abrami, Mevlüt Bagci, Leon Hammerla, and Alexander Mehler. 2022. German parliamentary corpus (gerparcor). In *Proceedings of the Language Resources and Evaluation Conference*, pages 1900–1906, Marseille, France. European Language Resources Association.

Tomaž Erjavec, Matyáš Kopp, Maciej Ogrodniczuk, Petya Osenova, Rodrigo Agerri, Manex Agirrezabal, Tommaso Agnoloni, José Aires, Monica Albini, Jon Alkorta, Iván Antiba-Cartazo, Ekain Arrieta, Mario Barcala, Daniel Bardanca, Starkaður Barkarson, Roberto Bartolini, Roberto Battistoni, Nuria Bel, Maria del Mar Bonet Ramos, María Calzada Pérez, Aida Cardoso, Çağrı Çöltekin, Matthew Coole, Roberts Dargis, Jesse de Does, Ruben de Libano, Griet Depoorter, Katrien Depuydt, Sascha Diwersy, Réka Dodé, Kike Fernandez, Elisa Fernández Rei, Francesca Frontini, Marcos Garcia, Noelia García Díaz, Pedro García Louzao, Maria Gavriilidou, Dimitris Gkoumas, Ilko Grigorov, Vladislava Grigorova, Dorte Haltrup Hansen, Mikel Iruskieta, Johan Jarlbrink, Kinga Jelencsik-Mátyus, Bart Jongejan, Neeme Kahusk, Martin Kirnbauer, Anna Kryvenko, Noémi Ligeti-Nagy, Nikola Ljubešić, Giancarlo Luxardo, Carmen Magariños, Måns Magnusson, Carlo Marchetti, Maarten Marx, Katja Meden, Amália Mendes, Michal Mochtak, Martin Mölder, Simonetta Montemagni, Costanza Navarretta, Bartłomiej Nitoń, Fredrik Mohammadi Norén, Amanda Nwadukwe, Mihael Ojsteršek, Andrej Pančur, Vassilis Papavassiliou, Rui Pereira, María Pérez Lago, Stelios Piperidis, Hannes Pirker, Marilina Pisani, Henk van der Pol, Prokopis Prokopidis, Vale-

[5] https://gitlab.com/Atom194/european-parliamentary-protocols

ria Quochi, Paul Rayson, Xosé Luís Regueira, Michał Rudolf, Manuela Ruisi, Peter Rupnik, Daniel Schopper, Kiril Simov, Laura Sinikallio, Jure Skubic, Minna Tamper, Lars Magne Tungland, Jouni Tuominen, Ruben van Heusden, Zsófia Varga, Marta Vázquez Abuín, Giulia Venturi, Adrián Vidal Miguéns, Kadri Vider, Ainhoa Vivel Couso, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, Rodolfo Zevallos, and Darja Fišer. 2023a. Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 4.0. Slovenian language resource repository CLARIN.SI.

Tomaž Erjavec, Matyáš Kopp, Maciej Ogrodniczuk, Petya Osenova, Darja Fišer, Hannes Pirker, Tanja Wissik, Daniel Schopper, Martin Kirnbauer, Michal Mochtak, Nikola Ljubešić, Peter Rupnik, Henk van der Pol, Griet Depoorter, Jesse de Does, Kiril Simov, Vladislava Grigorova, Ilko Grigorov, Bart Jongejan, Dorte Haltrup Hansen, Costanza Navarretta, Martin Mölder, Neeme Kahusk, Kadri Vider, Nuria Bel, Iván Antiba-Cartazo, Marilina Pisani, Rodolfo Zevallos, Xosé Luís Regueira, Adina Ioana Vladu, Carmen Magariños, Daniel Bardanca, Mario Barcala, Marcos Garcia, María Pérez Lago, Pedro García Louzao, Ainhoa Vivel Couso, Marta Vázquez Abuín, Noelia García Díaz, Adrián Vidal Miguéns, Elisa Fernández Rei, Sascha Diwersy, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Amanda Nwadukwe, Dimitris Gkoumas, Vassilis Papavassiliou, Prokopis Prokopidis, Maria Gavriilidou, Stelios Piperidis, Noémi Ligeti-Nagy, Kinga Jelencsik-Mátyus, Zsófia Varga, Réka Dodé, Starkaður Barkarson, Tommaso Agnoloni, Roberto Bartolini, Francesca Frontini, Simonetta Montemagni, Valeria Quochi, Giulia Venturi, Manuela Ruisi, Carlo Marchetti, Roberto Battistoni, Roberts Darģis, Ruben van Heusden, Maarten Marx, Katrien Depuydt, Lars Magne Tungland, Michał Rudolf, Bartłomiej Nitoń, José Aires, Amália Mendes, Aida Cardoso, Rui Pereira, Väinö Yrjänäinen, Fredrik Mohammadi Norén, Måns Magnusson, Johan Jarlbrink, Katja Meden, Andrej Pančur, Mihael Ojsteršek, Çağrı Çöltekin, and Anna Kryvenko. 2023b. Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 3.0. Slovenian language resource repository CLARIN.SI.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Vladislava Grigorova, Michał Rudolf, Andrej Pančur, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Henk van der Pol, Griet Depoorter, Jesse de Does, Bart Jongejan, Dorte Haltrup Hansen, Costanza Navarretta, María Calzada Pérez, Luciana D. de Macedo, Ruben van Heusden, Maarten Marx, Çağrı Çöltekin, Matthew Coole, Tommaso Agnoloni, Francesca Frontini, Simonetta Montemagni, Valeria Quochi, Giulia Venturi, Manuela Ruisi, Carlo Marchetti, Roberto Battistoni, Miklós Sebők, Orsolya Ring, Roberts Darģis, Andrius Utka, Mindaugas Petkevičius, Monika Briedienė, Tomas Krilavičius, Vaidas Morkevičius, Roberto Bartolini, Andrea Cimino, Sascha Diwersy, Giancarlo Luxardo, and Paul Rayson. 2021. Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1. Slovenian language resource repository CLARIN.SI.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2022. The parlamint corpora of parliamentary proceedings. *Language Resources and Evaluation*.

Miloš Jakubíček and Vojtěch Kovář. 2010. Czechparl: Corpus of stenographic protocols from czech parliament. In *Proceedings of Recent Advances in Slavonic Natural Language Processing 2010*, pages 41–46, Brno. Masaryk University.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, pages 7–36.

Maarten Marx, A Schuth, et al. 2010. Dutchparl. a corpus of parliamentary documents in dutch. *Proceedings Language Resources and Evaluation (LREC)*, pages 3670–3677.

Maciej Ogrodniczuk. 2018. Polish parliamentary corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).