Data Gathered with Automatic Tools from European Parliamentary Chambers

Ota Mikušek[1,2]

[1] Faculty of Informatics, Masaryk University
xmikusek@fi.muni.cz
[2] Lexical Computing, Brno, Czech Republic
ota.mikusek@sketchengine.eu

Abstract. This paper reflects on the set of tools developed in my bachelor's thesis, titled "Continuous Automatic Development of European Parliamentary Corpora." Despite the existence of numerous corpora offering speeches from the parliaments of the European Union, the developed toolset is designed to gather and build such corpora with minimal human intervention. With nine months of practical application, this paper presents insights into the faced challenges and their respective solutions, providing an overview since the initial release of the toolset.

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

1 European Parliamentary Corpora

Between July 2020 and May 2021, the ParlaMint I [4] 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, which helped the overall quality of the resulting corpora.

In December 2021, the ParlaMint II [3] 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 [2] began as a follow-up to ParlaMint II. ParlaMint 3.0 added new metadata information for bicameral parliaments if data was provided from the upper or the lower house of parliament. New corpora were introduced in ParlaMint 3.0, namely corpora of Austria, Bosnia, Catalonia, Galicia, Greece, Norway, Portugal, Serbia, Sweden, and Ukraine. Two corpora (Spanish and Lithuanian) were removed.

The ParlaMint projects provide unified metadata for all corpora, consisting of 24 types of information, including timestamps, speaker details, transcriber

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2023, pp. 107–112, 2023. © Tribun EU 2023

notes, and source URLs for documents. However, it's important to note that despite their rich metadata, only 19 out of the 27 current EU states are covered by ParlaMint. Expanding coverage to include these missing 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 [7], which covers debates from 1919 to the present, and the German Parliamentary Corpus (GerParCor) [1], which includes transcripts from Germany, Liechtenstein, Austria, and Switzerland up to 2021, with plans for continuous development. The Czech Parliamentary Corpus (CzechParl) [5] is based on Czech parliament stenographic protocols from the 1990s. The Dutch Parliamentary Corpus (DutchParl) [6] aims to collect Dutch parliamentary documents and has different sized corpora for Belgium, Flanders, and the Netherlands, with ongoing development efforts.

2 Automatic tools

The outcome of my thesis, titled ″Continuous Automatic Development of European Parliamentary Corpora,″ is a Python-based toolset designed to facilitate the ongoing automatic development of corpora derived from transcriptions of parliamentary sessions involving selected members of the European Union. The toolset employs scripts that gather protocols from suitable sources on chamber websites, accommodating various formats and unifying them into a standardized prevertical format. The prevertical[3] format is a file format containing plain text and structures. The structures enclose the text and provide metadata about the text.

The scripts are designed to operate independently of each other, functioning autonomously, automatically, and atomically. Each script comprises three main components: shared code, a tool for discovering and downloading new protocols, and a tool for processing the downloaded protocols into prevertical files. In the event of an error, the scripts have the capability to log the error, notify the script administrator, and revert to the last consistent state.

The source code of all the tools is licensed under GNU Lesser General Public License 3.0 and available in a GitLab repository.[4]

2.1 Downloading of data

To secure reliable sources of protocols, a search was conducted on official parliamentary websites. To be deemed reliable, a source must originate directly from the parliament, offer a mechanism to identify newly added protocols, and refrain from dependence on website-provided scripts, particularly those depending on 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

[3] `https://www.sketchengine.eu/my_keywords/prevertical/`

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

may cause errors during the automatic download process. Such dependency is unwanted because it increases maintenance difficulty.

The identified sources presented data in various formats, including plain text, HTML, JSON, CSV, XML, XLSX, and DOCX. Additionally, some of the chambers provided PDF files with transcriptions. However, challenges arose with the PDF format, specifically regarding the ordering of paragraphs and text extraction, especially when words were hyphenated at the end of a line using the "-" character. In instances where the source was not available on the parliament website, the parliament was connected through email.

The developed scripts automatically and atomically download protocols from designated sources. In the event of a protocol download failure, the error 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.

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 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.

3 Tools maintenance

During the continuous nine-month operation, the tools underwent several modifications to accommodate changes in the source data. These adjustments primarily focused on adapting the prevertical creation process to handle new elements, structures, and attributes in the sources.

For instance, a change emerged within Slovenia's parliament, where changes in month naming conventions were made after the first tool deployment. The updated month names differ from the previous ones in inflection of the month names. The solution to this change involved adding records to the month name to month number dictionary as errors arose from unknown month names. Due to a lack of knowledge in Slovenian inflection, this approach proved more manageable than attempting to add all new month names simultaneously, as errors were prone to occur in that process.

[5] The chamber releases new transcriptions yearly.

[6] The chamber releases new transcriptions yearly.

Table 1: Comparison of processed data from May 2023 to November 2023

corpus name	words	words now	change	from year
bg_deputies	5.40M	5.82M	+0.42M	2022
cz_deputies	18.41M	20.71M	+2.30M	2018
cz_senate	11.32M	11.51M	0.19M	2010
dk_deputies	79.00M	79.55M	+0.55M	2007
nl_deputies	71.20M	80.20M	+9.00M	2013
nl_senate	9.99M	11.01M	+0.02M	2019
ir_deputies	40.70M	87.28M	+46.58M	2022
ee_deputies	9.04M	10.47M	+1.43M	2020
fi_deputies	21.09M	21.09M	0[5]	2015
be_deputies	54.94M	56.70M	+1.76M	2007
be_senate	0.06M	0.69M	+0.63M	2019
fr_deputies	21.09M	59.55M	+38.46M	2015
fr_senate	169.08M	173.52M	+4.44M	2004
at_deputies	6.94M	7.19M	+0.25M	2022
at_senate	2.73M	2.87M	+0.14M	2019
de_deputies	125.03M	125.53M	+0.50M	1950
gr_deputies	58.31M	59.47M	+1.16M	2015
hu_deputies	3.08M	3.93M	+0.85M	2022
it_deputies	3.32M	5.15M	+1.83M	2022
it_senate	13.31M	14.61M	+1.30M	2018
pl_senate	20.08M	20.25M	+0.17M	2011
pt_deputies	141.10M	154.36M	+13.26M	1976
ro_deputies	14.02M	14.86M	+0.84M	2016
ro_senate	26.36M	26.88M	+0.52M	2001
sk_deputies	6.76M	8.73M	+1.97M	2022
si_deputies	15.49M	23.69M	+8.20M	2018
es_deputies	66.66M	68.73M	+2.07M	2019
se_deputies	131.74M	131.74M	0[6]	1994
sum	1,146.25M	1,286.09M	+139.84M	-

Changes also happened in the Parliament of Bulgaria, which implemented specific measures to block requests not containing a 'User-Agent' header. This change caused the tool to be unable to download any protocol. The tool was modified to use 'User-Agent': 'curl/7.82.0' header, which resolved the problem.

Sometimes, when a protocol is being downloaded, the connection fails, and the tool ends up in an error state. This is the most common type of error in the toolset. Out of 299 errors encountered during past nine months, 72 were caused by connection failure. The tools feature robust error recovery mechanisms, allowing them to seamlessly roll back to the last stable state in the event of any encountered errors. In such cases, the problematic protocol is automatically reattempted for download during the subsequent execution of the tool.

4 Gathered data

The resulting preverticals underwent a thorough error check. Corpora were then generated from all preverticals, and an analysis was conducted on the top 100 keywords, as well as the most frequently occurring 500 words in each corpus. This analysis aimed to identify any potential presence of source metadata that might not be part of the protocol text.

As of now, the entire toolset has compiled a total of 1,286.09 million words sourced from 28 chambers within the EU parliaments, out of the 38 chambers available. This collection spans across 17 languages, namely Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Hungarian, Italian, Portuguese, Romanian, Slovak, Slovenian, Spanish, and Swedish. Statistics of each parliamentary chamber can be found in Table 1.

One notable property of some chambers is their grammatical correctness in transcriptions, even though the speaker does not speak grammatically correctly. Therefore, the gathered data are also grammatically correct. This property can be found in chambers such as chambers of the Czech Republic, Slovakia, Ireland, and possibly others, depending on the internal policy of the chamber.

For instance, in the lower chamber of the Czech Republic, transcriptions are transcribed into grammatically correct language, even though transcribed speech contains ungrammatical language. An exception is made for instances of a speaker delivering a strongly emotional speech. Corrections are applied in cases involving incorrect endings or inflection, addressing obvious errors in verbosity, stuttering in speech, and similar linguistic inaccuracies. Obvious mispronunciations are corrected, unless subsequently addressed in the following speeches. Corrections also include addressing the excessive use of personal and demonstrative pronouns, as well as repetition of words, unless such repetition serves an emphatic purpose. It is important to note that there are no corrections made for factual errors or instances of offensive or obscene language.

5 Conclusions

The size of gathered data is continuously growing. In addition to collecting textual data, these tools gather metadata associated with the texts. Common metadata across all sources include the names of the speaker and the date of the speech. Additional metadata is provided for specific chambers, such as notes from the transcriber, party affiliation, the role of the speaker in the chamber, and other relevant details.

However, it is crucial to acknowledge that the quality of the extracted metadata depends on the quality and formatting of the source. Consequently, errors may occur in both the metadata and texts due to the inability to autonomously distinguish between text and metadata in the source. For example, some of the older transcriptions of the German parliament were gathered by OCR, and the resulting scans are sometimes missing a separator of speaker and speech. In the Romania upper chamber of parliament, the role and name of the speaker are sometimes used as the name of the speaker.

References

1. Abrami, G., Bagci, M., Hammerla, L., Mehler, A.: German parliamentary corpus (gerparcor). In: Proceedings of the Language Resources and Evaluation Conference. pp. 1900–1906. European Language Resources Association, Marseille, France (June 2022), https://aclanthology.org/2022.lrec-1.202
2. Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Fišer, D., Pirker, H., Wissik, T., Schopper, D., Kirnbauer, M., Mochtak, M., Ljubešić, N., Rupnik, P., Pol, H.v.d., Depoorter, G., de Does, J., Simov, K., Grigorova, V., Grigorov, I., Jongejan, B., Haltrup Hansen, D., Navarretta, C., Mölder, M., Kahusk, N., Vider, K., Bel, N., Antiba-Cartazo, I., Pisani, M., Zevallos, R., Regueira, X.L., Vladu, A.I., Magariños, C., Bardanca, D., Barcala, M., Garcia, M., Pérez Lago, M., García Louzao, P., Vivel Couso, A., Vázquez Abuín, M., García Díaz, N., Vidal Miguéns, A., Fernández Rei, E., Diwersy, S., Luxardo, G., Coole, M., Rayson, P., Nwadukwe, A., Gkoumas, D., Papavassiliou, V., Prokopidis, P., Gavriilidou, M., Piperidis, S., Ligeti-Nagy, N., Jelencsik-Mátyus, K., Varga, Z., Dodé, R., Barkarson, S., Agnoloni, T., Bartolini, R., Frontini, F., Montemagni, S., Quochi, V., Venturi, G., Ruisi, M., Marchetti, C., Battistoni, R., Darģis, R., van Heusden, R., Marx, M., Depuydt, K., Tungland, L.M., Rudolf, M., Nitoń, B., Aires, J., Mendes, A., Cardoso, A., Pereira, R., Yrjänäinen, V., Norén, F.M., Magnusson, M., Jarlbrink, J., Meden, K., Pančur, A., Ojsteršek, M., Çöltekin, Ç., Kryvenko, A.: Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 3.0 (2023), http://hdl.handle.net/11356/1488, slovenian language resource repository CLARIN.SI
3. Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Grigorova, V., Rudolf, M., Pančur, A., Kopp, M., Barkarson, S., Steingrímsson, S., van der Pol, H., Depoorter, G., de Does, J., Jongejan, B., Haltrup Hansen, D., Navarretta, C., Calzada Pérez, M., de Macedo, L.D., van Heusden, R., Marx, M., Çöltekin, Ç., Coole, M., Agnoloni, T., Frontini, F., Montemagni, S., Quochi, V., Venturi, G., Ruisi, M., Marchetti, C., Battistoni, R., Sebők, M., Ring, O., Darģis, R., Utka, A., Petkevičius, M., Briedienė, M., Krilavičius, T., Morkevičius, V., Bartolini, R., Cimino, A., Diwersy, S., Luxardo, G., Rayson, P.: Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1 (2021), http://hdl.handle.net/11356/1431, slovenian language resource repository CLARIN.SI
4. Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M.C., de Macedo, L.D., Navarretta, C., Luxardo, G., Coole, M., Rayson, P., Morkevičius, V., Krilavičius, T., Darģis, R., Ring, O., van Heusden, R., Marx, M., Fišer, D.: The parlamint corpora of parliamentary proceedings. Language Resources and Evaluation (Feb 2022). https://doi.org/10.1007/s10579-021-09574-0, https://doi.org/10.1007/s10579-021-09574-0
5. Jakubíček, M., Kovář, V.: Czechparl: Corpus of stenographic protocols from czech parliament. In: Proceedings of Recent Advances in Slavonic Natural Language Processing 2010. pp. 41–46. Masaryk University, Brno (2010)
6. Marx, M., Schuth, A., et al.: Dutchparl. a corpus of parliamentary documents in dutch. Proceedings Language Resources and Evaluation (LREC) pp. 3670–3677 (2010), https://pure.uva.nl/ws/files/990556/88437_332665.pdf
7. Ogrodniczuk, M.: Polish parliamentary corpus (2018), http://hdl.handle.net/11321/467, CLARIN-PL digital repository