Proceedings of ParlaCLARIN II Workshop, pages 18–22
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Compiling Czech Parliamentary Stenographic Protocols into a Corpus

Barbora Hladká, Matyáš Kopp, Pavel Straňák
Charles University
Faculty of Mathematics and Physics
Prague, Czech Republic
{hladka, kopp, stranak}@ufal.mff.cuni.cz

Abstract

The Parliament of the Czech Republic consists of two chambers: the Chamber of Deputies (Lower House) and the Senate (Upper House). In our work, we focus on agenda and documents that relate to the Chamber of Deputies. Namely, we pay particular attention to stenographic protocols that record the Chamber of Deputies' meetings. Our overall goal is to continually compile the protocols into the TEI encoded corpus ParCzech and make the corpus accessible in a more user friendly way than the Parliament publishes the protocols. In the very first stage of the compilation, the ParCzech corpus consists of the 2013+ protocols that we make accessible and searchable in the TEITOK web-based platform.

Keywords: Parliament of the Czech Republic, Chamber of Deputies, stenographic protocols, TEI encoding, TEITOK

1. Motivation

Parliamentary data is interesting for social and political scientists, data scientists, historians, linguists, journalists and citizens in general. For a wide range of tasks parliamentary data must be easily findable and accessible, encoded according to international standards and, if possible, with rich and correct annotations and metadata. In the fields of Natural Language Processing and Corpus Linguistics, the CLARIN ERIC infrastructure plays a leading role in the task of compilation of parliamentary data into language corpora. They organized the CLARIN-PLUS cross-disciplinary workshop "Working with parliamentary records" held in Sofia, Bulgaria in 2017 that clearly indicated a need for discussion on processing parliamentary data in a wider community.[1] In 2018 the ParlaCLARIN workshop was organized in Miyazaki, Japan and it means a significant step forward in the given discussion (Fišer et al., 2018).

The CLARIN ERIC infrastructure provides the most comprehensive overview of the existing parliamentary corpora and related publications.[2] As of March 2020, there are 34 parliamentary corpora in the overview and its description, size, licence and availability are provided for each of them.

In our work we focus on stenographic protocols published by the Chamber of Deputies of the Parliament of the Czech Republic. In the past, two Czech parliamentary corpora have been published: (1) CzechParl is a corpus of stenographic protocols recorded during the meetings of both chambers of the Parliament of the Czech Republic between 1993–2010 (Jakubíček and Kovář, 2010). This corpus contains 82 million tokens and it is available for on-line searching in SketchEngine;[3] (2) The Czech Parliamentary Meetings corpus consists of the recordings from the Chamber of Deputies of the Parliament of the Czech Republic made between February – August 2011. This corpus contains 88 hours of speech data and their transcriptions. Both the spoken and written data are available for download (Pražák and Šmídl, 2012) and for on-line searching in the KonText concordancer.[4]

Our goal is to make the protocols accessible and searchable in a more user friendly way than the Czech Parliament does. We make them available in the ParCzech corpus that we, in contrast with other corpora in the CLARIN ERIC overview, approach as a live text collection rather than a static collection. This provides interesting aspects of a workflow design, especially into a procedure of regular updates. Regarding data encoding, the works (Erjavec and Pančur, 2019) and (Pančur et al., 2019) inspired us the most. At the same time we looked at the CLARIN ERIC overview from a different angle and created Table 1 including ParCzech.

This paper is organized as follows. In Section 2. we describe the digital repository of the Czech Parliament and the agenda that is available online on the website of the Czech Chamber of Deputies. The details on recording, editing, and publishing the stenographic protocols by the Czech Parliament are explained in Section 3. In Section 4. we describe our procedure to compile the protocols into ParCzech. Section 5. shows how to search ParCzech in the TEITOK web-based platform.

Terminological note The following terms in parliamentary procedures are relevant for our topic. During a *term*, there are *meetings* which are a group of *sittings* and which typically take place in more than one day. For illustration, the 30th meeting in the 8th term of the Czech Chamber of Deputies was a group of 12 sittings.[5] Each meeting has its own agenda and an *agenda item* is discussed in *speeches* that can be made at more than one sitting.

[1] https://www.clarin.eu/event/2017/clarin-plus-workshop-working-parliamentary-records

[2] https://www.clarin.eu/resource-families/parliamentary-corpora

[3] https://www.sketchengine.eu/czechparl-corpus-of-czech-parliament

[4] https://lindat.mff.cuni.cz/services/kontext/first_form?corpname=czechparl_2012_03_28_cs_w

[5] On 28, 29, 30, 31 May and 4, 5, 6, 7, 18, 19, 20, 21 June 2019, see https://www.psp.cz/eknih/2017ps/stenprot/030schuz/index.htm

language	corpus (url if cannot be downloaded)	concordancer						format	down
		K	**SE**	**Kr**	**N**	**C**	**w**		load
Bulgarian	Corpus of Bulgarian Political and Journalistic Speech (↗)						●	?	
Croatian	Croatian parliamentary corpus ParlaMeter-hr 1.0	●	●				●	TEI	⤓
Czech	Czech Parliamentary Meeting	●						XML	⤓
	CzechParl (↗)		●					?	
	ParCzech	**+T**						**TEI**	⤓
Danish	The Danish Parliament Corpus 2009–2017, v1							TEI	⤓
Dutch	DutchParl						●	XML	⤓
English	HanDeSeT: Hansard Debates with Sentiment Tags							CSV	⤓
	Hansard corpus						●	?	⤓
	Parliamentary Debates on Europe at the House of Commons 1998–2015 (↗)							TEI	
	UKParl Dataset							?	⤓
Estonian	Transcripts of Riigikogu (Estonian Parliament)						●	TEI	⤓
Finnish	Plenary Sessions of the Parliament of Finland			●				?	
French	Archives Parlementaires (↗)							TEI	
	Parliamentary Debates on Europe at the Assemblée nationale 2002–2012							TEI	⤓
German	Korpusbasierte Analyse österreichischer Parlamentsreden							XML	⤓
	ParlAT beta							CSV	
	Parliamentary Debates on Europe at the Bundestag 1998–2015							TEI	⤓
	polmineR corpus						●	TEI	⤓
Greek	Hellenic Parliament Minutes 1989–1994, 1997–2018							text	⤓
	Speeches of Politicians in the Greek Parliament							TXT	⤓
Icelandic	The Icelandic Parliamentary Corpus			●					⤓
Latvian	LinkedSAEIMA (↗)		●					RDF, CoNLL-U	
Lithuanian	Lithuanian Parliament Corpus for Authorship Attribution							CSV	⤓
Norwegian	Proceedings of Norwegian Parliamentary Debates (↗)					●		?	
	Talk of Norway							CSV	⤓
Polish	Polish Parliamentary Corpus				●			TEI	⤓
Portuguese	PTPARL Corpus							TXT	⤓
Slovenian	Slovenian parliamentary corpus ParlaMeter-sl 1.0	●	●					TEI	⤓
	Slovenian parliamentary corpus siParl 1.0	●	●					TEI	⤓
	Slovenian parliamentary corpus SlovParl 2.0	●	●					TEI	⤓
Swedish	Riksdag's Open Data			●				XML	⤓
7 lang.	The ParlSpeech V2 data set								⤓
21 lang.	Europarl: European Parliament Proceedings Parallel Corpus 1996–2011							HTML	⤓

Table 1: A different view on the overview of parliamentary corpora published by the CLARIN ERIC infrastructure on `https://www.clarin.eu/resource-families/parliamentary-corpora` as of 27 March, 2020. For each corpus we provide concordancers through which it is available (**K**-KonText, **SE**-(no)SketchEngine, **Kr**-Korp, **N**-NKJP, **C**-Corpuscle, **w**-dedicated website), its internal format, and an url link if a corpus is available to download. **+T** by ParCzech stands for KonText+TEITOK.

2. Digital repository of the Czech Parliament

Digital repository of the Parliament of the Czech Republic https://public.psp.cz/en/sqw/hp.sqw?k=82 contains recording of the Assemblies since the earliest time of their existence until the last sitting of parliament. It consists of two parts: Bohemian Diet from its first reported (not directly recorded) acts in 1039 until 1848. Various historical periods have variable recordings, but many do contain transcripts. E.g. for the period 1526–1611 we can see by looking for the first period (1526–1545) (https://public.psp.cz/eknih/snemy/v010/) that there is the first correspondence of the Diet and records of the most important acts, mostly elections of Czech kings. After that the content of each Diet follows. The contents are in form of letters, but they are rather detailed and for most assemblies they consists of dozens of documents, arguments and replies, rather well documenting issues of the assembly. There are excuses for not participating due to sickness, there are king's proposals for the diet, diet's replies, e.g. this one concerning help fighting Turks: https://public.psp.cz/eknih/snemy/v010/1545/t032600.htm. For this period of 19 years there are 336 documents. Later diets are documented progressively better. For a diet of February 4–19, 1605 there are 46 documents (https://public.psp.cz/eknih/snemy/v11a/) Between the years 1611 and 1847 very few documents have been digitised, although the diet was active for the whole time.

From 1848, when the Austrian parliament was reformed and first members of the Bohemian Diet were elected also from citizens, the parliaments and their chambers of first Austrian, later Czechoslovak, and currently Czech Parliaments are available in the repository: https://public.psp.cz/eknih/index.htm. For all of these parliaments, protocols of each meeting are available in the repository. For the Austrian period the documents are often in German and are more similar to minutes rather than full transcript.[6] In general, form and quality of the Austrian' era transcripts are very variable, but they might become an interesting resource in future.

Since establishment of the first parliament of the new Czechoslovak Republic in 1918 the available documents are much more extensive. For every sitting, there is a "nest"-style site which has not only full transcripts, but there are also registries of all members of parliament (MOP) and its organs, for each MOP there is a list of their activities in the meetings[7], registry of "parliamentary prints", i.e. documents submitted to the parliament for discussion and vote, etc. lists of committees, and lists of topics in the prints and transcripts. All of these documents are published basically in plain text.[8] This structure remains in general constant all the way until 1989, with only a minor addition of additional documents like invitations to parliamentary sessions.

A substantial improvement of the proceedings has occured with the newly elected House of Deputies for 2006–2010. From the first sitting of this house in addition to transcripts[9], which are necessarily edited at least for fluency of spoken language, also the unedited audio recordings of the sessions are available.[10] Then from the sitting Senate of 2010–2012 also the Senate has improved their data and their transcripts are available in XML (XHTML) with linked votes, audio and even video recordings.[11] A small problem is in the form of published audio and video, which is available for streaming, not for a simple download. However in general we can say that from 2010 all the proceedings of the Parliament of the Czech Republic are available in the rich form of agenda, documents, transcripts of the proceedings, and votes, together with audio recording of the proceedings. All of this can be downloaded from the Digital repository of the Parliament.

3. Stenographic Protocols

The Czech Chamber of Deputies uses stenography to record its meetings like few other countries (Torregrossa, 2016). Stenography allows reporters to take notes during sittings and then they need time to transcribe them listening to the audio recording. The Czech Chamber of Deputies reporters take 10 minute shifts and then they have 80–90 minutes to transcribe their records. The draft versions of stenographic protocols are published online on the same day and it takes several days to do language revisions and get the final versions. Finally the speakers have 2–3 weeks for authorization. Figure 1 presents the protocol that has been already published online but neither correction nor authorization has been done yet.

The language revisions respect differences between spoken and written language. Reporters put focus on incorrect endings and cases, apparently incorrect word order, stuttering, evident slip of the tongue, if not further repaired, excessive use of personal and demonstrative pronouns, word repetition unless it is an intention. On the other hand, editing of factual errors and mistakes is not acceptable. In addition, notes important for capturing the atmosphere of the meeting and the events in the meeting hall are added to the text in brackets only to the extent strictly necessary and as objective as possible. Minor modifications for the purpose of text formatting are permitted. The reporters can neither correct nor replace offensive and indecent words. The speeches

[6] Some protocols, e.g. 1866, have not been processed via OCR because they are typeset in fraktur (gothic) font. Interestingly, they are detailed stenographic protocols and have been published in parallel in German and Czech.

[7] C.f. activities of senators (MOP from the upper chamber) with names starting with 'E' in 1925: https://public.psp.cz/eknih/1925ns/se/rejstrik/jmenny/e.htm

[8] This level of detail is true for the pre-WWII Czechoslovak parliament and post-war federal parliament. National – Czech and Slovak – assemblies only have transcripts available.

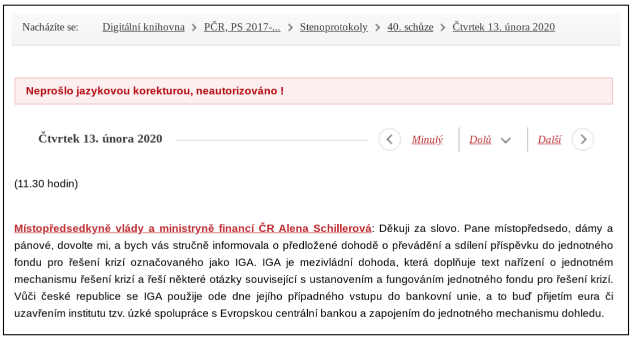
[9] https://public.psp.cz/eknih/2006ps/stenprot/

[10] https://public.psp.cz/eknih/2006ps/audio/2006/

[11] The first XHTML transcript: https://www.senat.cz/xqw/webdav/pssenat/original/66197/55769 and linked vote: https://www.senat.cz/xqw/xervlet/pssenat/hlasy?G=11178&O=8

must remain undistorted and authentic regardless of their content and political affiliation of their speakers.

Spoken and written language differ in many ways, e.g. speech can use timing, tone, volume, and timbre to add emotional context. In order to provide a complete picture of the event, a corresponding part of the audio recording is available for each stenographic protocol.



Nacházíte se: Digitální knihovna > PČR, PS 2017-... > Stenoprotokoly > 40. schůze > Čtvrtek 13. února 2020

Neprošlo jazykovou korekturou, neautorizováno !

Čtvrtek 13. února 2020

(11.30 hodin)

Místopředsedkyně vlády a ministryně financí ČR Alena Schillerová: Děkuji za slovo. Pane místopředsedo, dámy a pánové, dovolte mi, a bych vás stručně informovala o předložené dohodě o převádění a sdílení příspěvku do jednotného fondu pro řešení krizí označovaného jako IGA. IGA je mezivládní dohoda, která doplňuje text nařízení o jednotném mechanismu řešení krizí a řeší některé otázky související s ustanovením a fungováním jednotného fondu pro řešení krizí. Vůči české republice se IGA použije ode dne jejího případného vstupu do bankovní unie, a to buď přijetím eura či uzavřením institutu tzv. úzké spolupráce s Evropskou centrální bankou a zapojením do jednotného mechanismu dohledu.

Figure 1: Neither correction nor authorisation of the given stenographic protocol published online has been done yet, see the information *Neprošlo jazykovou korekturou, neautorizováno!* at the top of the screenshot.

4. Compiling the Protocols into a Corpus

We take the following steps in order to compile the stenographic protocols of the Czech Chamber of Deputies into the ParCzech corpus:

Study source data We identified *what works* and *what does not work* in the protocols of each Chamber of Deputies published during the eight parliamentary terms.[12] Namely we focused on the features of author identification, links to the authors, links to the agenda items, spoken interpellation, availability of audio recordings, browsing the data. Since most of the features work for the protocols between 2013–present we have decided to compile this subset of the protocols first.

Get and encode source data We downloaded the 2013+ protocols and converted them into the ParlaCLARIN TEI based format.[13] Since that moment we call this collection the ParCzech corpus.

One TEI document corresponds to one agenda item. We label the documents in a way that describes a hierarchy of terms, meetings, sittings, and agenda items. All meetings are numbered from 001 onwards for each term, sittings from 01 onwards for each meeting, agenda items from 001 onwards for each meeting. For illustration, the document `2013-001-01-005` is a protocol of speeches on the fifth agenda item (`005`) made in the first sitting (`01`) of the first meeting (`001`) of the term that started in 2013 (`2013`). The document `2013-001-01-003b.u` is a protocol of speeches on the third agenda item made in multiple parts and `b` stands for the second part; the suffix `u` stands for an unauthorized version.

It may happen that one agenda item is being discussed more than once during a current sitting. In other words, an agenda item discussion can be interrupted with a discussion on a different agenda item. But it does not affect our strategy to store one agenda item in a single TEI document. Sitting openings are stored in single TEI documents. The data scrapper is implemented as a Perl script downloading both the newly published protocols (i.e. not authorized yet) and the authorized protocols.

The version of 2013–present ParCzech consists of in 4,689 TEI documents containing 136,888 speeches, 1,312,897 sentences and 23,360,798 tokens.

The Czech Chamber of Deputies publishes audio recordings as mp3 files on its website. We have not aligned these audio files with the TEI documents yet.

Process ParCzech We enrich ParCzech automatically by morphological and named-entity annotations using the procedures MorphoDita[14] and NameTag[15], resp. (Straková et al., 2014). We run MorphoDita with the model MorfFlex CZ (Straka and Straková, 2016) and NameTag with the CNEC model (Straka and Straková, 2014). NameTag classifies entities into a set of 42 classes (called "types") with a very detailed characterization and these fine-grained classes are merged into 7 super-classes (called "supertypes"). In comparison with the ParlaCLARIN TEI elements, the repertoire of the NameTag classes is richer and therefore we introduce new TEI elements not included in the ParlaCLARIN TEI format recommendations.

5. ParCzech in TEITOK

TEITOK is a web-based platform for viewing, creating, and editing corpora with both rich textual mark-up and linguistic annotation.[16] It communicates with the KonText search engine allowing evaluation of simple and complex queries, displaying their results as concordance lines, computing frequency distribution and further work with language data.[17] The ParCzech corpus is downloadable and accessible in TEITOK at

`http://hdl.handle.net/11234/1-3174`

Figure 2 illustrates four different options over which users can browse ParCzech (sitting date, meeting, term, authorized). For example, when browsing over the sitting date users can see that four items of the fifth meeting in the term 2013–2017 were on the agenda on 21 January 2014.

TEITOK uses the Corpus Query Processor (CQP) to query corpora in the CQP query language (CQL).[18] Figure 3 illustrates a query builder that provides an easy way to define queries in CQL. At present, users can formulate queries on words, lemma, part-of-speech tags, named entities, and speakers.

[12] 1993–1996, 1996–1998, 1998–2002, 2002–2006, 2006–2010, 2010–2013, 2013–2017, 2017–present

[13] `https://clarin-eric.github.io/parla-clarin/#sec-intro`

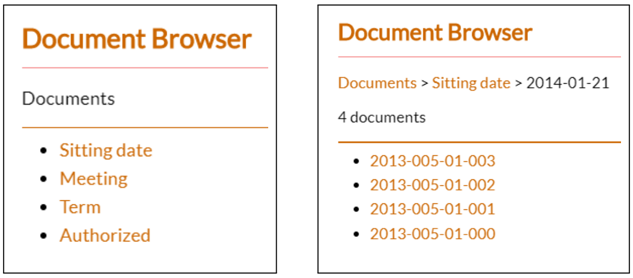
[14] `http://ufal.mff.cuni.cz/morphodita`

[15] `http://ufal.mff.cuni.cz/nametag`

[16] `http://teitok.corpuswiki.org`

[17] `https://lindat.mff.cuni.cz/services/kontext/corpora/corplist`

[18] `http://cwb.sourceforge.net/index.php`



Document Browser

Documents

- Sitting date
- Meeting
- Term
- Authorized

Document Browser

Documents > Sitting date > 2014-01-21

4 documents

- 2013-005-01-003
- 2013-005-01-002
- 2013-005-01-001
- 2013-005-01-000

Figure 2: Browsing the ParCzech corpus in TEITOK

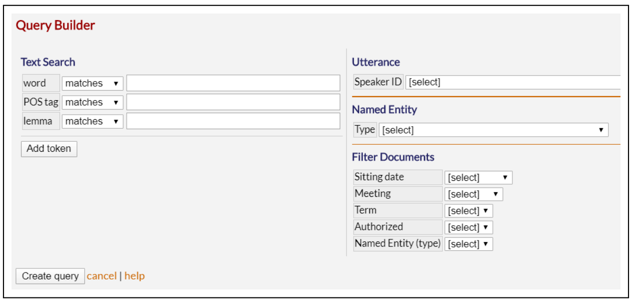


Figure 3: TEITOK interface to query ParCzech

6. Conclusion

Publishing the proceedings in the form of coherent and annotated dataset is also important from the perspective of data accessibility. While the Library of the Parliament of the Czech Republic has done a very good job in publishing all of the material, it is still available in a complicated and easily broken form. Making it all available as not only an online searchable service, but also a downloadable and citable collection available with a PID via a certified data repository will significantly improve the accessibility of the data and its availability for further research.

We have designed and implemented a procedure to compile the Czech stenographic protocols into a corpus which we call the ParCzech corpus. The corpus is accessible and searchable in the TEITOK tool and it is directly downloadable. However, our compilation pipeline is not fully tuned. Mainly we have to concentrate on studying the protocol flow in the Digital repository of the Parliament of the Czech Republic since it affects the procedure of ParCzech regular updates. Once we fix it, we will focus on interlinking ParCzech with other data sources.

Acknowledgements

We wish to thank Maarten Janssen, the author of TEITOK, for his advice on the stenographic protocol representation and on uploading the protocols into TEITOK. This work has been using language resources and tools developed and/or stored and/or distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101).

7. Bibliographical References

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

Darja Fišer, et al., editors. (2018). *Proceedings of LREC2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora*. European Language Resources Association (ELRA), Paris, France.

Jakubíček, M. and Kovář, V. (2010). CzechParl: Corpus of Stenographic Protocols from Czech Parliament. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2010*, pages 41–46.

Straková, J., Straka, M., and Hajič, J. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, June. Association for Computational Linguistics. http://www.aclweb.org/anthology/P/P14/P14-5003.pdf.

Torregrossa, G. (2016). The production of parliamentary reports – a research about the methods used in different countries. IPRS. https://issuu.com/iprs/docs/torregrossa2016.

8. Language Resource References

Pančur, A., Erjavec, T., Ojsteršek, M., Šorn, M., and Blaj Hribar, N. (2019). Slovenian parliamentary corpus siParl 1.0 (1990–2018). Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1236.

Pražák, A. and Šmídl, L. (2012). Czech Parliament Meetings. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11858/00-097C-0000-0005-CF9C-4.

Straka, M. and Straková, J. (2014). Czech models (CNEC) for NameTag. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11858/00-097C-0000-0023-7D42-8.

Straka, M. and Straková, J. (2016). Czech models (MorfFlex CZ 161115 + PDT 3.0) for MorphoDiTa 161115. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-1836.