Czech Parliamentary Data in a Corpus

Text Mining Parliamentary Data Seminar on “Practices of Parliament”

March 18, 2021

Barbora Hladká (hladka@ufal.mff.cuni.cz)
Institute of Formal and Applied Linguistics
Charles University, Prague, Czech Republic

Czech Parliamentary Data in a Corpus

A corpus is a collection of language data compiled from either written texts 
(or transcriptions of recorded speech) or recorded speech  .

An annotated corpus  is a text corpus enriched with linguistic information.

We do prefer to have corpora

- available as datasets for further processing 
- accessible through user friendly systems 
- searchable through user friendly systems 

Czech Parliamentary Data

The Parliament of the Czech Republic (PCR)

- Senate (The Upper House), Chamber of Deputies (The Lower House)

Digital repository of PCR

 10th century - present

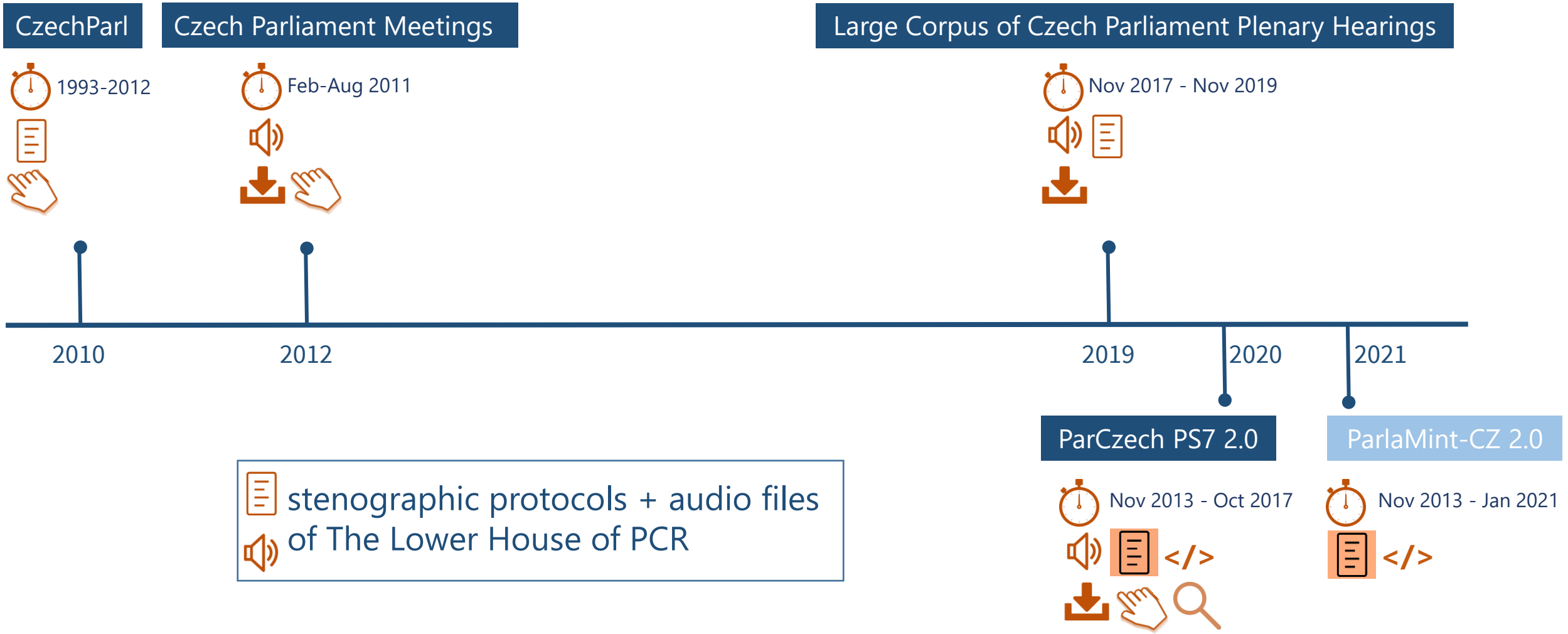
- Stenographic protocols (data)
- Audio files (data)
- Voting (metadata)
- Members of Parliament (metadata)
- …

Digital repository of PCR https://www.psp.cz/eknih

The data are accessible but

- in user/machine unfriendly way

Corpora of Czech Parliamentary Data as of March 18, 2021



ParCzech

http://ufal.mff.cuni.cz/parczech

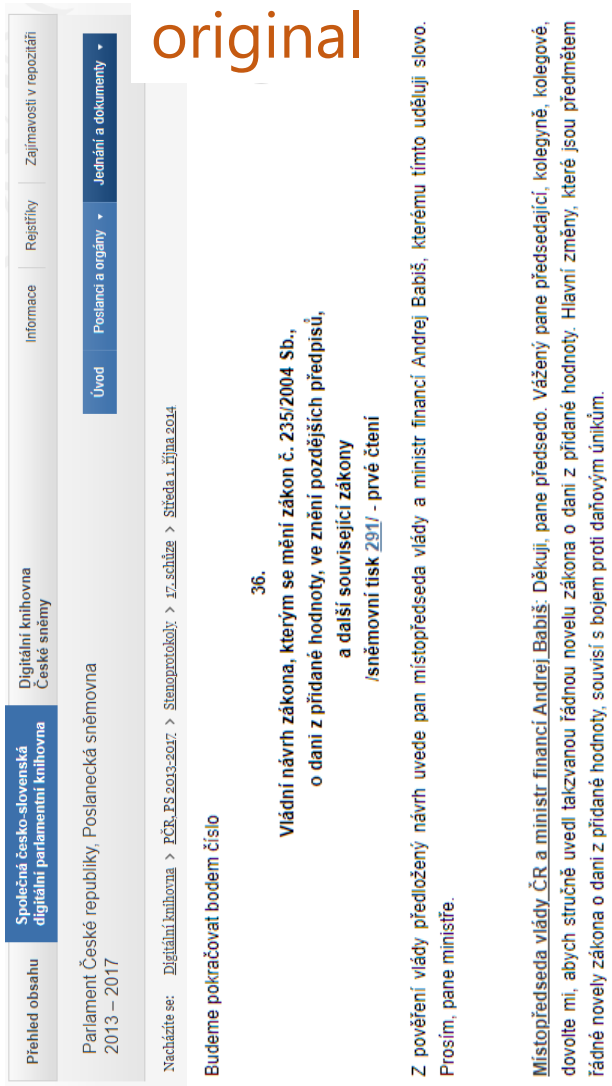
ParCzech = a project of compiling Czech Parliamentary Data into Corpora, started in 2020

- coordination (fragmented activities so far)
- creation of
  - static corpora for limited time periods 
  - live corpora, ideally of all the Czech parliamentary data 
  - using a pipeline
    1. data scraping
    2. format conversion
    3. checking scripts
    4. manual checks
    5. linguistic annotation (UDPipe)
    6. adding metadata

ParCzech PS7 2.0

http://lindat.cz/services/teitok/parczech-ps7-2.0

original



ParCzech

TEITOK

Login

Available Corpora

ParCzech PS7 2.0

Browse

CQL Search

Search in KonText

Download

Older Version

ParCzech PS7 1.0

Powered by TEITOK
Maarten Janssen, 2014-

ps2013-017-09-003-036.tt

Parliament of the Czech Republic, Chamber of Deputies

Agenda Item Title	**36. Vládní návrh zákona, kterým se mění zákon č. 235/2004 Sb., o dani z přidané hodnoty, ve znění pozdějších předpisů, a další související zákony /sněmovní tisk 291/ - prvé čtení**
Date	2014-10-01
Meeting	ps2013/017
Agenda Item	ps2013/017/036
Authorized	yes
Source	https://www.psp.cz/eknih/2013ps/stenprot/017schuz/s017357.htm

View options

Text: Transcription | Written form - Show: Colors - Tags: PoS | Tag | Features | Lemma

index Page 1 > 2

0:00 / 13:58

Předseda PSP Jan Hamáček

36. Vládní návrh **zákona**, kterým se mění zákon č. 235/2004 Sb., o dani z přidané hodnoty, ve znění pozdějších předpisů, a další související zákony /sněmovní ti

Z pověření vlády edseda vlády a ministr financí Andrej Babiš, kterému tímto uděluji slovo. Prosím, pane ministře.

zákona

PoS	NOUN
Tag	NNIS2-----A----
Features	Animacy=Inan, Case=Gen, Gender=Masc, Number=Sing, Polarity=Pos
Lemma	zákon

Místopředseda vlády

Děkuji, pane před egyně, kolegové, dovolte mi, abych stručně uvedl takzvanou řádnou novelu zákona o dani z přidané hod tem řádné novely zákona o dani z přidané hodnoty, souvisí s bojem proti daňovým únikům.

Target audience

Monolingual (ParCzech, ParlaMint) data available and accessible to

- researchers
  - (corpus) linguistics
  - political science, history, …
  - speech recognition - more data for training and evaluating ASR (Kratochvíl et al, 2020)
- data journalists
- organizations on transparency in public administration, e.g., https://www.hlidacstatu.cz/

- We ask for feedback from The Parliament of the Czech Republic

References

- Hladká B., Kopp M., Straňák P. Compiling Czech Parliamentary Stenographic Protocols into a Corpus. In: *Proceedings of the LREC 2020 Workshop on Creating, Using and Linking of Parliamentary Corpora with Other Types of Political Discourse* (ParlaCLARIN II), pp. 18-22, European Language Resources Association (ELRA), Paris, France. pdf

- Kratochvíl J. , Polák P., Bojar O. Large Corpus of Czech Parliament Plenary Hearings. In: *Proceedings of the 12th Conference on Language Resources and Evaluation* (LREC 2020), pages 6363–6367 Marseille, 11–16 May 2020. pdf

The Parliament of the Czech Republic (PCR)

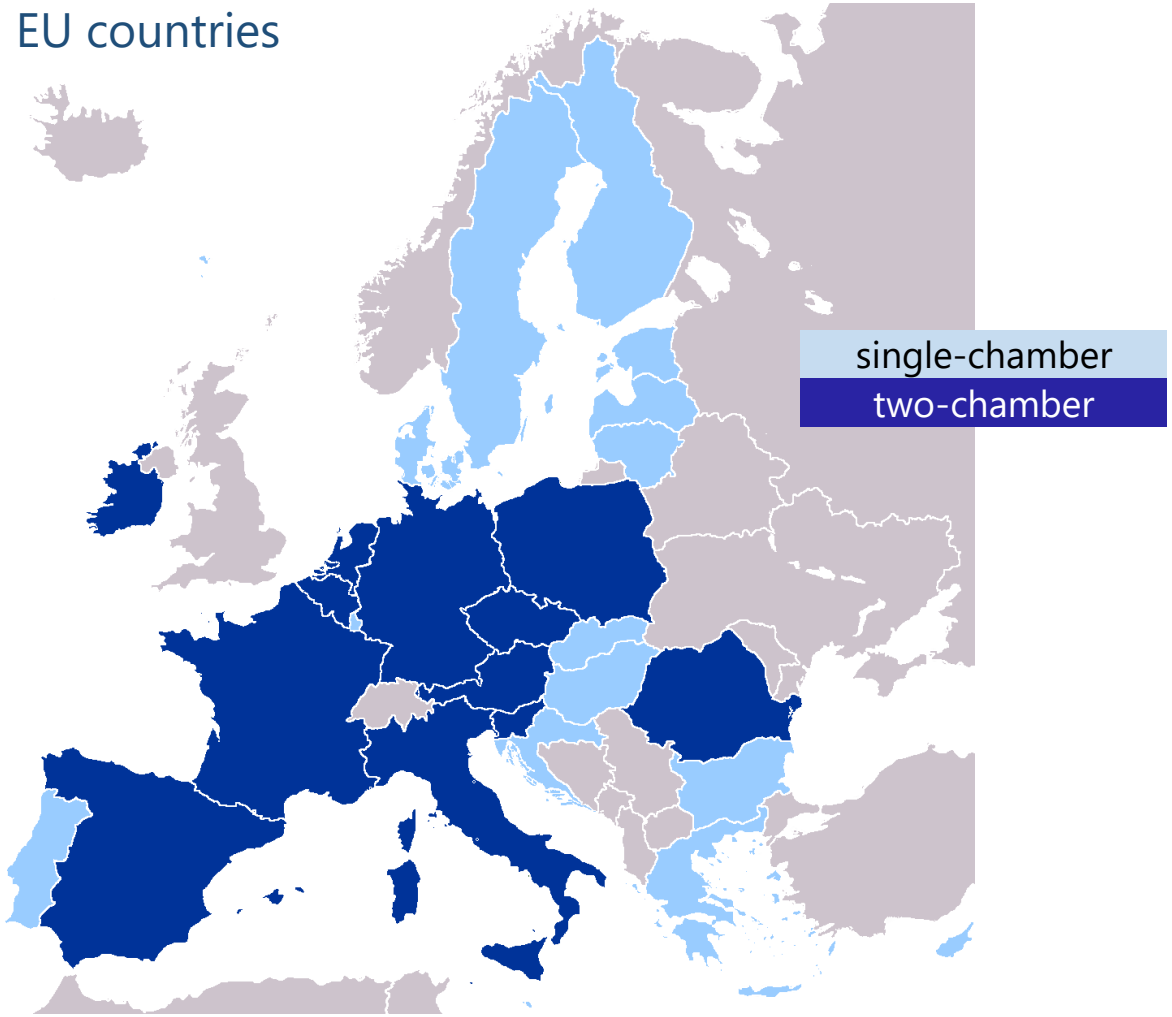
The Upper House: Senate

SENÁT
PARLAMENTU ČESKÉ REPUBLIKY

The Lower House: Chamber of Deputies

POSLANECKÁ SNĚMOVNA
PARLAMENTU ČESKÉ REPUBLIKY

EU countries



Digital repository of PCR

- Bohemian Assemblies and their development until the start of the 15th Century
- Bohemian Assemblies during the Hussite period until mid-15th Century
- Bohemian Assemblies during the Estate Monarchy until the start of the 17th Century
- Bohemian Assemblies during the Habsburg Absolutism until 1848
- Austrian Empire Council until 1918
- National Assembly of the Czechoslovak Republic until 1992
- Parliament of the Czech Republic