ParCzech 3.0: A Large Czech Speech Corpus with Rich Metadata

Matyáš Kopp[0000−0001−7953−8783], Vladislav Stankov[0000−0002−2034−6485], Jan Oldřich Krůza[0000−0002−7510−6470], Pavel Straňák[0000−0002−6895−8536], and Ondřej Bojar[0000−0002−0606−0050]

Charles University, Faculty of Mathematics and Physics, ÚFAL
Malostranské nám. 25, Praha 1, CZ-11800, Czech Republic
surname@ufal.mff.cuni.cz

Abstract. We present ParCzech 3.0, a speech corpus of the Czech parliamentary speeches from The Czech Chamber of Deputies which took place from 25th November 2013 to 1st April 2021.
Different from previous speech corpora of Czech, we preserve not just orthography but also all the available metadata (speaker identities, gender, web pages links, affiliations committees, political groups, etc.) and complement this with automatic morphological and syntactic annotation, and named entities recognition. The corpus is encoded in the TEI format which allows for a straightforward and versatile exploitation.
The rather rich metadata and annotation make the corpus relevant for a wide audience of researchers ranging from engineers in the speech community to theoretical linguists studying rhetorical patterns at scale.

Keywords: Czech speech corpus · TEI · speech corpora · speech recognition · parliamentary debates · Parliament of the Czech Republic.

1 Introduction

Public sessions such as parliamentary hearings have been a great source of data for natural language processing ever since [10] and the interest to use them is still steadily growing, see e.g. the ParlaMint initiative.[1] For the data collection to be readily usable in research, considerable effort has to be spent on its processing – and such processing often differs depending on the intended use.

The aim of this work is to unify the efforts in collecting Czech corpora from The Czech Chamber of Deputies and make the data usable for both training of automatic speech recognition systems (ASR) as well as for corpus studies at various layers of linguistic description. Specifically, we process all the hearings which took place from 25$^{\text{th}}$ November 2013 to 1$^{\text{st}}$ April 2021, covering the whole 7$^{\text{th}}$ and the majority of the 8$^{\text{th}}$ term of the chamber.

[1] https://www.clarin.eu/event/2020/clarin-cafe-join-our-parliamentary-flavoured-coffee-parlamint

Being short of extensive resources for manual annotation, we have to rely on what has been provided by the chamber of deputies, and on our automatic tools. The recorded speeches were published with manually revised transcripts, which however sometimes correct or improve the actual wording in the sound. Occasionally, errors in the collection lead to mismatching audio and transcript files and other processing errors. We have to work with such an input and we at least attempt to automatically identify which recordings are flawless and in which more errors are to be expected.

Section 2 briefly reviews existing corpora derived from Czech parliamentary data. Section 3 details our methodology, including the alignment between the transcripts and the speech. Section 4 presents the corpus in its TEI and speech formats, including corpus statistics. Section 5 concludes the paper.

2 Related work

Czech parliamentary data is being used in speech recognition since [5] and [9] in 2010. The first unaligned mp3 files were released on the web of the Chamber of Deputies 2006.[2] Some of the derived corpora are motivated exclusively by training of speech recognition systems. They provide as many aligned short text and speech segments as possible, focusing on the verbatim match between the sound and the transcription and dropping segments aligned less reliably. Aspects important to other uses, including letter case and punctuation are often disregarded. General corpora, on the other hand, lack the necessary segmentation of speech and are thus convenient for search or linguistic analyses, but not for training of speech processing systems.

The 7^{th} term of The Czech Chamber of Deputies has been published in ParCzech PS7 1.0 [2] and ParCzech PS7 2.0 [3]. They both cover the period from 25^{th} November 2013 to 16^{th} October 2017.

ParCzech PS7 1.0 is encoded in TEI-based XML format suitable for TEITOK [6] document search and visualisation platform. Annotations on the morphological layer are done with MorphoDiTa and automatic named entity recognition with NameTag [12].

ParCzech PS7 2.0 covers the same data, but it improves the cleaning process, keeps links to the original data and hypertext links in the text. Morphological and syntactic analyses, and named entity recognition are improved by using UDPipe 2 [11] and NameTag 2 [13], resp. Data is additionally distributed in TEI format (Text Encoding Initiative, [14]).

Both corpora provide the original audio data, but no alignment between the sound and text was done at the time.

Recently, a large speech corpus based on Czech parliament plenary sessions was described in [7]. The corpus uses data from November 2017 till November 2019 and consists of approximately 444 hours of audio and corresponding transcription and speaker information.

[2] https://www.psp.cz/eknih/2002ps/audio/2006/01/17/index.htm

An attempt to use the Czech parliamentary meeting recordings as training data for ASR has been also published at 2019's FedCSIS [8]. The work presents an alignment system that exploits GMM-based speech recognition featuring word-level alignment, which is then used to align the original stenographs to the recordings by means of comparing the stenograph with the ASR output with Levenshtein distance. [8] presented primarily a system intended to gather training data for another speech recognition system, hence the ambition is different than in our case. Whereas we attempt to provide the parliament data as a compact, annotated corpus, [8] have no need to align the whole corpus. Significant parts of the data, actually the majority of sentences, are discarded in favour of precise phoneme-grapheme matching which is essential for ASR training data.

3 Methodology

Our work builds upon [4] and improves it by using the more standardized TEI format [1], more detailed speaker metadata, speaker affiliations to different parliamentary groups (political groups, committees, commissions, delegations, etc.) or other institutions, and finally audio alignment.

3.1 Data gathering

The official web of The Czech Chamber of Deputies contains stenographic texts from the chamber of deputies' sittings that are structured into terms, meetings, sittings, and agenda items.

There are two possible ways to get the texts. The first option is downloading the official zip archives[3] and the second one is scraping directly the web.[4] We have chosen the second option because the first one misses some metadata, e.g. links to audio, voting, and parliamentary prints.

The scraped data are unfortunately not encoded consistently across the whole web. Our downloading procedure is thus based on an ideal structure of the web page that should contain speaker's name and a hypertext link to the profile web page, sitting date, and the link to the corresponding audio recording in the side menu. If some data is missing on a real page, downloading procedure tries to find it on other web pages. The missing audio link can be taken from the list of audio files.[5] If it is not listed there, our downloading script guesses the missing URL from the date and time mentioned on the imperfect web page. In this way, we gain the maximum number of audio files, even those that are not linked.

3.2 Data processing and annotation

We use a Perl script to encode stenographic protocols directly to TEI format. Speaker names, personal web links, and hypertext links in texts are preserved.

[3] https://psp.cz/eknih/2013ps/stenprot/zip/index.htm
[4] https://psp.cz/eknih/2013ps/stenprot/index.htm
[5] e.g. https://psp.cz/eknih/2013ps/audio/index.htm

Furthermore, dates and times stored in stenographic notes are decoded. Notes are emphasized and subsequently annotated as phenomena or occurrences. For example *(Ministr Babiš přinesl kopie účtenek k řečnickému pultu. Veselost v sále pokračuje.)* is annotated as `<kinesic type="kinesic"><desc>(Ministr Babiš přinesl kopie účtenek k řečnickému pultu. Veselost v sále pokračuje.)</desc></kinesic>`

Persons are not identified with a unique identifier in the protocols. Luckily, members of parliament and government usually have assigned hypertext links that can help person identification. So we draw on two sources of data: the Czech government web (www.vlada.cz) which contains short bibliography of members of government, and the database of the Chamber of Deputies[6] which contains personal data of not only members of the lower chamber, but also some senators.

We join these source data based on the first name, surname, and birth date and then generate a unique ID for each person. We include other personal data if available: sex, website, Facebook link, profile photo URL, and affiliation to different organizations with a role: president, minister, member, observer, verifier.

At this point of corpus building procedure, we have text data in the TEI format segmented to utterances and paragraphs.[7]

For ease of use in corpus and other studies, we process all the data with UDPipe 2 to provide tokenization and morphological and syntactic analysis. Furthermore, named entities are automatically annotated with NameTag 2.

3.3 Sentence- and word-level alignment with speech

For the purposes of the training of speech recognition systems, a speech corpus needs to be broken into segments of not more than a few dozens of seconds, each equipped with the transcription. Such a segmentation can be achieved with relatively simple detection of silence.

Given the proper segmentation of the text into paragraphs and sentences, we prefer to cast this segmentation also on the speech signal. We automatically identify time spans for each word and then follow the segmentation into sentences as given in the transcript to break the sound.

To extract word timings from the audio, a GMM-based ASR system [8] was used. This system outputs recognized words with their timestamps, given the original stenographic transcripts. Since spoken language differs from the written, sometimes recordings do not match the transcription exactly. Our procedure is not a 'forced decoding' per se, which would find the best alignment and strictly adhere to the words in the transcript. Instead, it proposes a (time-stamped) sequence of words which is close but not necessarily identical to the expected transcript. We find the alignment between the sound and the transcript by matching the two sequences of words, the transcript and the predicted ones, taking each word as an atomic unit. If words do not match, a penalty is given, multiplied by character-level edit distance between these words. One possible choice of the

[6] https://psp.cz/sqw/hp.sqw?k=1300

[7] Paragraphs are made by stenographers and can be revised by speaker.

algorithm for sequence alignment is Needleman–Wunsch algorithm with affine gap penalties. The reward for match is the length of the matched words; for mismatch, a multiple of the edit distance of the two words is subtracted. Experiments showed that the multiple can be set to 3. The algorithm utilizes two parameters: start gap penalty and extend gap penalty. These were set to -5 and -4 based on our experiments with a small subset of the data.

Known issues For an unknown reason, the ASR system may fail to recognize a small part in the original audio file, but in most cases this can be solved by providing the failed audio segment again with some silence added at the beginning. Using a greedy approach, the best recognition can be selected by maximizing the alignment score. Not more than 7% of the corpus (in terms of duration) was corrupted by the failure and this simple technique can recover more than 70% of the failed part. As observations showed, in most cases the unresolved failed segments are true silences in the recordings.

Another issue is that the ASR system may mishear some words, skip them or output phonetically similar words. Thus, once the alignment is done and visualised, one can notice that some words were recognized as an n-tuple of words which as a whole is phonetically similar to the original word. The fix itself boils down to detecting when the original word is surrounded by empty strings and then gluing up transcribed words together, minimizing the edit distance.

After the post-processing, the original audio file can be split into segments, where each segment will have its corresponding transcription. The segments are usually sentences, but sometimes one segment may contain more than one sentence. This happens when the last word of the sentence and the first word of the next sentence are not recognized. In this situation, the sentences are merged into a larger segment.

4 Corpus description

Here we describe the final layout of the corpus as released. We distribute the corpus in three variants: (1) the original HTML and audio files, (2) a large collection of individual files, short segments of audio and transcript useful for training of speech systems, and (3) TEI-encoded texts with explicit references to the audio files and the annotation described in Section 3.2.

4.1 Source HTML and audio files

To allow for a complete revisit of our extraction procedures, we provide all HTML files that our downloading script visited, in the original tree structure. There are two main sub-trees. The first one contains all stenographic protocols with original pages, and the second one contains the directory structure of a list of audio files.

The original audio files are referenced from the TEI version of the corpus and thus should be seen as an inherent part of our corpus. This is important

because many older audio files are no longer available for download even if the audio link still exists on the parliamentary website.

4.2 Version for ASR

In the ASR version of the corpus, each original recording is represented by a folder, containing segment folders and a file with global statistics. These statistics are about alignment of the whole stenographic transcription to the recognized words. Each segment folder contains the following files:

- **Audio file** This is a `.wav` file corresponding to the segment.
- **Pretty transcript** Stenographic transcription of the audio file with the original letter casing and punctuation.
- **ASR transcript** Stenographic transcription of the sound file in upper case and with no punctuation.
- **Words information** Information about words inside the segments in a `.tsv` file. This detailed format provides each word with its starting and ending time (both can be -1 in case if the word was not recognized), normalized Levenshtein distance between the stenographic transcript (which appears in this `.tsv` file) and the recognized word (which is no longer shown but it served as the basis for the estimation of the timestamps), normalized duration that is computed as duration of the word in seconds divided by the number of characters, and also the speaker information. Additionally, there is the ID for each word, so one can find any word in TEI data (see below) and extract additional morphological and syntactic information.
- **Speakers** This files lists speakers' IDs in the segment. These IDs then can be used to extract additional information about speakers like age or sex.
- **Statistics** The file contains statistics that describe the segment and can be useful for data analysis, e.g.: the number of words, number of characters, percentage of missed words, or percentage of missed characters. We also report sound coverage, computed as percentage of sound where some word was recognized, divided by the duration of the segment, and "end correctness" which signalizes whether the end of the segment was not recognized; this is only an issue for the last segment. We note that the percentage of missed characters is the sum of the lengths of the missed words normalized by the lengths of all words in the segment, hence missing shorter words will not influence this statistic much.

Sometimes the alignment between the recognized audio file and the stenographic transcript fails. We provide global statistics for each original audio file to help identifying these cases, for example: percentage of missed words (words that are aligned to empty strings), median normalized edit distance and also 80^{th} percentile of the normalized edit distance. Normalized edit distance is 1 if the true word is aligned to the empty string, meaning that original word is not recognized. To avoid overoptimistic edit distance results, words of length less than 3 are ignored in this calculation. To detect transcripts that poorly, if at all,

match the audio recordings, we created a statistic that counts each sequence of words missed in the recognized audio as one gap, instead of counting each word separately. The intuition is to ignore the gap size since bad alignment is detected better from gap frequency than from the gap size. We normalize the number of continuous gaps to the interval $[0; 1]$ by dividing it by the number of words in the stenographic transcription plus the number of continuous gaps.

Using the provided file-level and segment-level statistics one can filter the data, trading their size for quality. We make a recommendation and provide a filtered version of the corpus. For the filtered version, 2% of the audio recordings were first removed based on the statistic of normalized continuous gaps. Then we preserved only segments with correct endings, percentage of missed characters lower than 6.5%, sound coverage above 62.5%, 80^{th} percentile of the normalized edit distance below 30% and a small enough standard deviation of the normalized edit distance. Additionally, too short (duration lower than 0.82 seconds) and too long (duration higher than 54 seconds) segments were discarded.

Because we release full data, too, one can create custom filtering.

4.3 TEI encoding

As mentioned above, one of the provided corpus variants is in ParlaCLARIN TEI format [1]. ParlaCLARIN is a set of guidelines that provide recommendations on the encoding of various issues, but there are still many possibilities of how a single issue can be handled. In this section, we describe the encoding choices in our corpus.

Directories and files. We compiled two versions of TEI data. The first one is a raw text version that contains all persons and organizations' metadata, categorized stenographers' notes, hypertext links from the source text, and links to source data. The second version extends the raw version with linguistic annotations and the alignment to audio.

There is one main file in each corpus version, that glues all related files together. We refer to the file as the "teiCorpus file". The two teiCorpus files are `ParCzech.xml` and `ParCzech.ana.xml` for raw and annotated versions, respectively. The header of the file contains definitions of taxonomies used in particular TEI files and lists of persons and organizations. A sequence of XML `<include>` elements comes after the header. Files are sorted by dates. Included TEI files are structured into directories – each directory represents a meeting in a term, e.g. `ps2013-001` contains data from the first meeting in the term that starts in the year 2013.

Users who want to only work with data from the complete 7^{th} term can use teiCorpus files `ParCzechPS7.xml` or `ParCzechPS7.ana.xml`. These files include only TEI files related to the 7^{th} term.

Every TEI file contains continuous stenographic notes from agenda items or initial sitting speech. Filenames consist of dash-separated parts: the first part contains the starting year of given term (`ps2013`), and then follow meeting number (`001`), sitting number (`02`), order of agenda item within sitting day

(003), and agenda item (008). So the full pathname of the given example is ps2013-001/ps2013-001-02-003-008.xml for the raw version of TEI files. The annotated version has the suffix .ana.xml. Single agenda items can be discussed multiple times in a sitting day, so the last part of the filename (008) shouldn't be unique. If a user of the corpus wants to follow one particular topic, i.e. the 8th agenda item of the 1st meeting in term starting at the year 2013, they can use this pattern ps2013-001/ps2013-001-??-???-008.xml to filter all TEI files containing the discussed topic. The disadvantage of this solution is that the chairman's speech is split at the topic change point.

Encoding. Each TEI file contains two parts. The first one stored in <teiHeader> element contains metadata about file content and the second one (<text>) contains real stenographic protocols and timelines for each audio file, see Figure 1.

```
<?xml version="1.0" encoding="utf-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="ps2013-001-02-003-008.ana" (···) >
  <teiHeader>
    <fileDesc>
      [···] <!-- title, version, measurements, publisher, license and release date -->
      <sourceDesc>
        ⊞<bibl> <!-- source description, source url, and date -->
        <recordingStmt>
          <recording type="audio"> <!-- list of media files -->
            <media xml:id="ps2013-001-02-003-008.audio1" mimeType="audio/mp3"
               source="https://www.psp.cz/eknih/2013ps/audio/2013/11/27/2013112711581212.mp3"
               url="2013ps/audio/2013/11/27/2013112711581212.mp3"/>
          </recording>
        </recordingStmt>
      </sourceDesc>
    </fileDesc>
    [···] <!-- description of file, elements statistics, where and when sitting sets -->
  </teiHeader>
  <text>
    <body>
      ⊞<div type="debateSection"> <!-- steno -->
      <!-- list of timings for each audio file: -->
      ⊞<timeline (···) corresp="#ps2013-001-02-003-008.audio1" cert="0.891">
    </body>
  <text>
<TEI>
```

Fig. 1. TEI file structure example

The part of the TEI file with stenographic protocols is further divided into speeches by individual members of parliament, government and guests. A single speech is represented with the <u> element and annotated with attributes that determine the speaker's identity, role (chair, regular, or guest), and speech identification with XML ID. Speeches then contain segments (<seg>). Each segment has one or more sentences <s> with words <w> and punctation <pc>.

TEI does not allow start and end synchronisation attributes in word elements so we use anchors <anchor> to encode word-level timing.

Figure 2 illustrates the synchronization of the first two words of sentence "Informace mandátového a imunitního výboru o ověření platnosti volby poslanců.".

```
<s xml:id="ps2013-001-02-003-008.u1.p1.s2">
  <anchor synch="#ps2013-001-02-003-008.u1.p1.s2.w1.ab"/>
  <w xml:id="ps2013-001-02-003-008.u1.p1.s2.w1" (···)>Informace</w>
  <anchor synch="#ps2013-001-02-003-008.u1.p1.s2.w1.ae"/>
  <anchor synch="#ps2013-001-02-003-008.u1.p1.s2.w2.ab"/>
  <w xml:id="ps2013-001-02-003-008.u1.p1.s2.w2" (···)>mandátového</w>
  <anchor synch="#ps2013-001-02-003-008.u1.p1.s2.w2.ae"/>
  [···]
</s>
```

Fig. 2. TEI words synchronization example

```
<timeline unit="ms" origin="#ps2013-001-02-003-008.audio1.origin"
     corresp="#ps2013-001-02-003-008.audio1" cert="0.891">
  <when xml:id="ps2013-001-02-003-008.audio1.origin" absolute="2013-11-27T11:58:00"/>
  <when xml:id="ps2013-001-02-003-008.u1.p1.s2.w1.ab" interval="388290.0"
        since="#ps2013-001-02-003-008.audio1.origin"/>
  <when xml:id="ps2013-001-02-003-008.u1.p1.s2.w1.ae" interval="388900.0"
        since="#ps2013-001-02-003-008.audio1.origin"/>
  [···]
</timeline>
```

Fig. 3. TEI timeline example

The first `<anchor>`, referring to ID `ps2013-001-02-003-008.u1.p1.s2.w1.ab`, points to the time the word "Informace" starts (the suffix `ab` stands for "audio begin") and the second anchor (referring to `ps2013-001-02-003-008.u1.p1.s2.w1.ae`) determines the word's ending time (the suffix `ae` stands for "audio end"). If the alignment algorithm failed to find a good match, the anchor is missing. Sentence-level timing is determined from the first and the last anchor in the sentence. An advantage of the representation using anchors instead of tag attributes is the possibility to add data from other speech-text aligners to the same TEI file.

Formally, the interesting times for each audio file are defined an accompanying `<timeline>` (Figure 3). Here the `origin` attribute refers to the `<when>` element that contains the exact absolute time and date of the recording. Further `<when>` elements define targets for the `ab` and `ae` anchors relative to this origin.

The attribute `cert` with values from interval $[0, 1]$ determines the level of certainty based on the 80^{th} percentile of normalized edit distance subtracted from 1 or `cert` has been set to 0 for timelines corresponding to 2% of statistic of normalized continuous gaps filtered audio files, described in Section 4.2.

4.4 Division into training, development and test sets

For the purposes of ASR training, we extract three pairs of development (dev) and evaluation (test) sets, and one common training (train) set. All these sets were created from segments which passed our filters. Technically, this is realized using a simple filelist which specifies the destination set for each segment. Consequently, all these sets are disjoint.

The dev and test sets were created for three different purposes:

Speakers Dev and Test were extracted from the clean data first, taking all utterances of a few speakers. This dev and test are thus useful in experiments, where you want to assess system performance on unseen speakers. The

proportion of men and women in this dev and test set is artificially balanced, oversampling women compared to the corpus average.

Context Dev and Test were formed in a way that preserves partitioning from original audio recordings. Thus, few audio recordings were taken out from the clean data and all their segments put into the context dev or test set. This way, the context of each utterance is available and discourse phenomena can be studied up to the level of the original division info files (and subject to filtering). n

Segments Dev and Test were created from the rest of filtered data by sampling random segments.

4.5 Corpus statistics

Table 1 shows statistics counted on the annotated TEI files. The corpus contains all speakers in the focused period and in addition also the members of parliament of the 7^{th} and 8^{th} term who did not have a speech. The number of source audio files and source web pages do not match, because some of the audio files are not available.

Table 1. ParCzech 3.0 statistics for the TEI format

Number of TEI files	5 409
Number of utterances	154 460
Number of sentences	1 479 990
Number of words	22 546 417
Number of unique persons	486
Number of source audio files	20 674
Number of steno source web pages	20 775
Source audio length (hours incl. overlaps)	4 815.31
Time period	25^{th} Nov 2013 – 1^{st} Apr 2021

Table 2 compares the cleaned corpus for training and testing ASR models with original data. After applying the filtering as described in Section 4.2, the correctly aligned audio files amount to 1 332.38 hours and 606 540 segments. The average duration of each filtered segment is 7.90 seconds with the standard deviation of 7.14 seconds. Each segment consists of 16.72 words on average (with a standard deviation of 13.73 words); punctuation is not included in these word counts. In total, the corpus contains 10 146 591 words. After filtering the percentage of words aligned to the sound increased from 89.6% to 96.3%. We can also see that duration range and segment size range are smaller after filtering.

Table 3 summarizes the sizes of our divisions of the filtered data.

5 Conclusion

We presented ParCzech 3.0, a sizeable speech corpus of Czech which preserves and formalizes as much metadata as possible (speakers and their gender, struc-

Table 2. ParCzech 3.0 statistics for the ASR format before and after data cleaning

	Original data	Filtered data
Hours	3 071.57	1 332.38
Segments	1 391 785	606 540
Average segment duration in seconds	7.94±11.53	7.90±7.14
Average number of words in a segment	15.91±16.32	16.72±13.73
Words	22 153 778	10 146 591
Aligned words percentage	89.6%	96.3%
Unique Speakers	475	474
Segment size range	[1, 1058]	[2, 138]
Duration range	[0.0, 720.76]	[0.82, 53.99]

Table 3. Statistics for the ASR train, test, dev sets. "Files" shows the number of original mp3 recordings that contributed to the given set; some files have contributed to more than one set. "Segment Duration" in seconds (average and std. dev.).

	Total			Segment	Words		Unique
Set	Segments	Files	Hours	Duration	Per Segment	Total	Speakers
Train	579 169	19 931	1 271	7.9±7.1	16.7±13.7	9 679 268	417
Speakers Dev	4 596	744	10.7	8.4±7.1	17.8±13.4	81 708	30
Speakers Test	4 261	689	10.6	9.0±7.0	18.4±12.9	78 277	30
Context Dev	4 556	149	10.0	7.9±7.0	16.8±13.5	76 512	186
Context Test	4 868	149	10.0	7.4±6.8	16.1±13.5	78 360	186
Segment Dev	4 575	4 020	10.0	7.9±7.3	16.7±14.0	76 243	301
Segment Test	4 515	3 986	10.0	8.0±7.2	16.9±13.8	76 223	291

ture of the meetings and more) and adds also automatic annotation: morphological tags, syntactic structure and named entities.

The corpus comes in three data formats: the original HTML, TEI XML with rich metadata and annotation, and simple segmented plain texts with sound files directly usable for training of speech recognition systems.

ParCzech 3.0 corpus is available in the LINDAT repository under CC0 Public Domain waiver: http://hdl.handle.net/11234/1-3631.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 825460 (ELITR) and the grant 19-26934X (NEUREM3) of the Czech Science Foundation, and Project No. LM2018101 LINDAT/CLARIAH-CZ of the Ministry of Education, Youth and Sports of the Czech Republic.

References

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

2. Hladká, B., Kopp, M., Straňák, P.: ParCzech PS7 1.0 (2020), http://hdl.handle.net/11234/1-3174, LINDAT/CLARIAH-CZ digital library at ÚFAL, Faculty of Mathematics and Physics, Charles University
3. Hladká, B., Kopp, M., Straňák, P.: ParCzech PS7 2.0 (2020), http://hdl.handle.net/11234/1-3436, LINDAT/CLARIAH-CZ digital library at ÚFAL, Faculty of Mathematics and Physics, Charles University
4. 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. ELRA, Paris, France (2020)
5. Jakubíček, M., Kovář, V.: CzechParl: Corpus of Stenographic Protocols from Czech Parliament. In: RASLAN 2010. pp. 41–46 (2010), http://nlp.fi.muni.cz/raslan/2010/paper11.pdf
6. Janssen, M.: TEITOK: Text-faithful annotated corpora. In: Proceeding of LREC 2016). pp. 4037–4043 (2016)
7. Kratochvíl, J., Polak, P., Bojar, O.: Large corpus of czech parliament plenary hearings. In: Proceedings of LREC 2020. pp. 6363–6367. ELRA (2020), https://www.aclweb.org/anthology/2020.lrec-1.781/
8. Krůza, J.O.: Czech parliament meeting recordings as ASR training data. In: Proceedings of the 2020 FCCSIS. Annals of Computer Science and Information Systems, vol. 21, pp. 185–188. IEEE (2020), http://dx.doi.org/10.15439/2020F119
9. Pražák, A., Šmídl, L.: Czech parliament meetings (2012), http://hdl.handle.net/11858/00-097C-0000-0005-CF9C-4, LINDAT/CLARIAH-CZ digital library at ÚFAL, Faculty of Mathematics and Physics, Charles University
10. Roukos, S., Graff, D., Melamed, D.: Hansard French/English LDC95T20 (1995), https://doi.org/10.35111/jhgn-rv21
11. Straka, M.: UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In: Proceedings of the CoNLL 2018 ST: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 197–207. ACL (2018), http://dx.doi.org/10.18653/v1/K18-2020
12. Straková, J., Straka, M., Hajič, J.: Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In: Proc. of ACL System Demonstrations. pp. 13–18 (June 2014), http://dx.doi.org/10.3115/v1/P14-5003
13. Straková, J., Straka, M., Hajič, J.: Neural Architectures for Nested NER through Linearization. In: Proc. of ACL. pp. 5326–5331 (2019)
14. TEI Consortium, e.: TEI P5: Guidelines for Electronic Text Encoding and Interchange. 4.2.1. 2021-03-01. TEI Consortium, http://www.tei-c.org/Guidelines/P5/