Making Parliamentary Debates More Accessible: Aligning Video Recordings with Text Proceedings in Open Parliament TV

Olivier Aubert, Joscha Jäger
LS2N - Nantes Université, Open Parliament TV
contact@olivieraubert.net, joscha.jaeger@openparliament.tv

Abstract
We are going to describe the Open Parliament TV project and more specifically the work we have done on alignment of video recordings with text proceedings of the German Bundestag. This has allowed us to create a comprehensive and accessible platform for citizens and journalists to engage with parliamentary proceedings. Through our diligent work, we have ensured that the video recordings accurately correspond to the corresponding text, providing a seamless and synchronised experience for users. In this article, we describe the issues we were faced with and the method we used to solve it, along with the visualisations we developed to investigate and assess the content.
Keywords: video, text proceedings, alignment, data

1. Introduction

While parliamentary discourse analysis has traditionally been text-based, over the last 5 years the research community has seen a slow shift towards incorporating audiovisual information into parliamentary datasets.

Enriching parliamentary datasets with multimodal information allows new methods of analysis, like non-verbal cues, gestures/mimical information eg. to gain insights into their influence on perceived trust and/or confidence in politicians.

Additionally the audio information can help identify important events that were not transcribed or can be used as supplementary cues, e.g. for sentiment analysis.

Beyond the academic realm, video recordings of parliamentary debates hold great untapped potential for digital democracy. They serve as a tangible and contemporary interface to the daily work of parliaments. The recordings and live streams are not just video collections for journalists or corpora for scientific research but a direct application of the guiding principle “the parliament negotiates in public“.

Open Parliament TV uses this potential by developing a search engine and interactive video platform, in which speeches are searchable, linkable, citable and shareable beyond the boundaries of single parliaments.

1.1 Background

Almost every parliament publishes video recordings and text proceedings of sessions. But despite comparable structures and similar workflows, parliamentary proceedings are published in various, incompatible formats and parliament tv contents are only accessible via proprietary platforms. With Open Parliament TV we are developing a parliament independent open source solution which makes the video recordings searchable, shareable and citable via an automatic synchronisation of video recordings and text proceedings.

Our work is thereby focused on live data, which is made accessible via an easy to use platform interface[1], shown in figure 1, as well as a standardised and well documented open data api[2]. By implementing parliament independent data processing workflows we aim to interconnect political discourse between parliaments on national, regional as well as supranational (eg. EU Parliament) levels.

We have created a reference implementation with data from the German Bundestag, through which more than 60k speeches spanning over 10 years of parliamentary history are accessible (from 2013 until today).

In contrast to efforts like Open Discourse (Richter et al, 2023) and GermaParl (Blätte & Blessing, 2018; Blätte et al, 2022) who also work with a German Bundestag corpus, we focus on the audiovisual representation of speeches and work with archived and live data. The proceedings are hereby a means of making the videos more accessible, not vice versa.

1.2 Web Platform

Via the automated synchronisation of video recordings and official text proceedings we enable a full text search of the videos on the Open Parliament TV platform. By force aligning text fragments in the proceedings with specific points of time in the video recordings, we can additionally provide

- Interactive Transcripts
  (click on a sentence > jump to point of time in the video)
- Additional Information
  (show relevant documents and links at specific points of time in the speech)
- Improved means of participation
  (cite, embed and share video segments in the context of the full speech)

The platform significantly simplifies finding, sharing, embedding and citing specific video segments of political speeches and thus makes

[1] https://de.openparliament.tv
[2] https://de.openparliament.tv/api

parliamentary processes more transparent and accessible.

By providing an easy-to-use platform interface we make parliamentary work more accessible for researchers but also for journalists, political activists, educational institutions and the general public.

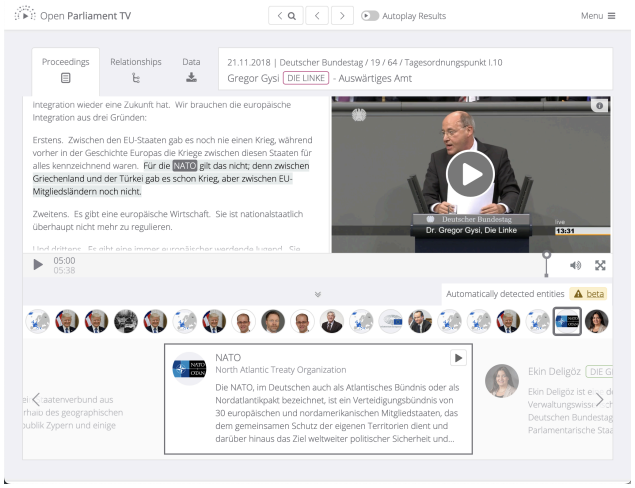

Figure 1: the Open Parliament TV platform

The Open Parliament TV platform additionally serves as a tangible open data use case, which is regularly used to advocate for better open data policies, standards and implementations on parliament level.

Linking platform contents directly with the respective parliamentary sources provides an additional layer of transparency and makes it easier for users to track and cite the original source as well as the context of quotes by politicians.

Especially in the context of citing quotes from political speeches, the official proceedings are a valuable source. The fact that the stenographic protocols don't exactly match the actual spoken word is in our case a feature, as the proceedings function as an immediately citable trusted source and will always be more reliable than transcripts generated by automatic speech-to-text (ASR) systems.

We do however use automated processes in order to annotate and enrich platform content with additional information. Based on the proceedings text, we extract Wikidata entities for people, organisations, laws and specific terms via Named Entity Recognition and provide information like Wikipedia abstracts or links to additional sources right inside the video player.

2. Related Work

In the ParlaMint community, several efforts have been made to use a combination of audio plus proceedings or transcripts to train ASR models (Ogrodniczuk et al, 2022).

A common issue is the alignment of incoherent source data for text proceedings and video recordings. This issue can be broken down into 2 main challenges:

- Finding a common identifier for both sources
- Determining common speech boundaries

Parliaments usually publish proceedings and video recordings via separate platforms, managed by different departments. In some cases the publication of video recordings is even outsourced to third party companies or media partners. This leads to differences in naming conventions of speakers and agenda items, making it difficult to identify the correct video resource for a specific speech in the proceedings (Ljubešić et al, 2022; Kulebi et al 2022).

One approach to deal with these inconsistencies is applying fuzzy match algorithms for the names of speakers (Kulebi et al, 2022). Beyond naming conventions both modalities sometimes have a different segmentation of speeches (specifically regarding speech items by the president), making it difficult to apply the otherwise feasible solution of comparing the two sources by the order / indices of items (Kulebi et al, 2022). In the ParlamentParla project, the two speaker sources have additionally been aligned using the Smith-Waterman sequence matching algorithm (Smith & Waterman, 1981).

Subsequently there is no common understanding of the beginning and end of speeches and agenda items. This is specifically relevant when using a combination of audio streams and proceedings in order to train ASR models (Ljubešić et al, 2022) as well as with automatic video subtitling systems (Alkorta, J., & Quintian, 2022). To determine common boundaries, some use automatic transcripts derived from speech-to-text systems and compare those with the official proceedings text via a forced alignment process (Hladká et al, 2020).

One solution to the problem of incoherent sources are machine readable proceedings, which contain references to the respective audiovisual resources in the metadata, as can be found (in non-standardised formats) in some parliament's data, like the Czech parliament (Hladká et al, 2020), the French Assemblée Nationale[3] or more recently the Austrian Parliament[4].

In recent years the extension of proceedings data with video recordings and the subsequent publication of multi-modal aligned (research) corpora has increasingly been mentioned as future work (Ogrodniczuk et al, 2022; Agnoloni et al,

[3] https://www.assemblee-nationale.fr/dyn/16/comptes-rendus/seance

[4] https://www.parlament.gv.at/recherchieren/protokolle/index.html

2022). This would allow annotating corpora with physical communicative features like gestures and facial expressions (Ogrodniczuk et al, 2022; Ménard & Aleksandrova, 2022).

3. Automating AV Alignment

Access to video material depends on some kind of discretization to facilitate indexing and navigation. In the case of parliaments, the proceedings are an official source of textual data that should match the video feeds. There is not yet any standard shared by all parliaments, therefore the Open Parliament TV has to conceive an ingest infrastructure dedicated to handle the specificities of each parliament and convert its data into its own common model.

3.1 Context: the Bundestag Plenary Sessions

In the Bundestag case, the video stream is broadcasted live on the https://www.bundestag.de/ website. Some textual metadata is associated with it before the recording, based on the agenda of the session. The interface displays the title of each intervention - current and forecoming - as well as the speaker name (with the planned time of speaking), and features references to additional material.

In addition to the frontend web interface, the video feed is provided as a video podcast, i.e. a RSS stream of mp4 files. Each item features the session identification with its date, the intervention title and the speaker name.

Official minutes are provided through the website as well, with a delay of 2 to 3 days. The main web interface features links to related documents, as well as a link to a summary of each part. The official plenary minutes[5] are provided as PDF files. A session is divided into multiple agenda items. Each item provides a link to the corresponding PDF file and to the video of the session. API-wise, the Bundestag proposes the Documentation and Information system for Parliamentary materials (DIP) which provides structured access to a query interface into the document material (as text fragments or PDF documents), but does not give access to proceedings themselves in a structured format.

An additional opendata API is also available[6]. It provides a stream of plenary proceedings in XML format, structured using a dedicated dbtplenarprotokoll DTD.

The goal of the data ingestion phase is to provide for the Open Parliament TV platform a unified data source combining both video streams and text proceedings, enriched with Wikidata IDs, so that they can be presented in meaningful ways through the platform interface. This process has to be able to process both old data, but also to run unattended to provide an as-live-as-possible experience to the users: the video data is presented as soon as it is available, and later enriched with the text proceedings when the data becomes available.

The data is organised in electoral terms. The current one, the 20th, started on 26/09/2021 and is still ongoing. In order to give an idea of the corpus dimensions, we will focus on the preceding term, for which we have the complete data. The 19th electoral term ran from 24 October 2017 until 26 October 2021. The 736 representatives attended 239 plenary sessions, which produced 2151 hours of video.

3.2 Architecture of the Code

The data processing code is published on github[7]. It is free software, licensed under the General Public License.

It is divided into fetcher modules that download updated data (media and proceedings) in raw XML or json format, and parser modules that massage the data into the unified model of the Open Parliament TV platform. Then a merger module, which we will more precisely describe in this article, matches data from both sources in order to produce a unified format mixing both video and textual information.

Once the media and proceedings items are aligned, additional processing takes place. Speaker names are linked with their corresponding Wikidata id (in nel module) and forced alignment is applied on the video fragments and transcript in order to provide a more fine-grained association of the text proceedings.

The different modules (fetcher, parser, nel, forced alignment…) can be used independently, and their orchestration is implemented in a workflow script.

3.3 Identified Alignment Issues

The main key for aligning items between the video feed and the OpenData XML proceeding feed is the title of the item and the speaker name. However, similarly to [Kulebi et al, 2022], a number of mismatches plague the data. They may come from human errors or the application of transcription conventions that remove some text for the sake of clarity. In the German Bundestag , speakers are also given the opportunity to amend the transcribed version, as described in rules 116-119 of (Deutscher Bundestag, 2022).

First, there are small transcription errors in speaker names and speech titles, e.g. putting the title (Dr., Prof.) in front of the name in media data but not in

[5] Official plenary minutes https://www.bundestag.de/dokumente/protokolle/plenarprotokolle

[6] Opendata API https://www.bundestag.de/services/opendata

[7] Processing tools repository https://github.com/OpenParliamentTV/OpenParliamentTV-Tools

proceeding data. Then, there are larger and more systematic errors, often occurring in batches, where a whole agenda item title is wrongly assigned. Similar issues can be found in the time segmentation: speech boundaries do not always match, the session president introduction being sometimes included in the preceding speech in the video capture. More importantly, completely different segmentations may occur, resulting in different amounts of media and proceeding items for the same session, and increasing the difficulty for the matching process.

3.4 First Approach based on Speaker/Title Similarity

A first naive matching approach was first used, based on using speaker and title - after a small normalizing process - for generating a key identifying each media and proceedings item. Collisions were handled by adding an incremental index.

To try to alleviate small transcription errors, common similarity measures like the Levenshtein distance were experimented to compare the generated keys, but the nature of the underlying data, where agenda items can often share the same base title and differ only with an index or a reference number, made this approach inappropriate.

Moreover, the discrepancy between the number of media items and the number of proceeding items made this approach inherently fragile. Figure 1 presents a scatter plot of each session in the 19th term with its number of proceeding and media items on the horizontal and vertical axis. We can see that the majority of sessions have the same number of media and proceeding items, being concentrated on the diagonal, but the number of non-matching sessions is important.

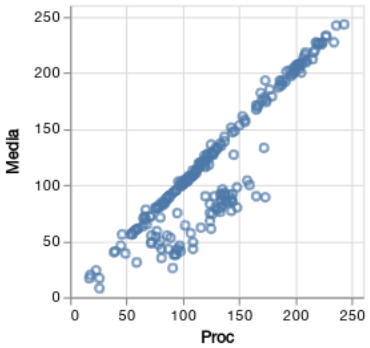

Figure 2: representation of proceeding items count vs media items count

To investigate further, we built a new visualisation, using small multiple scatter plots, as presented in figure 2 : x-axis represents the index of the media item in the "media" sequence. The y-axis represents the index of the corresponding proceeding item in the proceedings sequence. In the ideal case, both axes have the same length (same number of media items wrt. proceeding items) and the representation should be a diagonal, meaning that media item number N matches proceeding item number N for all items. This allows to visually quickly discriminate against misaligned sessions. Additionally, this gives information about the alignment symptoms. For instance, for the first session in the second line, we see a horizontal line with a diagonal starting approximately in the middle x axis, with a low media index. The interpretation, corroborated by examination of the actual data, is that the proceeding segmentation has been more fine-grained than the media segmentation. Hence, the same media item (first one) has been aligned with multiple proceeding items, which gives a horizontal line. Upon examination of the data, this shape often occurs in sessions of questions to the government, which generates a single, long media item with the president of parliament as indicated speaker, while the proceedings have split the questions by speaker, producing multiple shorter items in proceedings.

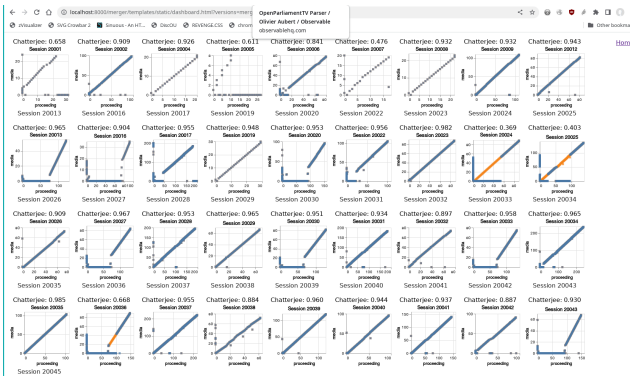

Figure 3: small-multiple scatter plot visualisation of sessions

This visualisation moreover allows comparison between results of different alignment algorithm/parameters: respective outputs are plotted in different colors (here, yellow and blue), which allows to quickly identify the sessions where changes occurred. For instance in the image, the alignment for the last 2 sessions of the second line improved greatly, giving an adequate diagonal.

This lead us to the conclusion that the problem was not simply tackable through simple index matching, especially because of the segmentation difference. It occurred regularly in specific groups of sessions, like the questions to parliament. The index matching approach also had the inconvenience of ignoring the sequencing of items, while the data, being the recording of an event, implies that item sequences must be preserved.

In essence, we have 2 sequences of "alphabet" that should somehow match. There can be cases of insertion of sequences in one side (shorter segments for instance) or of deletion (longer segments). There can also be complete changes

(like in human transcription errors), which we can call mutation.

This similarity to DNA-alignment led us to investigate this direction as a new approach.

3.5 Needleman-Wunsch Algorithm

We investigated with colleagues from a bioinformatics team, in order to explain the issue and find similarities and solutions that could be transferred from this domain. Indeed,the Needleman-Wunsch algorithm, a classical algorithm from the 1970s, can be used to align DNA sequences, trying to preserve global order, with parameterized costs for insertion or deletion. Another algorithm, the Smith-Waterman Algorithm (Smith & Waterman, 1981) has been used in (Kulebi et al, 2022) for similar purposes, but focused on local sequences.

In our implementation the algorithm has 4 parameters. Two, *speaker_weight* and *title_weight*, are related to the similarity measure between 2 items. As with our previous experiments, we noticed that the data specificities made common string approximations like Levenshtein inappropriate, and we chose to do basic string comparison, weighted by parameters. The other two parameters, *merge_penalty* and *split_penalty*, are used by the algorithm itself to parameterize.

The Needleman-Wunsch algorithm is a dynamic programming algorithm used in bioinformatics to perform global sequence alignment between two sequences. It starts by creating a matrix, typically called the scoring matrix, where the columns represent items from proceedings and the rows represent items from the media source. A similarity function is defined, as the ponderated sum of the string similarities of speaker and titles. The matrix is initialized along the first row and column with the similarity between corresponding items.

To fill the matrix, we recursively calculate scores, starting from the first cell, and computing the score of the neighboring cells, comparing the hypotheses of moving to one of the horizontal, vertical or diagonal neighbors, and keeping the hypothesis with the highest valued. The horizontal neighbor hypothesis adds an increment of *merge_penalty*, since it represents the cell that would be reached by merging two proceeding items. The vertical neighbor hypothesis adds an increment of *split_penalty*, since it represents the fact that a proceeding should be split between two media items. The diagonal hypothesis provides an increment of the similarity score between its items, representing the "normal" hypothesis. We iterate through the matrix, calculating scores for each cell based on the recurrence relation until the entire matrix is filled.

Once the matrix is filled, as presented in figure 3, traceback is performed to find the optimal alignment between the two sequences. We start at the highest score in the top-right corner of the matrix, and trace back to the bottom-left corner, following the path of highest scores. This traceback process identifies the alignments that maximizes similarity between the sequences.

As a result, the algorithm outputs the optimal item alignments along with their corresponding scores. The graphical and interactive representation, linked with the transcript and the video, allowed us to validate the efficiency of the approach.

Overall, the Needleman-Wunsch algorithm efficiently finds the optimal alignment(s) between two sequences by considering all possible alignments and scoring them based on a defined scoring scheme, making it a fundamental tool in bioinformatics for sequence analysis and comparison.

3.6 Dashboard and Visualisations

A dashboard presenting visualisations of the processed corpus is available at https://openparliamenttv.github.io/OpenParliamentTV-Tools/optv/parliaments/DE/dashboard/dashboard.html

The data hosted on github is subjected to download rate limits though, and also does not have the intermediary parsed media and proceeding files, limiting the use of some visualisations. Hence, we are also hosting the same dashboard on a dedicated server at https://optv.olivieraubert.net/ with the whole data.

The dashboard proposes to select subsets of the whole corpus, incrementally loading their data to present it. Once a group is selected, the different scatter plot visualisations are presented. Clicking on the title of each graph leads to more precise visualisations, in order to provide better context for exploring data and issues. The Session word is linked to the "block view" described below, while the session number is linked to the "transcript view".

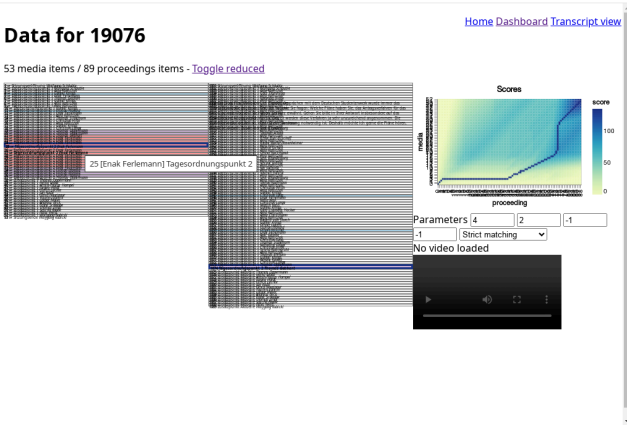

Figure 4: "block visualisation" with dynamic parameterization of the matching algorithm

Figure 4 presents an interactive visualisation, called "block view", that was built to validate and fine-tune parameters of the algorithm. It takes as

input a session identifier and gets its data from the media and processing parsed files. It presents in the first two columns a visualisation of the media and item blocks, with their information (index, speaker and title) readable on mouse over. As a synchronized view, it highlights in the other column the items having a matching speaker name, speech title or having both. This view implements a javascript version of the Needleman-Wunsch algorithm, and presents a live-generated matrix of its output, with the ability to dynamically tune the 4 algorithm parameters and the string similarity method, in order to assess their influence.

Figure 5 presents a second interactive visualisation that was developed to explore and evaluate the actual output of the processing and merging workflow. It uses the merged output file data as input. On the left of the page, the transcript - generated from the proceeding data - is presented, along with an affordance to play the aligned video. It also offers a visualisation of the path produced by the algorithm in the right-hand side, along with a scatter-plot visualisation of the word count (from proceedings) vs duration (from media) of the aligned items, in order to explore other indicators.

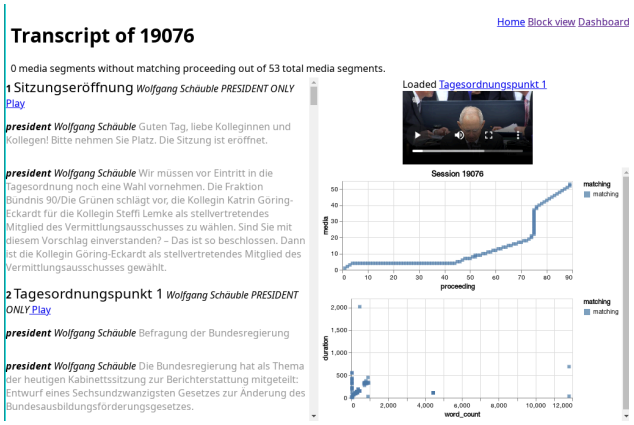

Figure 5: transcription view, presenting the result of the alignment process

3.7 Contributions

The code and live-updated data are publicly available as repositories hosted on Github[8]. They can be interacted with through the Open Parliament TV platform, and assessed through a dashboard. The fetching and parsing code are specific to the concerned parliament, but the whole suite has been designed to be also used as much as possible on other parliament's data.

As a methodological contribution, we identified a number of issues like transcription errors and segmentation issues that will be common to other similar projects, as can be seen in (Kulebi et al, 2022). This led us to implement and evaluate the adequacy of the Needleman-Wunsch sequence alignment algorithm.

Moreover, we produced a number of interactive visualisations for the data, either global (the dashboard view) or more specific (the block and transcript view), which could also be used as an inspiration in other projects.

4. Conclusion / Future Work

The Open Parliament TV project proposes a user-oriented interface for making parliamentary debates more accessible to the public and the media. By unifying video recordings with text proceedings, we have created a valuable resource for understanding the intricacies of legislative discussions. This work paves the way for expansion to other parliaments, thereby improving the parliament independent workflow and interconnecting discourse beyond national borders. While we have been using our own unified data model, we would like to move towards more standardised models in order to foster interoperability. Collaborating with organisations such as Open Discourse or GermaParl would offer an opportunity to integrate historical debates into the platform, extending its reach and value further.

While alternative approaches, such as complete transcription through speech-to-text algorithms and automatic translation, could have been considered, the availability of performant and robust models at the time of our research necessitated a different approach, and we wanted to be able to process data on standard computers. Today, however, advancements in this technology make it an exciting prospect for future exploration and further refining the alignment results.

In addition to checking discrepancies between official proceedings and actual discourse, increased accessibility through automatic speech translation into multiple languages would open new possibilities for users following debates in other parliaments. Furthermore, the potential to interconnect parliamentary discourse beyond language and parliamentary boundaries enables more comprehensive search and analysis capabilities. Overall, the Open Parliament TV project signifies a crucial advancement towards making parliamentary proceedings more accessible, transparent, and globally interconnected.

5. Acknowledgements

The authors wish to thank Guillaume Fertin and Geraldine Jean from LS2N, Nantes Université, for their advice in bioinformatics inspiration.

6. Bibliographical References

Agnoloni, T., Bartolini, R., Frontini, F., Montemagni, S., Marchetti, C., Quochi, V., Ruisi, M.. & Venturi, G. (2022, June). Making Italian Parliamentary

[8] Data repository (see above footnote for tools repository): https://github.com/OpenParliamentTV/OpenParliamentTV-Data-DE/

Records Machine-Actionable: The Construction of the ParlaMint-IT Corpus. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference* (pp. 117-124).
Alkorta, J., & Quintian, M. I. (2022, June). Adding the Basque Parliament Corpus to ParlaMint Project. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference* (pp. 107-110).
Blätte, A., & Blessing, A. (2018, May). The germaparl corpus of parliamentary protocols. In proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).
Blätte, A., Rakers, J., & Leonhardt, C. (2022, June). How germaparl evolves: Improving data quality by reproducible corpus preparation and user involvement. In Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference (pp. 7-15).
German Bundestag (2022). Rules of Procedure of the German Bundestag and Rules of Procedure of the Mediation Committee. https://www.btg-bestellservice.de/pdf/80060000.pdf (visited on 2024.03.29)
Hladká, B., Kopp, M., & Straňák, P. (2020, May). Compiling Czech parliamentary stenographic protocols into a corpus. In *Proceedings of the Second ParlaCLARIN Workshop* (pp. 18-22).
Kulebi, B., Armentano-Oller, C., Rodríguez-Penagos, C., & Villegas, M. (2022, June). ParlamentParla: A speech corpus of catalan parliamentary sessions. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference* (pp. 125-130).
Ljubešić, N., Koržinek, D., Rupnik, P., & Jazbec, I. P. (2022, June). ParlaSpeech-HR-a freely available ASR dataset for croatian bootstrapped from the parlaMint corpus. In Proceedings of the workshop ParlaCLARIN III within the 13th language resources and evaluation Conference (pp. 111-116).
Ménard, P. A., & Aleksandrova, D. (2022, June). A French Corpus of Québec's Parliamentary Debates. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference* (pp. 25-32).
Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, *48*(3), 443-453.
Ogrodniczuk, M., Osenova, P., Erjavec, T., Fišer, D., Ljubešić, N., Çöltekin, Ç., Kopp, M. & Katja, M. (2022, June). ParlaMint II: The Show Must Go On. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference* (pp. 1-6).
Richter, F., Koch, P., Franke, O., Kraus, J., Warode, L., Kuruc, F., Heine, S., Schöps, K. (2023, January 21). Open Discourse: Towards the first fully Comprehensive and Annotated Corpus of the Parliamentary Protocols of the German Bundestag. https://doi.org/10.31235/osf.io/dx87u
Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, *147*(1), 195-197.