

Aligning Video Recordings with Text Proceedings in Open Parliament TV

Olivier Aubert - Nantes Université
Joscha Jaeger - Open Parliament TV

Summary

- Open Parliament TV project: goals and architecture
 - "Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora"
- The Bundestag context
- The data processing pipeline
 - Official text proceedings and video feed
- Identified issues in the data
 - Mislabelling
 - Segmentation
- Alignment approaches
- Data visualisations

The Open Parliament TV project

Started in 2019

Goal: Making parliamentary debates more transparent and accessible

- User-accessible interface and search engine
- Enriched video and transcripts (with entities and documents)
- Ability to quote/link parts of speech
- Connect political discourse beyond the boundaries of single parliaments

Search engine

Open Parliament TV

Alle Parlamente

Rente* x Suchbegriff eingeben

Fraktion

CDU/CSU SPD AfD FDP DIE LINKE DIE GRÜNEN fraktionslos

Zeitraum: 05.01.2018 - 03.04.2021

te verbinden. [...] Für uns ist in dem Kontext die **Rentengarantie** das oberste Gebot, das Versprechen, das wir gegeben haben. [...] Mit dem **Rentenpaket** sind zwei Ziele, zwei Gedanken, zwei Garantien verbunden, die doppelte Haltrahlinie: so haben Sie, Herr Kan-

[...] Fast 30 Jahre nach der Wiedervereinigung sollte es keine unterschiedliche **Rentenberechnung** in Ost und West mehr geben. [...] Ich will mal etwas tun, was ich hier selten tue, nämlich die SPD, insbesondere die ehemalige Arbeits- und Sozialministerin Andrea

[...] Uns liegt nun ein Gesetzentwurf der Bundesregierung zur Leistungsverbesserung und Stabilisierung in der gesetzlichen **Rentenversicherung** vor. [...] Herr Minister Heil, Sie belasten die heutigen Beitragszahler, indem der Beitragssatz nicht so gesenkt wird wie ei-

lösung der **Rentenerinnen** und **Rentener** ist die im Jahr 2004 von Rot-Grün beschlossene Umstellung auf die sogenannte nachgelagerte Besteuerung der **Renten**. [...] Seit 2005 steigt der Anteil der **Rente**, der besteuert wird, Jahr für Jahr. [...] Wer 2040 in **Rente** gehen

Die neue **Rentenformel** stellt sicher: Die **Renten** steigen wie die Löhne. [...] Wir sichern damit ein **Rentenniveau** auf dem heutigen Level. [...] Denn die gesetzliche **Rentenversicherung** ist und bleibt die zentrale Säule im deutschen **Rentensystem**. [...] Die **Renten**

08.11.2018 07:38
Hermann Gröhe (CDU/CSU)
Gesetzliche Rentenversicherung
te **Rentenkassen**. [...] Wir konnten in den letzten Jahren übrigens auch einen leichten Anstieg des **Rentenniveaus** beobachten. [...] Die gesetzliche **Rentenversicherung** ist ein starkes Stück Sozialstaat. [...] Ich erlaube mir ange-

12.10.2018 08:01
Kai Whittaker (CDU/CSU)
Gesetzliche Rentenversicherung
würde man nämlich erkennen, dass Sie zwei **Rentenkonzepte** haben [...] Werte Kollegen, mir geht in dieser Debatte auf den Zeiger, wie dieses **Rentenpaket** den Menschen in diesem Land madig gemacht wird. [...] Wir wollen die Er-

06.06.2019 07:25
Olav Gutting (CDU/CSU)
Rentenbesteuerung
stellen: Die Steuerbelastung von **Rentnern** ist grundsätzlich erst mal nicht zu kritisieren. [...] Vielmehr ist sie eine Folge der höheren **Rentenauszahlungen**. [...] Die **Rentenerhöhungen** der letzten Jahre machen sich hier bemerkbar und

15.05.2020 03:11
Leni Breymaier (SPD)
Versorgungsausgleich
Rentensanspruch bzw. den entsprechenden **Rentenpunkten** auf dem Konto ihrer weiteren Wege. [...] Dann heißt der Vorgang **Rentensplitting**. [...] **Rentensplitting** finden wir Diplomfeministinnen toll. [...] Ich glaube aber, das im

04.03.2021 06:53
Olav Gutting (CDU/CSU)
Doppelbesteuerung bei Renten
über die **Rentenbesteuerung** verbreitet. [...] Da das Einkommen im **Rententalter** regelmäßig geringer ist als während des Erwerbslebens, führt das dazu, dass die **Rentenzahlungen** aufgrund der Steuerprogression mit einem nied-

Enriched video presentation

Open Parliament TV Menu ☰

Protokoll Querverweise 21.11.2018 | Deutscher Bundestag / 19. WP / Sitzung 64 / Tagesordnungspunkt I.10
Gregor Gysi DIE LINKE - Auswärtiges Amt

☞ (Beifall bei der LINKEN – Christian Schmiot [Fürth] [CDU/CSU]: Herr Gysi, Sie wissen es besser!)

Die EU muss grundsätzlich neu gestaltet werden, damit die europäische Integration wieder eine Zukunft hat. Wir brauchen die europäische Integration aus drei Gründen:


Erstens. Zwischen den EU-Staaten gab es noch nie einen Krieg, während vorher in der Geschichte Europas die Kriege zwischen diesen Staaten für alles kennzeichnend waren. Für die NATO gilt das nicht, denn zwischen Griechenland und der Türkei gab es schon Krieg, aber zwischen EU-Mitgliedsländern noch nicht.

Zweitens. Es gibt eine europäische Wirtschaft. Sie ist nationalstaatlich überhaupt nicht mehr zu regulieren.

Und drittens. Es gibt eine immer europäischer werdende Jugend. Sie spricht ganz gut Englisch und erobert sich den ganzen Kontinent. Wenn wir denen sagen: „Zurück zu alten Nationalstaaten mit Pass und vielleicht noch Visum“, denken die ja, wir haben eine Meise. Deshalb sind wir Alten verpflichtet, für die Jugend die europäische Integration zu retten, aber ganz anders, als das bisher geschieht.


05:02 05:38

Automatisch erkannte Entitäten (beta)




Gunther Krichbaum
CDU/CSU
deutscher Politiker, MdB
<http://www.gunther-krichbaum.de/>

Open Entry in New Tab




NATO
Die NATO, im Deutschen auch als Atlantisches Bündnis oder als Nordatlantische Allianz bezeichnet, ist ein Verteidigungsbündnis von 30 europäischen und nordamerikanischen Mitgliedsstaaten
<https://www.nato.int/>

Open Entry in New Tab



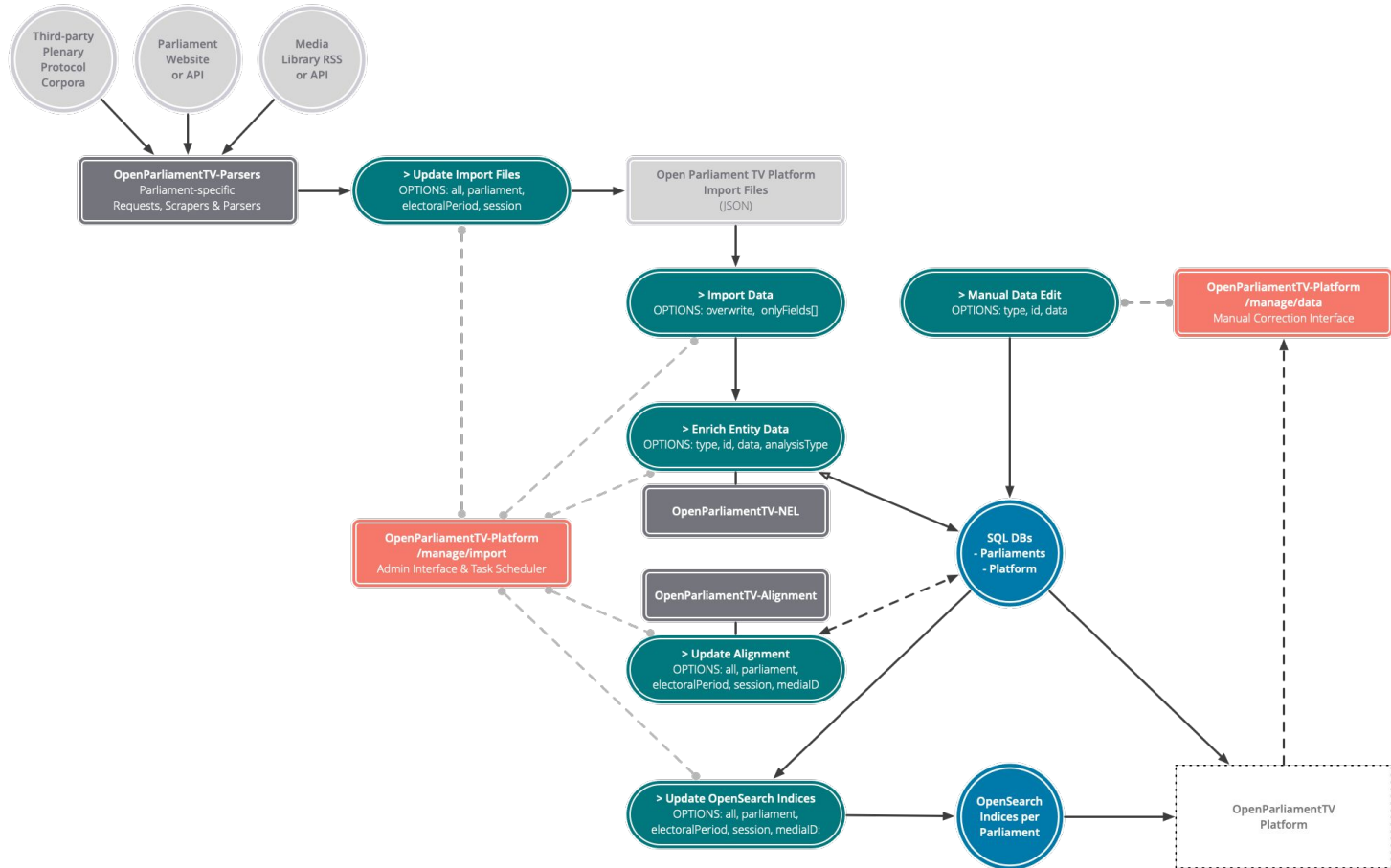
Bündnis 90/Die Grünen
grüne politische Partei in Deutschland
<https://www.gruene.de/>

Open Entry in New Tab



live **13:31**

Open Parliament TV architecture



The Bundestag context

- Video feed for plenary sessions
 - live, and available afterwards as a RSS feed
 - features title and speaker name
 - segmented by agenda item
 - 1483 hours for the current parliamentary session (20) started in nov. 2021
- Official text proceedings
 - provided after a 2-3 days delay
 - in PDF format and XML format (using the dbtplenarprotokoll DTD)
- Goal : align official text proceedings and video feed
- Existing projects (OpenDiscourse and GermaParl) not quite fit
- Development funded by Prototype Fund program

Data processing goals

- Download data for video and text documents
- Parse to provide a unified data model from both sources
- Merge video items and text items
- Enrich the merged model with
 - Named-Entity Linking (for explicitly structured data)
 - Named-Entity Recognition (for other entities)
 - text sentences forced alignment with audio/video
- Have a system that can run mostly unattended on low-end servers

Data processing pipeline

Individual tools/modules orchestrated in a `workflow.py` :

- Scrapers (video and text) (`scraper`)
- Parsers (video and text) (`parser`)
- Video and text alignment/merging (`merger`)
- Named-Entity Linking with Wikidata entities (curated source) (`nel`)
- Named-Entity Recognition with Wikidata/*spacy* (`ner`)
- Forced text alignment with *aeneas* (`aligner`)
- Publication (push to git) (`workflow.py`)

See <https://github.com/OpenParliamentTV/OpenParliamentTV-Tools>

Issues

- Technical issues

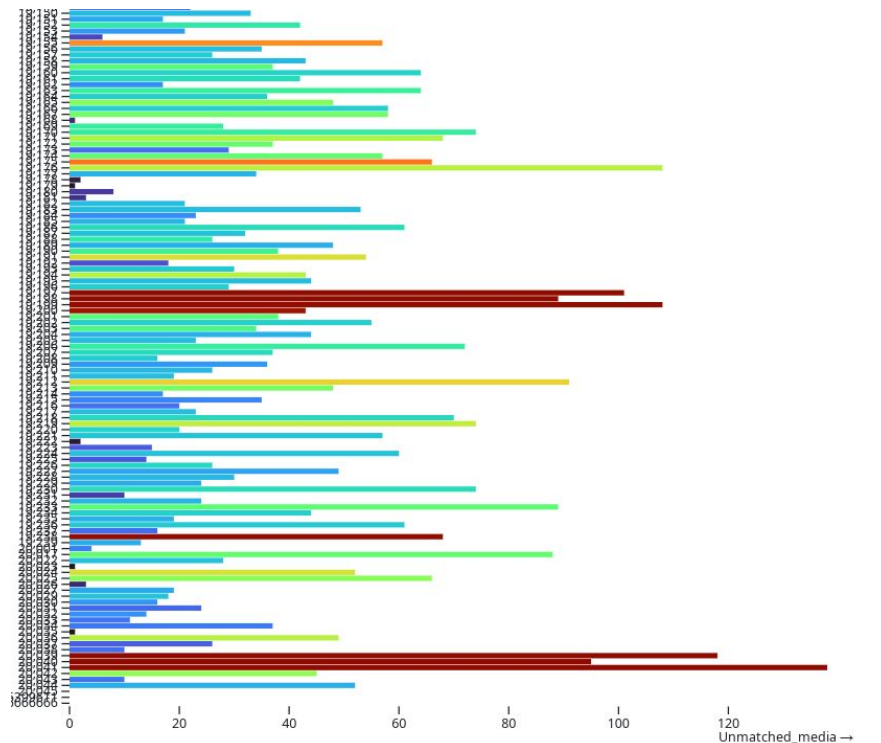
- servers sometimes unreliable -> implement a retry mechanism
- non-monotonous session numbering (e.g. 20082 -> 20904 -> 20083)

- Main data issues

- non-formalized formats and inconsistent entries : the DTD specifies a schema, but not all content form (e.g. speaker + faction identification)
- plain human entry errors (wrong speak title, wrong speaker name...)
- segmentation differences between video feed and text proceedings

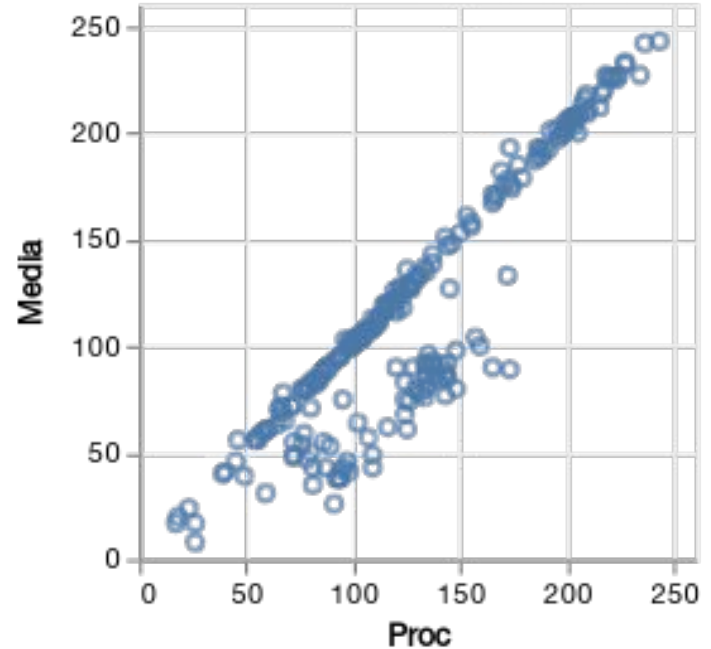
Initial key-based alignment approach

- key for item based on speaker name + title + index in case of duplicates
 - local order only
 - very fragile wrt human errors and segmentation differences
- highlights some error patterns



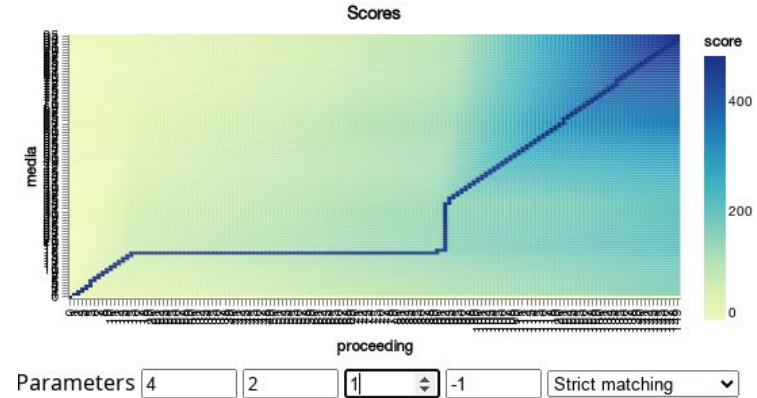
Segmentation issue

- Plot proceeding item count vs media item count - mismatch
- Often due to "Fragestunde" (Questions to the Parliament)
- Other corner cases (arbitrary segmentation)

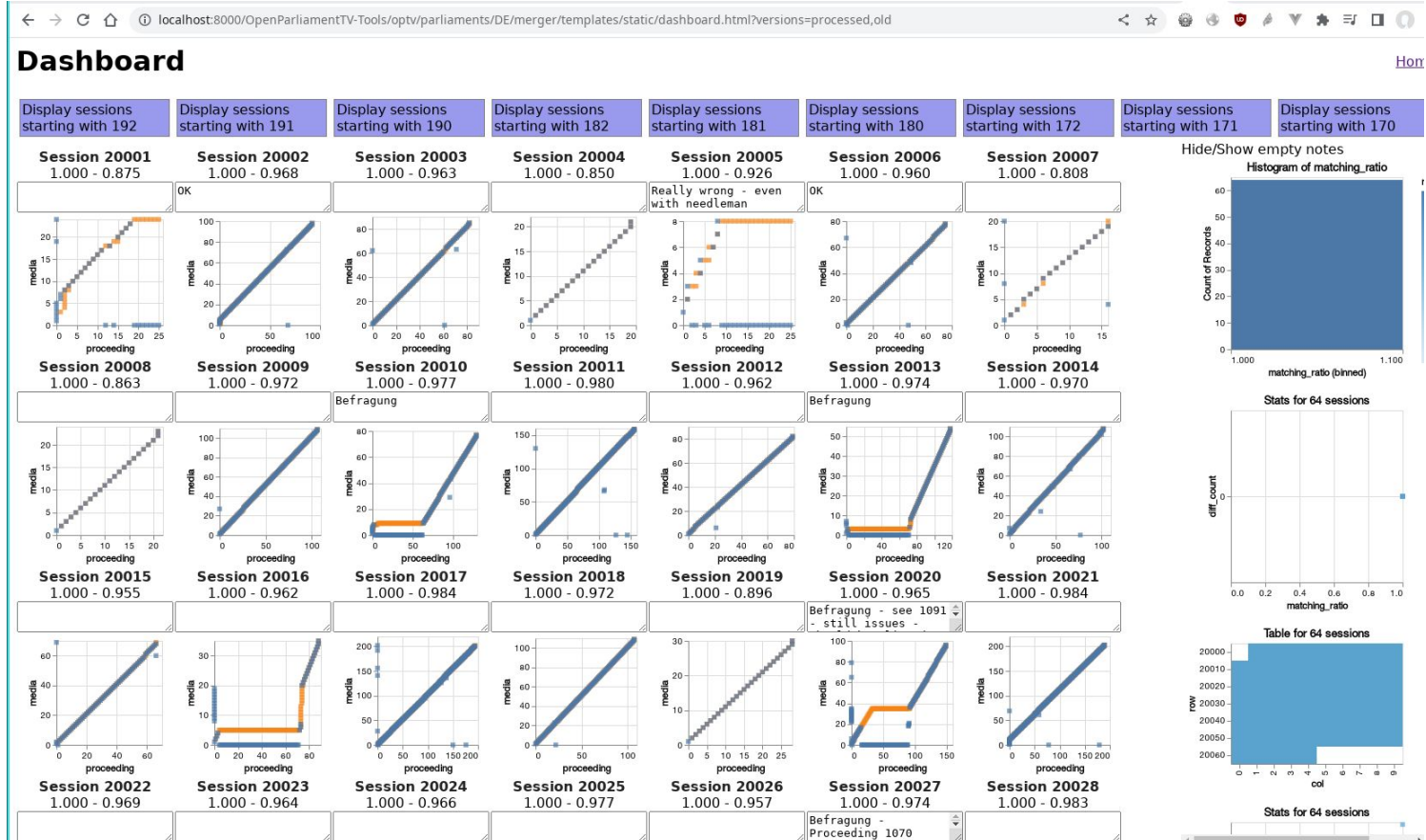


Needleman-Wunsch alignment

- Analogy with DNA sequence alignment, with global order constraint
- Speak items with title + speaker name (with possible mutations/deletions/insertions)
- Parameters
 - **weights** for speaker name and item title for item similarity
 - **merge_penalty** and **split_penalty** to express cost of merging/splitting a media item in multiple proceeding items
- Principle
 - build a score matrix from start items
 - follow the highest score path starting from the end



General dashboard

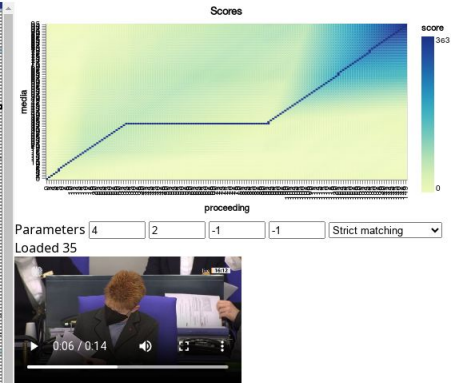


"Block" visualization - dynamic parameters

[Home Dashboard Transcript view](#)

Data for 2020

96 media items / 150 proceedings items - [Toggle reduced](#)



<https://optv.olivieraubert.net/OpenParliamentTV-Tools/optv/parliaments/DE/dashboard/block.html?session=20020#>

Transcript visualisation - alignment result

[Home](#) [Block view](#) [Dashboard](#)

Transcript of 2020

0 media segments without matching preceding out of 96 total media segments.

1 Sitzungseröffnung *Katrin Göring-Eckardt* **PRESIDENT ONLY** [Play](#)

vice-president Katrin Göring-Eckardt Einen schönen guten Tag, liebe Kolleginnen und Kollegen! Die Sitzung ist eröffnet.

vice-president Katrin Göring-Eckardt Bevor wir zu unserer Tagesordnung kommen, will ich von hier aus sehr herzlich die Präsidentin des Bundestages und andere Präsidiumsmitglieder grüßen, die erkrankt sind, sowie eine ganze Reihe von Kolleginnen und Kollegen aus dem Parlament, die ebenfalls erkrankt sind. Ich wünsche allen gute Besserung.

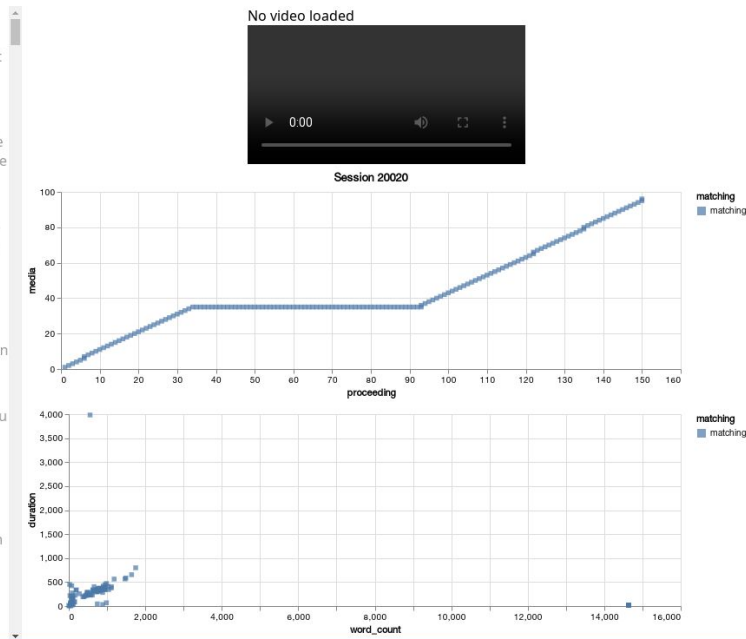
vice-president Katrin Göring-Eckardt Sie werden in dieser Woche mit Petra Pau und mir – immer im Wechsel – vorliebnehmen müssen. Wir freuen uns sehr darauf, das machen zu können. Wir freuen uns auch, wenn Sie heute, morgen und übermorgen bei allen Zwischenfragen und -bemerkungen, die sie im Kopf haben, überlegen, ob Sie diese im Parlament vortragen müssen. Aber das ist natürlich nur eine herzliche Bitte, damit wir eine straffe Sitzungsleitung hinbekommen und alles gut bewältigen.

vice-president Katrin Göring-Eckardt Jetzt kommen wir zur Sitzung heute. Ich habe Ihnen eine ganze Reihe von Dingen mitzuteilen, bevor wir in die Tagesordnung eintreten.

vice-president Katrin Göring-Eckardt Mit Wirkung zum 14. März hat die Präsidentin die Allgemeinverfügung zu Corona-Schutzmaßnahmen im Deutschen Bundestag geändert.

vice-president Katrin Göring-Eckardt Für das Plenum bedeutet das Folgendes: Für den Zutritt zu Plenarsitzungen gilt fortan wieder eine 3-G-Regel. Die 3-G-Regel bedeutet, dass Zutritt zu den Plenarsitzungen ausschließlich diejenigen Personen erhalten, die vollständig gegen das SARS-CoV-2-Virus geimpft, von einer Coronaerkrankung genesen oder aktuell negativ getestet sind. Dieser 3-G-Status ist nach Maßgabe der Allgemeinverfügung als Zutrittsberechtigung zum Plenarsaal sowie zur Ost- und Westlobby einschließlich der Abgeordnetenlobby nachzuweisen.

vice-president Katrin Göring-Eckardt Die FFP2-Maskenpflicht für die Plenarsitzung gilt wie zuvor. Danach ist die FFP2-Maske auch am Platz zu tragen. Ausgehende sind allein für die antwortende Präsidentinnen und



<https://optv.olivieraubert.net/OpenParliamentTV-Tools/optv/parliaments/DE/dashboard/transcript.html?session=2020>

Results

Term	Period	# Sessions	# Items	Video duration (h)
20	2021-	173	20295	1483
19	2017-2021	239	27603	2151
18	2013-2017	245	19382	1866

Conclusion and perspectives

- Data ingestion and processing code available at <https://github.com/OpenParliamentTV/OpenParliamentTV-Tools>
- Data (origin and aligned/merged) available at <https://github.com/OpenParliamentTV/OpenParliamentTV-Data-DE/>
- Still some QA to do
- Extension to other parliaments

Feedback welcome!