From parliamentary history to digital and computational history: a NLP-friendly TEI model for historical parliamentary proceedings

Marie Puren, Fanny Lebreton, Aurélien Pellet, Pierre Vernus

▶ **To cite this version:**

Marie Puren, Fanny Lebreton, Aurélien Pellet, Pierre Vernus. From parliamentary history to digital and computational history: a NLP-friendly TEI model for historical parliamentary proceedings. Digital Scholarship in the Humanities, In press, 10.1093/llc/fqae071 . hal-04104205

HAL Id: hal-04104205
https://hal.science/hal-04104205v1

Submitted on 23 May 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

From parliamentary history to digital and computational history : a NLP-friendly TEI model for historical and contemporary parliamentary proceedings

Abstract

This paper introduces a new method for the digital and computational analysis of historical and contemporary parliamentary proceedings. It addresses the dichotomy in the utilization of these resources between historians and other disciplines, and emphasizes the significance of continuity in studying long-term phenomena. The paper presents an XML-TEI model specifically designed for encoding parliamentary documents from diverse temporal and regional contexts. This model is exemplified through the analysis of parliamentary debates from the French Chamber of Deputies (1889-1893). The first part of the paper discusses the motivations behind the model's development. The second part outlines the methodological choices in constructing the model and the need for schema adaptation. We subsequently detail our method for automatic encoding of such extensive corpora. Finally, we propose an approach to annotate parliamentary debates using natural language processing analyses, focusing on topic modeling. This study aims to enhance computational research in humanities, especially historical and political studies, by providing an efficient tool to harness the potential of the massive digitized parliamentary data.

Introduction

Parliamentary debates have been a valuable source for research in the humanities and social sciences (Chester et al., 1962; Franklin et al., 1993; Mela et al., 2022). Sociology (Cheng, 2015), political science (Van Dijk, 2010) or linguistics (de Galembert et al., 2013; Hirst et al., 2014; Rheault et al., 2016) have used them as research material. These are also analyzed by political historians - in a national (Ouellet et al., 2003; Bouchet, 2018) or transnational (Ihalainen et al., 2006; Ihalainen et al., 2018) perspective - but also of social life, economics and religion (Marnot, 2000; Lemercier, 2021), or even law (Fournier et al., 1991). From a mere disciplinary point of view, a clear dichotomy emerges for the exploitation of parliamentary debates with historians working on older periods, and actors from other disciplines on contemporary proceedings. However, especially when it comes to studying long-term phenomena[1], the continuity between past and present data is of crucial importance.

[1] We have found that discussions can start several years before a law is actually passed (Bourgeois et al., 2022).

Working on a transnational corpus from 1803 to 2005, Helen Baker and her co-authors (Baker et al., 2017) have for instance shown how bridging political history and political science could be a fertile research avenue.

At the same time, contemporary parliamentary data is increasingly available on the web, as are historical parliamentary data (Bonin, 2019). Pasi Ihalainen even speaks of the "parliamentary turn" that conceptual and intellectual history has taken as a result of the massive digitisation of debates and the rise of comparative studies based on these documents (Ihalainen, 2021). Parliamentary proceedings constitute indeed a "remarkable corpus" for political and intellectual history, and a perfect material for computational methods (Bonin, 2019). Recent interest in debates held in transnational institutions and the beginning of computational diplomacy (Cafiero, 2023) make us anticipate that more data will soon be available. Yet, this new abundance of parliamentary data has not generated in history and political science as much work as hoped for (Blaxill, 2020). And while it is becoming increasingly easy to collect transcripts of old parliamentary debates, to our knowledge there is no corpus of historical parliamentary debates encoded in XML-TEI.

In this paper, we present processes developed to allow for that kind of trans-period or transnational research, thinking of the necessary continuity between periods, and allowing for computational methods annotations. We propose an XML-TEI model specifically adapted to the encoding of these documents, be they proceedings from regional and national parliaments or international institutions. To illustrate our model, we work here on the parliamentary debates held in the French Chamber of Deputies[2], transcribed and published between 1889 and 1893 in the *Journal officiel de la République française. Débats parlementaires*[3].

In the first part, we explain the reasons that led us to develop such a model. In the second part, we outline the choices we made to build this model and explain why we had to adapt existing schema. Then, we present the method we adopted to automatically encode that kind of large corpus. Finally, we formulate a proposal to annotate parliamentary debates in TEI with natural language processing analyses, in particular topic modeling.

1. Interoperable historical data for a transperiod and transnational history

National parliamentary data are increasingly available online, with the growing digitisation of parliamentary, legislative and judiciary publications. In France, for example, the debates of the

[2] The Chamber of Deputies was the lower house of the French Parliament during the Third Republic (1870-1940).
[3]The chosen corpus includes 10418 images in JPG format. Images have been downloaded via the Gallica Images API: https://api.bnf.fr/fr/api-iiif-de-recuperation-des-images-de-gallica.

National Assembly and the Senate[4] are only available online. The need to develop a format for the exchange of "machine-readable" parliamentary data quickly became apparent. One example is the OASIS standard Akoma Ntoso, which developed an XML format for encoding legislative and judiciary documents, including parliamentary proceedings (Ervajec et al., 2022b). After noticing the lack of standardization in XML-TEI for encoding parliamentary debates, the CLARIN research infrastructure initiated ParlaClarin in 2019. This project developed a data structuring schema limiting the TEI encoding options to options specifically applicable to parliamentary debates. ParlaMint was formed as a result of the ParlaClarin initiative and has so far developed 27 corpora of contemporary parliamentary debates in XML-TEI, covering the COVID pandemic, with 16 main languages (Ervajec et al., 2022b). Significant efforts have thus been made to publish comparable and multilingual Parliamentary Proceedings Corpora (PPCs) in XML-TEI (Ervajec et al., 2021; Ervajec et al., 2022c) by these two initiatives (Fišer et al., 2018; Erjavec et al., 2023).

The development of an encoding model in TEI for historical proceedings also addresses the need to produce parliamentary data in a machine-readable and interoperable format.

Fig. 1 A page from a parliamentary debate of 26 November 1889[5]

The standards for transcribing parliamentary debates have changed little since they were first introduced in the nineteenth century France (Gardey, 2010). The transcripts of these debates are still published today in a dedicated edition of the *Journal officiel*[6]. From the outset, we assumed that the schema proposed by CLARIN should be easily adaptable for historical proceedings of parliamentary debates. It also seemed appropriate to produce data in a format

[4] The National Assembly is the lower house of the French parliament, the upper house being the Senate.
[5] https://gallica.bnf.fr/ark:/12148/bpt6k63959929/f569.item#
[6] The most recent transcripts of the parliamentary debates can be found in the digital edition of the *Journal officiel*: https://www.legifrance.gouv.fr/liste/debatsParlementaires.

that becomes standard not only for parliamentary proceedings, but more generally for the management and exchange of digital data in the human sciences (Burnard, 2014) and for natural language processing (Piotrowski, 2012).

TEI also offers the possibility to select sub-corpuses more easily. Historians often need to work on smaller corpora of data. For example, they may want to work on a particular speaker and select only his or her speeches; they may want to define a specific time period; or they may be interested in a particular topic and collect only debates on it. It seemed to us that well-chosen annotations should facilitate the creation of such sub-corpuses: for example, we will select the <persName> whose value corresponds to the name of the searched deputy.

TEI has also proved to be a particularly interesting technical choice for the project. One of our objectives is to publish a TEI-compliant corpus of historical documents stored in an eXist-db[7] database and published via TEI Publisher[8]. To publish this data via TEI Publisher, it is indeed necessary to encode it in XML-TEI. Since we respect the standard set by ParlaClarin and ParlaMint, it would therefore be possible to integrate the most recent debates into the platform and thus realise our ambition to offer a corpus of debates covering, in time, the whole of the nineteenth and twentieth centuries.

2. Between modification and adaptation: creating an ODD for historic parliamentary proceedings

From a "proof of concept" perspective, we ourselves worked on a sub-corpus, namely the debates of the 1889-1893 parliamentary cycle[9]. Available on *Gallica*, the digital library of the Bibliothèque nationale de France, the chosen corpus includes 10418 images in JPG format. Images have been downloaded via the Gallica Images API[10].

We designed an encoding model in XML-TEI, formalizing it in a Relax NG schema and documenting it in a *One Document Does it all* (ODD)[11] (Rahtz et al., 2013). While we drew on the ODD produced by ParlaClarin[12] and the work carried out by ParlaMint[13] (Erjavec et al., 2022a; Erjavec et al., 2022c) to create it, an important analysis of this ODD was however

[7] http://exist-db.org/exist/apps/homepage/index.html
[8] https://teipublisher.com/index.html
[9] This parliamentary cycle or "Vème législature / 5th legislature" began on 12 November 1889 and ended on 14 October 1893.
[10] https://api.bnf.fr/fr/api-iiif-de-recuperation-des-images-de-gallica
[11] In order to generate the Relax NG schema and the ODD, we used TEI Roma: https://roma.tei-c.org/. This tool proposed by the TEI Consortium is a web interface for manually creating the specifications of an encoding model.
[12] https://clarin-eric.github.io/parla-clarin/
[13] https://clarin-eric.github.io/ParlaMint/

necessary, both on the textual particularities of our source and on the recommendations of the Text Encoding Initiative. If the proceedings of contemporary debates are very close to those of the 19th century, the annotation rules must nonetheless be adapted to the source being treated[14].

Here we propose to focus on the aspects that we felt were most important in the creation of the ODD and the related schema. One of the challenges was to annotate the names of the speakers, and more broadly the named entities in the proceedings. To recognize these named entities, we use the OCR tool developed in the OCR tool developed within the SoDuCo project[15] which also offers a very efficient NER module on our documents. Currently in private alpha version, this tool has been used, for example, to prepare the data used in Abadie, 2022. It allows the recognition of named entities using the French language model CamemBERT (Martin et al., 2020) fine-tuned on data from trade directories. This functionality is an essential asset for us because it allows, at the cost of a low-cost training[16], to recognise the names of the deputies in the corpus, and thus to consider annotating them and linking them to repositories such the database of French deputies since 1789[17].

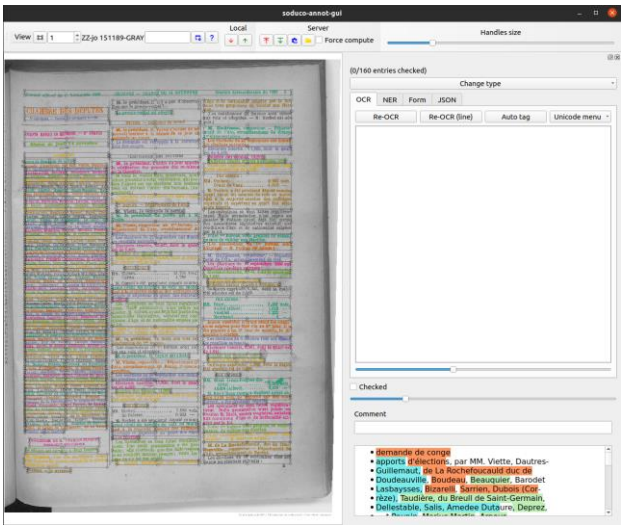


Fig. 2 View of the interface

Another challenge was to annotate the structure of the source. The proceeding always starts with a summary, directly followed by the speeches. It concludes with the next agenda and is completed by additional parts: annexes (results of votes, names of voters, their positions, etc.), reports, erratum, rectifications. The body of the text (the debates themselves) is divided

[14] For example, we have removed the <recordingStmt> element which is unnecessary for the debates of the Third Republic.
[15] https://soduco.github.io/, https://anr.fr/Projet-ANR-18-CE38-0013
[16] CamemBERT was found to require little training data to perform very well on directories (Abadie, 2022).
[17] https://www2.assemblee-nationale.fr/sycomore/recherche

into parts - corresponding to the different topics discussed - introduced with a title. If we want to use TEI annotation to facilitate the selection of sub-corpuses, it is necessary to annotate this structure. For example, many users will want to retrieve only the text of the debates, without the summary or annexes.

ParlaClarin and ParlaMint did not process appendices similar to the ones we were facing. Should they be embedded within the text in the <body> element, or distinguished from the text by including them in the <back> element? We could not determine whether they represented what was said in the sessions, and therefore whether they were an integral part of the spoken discourse, or whether they were included afterwards, as a complement to the proceedings. After consulting several references on the conception of minutes (Martin, 1867; Coniez, 2010; Gardey, 2010), we included each of the complementary parts in the <back> element.

```
<!-- APPENDICES -->
<back>
   <head>Annexes au procès-verbal de la séance du <date when="1889-11-26">mardi 26 novembre
         1889</date>.</head>

   <div xml:id="CR_1889-11-26_vot">

      <!-- VOTE 1 -->
      <div xml:id="CR_1889-11-26_vot1" type="voting" corresp="#discussion7ebureau">
         <head>
            <label>SCRUTIN</label>
            <note xml:id="CR_1889-11-26_n5"><seg xml:id="CR_1889-11-26_n5.1">Sur les
                  conclusions du <num>7e</num> bureau tendant à l'annulation des opérations
                  électorales de la <placeName ref="#lieu_ID"><num>1re</num> circonscription
                     de l'arrondissement de Lorient (Morbihan)</placeName>.</seg></note>
         </head>

         <desc>
            <measure type="nbvoters" quantity="506">Nombre des votants
               <num>506</num></measure>
            <measure type="majority" quantity="254">Majorité absolue <num>254</num></measure>
            <measure type="ayes" quantity="330">Pour l'adoption <num>330</num></measure>
            <measure type="noes" quantity="176">Contre <num>176</num></measure>
         </desc>
```

Fig. 3 <back> element used for the appendices - and especially the votes

The proposal we make for encoding these appendices could be reused to encode the appendices of contemporary proceedings. Indeed, Fig. 4 shows that the appendices recording the results of votes in 2023 are particularly similar to the appendices dating from the 1880s.

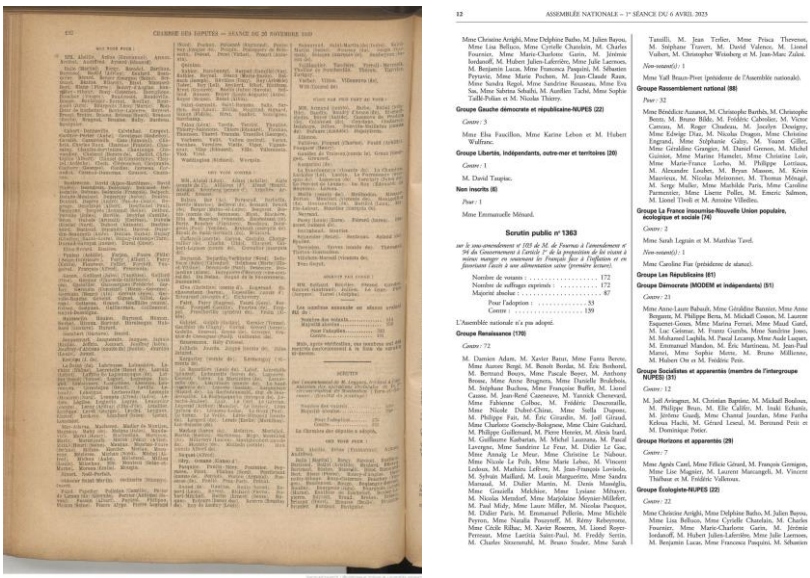
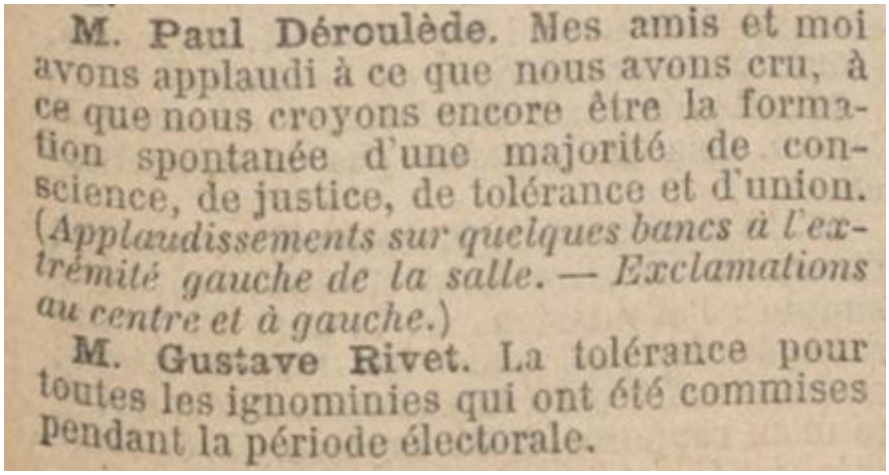


Fig 4 On the left, a page with the results of the vote of 26 November 1886; on the right, a page with the results of the vote of 6 April 2023

Although the proceedings have a homogeneous structure, their content is heterogeneous, and requires the consideration of many different elements[18]. The role of annotation is to highlight the structural and semantic elements of the debates (e.g. the named entities) ; but it was very complicated to make all these scientific objectives coexist while respecting the TEI Guidelines. For example, the Transcriptions of Speech module declares the TEI element <incident> to encode descriptions of incidental sounds (applause, exclamations, etc.). But the TEI does not allow <lb>, <pb> or <cb> elements to be put in <incident>, which does not allow line breaks, or page or column changes to be reported. The Transcriptions of speech module was indeed created to directly transcribe spoken speech recorded using contemporary tools (such as a dictaphone), and not to process older transcriptions of spoken discourse.

[18] The proceedings cover all the constituent elements of a sitting, including speeches, votes and their outcome, as well as the opening and closing of the sitting, interruptions and even the atmosphere.



M. Paul Déroulède. Mes amis et moi avons applaudi à ce que nous avons cru, à ce que nous croyons encore être la formation spontanée d'une majorité de conscience, de justice, de tolérance et d'union. (*Applaudissements sur quelques bancs à l'extrémité gauche de la salle. — Exclamations au centre et à gauche.*)

M. Gustave Rivet. La tolérance pour toutes les ignominies qui ont été commises pendant la période électorale.

(a) Digitized source

```
<lb/><u who="#vot1" ana="speaker">
   <seg>
      M. <persName>Paul Déroulède</persName>. Mes amis et moi
      <lb/>avons applaudi à ce que nous avons cru, à
      <lb/>ce que nous croyons encore être la forma-
      <lb/>tion spontanée d'une majorité de con-
      <lb/>science, de justice, de tolérance et d'union.
      <lb/><incident><desc>(Applaudissements sur quelques bancs à l'ex-
         <!-- Element <lb> cannot be contained in <incident> --> trémité gauche de La salle. — Exclamations
         <!-- Element <lb> cannot be contained in <incident> --> au centre et à gauche.)</desc></incident>
   </seg>

</u>

<lb/><u who="#vot1" ana="speaker">
   <seg>
      <persName>Gustave Rivet</persName>. La tolérance pour
      <lb/>toutes Les ignominies qui ont été commises
      <lb/>Pendant la période électorale.
   </seg>
</u>
```

(b) Encoding model illustrating the problematic case of the placement of <lb> and <incident> elements

Fig. 5 Extract from a speech with an incident at the parliamentary sitting of 26 November 1889

We decided not to keep the <lb> and <cb> elements, but to only annotate the page changes, in order to retain the page number. It is indeed essential for historical analysis to be able to refer to page numbers - for example, to quote the verbatim of a speech. With the help of the TEI list[19], we have chosen to encode page changes occuring in an <incident> as follows:

[19] We especially thank Lou Burnard for his advice.

```
<!-- Change of page within in <incident> -->
<incident>
   <floatingText>
      <body>
         <div>
            <pb n="175"/>
         </div>
      </body>
   </floatingText>
</incident>
```

Fig. 6 Change of page occurring in <incident>

3. Developing an automated annotation strategy

As the corpus represents a very large set of texts to be processed, we envisaged automating TEI tagging from the start. We created Python scripts that would transform the texts OCR'd in JSON format into textual data annotated in XML-TEI format, according to the predefined encoding model[20].

To apply all the tags to the OCR'd data, we needed to find features within the texts. For example, if we wanted to encode speakers' speeches using the <u> element, we needed to be able to find a feature at the beginning and end of the speech in order to locate the opening and closing tags of the <u> element. However, we realized that the proceedings had too few recurring text features and were therefore complex to encode automatically. This led to the creation of an annotation guide[21], which was the main instrument to help with data conversion. Its aim was to guide a manual pre-annotation within the OCR tool, by providing a list of tags and defining their context of use so that they could be employed according to a common standard[22].

[20] Scripts available on Github: https://github.com/FannyLbr/Memoire-AGODA-TNAH2022/tree/f18284d45b83248c4d7938b2bf7984b2b006a832/C%20-%20Encodage%20automatique/C2%20-%20Scripts%20Python

[21] Annotation guide available online: https://github.com/FannyLbr/Memoire-AGODA-TNAH2022/blob/f18284d45b83248c4d7938b2bf7984b2b006a832/C%20-%20Encodage%20automatique/C1%20-%20Guides/guide_annotations_agoda.pdf

[22] The annotation guide we have set up indicates the name of the tags, their context of use, and the modelling of the XML-TEI result of the conversion of the tagged area.

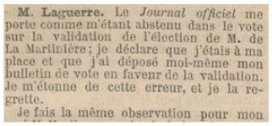
Tags	Uses	XML-TEI	Examples
u seg	Speech corresponding to a single paragraph	<u> <seg>text</seg> </u>	
u-beginning seg	The first paragraph of a speech that extends over a paragraph or more.	<u> <seg>text</seg>	
u-end seg	Last paragraph of a speech that started with one or more paragraphs	<seg>text</seg> </u>	

Fig. 7 Extract from the annotation guide for the <u> element

Transformation rules in Python were written to automatically apply the XML-TEI tags. The Python scripts were used to obtain a compliant and valid XML-TEI file, containing the document's metadata and all the textual content encoded according to the predefined model.


```
{
 "activities": [],
 "addresses": [],
 "box": [
  57.6116701150507,
  1710.0,
  560.3883298849495,
  50.07766597698992
 ],
 "checked": true,
 "comment": "u seg",
 "id": 305,
 "key": [
  0,
  1735
 ],
 "ner_xml": "<PER>M. Paul Déroulède</PER>. Je demande la pa-<0x2029>Tole",
 "origin": "computer",
 "parent": 269,
 "persons": [
  "M. Paul Déroulède"
 ],
 "text_ocr": "M. Paul Déroulede. Je demande la pa-\nrole.",
 "type": "ENTRY"
},
```


(a) JSON result containing a speech and its associated tags

```
def add_utterance(data):
    """
    Add TEI element "u" for each box labelled "u" or "u-beginning" and "u-end
    :param data: dictionary containing all data from JSON
    """
    for i in range(len(data)):
        if "comment" in data[i]:
            if re.search(r"\bu(?!-)\b", data[i]["comment"]):
                data[i]['text_ocr'] = "".join(['<u>', data[i]['text_ocr'], '</u>'])
            elif re.search(r"u-beginning", data[i]["comment"]):
                data[i]['text_ocr'] = "".join(['<u>', data[i]['text_ocr']])
            elif re.search(r"u-end", data[i]["comment"]):
                data[i]['text_ocr'] = "".join([data[i]['text_ocr'], '</u>'])
            else:
                pass
    return data
```

(b) Script of the "add_utterance" function allowing the addition of the <u> element

```
<u><seg>M. Paul Déroulede. Je demande la parole.</seg></u>
```

(c) Result of applying the "add_utterance" function

Fig. 8 Example of automated application of the <u> element from OCR'd data in JSON format

4. Annotate the proceedings with the results of the topic modeling

The parliamentary debates of the Third Republic constitute a massive corpus[23]. Faced with such a large quantity of documents, it is necessary to develop new ways of reading these sources (Clavert, 2014) based on "distant reading" as advocated by Franco Moretti (Moretti, 2013). Distance is here a "condition of knowledge" (Moretti, 2000) because it makes it easier to identify anomalies or patterns that a close reading would not have detected. Parliamentary debates lend themselves particularly well to "computer-assisted analyses", in the sense that the computer facilitates changes of scale (Ihalainen, 2020) with the passage from distant to close reading and vice versa.

To analyze parliamentary proceedings, the use of computers is becoming an increasingly indispensable prerequisite (Bonin, 2019; Ihalainen, 2020; Blaxill, 2022), even though its use in the humanities has decreased since the 1970s, particularly in history

[23] There were 14 parliamentary cycles or legislatures between 1881 and 1940; the debates of a parliamentary cycle are about 10-12,000 pages long (208532 images).

(Lemercier et al., 2019; Kemman, 2021; Salmi, 2021). If, like the authors of *Exploring Big Historical Data*, we believe that "Big Data analysis skills are on the verge of no longer being a 'nice to have' for historians but nearly a necessity" (Graham et al., 2014), we also believe that it is essential to provide turnkey tools to facilitate the reading of these large corpora by the general public, as well as researchers whose needs are limited to consulting the documents. One avenue is to develop interfaces that make it easy to query these corpora (Ihalainen et al., 2022). This type of platform is essential if historians are to become aware of the research potential offered by this digitized source. This is what we propose to develop in the framework of the AGODA project (Puren et al., 2021). It seems essential to us to propose such a platform in the French context, where digital history suffers from a disaffection for quantitative methods (Karila-Cohen et al., 2018; Lemercier et al., 2019), and from a lack of training in this field (Ruiz, 2022).

This platform will offer two levels of reading: on the one hand by creating a corpus that is searchable and readable by humans (close reading); on the other hand, by producing data that can be easily exploited with computational techniques (distant reading). A pitfall of "distant reading" methods is the imposition of inappropriate categories (Karila-Cohen et al., 2018). Rather than imposing manually defined categories, we chose to adopt an inductive approach based on the results of topic modeling. This method allows topics to emerge from the texts themselves, without human intervention (Lemercier et al., 2019; Graham et al., 2022). Laurent Klein and his co-authors propose to use the topics generated with the topic modeling to propose a "new way of reading" newspaper collections (Klein et al., 2015). Exploring parliamentary debates by their subjects has already been widely deployed (Abercrombie et al., 2020). In a first study (Bourgeois et al., 2022), we found that Latent Dirichtet Allocation (LDA) (Blei et al., 2003) performs well on the corpus. It therefore seems relevant to us to create an interface that would allow users to browse the corpus according to the topics that cross it.

We propose to use the topics generated by LDA to enrich the data by annotating the debates with them. Using the <standOff> element, we can store these annotations in XML files. The topic name is chosen by hand; to attach this semantic annotation to the corresponding word list, we chose to use the element which allows us to attach an analytical note to parts of the text. Each word in the text is annotated by a <w> element accompanied by a unique identifier consisting of the document identifier (formed by the prefix "ps" for parliamentary sitting, followed by the date of the sitting) and a number corresponding to the place of the word in the text. A "ref" attribute then associates the topic with the corresponding words. We have also chosen to group the semantic annotations in the <standOff> element to facilitate their

management. These tags are also grouped in a <spanGrp> element associated with a "type" attribute:

Fig. 9 Example of the use of the <standOff> element

Conclusion

In conclusion, this study has demonstrated the value of a novel XML-TEI model for the encoding of historical and contemporary parliamentary proceedings. The model, with its ability to handle diverse temporal and regional contexts, serves as an efficient tool for computational research in the humanities, particularly in historical and political studies. It promotes the continuity of analysis across different periods and political systems, thereby addressing the dichotomy that currently exists in the study of parliamentary debates.

Through the use of our model, as demonstrated in the analysis of debates from the French Chamber of Deputies (1889-1893), we have shown that the massive digitized parliamentary data now available can be made more accessible and analytically tractable. Furthermore, by incorporating natural language processing techniques, especially topic modeling, the model allows for a more comprehensive exploration of the themes and patterns present in the data.

We believe that our XML-TEI model will inspire further advancements in the field, encouraging more scholars to engage with digital and computational approaches in their research. By connecting past and present, and enabling more effective transnational and trans-period research, this model opens up a new research avenue that will enrich our understanding of parliamentary history in its broader context.

While our work represents a significant step forward, we acknowledge that there are further avenues to explore and challenges to tackle. Future research could refine the model's ability to handle linguistic and semantic complexities inherent in the parliamentary data, or explore ways to incorporate additional NLP techniques to enrich the analysis.

Funding

This work was supported by the DataLab of the Bibliothèque nationale de France.

References

Abadie, N., Carlinet, E., Chazalon, J., & Duménieu, B. (2022). A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories. In S. Uchida, E. Barney, & V. Eglin (Eds.), Document Analysis Systems (pp. 445–460). Springer International Publishing. https://doi.org/10.1007/978-3-031-06555-2_30

Abercrombie, G., & Batista-Navarro, R. (2020). Sentiment and position-taking analysis of parliamentary debates: A systematic literature review. Journal of Computational Social Science, 3(1), 245–270. https://doi.org/10.1007/s42001-019-00060-w

Baker, H., Brezina, V., & McEnery, A. (2017). Ireland in British parliamentary debates 1803–2005. Plotting changes in discourse in a large volume of time-series corpus data. Exploring Future Paths for Historical Sociolinguistics. Advances in Historical Sociolinguistics . John Benjamins, Amsterdam, pp. 83-107.

Blaxill, L. (2020). The War of Words : The Language of British Elections, 1880-1914. The Boydell Press, Woodbridge.

Blaxill, L. (2022). Parliamentary Corpora and Research in Political Science and Political History. Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference, 33–34. https://aclanthology.org/2022.parlaclarin-1.5

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. **The Journal of Machine Learning Research**, 3(Jan), 993–1022.

Bonin, H. (2020). From antagonist to protagonist: 'Democracy' and 'people' in British parliamentary debates, 1775–1885. Digital Scholarship in the Humanities, 35(4), 759–775. https://doi.org/10.1093/llc/fqz082

Bouchet, T. (2018). French Parliamentary Discourse, 1789–1914. In P. Ihalainen, C. Ilie, & K. Palonen (Eds.), Parliament and Parliamentarism. A Comparative History of a European Concept (pp. 162–175). Berghahn Books.

Bourgeois, N., Pellet, A., & Puren, M. (2022). Using Topic Generation Model to explore the French Parliamentary Debates during the early Third Republic (1881-1899). Proceedings of the Digital Parliamentary Data in Action (DiPaDA 2022) Workshop (Vol. 3133). CEUR. https://ceur-ws.org/Vol-3133/paper03.pdf

Burnard, L. (2014). What is the Text Encoding Initiative?. OpenEdition Press. https://doi.org/10.4000/books.oep.426

Cafiero, F. (2023). Datafying diplomacy: how to enable the computational analysis and support of international negotiations, Journal of Computational Science, 2023, 102056, https://doi.org/10.1016/j.jocs.2023.102056.

Cheng, J. E. (2015). Islamophobia, Muslimophobia or racism? Parliamentary discourses on Islam and Muslims in debates on the minaret ban in Switzerland. Discourse & Society, 26(5), 562–586. https://doi.org/10.1177/0957926515581157

Chester, D. N., & Bowring, N. (1962). Questions in Parliament. Clarendon Press.

Clavert, F. (2014). Vers de nouveaux modes de lecture des sources. Le temps des humanités digitales. FYP EDITIONS. http://orbilu.uni.lu/handle/10993/34980

Coniez, H. (2010). L'Invention du compte rendu intégral des débats en France (1789-1848). Parlement[s], Revue d'histoire politique, 146-158. https://www.cairn.info/revue-parlements1-2010-2-page-146.htm

Duménieu, B., Carlinet, E., Abadie, N., & Chazalon, J. (2023). Entry Separation using a Mixed Visual and Textual Language Model: Application to 19th century French Trade Directories (arXiv:2302.08948). arXiv. https://doi.org/10.48550/arXiv.2302.08948

Erjavec, T. and Pančur, A. (2021). Parla-CLARIN: a TEI schema for corpora of parliamentary proceedings.https://clarin-eric.github.io/parla-clarin/

Erjavec, T., Kopp, M., Rebeja, P., de Joes, J., and Longejan, B. (2022a). Parla-CLARIN. https://github.com/clarin-eric/parla-clarin

Erjavec, T., Ogrodniczuk, M., Osenova, P. et al. (2023). The ParlaMint corpora of parliamentary proceedings. Lang Resources & Evaluation 57, 415–448. https://doi.org/10.1007/s10579-021-09574-0

Erjavec, T. and Pančur, A. (2022b). The Parla-CLARIN Recommendations for Encoding Corpora of Parliamentary Proceedings. Journal of the Text Encoding Initiative [Online], Issue 14 | April 2021- March 2023, Online since 28 April 2022. URL: http://journals.openedition.org/jtei/4133

Erjavec, T., Pančur, A., and Kopp, M. (2022c). ParlaMint: Comparable parliamentary corpora. https://github.com/clarin-eric/ParlaMint

Fišer, D., Eskevich, M., & de Jong, F. (Eds.). (2018). Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Paris, France. European Language Resources Association (ELRA).

Fišer, D., Eskevich, M., & de Jong, F. (Eds.). (2010). Proceedings of the Second ParlaCLARIN Workshop, Marseille, France. European Language Resources Association (ELRA).

Fišer, D., & Lenardič, J. (2018). CLARIN resources for parliamentary discourse research. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (pp. 2–7). European Language Resources Association (ELRA).

Fišer, D., & Maiti, K. P. de. (2021). »Prvič, sem političarka in ne politik, drugič pa …«: Contributions to Contemporary History, 61(1), Article 1. https://doi.org/10.51663/pnz.61.1.07

Fournier, B., & Pépratx, F. (1991). La majorité politique: Étude des débats parlementaires sur la fixation d'un seuil. In A. Percheron & R. Rémond (Eds.), Âge et vie politique (pp. 85–110). Economica.

Franklin, M. (1993). Parliamentary questions. Clarendon Press.

Galembert, C. de, Rozenberg, O., Vigour, C., & Réseau européen droit et société (2013). Faire parler le parlement: Méthodes et enjeux de l'analyse des débats parlementaires pour les sciences sociales. LGDJ-Lextenso.

Gardey, D. (2010). Scriptes de la démocratie: Les sténographes et rédacteurs des débats (1848–2005). Sociologie du travail, 52(2), Article 2. https://doi.org/10.4000/sdt.13695

Graham, S., Milligan, I., Weingart, S., & Martin, K. (2022). Exploring big historical data: The historian's macroscope (Second edition). World Scientific.

Hirst, G., Feng, V., Cochrane, C., & Naderi, N. (2014). Argumentation, Ideology, and Issue Framing in Parliamentary Discourse. ArgNLP. Frontiers and Connections between Argumentation Theory and Natural Language Processing. Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing. (Vol. 1341). CEUR. https://ceur-ws.org/Vol-1341/paper6.pdf

Ihalainen, P. (2020). European History as a Nationalist and Post-Nationalist Project. Why Europe, Which Europe? A Debate on Contemporary European History as a Field of Research. https://europedebate.hypotheses.org/353

Ihalainen, P. (2021). Parliaments as Meeting Places for Political Concepts. CIH Blog. https://jyx.jyu.fi/handle/123456789/78526

Ihalainen, P., Ilie, C., & Palonen, K. (2018). Parliament and parliamentarism: A comparative history of a european concept. Berghahn Books.

Ihalainen, P., Janssen, B., Marjanen, J., & Vaara, V. (2022). Building and Testing a Comparative Interface on Northwest European Historical Parliamentary Debates: Relative Term Frequency Analysis of British Representative Democracy. Proceedings of the Digital Parliamentary Data in Action (DiPaDA 2022) Workshop (Vol. 3133). CEUR. https://ceur-ws.org/Vol-3133/

Karila-Cohen, K., Lemercier, C., Rosé, I., & Zalc, C. (2018). Nouvelles cuisines de l'histoire quantitative. Annales. Histoire, Sciences Sociales, 73(4), 771–783.

Kemman, M. (2021). Trading zones of digital history. De Gruyter Oldenbourg.

Klein, L. F., Eisenstein, J., & Sun, I. (2015). Exploratory Thematic Analysis for Digitized Archival Collections. Digital Scholarship in the Humanities, 30(suppl_1), i130–i141. https://doi.org/10.1093/llc/fqv052

Lebreton, F. (2022). Vers l'ouverture et l'exploration des débats parlementaires: Étude d'une méthodologie de structuration et d'enrichissement automatique des données. L'exemple des débats à la Chambre des députés durant la Ve législature de la IIIe République (1889-1893). [Mémoire de stage]. https://github.com/FannyLbr/Memoire-AGODA-TNAH2022/blob/main/memoire_TNAH2022_Fanny_LEBRETON.pdf

Lemercier, C. (2021). Un catholique libéral dans le débat parlementaire sur le travail des enfants dans l'industrie (1840). Parlement[s], Revue d'histoire politique, 33(1), 195–206.

Lemercier, C., & Zalc, C. (2019). Quantitative methods in the humanities: An introduction. University of Virginia Press.

Marnot, B. (2000). Les ingénieurs au Parlement sous la IIIe République. CNRS Editions.

Martin, H.-M. (1867). Variétés. Nouveau manuel de sténographie ou Art de suivre la parole en écrivant. Le Constitutionnel : journal du commerce, politique et littéraire. https://gallica.bnf.fr/ark:/12148/bpt6k674517j

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., & Sagot, B. (2020). CamemBERT: A Tasty French Language Model. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7203–7219. https://doi.org/10.18653/v1/2020.acl-main.645

Mela, M. L., Norén, F., & Hyvönen, E. (2022). Proceedings of the Digital Parliamentary Data in Action (DiPaDA 2022) Workshop (Vol. 3133). CEUR. https://ceur-ws.org/Vol-3133/

Moretti, F. (2000). Conjectures on World Literature. New Left Review, 1, 54–68.

Moretti, F. (2013). Distant reading. Verso.

Ouellet, J., & Roussel-Beaulieu, F. (2003). Les débats parlementaires au service de l'histoire politique. Bulletin d'histoire politique, 11(3), 23–40. https://doi.org/10.7202/1060736ar

Piotrowski, M. (2012). Natural language processing for historical texts. Morgan and Claypool.

Puren, M., Pellet, A., Bourgeois, N., Vernus, P., & Lebreton, F. (2022). Between History and Natural Language Processing: Study, Enrichment and Online Publication of French Parliamentary Debates of the Early Third Republic (1881-1899). ParlaCLARIN III at LREC2022 - Workshop on Creating, Enriching and Using Parliamentary Corpora. http://www.lrec-conf.org/proceedings/lrec2022/workshops/ParlaCLARINIII/pdf/2022.parlaclariniii-1.3.pdf

Puren, M., & Vernus, P. (2021). AGODA: Analyse sémantique et Graphes relationnels pour l'Ouverture et l'étude des Débats à l'Assemblée nationale. Inauguration du BnF DataLab. https://hal.science/hal-03382765

Rahtz, S., & Burnard, L. (2013). Reviewing the TEI ODD system. Proceedings of the 2013 ACM Symposium on Document Engineering, 193–196. https://doi.org/10.1145/2494266.2494321

Rheault, L., Beelen, K., Cochrane, C., & Hirst, G. (2016). Measuring Emotion in Parliamentary Debates with Automated Textual Analysis. PLOS ONE, 11(12), e0168843. https://doi.org/10.1371/journal.pone.0168843

Ruiz, E. (2022). Former « au numérique » en sciences humaines et sociales? Propositions d'un historien. In C. Bardiot, E. Dehoux, & E. Ruiz (Eds.), La fabrique numérique des corpus en sciences humaines et sociales. Presses universitaires du Septentrion.

Salmi, H. (2021). What is digital history?. Polity Press.

Van Dijk, T. A. (2010). Political identities in parliamentary debates. European Parliaments under Scrutiny. In C. Ilie (Ed.), European Parliaments under Scrutiny: Discourse strategies and interaction practices (pp. 29–56). John Benjamins Publishing Company.