# Between History and Natural Language Processing: Study, Enrichment and Online Publication of French Parliamentary Debates of the Early Third Republic (1881-1899)

Marie Puren, Aurélien Pellet, Nicolas Bourgeois, Pierre Vernus, Fanny Lebreton

# Between History and Natural Language Processing: Study, Enrichment and Online Publication of French Parliamentary Debates of the Early Third Republic (1881-1899)

**Marie Puren**[1 2], **Aurélien Pellet**[1], **Nicolas Bourgeois**[1], **Pierre Vernus**[3 4], **Fanny Lebreton**[5]
[1]MNSHS-Epitech, [2]Centre Jean Mabillon, [3]LARHRA, [4]Université Lumière Lyon 2, [5] École nationale des chartes
[1]Le Kremlin-Bicêtre (France), [2 5]Paris (France), [3 4]Lyon (France)
{marie.puren, nicolas.bourgeois, aurelien.pellet}@epitech.eu, pierre.vernus@msh-lse.fr

**Abstract**

We present the AGODA (*Analyse sémantique et Graphes relationnels pour l'Ouverture des Débats à l'Assemblée nationale*) project, which aims to create a platform for consulting and exploring digitised French parliamentary debates (1881-1940) available in the digital library of the National Library of France. This project brings together historians and NLP specialists: parliamentary debates are indeed an essential source for French history of the contemporary period, but also for linguistics. This project therefore aims to produce a corpus of texts that can be easily exploited with computational methods, and that respect the TEI standard. Ancient parliamentary debates are also an excellent case study for the development and application of tools for publishing and exploring large historical corpora. In this paper, we present the steps necessary to produce such a corpus. We detail the processing and publication chain of these documents, in particular by mentioning the problems linked to the extraction of texts from digitised images. We also introduce the first analyses that we have carried out on this corpus with "bag-of-words" techniques not too sensitive to OCR quality (namely topic modelling and word embedding).

**Keywords:** Parliamentary debates, France, Third Republic, OCR, XML-TEI, Topic modelling, Word embedding

## 1. Introduction

In this paper, we present the objectives of AGODA[1] (2021-2022) (Puren and Vernus, 2021), one of the five pilot projects supported by the DataLab of the National Library of France[2]. It aims to create an online platform for consulting and exploring parliamentary debates in the Chamber of Deputies (1881-1940), transcribed in the *Journal officiel de la République française. Débats parlementaires. Chambre des députés : compte rendu in-extenso*, available online on *Gallica* (the digital library of the National Library of France[3]), in the form of structured and semantically enriched textual data.

This project has the particularity of bringing together computer scientists (in particular NLP specialists) and historians: the aim is not only to produce annotated data from digitised documents, but also to offer historians functionalities that allow them to explore these sources according to their research interests. The editorialization and enrichment of these data require the design of a workflow adapted to the production and analysis of such large corpora of historical documents. During this project, we try to develop such a workflow, while demonstrating its feasibility and reusability. This is why we have adopted a "proof of concept" approach: we are working on a test sub-corpus, namely the parliamentary debates of the early Third Republic, in order to test our hypotheses on a smaller data set. In the (limited) framework of the AGODA project, we are mainly interested in the parliamentary cycle from 1889 to 1893[4] which is of major interest for historians; but we apply topic modelling and word embedding on a larger corpus (1881-1899) because both methods (and especially word embedding) require a large amount of text.

From January 1881 and throughout the Third Republic[5], the debates in the lower house of the French Parliament were published in the *Journal Officiel* (this is still the case today). Since its establishment in the nineteenth century, the Chamber of Deputies has played a central role in French politics, especially during the Third Republic (1870-1940). Proclaimed on 4 September 1870, constitutionally founded in 1875 as a temporary solution, the Third Republic only became fully republican between 1876 and 1879 with the conquest of the Chamber of Deputies and the Senate by the Republicans. From then on, the Republicans established a parliamentary system of government that placed the Chamber of Deputies at the heart of the system. This is why the government paid particular attention to the reactions of this assembly throughout the Third Republic (Coniez, 2010). We are fortunate to have access to detailed transcripts of the debates held in the Chamber: in the context of the increasing publicisation of parliamentary debates (Lavoinne, 1999), a body of spe-

[1]Analyse sémantique et Graphes relationnels pour l'Ouverture et l'étude des Débats à l'Assemblée nationale.

[2]https://www.bnf.fr/fr/les-projets-de-recherche

[3]Available on Gallica

[4]This parliamentary cycle or "5th *législature*" took place between 12 November 1889 and 14 October 1893.

[5]The Third Republic was the republican system of government in effect in France from September 1870 to July 1940.

cialised civil servants was set up in 1847 to transcribe the debates in detail, while trying to render the naturalness of the discussions (Gardey, 2010).
In this article, we will introduce the rationale of the project, showing what it can bring to history but also to other different disciplines. We will then present the corpus, and the processing and publication chain that we are developing. Finally, we will discuss two examples of exploration - topic modelling and word embedding - that have been tested on the corpus.

## 2. Background and Aims

These sources were digitised and put online between 2008 and 2016, as part of the French digitisation program for legal science[6] (Alix, 2008). For the past sixty years, parliamentary debates have been a source used by the humanities, as shown for example by the work carried out on the British Parliament (Chester and Bowring, 1962; Franklin and Norton, 1993). They are indeed very valuable for historians, particularly those interested in political history (Ouellet and Rousselbeaulieu, 2003) and comparative history (Ihalainen et al., 2016) but also in social, economic or religious history (Lemercier, 2021; Marnot, 2000). We are also aware of the importance that this corpus may have for other disciplines such as the history of law (Fournier and Pépratx, 1991), political science(Van Dijk, 2010), sociology (Cheng, 2015) or linguistics (de Galembert et al., 2013; Hirst et al., 2014; Rheault et al., 2016).
This historical source is not unknown to historians, and its digitisation can be expected to give rise to new projects[7]. But it turns out that the French parliamentary debates in the Chamber of Deputies are still little known to the general public, and are still under-used by specialists (Coniez, 2010). Despite the fact that the entire historical French parliamentary debates are available online, it is still difficult to manipulate these digitised documents, which constitute a particularly large corpus of texts[8]. The apprehension of such material requires a good prior knowledge of the source, and very often, to know precisely what one is looking for. On the other hand, the online availability of Hansard[9], the Anglo-Saxon equivalent of the *Journal Officiel* for parliamentary debates, has led to the emergence of new works in history and political science (Bonin, 2020). More generally, access to digitised and ocerised debates seems to have a positive effect on the number of historical works using these documents (Mela et al., 2022). The same effect can be observed for other disciplines using textual data from contemporary debates (Fišer et al., 2018; Fišer et al., 2020).
The Hansard also allows its users to explore the debates in a very intuitive way and makes the textual data easily exploitable. We believe that building a similar platform for French parliamentary debates, especially historical ones, would increase the number of users and thus the exploitation of these documents. The AGODA project therefore aims to facilitate access to these documents for Internet users, whether they are researchers or the general public. AGODA is therefore not only about doing research, but also about promoting a little-known heritage collection.

## 3. Related Works

The AGODA project is part of a wider movement to improve knowledge and exploitation of parliamentary data. Two trends, which are not exclusive, can be distinguished: on the one hand, the production of corpora exploitable by researchers and, on the other, the desire to facilitate the exploration of these debates by the general public.
The ParlaClarin (Fišer and Lenardič, 2018) and ParlaMint (Erjavec et al., 2022b) projects propose to produce comparable and multilingual Parliamentary Proceedings Corpora (PPCs). These two projects have produced corpora according to the XML-TEI standard (TEI, 2017), accompanied by adapted XML schemas (Erjavec et al., 2022a; Erjavec et al., 2022c). As part of her dissertation, Naomi Truan also produced a corpus of parliamentary debates encoded in XML-TEI (Truan, 2019; Truan and Romary, 2021; Tóth-Czifra and Truan, 2021). The production of this type of resource facilitates the publication of works exploiting these textual data to better understand the French political discourse (Diwersy et al., 2018; Diwersy and Luxardo, 2020; Blaette et al., 2020). The online publication of Hansard is part of this line of work, but it also offers an interface for the visualisation of the debates, facilitating their exploration by various users (researchers and the general public). The *Fabrique de la loi* project[10] aims to propose a new way of exploring parliamentary debates by making it possible to follow the evolution of a law - from its proposal to its publication. AGODA is at the crossroads of these different projects, producing both new PPCs respecting the XML-TEI standard and giving access to these corpora via an online platform.
From a methodological point of view, parliamentary debates also constitute an excellent case study for the development of tools for publishing and exploring large historical corpora. While digitisation provides access to an increasingly large mass of textual data, especially for history, it requires the implementation of "new modes of reading sources" (Clavert, 2014). Building

[6]Cf. La Mission de recherche Droit et Justice et le programme national de numérisation concertée en sciences juridiques

[7]Such as Political Representation - Tensions between Parliament and the People from the Age of Revolutions to the 21st Century

[8]There were 14 parliamentary cycles or *législatures* between 1881 and 1940 ; and the debates for a parliamentary cycle consist of about 10-12,000 pages.

[9]For example, the British Hansard : https://hansard.parliament.uk/

[10]https://www.lafabriquedelaloi.fr

on the work carried out during the ANR project TIME-US (2018-2021)[11], we propose to make it easier to access to, to search into and to visualize parliamentary debates. TIME-US has indeed led to the publishing of a corpus of contemporary historical documents respecting TEI standards stored in an eXist-db database and queryable through TEI Publisher (Chagué et al., 2019; Le Fourner, 2019; Généro, 2020; Généro et al., 2021).

## 4. Data Set

The issues of the *Journal Officiel* available on *Gallica* have been digitised by the National Library of France and the archives of the National Assembly. Between 1881 and 1899, 2596 issues were published, or 50791 images[12]. For the parliamentary cycle 1889-1893, 10418 images are available. The digital images of the documents available in JPG format can be downloaded via the *Gallica* API. The debates are also downloadable in TXT format. *ABBY FineReader* automatic transcription (OCR) software was used to extract the text of the debates on the fly, as they were being digitised. The generated text was made available online, but without extensive post-correction.

We roughly measured the quality of the OCR by estimating the number of words present in the ocerised texts, from the French dictionnary (which consists of a word frequency list) provided with the Python library *pyspellchecker*[13]. In practice, we selected 20 documents at random between 1889 and 1892. We counted the number of unique words present in the text automatically extracted from these digitised documents. We then calculated the number of unique words present in both the OCR results and the dictionary. We then divided this result by the total number of unique words in the digitised texts. We used unique words so that our results would not be biased, for example, by the name of a speaker that would not be present in the dictionary.

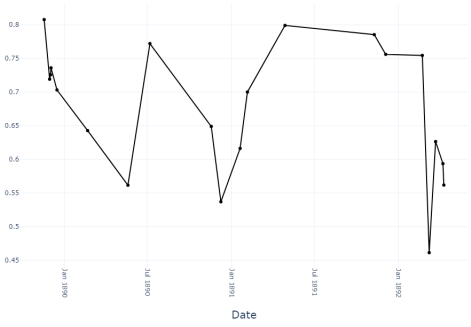


Figure 1: OCR quality evaluation (OCR retrieved from *Gallica*)

Figure 1 shows that the quality of the OCR varies greatly. Various factors explain the high variability of the quality of the ocerised texts. The curvature of the page, due to the binding, results in "curving" the text, sometimes even cutting off part of it or casting shadows on the pages. Besides the quality of the documents themselves (stains, overprinted text) is also at issue.

## 5. Processing and Publishing Chain

AGODA aims to offer users easier access to these debates (full-text search, navigation, selection of homogeneous sub-corpuses, etc.) and to allow them to explore and analyse this corpus with "distant reading" methods (Moretti, 2013).

### 5.1. Ocerisation and Postprocessing

Large-scale analysis of digitised historical sources requires the ability to extract data (in our case, textual data) from these documents. As we have seen, the quality of OCR is not sufficient to provide a satisfactory online browsing experience; it could also have a negative impact on the analyses conducted on these texts (van Strien et al., 2020). We chose to ocerise the text again, in order to obtain a better quality result. We first used Tesseract, an open source OCR engine [14], but the results obtained were somewhat mitigated on our corpus as shown in Figure 2. We used both default Tesseract method and Tesseract pre-trained on a French corpus. To estimate the quality of the OCR, we use the same method as described in section 4. Although the Tesseract French model greatly increases the quality of the OCR, the results are sometimes inferior to those obtained with *ABBY FineReader*.

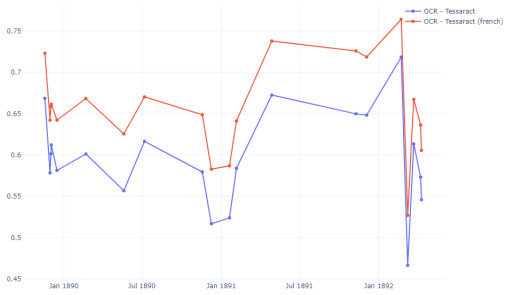


Figure 2: OCR quality evaluation (Tesseract)

As a result, we decided to use the OCR tool developed in the framework of the ANR SODUCO project[15]. Figure 3 shows a view of the tool. It should be noted that this tool not only ocerises texts but also recognises named entities (such as speakers' names). OCR is performed using the PERO OCR engine (Kišš et al., 2021;

[11] https://timeus.hypotheses.org/
[12] One image corresponding to one page.
[13] https://pyspellchecker.readthedocs.io/en/latest/index.html
[14] https://tesseract-ocr.github.io/
[15] https://soduco.github.io/, https://anr.fr/Projet-ANR-18-CE38-0013

Kodym and Hradiš, 2021; Kohút and Hradiš, 2021), which performs particularly well on historical printed texts. Currently in private alpha version, this tool was used, for example, to prepare the data used in (Abadie et al., 2022). This dataset, which will be freely available on Zenodo [16], consists of texts ocerised from a corpus of printed trade directories of Paris from the XIXth century.

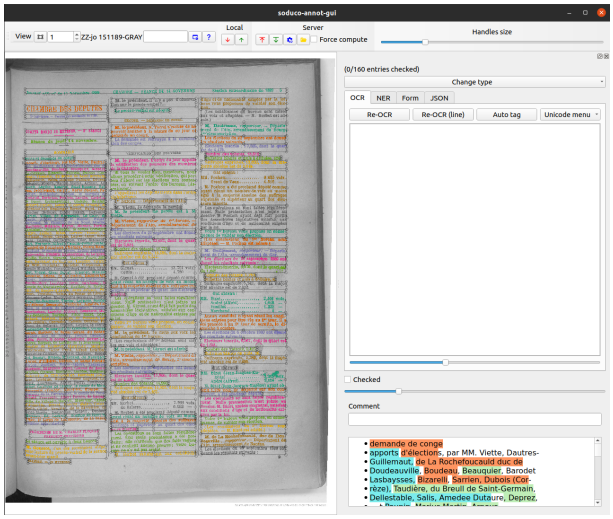


Figure 3: Interface of the OCR tool developed by SODUCO

With this tool, we obtain the ocerised texts in JSON format. But we are aware that the use of an OCR engine will not allow us to obtain error-free texts, and that it will be necessary to go through a post-correction phase. We used the Python library *pyspellchecker*, setting up a simple spell checking algorithm. *pyspellchecker* uses a dictionary based on a word frequency list extracted from film and TV subtitles (*OpenSubtitles*). To correct misspelled words, *pyspellchecker* uses the Levenshtein distance. This metric measures the distance between two words based on the characters that compose them. The distance is the shortest number of operations required to move from one word to another using three operations: insertions, deletions or substitutions. We used a distance of 1 as a post-correction rule, i.e. we allowed only one transformation to correct erroneous words. The results were disappointing, as it turns out that the French dictionary is not suitable for our corpus. It contains too much contemporary vocabulary, especially by integrating English words ("PC" for "*ordinateur*", "deal" instead of "*accord*", etc.). Some faulty words were thus corrected with words from the dictionary that had no relation to their original meaning: we can speak of "over-interpretation" of the correction. We are thus creating a dictionary based on French novels published between 1870 and 1920, texts from French press documents from the same period[17], and the list of names of MPs elected to the Chamber between 1889 and 1893[18].

## 5.2. Annotation in XML-TEI

These corrected texts will then be annotated in XML-TEI. The modelling in TEI will be formalised by using an adapted XML schema, created by means of a documented ODD (Rahtz and Burnard, 2013).
To create this ODD, we drew on the ODD produced by ParlaClarin (Erjavec and Pančur, 2021), and the work carried out by ParlaMint (Erjavec et al., 2022a; Erjavec et al., 2022c) on contemporary parliamentary debates. In the case of France, the rules governing the transcription of debates were set in the 19th century (cf. Section 1); the records of today's debates are therefore very similar to those produced under the Third Republic. A first version of the ODD as well as examples of encoding the parliamentary debates in TEI can be found on Github[19]
However, the annotation rules need to be adapted to the historical sources we are working on. For example, we have removed the <recordingStmt> element (and the elements contained in this element), as it will be useless in the context of AGODA. In addition, the presentation of the votes and their results is quite different: they are referred to in an appendix, and one finds there both numerical results and the names of the voters. This presentation of the results requires a modification of the proposed model for contemporary debates. There is also the question of layout: the TEI does not allow the use of <pb/>, <cb/> or <lb/> elements in the <incident> element that we use to encode all events disrupting the debates (usually interventions by other députies). We have adopted a middle ground: we do not retain the page layout (columns and line breaks), but we do retain the page number with a <pb/> element contained within a <floatingText>. .
We will focus on two types of annotations specific to our project. Firstly, we wish to keep track of the corrections made to the ocerised text. For this purpose, we propose to use the <corr> tag which allows us to mark the corrected text string and the nature of the correction:

```
[...]
<p>dans le scrutin sur
    <corr>
        <orig>lçi</orig>
        <reg>la</reg>
    </corr> motion de M. Millerand</p>
[...]
```

We also wish to add semantic annotations resulting from analyses performed on our corpus with topic mod-

[16] https://zenodo.org/record/6394464

[17] *OCR corrigé de documents de presse de Gallica*

[18] Extracted from the *Base de données des députés français depuis 1789*

[19] https://github.com/mpuren/agoda/tree/ODD

elling. Topic modelling is an unsupervised learning method that discovers the latent semantic structures of a text corpus, without using semantic and lexical resources (Blei et al., 2003)[20]. Basically, the result of topic modelling consists of weighted lists of words (each list corresponding to one topic). The topic name is not generated automatically, but chosen by hand according to the distribution of the vocabulary in the list of words considered. To attach this semantic annotation to the corresponding list of words, we chose to use the mechanism offered by the  element which allows to attach an analytical note to passages in the text. Each word in the text is encoded with a <w> element accompanied by a unique identifier composed of the document identifier[21] 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 with the value "topic":

```
[...]
<standOff>
    <spanGrp type="topic">
        <span target="#ps1895022_119">
        army</span>
        <span target="␣#ps1895022_123">
          colonisation</span>
</spanGrp>
</standOff>
<text>
    <body>
        <p> <w xml:id="ps1895022_116">
        une</w>
            <w xml:id="ps1895022_117">
            partie</w>
            <w xml:id="ps1895022_118">
            du</w>
            <w xml:id="ps1895022_119">
            matériel</w>
            <w xml:id="ps1895022_120">
            de</w>
            <w xml:id="ps1895022_121">
            guerre</w>
            <w xml:id="ps1895022_122">
            à</w>
            <w xml:id="ps1895022_123">
            Madagascar</w>.</p>
    </body>
</text>
[...]
```

Given the size of the corpus, we intend to use a method of automatic annotation of these documents. We are currently developing rule-based Python scripts to transform the files into JSON obtained with the OCR tool[22].

### 5.3. Online Publication with eXist-db and TEI Publisher

The TEI-encoded corpus will be stored in an eXist-db database[23]. TEI Publisher application is able transform the source data, stored in the XML database, into HTML web pages for publication[24]. The parliamentary debates will thus be available to users online as a digital edition, and integrated into an application context with the addition of navigation, full-text search and facsimile display. In addition, new functionality can be added as required using Web Components technology. The AGODA project will use components natively offered by TEI Publisher, implement components developed in the framework of TIME-US[25], and create new ones.

## 6. Topic Modelling and Word Embedding Applied to Parliamentary Debates

We also wish to facilitate the exploration of these debates by offering new ways to "reading" these documents, in particular by allowing users to use the results of the analyses carried out on the corpus. To gain an in-depth understanding of these documents, it is indeed necessary to adopt computational methods to analyse such a large corpus of sources (Pančur and Šorn, 2016; Bonin, 2020). As seen in section 5.2, we also plan to use the possibilities offered by the XML-TEI to add and make accessible the semantic annotations resulting from downstream NLP tasks such as Latent Dirichlet Allocation and Latent Semantic Analysis. In this section, we will present some examples of lexical analysis that have been performed on the parliamentary corpus (Bourgeois et al., 2022), using the original ocerised text provided by the Bibliothèque nationale de France.

### 6.1. LDA

Latent Dirichlet Allocation is a Bayesian model based on a strong hypothesis (Blei et al., 2003), that fits extremely well our corpus. The underlying model is that there exist hidden variables, namely the topics, which consist of weighted lists of words (the more significant, the higher their probability). Then, every text from the corpus is generated by (1) picking at random a limited number of topics and (2) selecting words from these topics, according to their probability distribution. The role of LDA is to revert this generation process in order

[20]The method is described in detail in section 6.1

[21]Formed by the prefix "ps" for parliamentary sitting, followed by the date of the sitting

[22]Following the example of the scripts developed for XML in the TIME US project, such as the LSE-OD2M script (https://github.com/TimeUs-ANR/LSE-OD2M), or those written by Victoria Le Fourner (https://gitlab.inria.fr/almanach/time-us/schema-tei).

[23]http://exist-db.org/exist/apps/homepage/index.html

[24]https://teipublisher.com/index.html

[25]The corpora produced by the project are accessible online via a TEI Publisher instance.

to retrieve the original topics, with the hope that their statistical coherence reflects some semantic homogeneity.

| Topic 8 | Topic 11 | Topic 15 |
|---|---|---|
| salaire | général | pari |
| question | commission | télégraphe |
| gouvernement | régiment | faire |
| jour | troupe | ingénieur |
| patron | monsieur | train |
| chambre | année | ligne |
| droit | jeune | chambre |
| syndicat | temps | personnel |
| délégué | faire | etat |
| monsieur | corps | administration |
| travail | soldat | employé |
| travaux | ministre | poste |
| ministre | homme | public |
| grève | loi | travaux |
| faire | an | service |
| mineur | guerre | agent |
| mine | service | ministre |
| loi | militaire | fer |
| compagnie | officier | chemin |
| ouvrier | armée | compagnie |

Table 1: Three topics among 40: the working class (8), the army (11) and the state infrastructures (15).

The difficulties of LDA include determining the number of topics, ensuring their coherence, naming them and aggregating those who are highly correlated (Newman et al., 2010). We can for instance produce a large number of topics (Table 1 shows only 3 of the 40 topics identified), then use an agglomerative clustering to build coherent classes and proof-check them with a qualitative survey. Hence we obtain 15 classes with each a strong identity and limited correlation (Figure 4).

The main drawback of this analysis is that a single parliamentary sitting is in fact a rather long text, in which a possibly large sequence of topics are addressed one after the other. It is therefore preferable to divide it into several smaller chunks of texts that better fit the hypothesis of the model. Theoretically, the structure of the document provides a perfect tool for this division, since the different parts of a parliamentary session are easily recognisable by their titles in capital letters. However, the recognition of these titles is very imperfect in the original OCR, so we resorted to fixed-length divisions. We hope that as the quality of the text improves, we will be able to use a semantic-based division instead of this arbitrary division.

### 6.2. Word Embedding

By definition, Latent Dirichlet Allocation builds a limited number of large semantic units and allow little

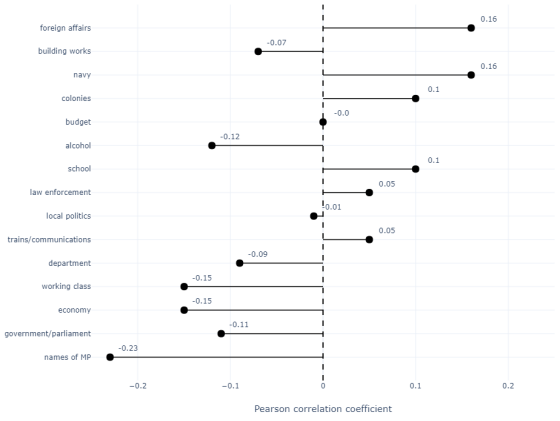


Figure 4: Correlation between the topic "army" and the other identified topics (by month).

control over the process. Alternatively, we may use word embedding to reduce the dimension of the original space from tenth of thousands of forms to a hundred of axes, and then apply classical data science tools such a clustering or correlation analysis on the reduced space (Mikolov et al., 2013). Word embedding has also shown its value in the study of parliamentary debates (Rheault and Cochrane, 2020).

We used a CBOW (Continuous Bag of Words) model for dimension reduction and an unsupervised classification algorithm - in this case DBSCAN (density-based spatial clustering of applications with noise) - to group words into clusters. This method worked well, mainly because the sample size is huge in terms of vocabulary and because similar patterns tend to occur regularly throughout the corpus (discussions on military service, taxes, colonies, etc.).

With word embedding, we obtain a large number (113) of highly coherent clusters (Table 2 shows three of the 113 clusters identified), which we can study in relation to each other, or in relation to other parameters such as time. We can recombine them, for example through agglomerative clustering (Figure 5): with some choices of linkage, we can find superclasses that are very similar to the topic models ; while with others, we get more detailed information about some aspects of the corpus.

However, word embedding is probably more sensitive than LDA to the quality of the OCR, since a clustering of documents requires that each text belongs to a single class, whereas several topics can be combined. Therefore, a good segmentation of the debate reports (according to the different parts of a parliamentary sitting) should have a significant impact on the results.

## 7. Conclusions

The AGODA project aims to produce a corpus of parliamentary debates, based on digitised old documents. The aim is to produce a resource that can be used

| Cluster 55 | Cluster 68 | Cluster 70 |
|---|---|---|
| victimes | divorce | enveloppes |
| inondations | epoux | timbres |
| secourir | mariage | poste |
| eprouvees | conjugal | postale |
| orages | divorces | timbre |
| sinistres | adultere | recepisses |
| grele | conjugale | postes |
| secours | remarier | postaux |
| venir | separation | telegraphes |
| infortunes | indissolubilite | colis |
| ravages | conjoints | fixe |
| miseres | mutuel | recouvrements |
| catastrophe | separations | graphes |
| evenements | mari | postales |
| repartition | mariages | taxe |
| incendies | femme | decide |
| soulager | conjoint | soit |

Table 2: Three clusters among 113: storms (55), divorce (68) and the post office (70).


Figure 5: t-SNE projection of the centroïds of the clusters.

both by historians and by researchers from other disciplines (in particular, linguists, science-politicians, sociologists). To this end, AGODA has three objectives: (1) to produce a new linguistic resource for French, respecting the TEI standard; (2) to create a processing chain that will be reusable in other contexts; (3) to make this corpus better known, by setting up an online consultation interface.

As in any project working with digitised ancient documents, one of the main obstacles is to extract (as clean as possible) text from images. We hope to achieve a very low error text at the end of the project, by combining both re-ocerisation of the documents and post-correction of the texts. We also hope to improve the results we obtain with topic modeling and word embedding. Although these "bag-of-words" techniques are not as sensitive to OCR quality as specific tasks such as name entity recognition, there is a strong incentive to use corrected text to perform natural language processing (van Strien et al., 2020; Mutuvi et al., 2018).

## 8. Acknowledgements

We would like to thank the National Library of France and Huma-Num for their support in the framework of the BnF DataLab.

## 9. Bibliographical References

Abadie, N., Carlinet, E., Chazalon, J., and Dumenieu, B. (2022). A Benchmark of Named Entity Recognition Approaches in Historical Documents. Application to 19th Century French Directories. DAS 2022 15th IAPR International Workshop on Document Analysis Systems `https://das2022.univ-lr.fr/`, La Rochelle.

Alix, Y. (2008). La numérisation concertée en sciences juridiques. *Bulletin des bibliothèques de France (BBF)*, 5:93–94.

Blaette, A., Gehlhar, S., and Leonhardt, C. (2020). The Europeanization of Parliamentary Debates on Migration in Austria, France, Germany, and the Netherlands. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 66–74, Marseille, France, May. European Language Resources Association.

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

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.

Bourgeois, N., Pellet, A., and Puren, M. (2022). Using Topic Generation Model to explore the French Parliamentary Debates during the early Third Republic (1881-1899). In *DHNB 2022 – Digital Humanities in Action - Workshop "Digital Parliamentary Data in Action*.

Chagué, A., Le Fourner, V., Martini, M., and Villemonte De La Clergerie, E. (2019). Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non-uniforme ?

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.

Chester, D. N. and Bowring, N. (1962). *Questions in parliament*. Clarendon Press, London.

Clavert, F. (2014). Vers de nouveaux modes de lecture des sources. In *Le temps des humanités digitales*. FYP EDITIONS.

Coniez, H. (2010). L'Invention du compte rendu intégral des débats en France (1789-1848). *Parlement[s], Revue d'histoire politique*, 2(14):146–159.

de Galembert, C., Rozenberg, O., and Vigour, C. (2013). *Faire parler le parlement: méthodes et enjeux de l'analyse des débats parlementaires pour les sciences sociales*. LGDJ-Lextenso éditions, Issy-les-Moulineaux.

Diwersy, S. and Luxardo, G. (2020). Querying a large annotated corpus of parliamentary debates. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 75–79, Marseille, France, May. European Language Resources Association.

Diwersy, S., Frontini, F., and Luxardo, G. (2018). The Parliamentary Debates as a Resource for the Textometric Study of the French Political Discourse. In *Proceedings of the ParlaCLARIN@LREC2018 workshop*, Miyazaki, Japan.

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., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Steingrímsso, S. B. Çağrı Çölteki, de Doeso and-Katrien Depuydto, J., Agnolonio, T., Venturio, G., Pérezo, M. C., de Macedoo, L. D., Navarrettao, C., Luxardoo, G., Cooleo, M., Raysono, P., Morkevičiuso, V., Krilavičiuso, T., Darġiso, R., Ringo, O., van Heusdeno, R., Marx, M., and Fišer, D. (2022b). The ParlaMint corpora of parliamentary proceedings. In *Language Resources and Evaluation*.

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

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

Darja Fišer, et al., editors. (2018). *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Darja Fišer, et al., editors. (2020). *Proceedings of the Second ParlaCLARIN Workshop*, Marseille, France, May. European Language Resources Association.

Fournier, B. and Pépratx, F. (1991). La majorité politique : Étude des débats parlementaires sur la fixation d'un seuil. In Annick Percheron et al., editors, *Age et politique*, La vie politique, pages 85–110. Economica, Paris.

Franklin, M. and Norton, P. (1993). *Parliamentary questions*. Oxford University Press, Oxford.

Gardey, D. (2010). Scriptes de la démocratie : les sténographes et rédacteurs des débats (1848–2005). *Sociologie du travail*, 52(2).

Généro, J.-D., Chagué, A., Le Fourner, V., and Puren, M. (2021). Transcribing and editing digitized sources on work in the textile industry. In *Rémunérations et usages du temps des hommes et des femmes dans le textile en France de la fin du XVIIe au début du XXe siècle*, Lyon, France. Manuela Martini.

Généro, J.-D. (2020). Le corpus des Ouvriers des deux mondes : des images et des URLs. Billet du carnet de recherche de l'ANR Time Us relatif aux fichiers XML-TEI de transcription des volumes des Ouvriers des deux mondes et au lien entre ceux-ci et les images numérisées d'origine.

Hirst, G., Feng, V. W., C. C., , and Naderi, N. (2014). Argumentation, ideology, and issue framing in parliamentary discourse. In A.Z. Wyner E. Cabrio, S. Villata, editor, *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*.

Ihalainen, P., Ilie, C., and Palonen, K. (2016). *Parliament and Parliamentarism: A Comparative History of a European Concept*. Berghahn Books, Oxford, NY.

Kišš, M., Beneš, K., and Hradiš, M. (2021). At-st: Self-training adaptation strategy for ocr in domains with limited transcriptions. In J. Lladós, et al., editors, *Document Analysis and Recognition – ICDAR 2021. ICDAR 2021. Lecture Notes in Computer Science*, pages 130–146. Cham: Springer.

Kodym, O. and Hradiš, M. (2021). Page layout analysis system for unconstrained historic documents. In J. Lladós, et al., editors, *Document Analysis and Recognition – ICDAR 2021: 16th International Conference*, page 492–506, Lausanne, Switzerland, september. Heidelberg: Springer-Verlag.

Kohút, J. and Hradiš, M. (2021). Ts-net: Ocr trained to switch between text transcription styless. In J. Lladós, et al., editors, *Document Analysis and Recognition – ICDAR 2021. ICDAR 2021. Lecture Notes in Computer Science*, pages 130–146. Cham: Springer.

Lavoinne, Y. (1999). Publicité des débats et espace public. *Études de communication*, 22:115–132.

Le Fourner, V. (2019). Étude de la structuration automatique et de l'éditorialisation d'un corpus hétérogène.

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*, pages 197–208.

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

Mela, M. L., Norén, F., and Hyvönen, E. (2022). Digital parliamentary data in action (dipada 2022). workshop co-located with the 6th digital humanities in the nordic and baltic countries conference (dhnb 2022). `https://dhnb.eu/conferences/dhnb2022/workshops/dipada/`.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *ICLR*.

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

Mutuvi, S., Doucet, A., Odeo, M., and Jatowt, A. (2018). Evaluating the Impact of OCR Errors on Topic Modeling. In *Maturity and Innovation in Digital Libraries. 20th International Conference on Asia-Pacific Digital Libraries, ICADL 2018, Hamilton, New Zealand, November 19-22, 2018, Proceedings*, pages 3 – 14.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic Evaluation of Topic Coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108.

Ouellet, J. and Roussel-Beaulieu, F. (2003). Les débats parlementaires au service de l'histoire politique. *Bulletin d'histoire politique*, 11(3):23–40.

Pančur, A. and Šorn, M. (2016). Smart big data : Use of slovenian parliamentary papers in digital history. *Contributions to Contemporary History*, 56(3):130–146.

Puren, M. and 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.

Rahtz, S. and Burnard, L. (2013). Reviewing the TEI ODD system. In *Proceedings of the 2013 ACM Symposium on Document Engineering*, DocEng '13, page 193–196, New York, NY, USA. Association for Computing Machinery.

Rheault, L. and Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, 28(1):112–133.

Rheault, L., Beelen, K., Cochrane, C., and Hirst., G. (2016). Measuring Emotion in Parliamentary Debates with Automated Textual Analysis. *PLoS ONE*, 11(12).

TEI Consortium, (2017). *TEI P5: Guidelines for electronic text encoding and interchange*.

Tóth-Czifra, E. and Truan, N. (2021). Creating and analyzing multilingual parliamentary corpora.

Truan, N. and Romary, L. (2021). Building, Encoding, and Annotating a Corpus of Parliamentary Debates in XML-TEI: A Cross-Linguistic Account. *Journal of the Text Encoding Initiative*.

Truan, N. (2019). Débats parlementaires sur l'Europe à l'Assemblée nationale (2002-2012). `https://hdl.handle.net/11403/fr-parl/v1.1`. ORTOLANG (Open Resources and TOols for LANGuage).

Van Dijk, T. A. (2010). Political identities in parliamentary debates. european parliaments under scrutiny. In Cornelia Ilie, editor, *European Parliaments under Scrutiny: Discourse strategies and interaction practices*, pages 29–56. John Benjamins Publishing Company.

van Strien, D., Beelen, K., Ardanuy, M., Hosseini, K., McGillivray, B., and Colavizza, G. (2020). Assessing the Impact of OCR Quality on Downstream NLP Tasks. *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, 21:484–496.