Using Topic Generation Model to explore the French Parliamentary Debates during the early Third Republic (1881-1899)

Nicolas Bourgeois, Aurélien Pellet, Marie Puren

HAL Id: hal-03526254
https://hal.science/hal-03526254v2

Submitted on 7 Mar 2022

Using Topic Generation Model to explore the French Parliamentary Debates during the early Third Republic (1881-1899)

Nicolas Bourgeois[a], Aurélien Pellet[a] and Marie Puren[a,b]

[a]*Méthodes Numériques pour les Sciences de l'Humain et de la Société (MNSHS), Epitech, 14-16 rue Voltaire, 94270 Le Kremlin-Bicêtre, France*

[b]*Centre Jean-Mabillon (CJM), École nationale des chartes, 65 rue de Richelieu, 75002 Paris, France*

Abstract

In this long paper, we use NLP techniques to explore two decades (1881-1899) of parliamentary debates of the French Third Republic (1870-1940), and more specifically to analyse the importance of the army in the political debate. We use Latent Dirichlet Allocation to partition the vocabulary into topics, and then study the distribution of the topic "army" over time. We also examine its connection with other topics, in relation to the main political and military events of the period.

Keywords

Natural Language Processing, Topic Modelling, Parliamentary Debates, France, Early Third Republic (1881-1899)

1. Introduction

In this paper, we present the preliminary work we have carried out on a set of parliamentary reports from the early years of French Third Republic (1881-1899), extracted from the *Journal officiel de la République française. Débats parlementaires. Chambre des députés : compte rendu in-extenso*, available online on the digital library of the National Library of France[1]. During the French Third Republic, which was the republican system of government in effect in France from September 1870 to July 1940, the debates in the lower house of French Parliament (the upper house being the Senate) have been carefully recorded and published in the *Journal Officiel*. Elected by universal male suffrage for four years, the Chamber of Deputies was created by the Constitutional Laws of 1875. We are working on debates held since 1881, and not 1876 (date of the first election of the new Chamber of Deputies), i.e. from the end of the second legislature[2]

[1]Available on Gallica

[2]Each parliamentary cycle is numbered and called *législature*.

onwards[3]. This choice is dictated by the document we are working on: it is from 1881 that the debates held in the Chamber of Deputies are recorded in a publication specifically dedicated to them[4].

The Chamber of Deputies played a considerable political role, especially in the nineteenth century. At that time, the government paid particular attention to this assembly [1]. We thus have access to the full report of the debates, written by a body of specialised civil servants set up in 1847, whose techniques aim to recreate the naturalness of the deliberations [2]. Parliamentary debates are therefore an essential historical source for political history [3], but also for other historical fields, since they make it possible to follow the major stages in the development of the legislative framework of various social, economic, religious or cultural fields of activity [4]. They are also of interest to other disciplines: political science, sociology, linguistics [5], or legal history [6].

However, while all parliamentary debates since the French Revolution were made available online between 2009 and 2016, this has not prompted a new wave of research. Although they constitute a fundamental democratic institution, debates are indeed little known by the general public and little studied by specialists [1]. On the other hand, the availability of its Anglo-Saxon counterpart, the *Hansard*[5], in the form of exploitable textual data, has stimulated new research in history and in political science [7] but also in linguistics and natural language processing [8]. The form of the French debates and the means made available to users to read them online, make them a difficult source to work with: to navigate through the digitised reports, it is best to already know what you are looking for (for example: to search for debates on a law carried out on a specific date). It is possible to do a full-text search within an issue (which corresponds to a parliamentary sitting), but this does not allow the user to explore the corpus as a whole, especially if he or she is interested in a major topic that has been debated over several years.

Fortunately it is possible to extract the text of these digitized documents. From a methodological point of view, parliamentary debates thus constitute an excellent case study for the computational exploration of large historical corpora. While digitisation provides access to an increasingly large amount of historical data, it requires the development of new ways of reading digitised ancient sources [9], such as the methods offered by "distant reading" as defined by Franco Moretti [10]. Within the framework of the AGODA[6] project, funded by the National Library of France [11], our team is working on the development of tools to facilitate the exploration of this corpus. As part of this work, we propose to use topic modelling, a method that is particularly appropriate for the study of large historical corpora [12].

Topic modelling has shown its value in analysing similar sources, in particular the press (such as in [13] or [14]). Such corpora, large in volume, serial, and crossed by many different topics that evolve over time, are well suited to a topic-based exploration. We wish to show the

---

[3]There were in fact two parliamentary cycles between 1876 and 1881 following the dissolution of the assembly in June 1877.

[4]Between 1876 and 1880, the *Annales du Sénat et de la Chambre des députés* recorded the debates in both chambers. Until 1880, they were printed by a private printer, Alfred Wittersheim. From January 1881 onwards, the French state took over the printing and then published the parliamentary debates in the *Journal officiel*

[5]This is the name given to the transcripts of the debates in Great Britain and and Commonwealth countries. Thomas C. Hansard (1776-1833) was the first official publisher.

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

interest in such a method to analyse and explore our corpus. Topic modelling indeed seems to us to be an interesting "entry point" into parliamentary debates. We start from the hypothesis that identifying the topics present in these debates makes it possible to better understand the evolution of political ideas and debates over time. We present here a first approach based on raw (uncorrected) data collected on a large scale. We have chosen to focus on the topic "army" and its co-occurring topics. The French army is indeed a stable institution during the period, which officially does not depend directly on political governments. However, discussions concerning the army were numerous and repeated in the Chamber, as the MPs had to decide on various issues related to its functioning (budgets, reforms, conscription, etc.), its activities (wars and conflicts, external operations, etc.) or political events (such as the infamous Dreyfus affair). Although soldiers did not have the right to vote from 1872 to 1945, and the army was supposed to remain politically neutral[7], the military were also surprisingly present on the French political scene in the nineteenth and twentieth centuries, even if they are still discreet in political history [15]. The Minister of War is the one who deals with Parliament, which keeps a close eye on him. In practice, the Minister proposes governmental projects, and Parliament chooses whether or not to support them. The centre of decision making in defence policy, particularly in regard to projects concerning the colonial army, is therefore in Parliament [16].

We hypothesise that topic generation model will allow us to better understand the action of the army over the identified period, what its fields of intervention were, and to grasp to what extent the French army (represented by the Minister of War) was able to participate in the elaboration and execution of the political decisions. To assess this hypothesis, our study is divided into two parts. First, we assess the consistency of the topics that our model identifies, with particular attention to the topic "army". We study the results obtained in the light of current historical knowledge, in order to verify the validity of the model. We then examine a few topics co-occurring with the topic "army", assuming that the validity of these correlations can be verified with the historical data at our disposal.

2. Data set

Digitised by the National Library of France and the archives of the National Assembly, the records of the French parliamentary debates are available online on *Gallica*, a freely accessible digital library, together with some precious metadata. Automatic transcription (OCR) have also been performed on these records, and the resulting texts have been made available online in ALTO-XML format and in raw text[8]. Transcription was generated on the fly at the time of digitisation by an OCR software (*ABBYY FineReader*), and put online without extensive post-correction.

A detailed analysis of the quality of this transcription - and how to improve it - is beyond the scope of this article. Let us just mention that while the current transcription is unfortunately not accurate enough for performing precise tasks such as named entities recognition, we believe it to be fit for the purpose of a broad analysis of the vocabulary. Most OCR errors are indeed

---

[7]The nickname of the French army was the "Grande Muette" at the time, which meant that soldiers remained "mute" on political issues, in order to avoid any risk of political destabilisation.

[8]They can be retrieved with the following API : https://api.bnf.fr/fr/api-document-de-gallica#/

located in specific parts of the text, namely in the binding of documents, where the pages can be very curved[9]. But luckily this does not represent a significant part of the corpus.

We are interested in the place of the army in the parliamentary debates of the early Third Republic, a period marked by significant military activity, particularly with the wars of colonisation (protectorate over Tunisia (1881), Tonkin campaign (1883-1886), exactions committed by the military (Voulet-Chanoine mission in 1899), etc.) and the resulting tensions with its European "competitors" (Fashoda Incident in 1898). The army also intervened within the metropolitan borders to suppress strike movements, ending sometimes in bloodshed (*Fusillade de Fourmies* (Fourmies shooting) in 1891). The trauma of the 1870 defeat also led to a reform of the army from 1871 onwards, which continued in the following years with the expansion of recruitment (Freycinet laws in 1889). It was also a period marked by various scandals and affairs such as the *Scandale des décorations* (Medals scandal) (1887), the Schaebelé Affair (1887), or the arrest and conviction of Captain Dreyfus (1894-1899). We have decided to limit our work to the years 1881-1899 in order to encompass these events without extending the size of the corpus beyond our reach. Over this period we dispose of 2597 reports in text format, almost 4 per week, and over 80 millions words.

A parliamentary sitting is a long and composite event, during which several unrelated issues are discussed in succession. For this reason, we have divided the reports according to their sections (a section corresponds to a single debate, which deals with a well-defined issue), which usually focus on a single topic. We processed this division automatically by identifying intermediate headings identified as isolated sentences written in capital letters. Thus, our corpus consists of 35891 small documents, with an average size of 2200 words.

3. Methodology

The Latent Dirichlet Allocation (LDA) topic generation model was first presented in 2003 [17]. It is based on a Bayesian probabilistic model, which is derived from the following theoretical assumption. Before any article is written, there are topics, this term designating semantic fields, i.e. sets of words linked by their meaning. Then, texts are produced by choosing words from a small subset of topics with a given probability distribution. In practice, this means that the texts are the observations derived from hidden variables, namely the topics, and that the statistical correlations in the texts are the direct results of the semantic similarities. We therefore hope to find the topics by reversing the generation process. In other words, we want to know the topics as word distributions and the texts as topic distributions, conditional on the observed word distribution. Unfortunately, the calculation of the universe probability is not feasible and so we have to approximate this quantity. Many algorithms have been introduced in the literature to deal with this issue; here we simply use the original algorithm of [17], namely the variational mean field method.

Topic modelling has been widely used in many areas of the Humanities and Social Sciences [18], as it is a very powerful tool for extracting information from a large corpus in an unsupervised context, i.e. when classes are not defined a priori. However, the results are particularly reliable when the assumptions of the model are satisfied by the study corpus. This includes: a large

---

[9]The following digitised image illustrates this problem.

number of texts; each text dealing with a limited number of topics; each topic being distributed over several texts in the corpus; a common conceptual framework shared by all authors. If newspaper articles are the paradigmatic example [13] [14], parliamentary debates also meet all these requirements. Once the topics are generated, they can be used as new variables for the study of vocabulary. This drastically reduces the size of the variable space (from more than 50000 forms to a few dozen topics) and thus makes visualisations possible - obviously with a significant loss of information. We can for example study the intensity of topics over time [19] or study the correlation between topics.

4. General Results

4.1. Structure and semantic coherence of the topics

The topics provided by the algorithm are remarkably coherent. If we consider the main keywords of some of them, it is easy to guess what they are representing (see Table 1). For instance, Topic 8 deals with the class struggle and the working class situation, with words like *salaire* (wages), *patron* (boss), *syndicat* (labour union), *grève* (strike) or *ouvrier* (worker). On the other hand, Topic 11 clearly relates to the army, with words like *général* (general), *régiment* (regiment), *troupe* (troop), *soldat* (soldier) or *guerre* (war).

Table 1
Four straightforward topics: the working class (8), the army (11), the voting process (13), the state infrastructures (15).

Topic 8	Topic 11	Topic 13	Topic 15
salaire	général	adoption	pari
question	commission	absolue	télégraphe
gouvernement	régiment	ouvert	faire
jour	troupe	votant	ingénieur
patron	monsieur	majorité	train
chambre	année	nombre	ligne
droit	jeune	secrétaire	chambre
syndicat	temps	député	personnel
délégué	faire	mets	etat
monsieur	corps	millimètre	administration
travail	soldat	article	employé
travaux	ministre	adopté	poste
ministre	homme	dépouillement	public
grève	loi	vote	travaux
faire	an	amendement	service
mineur	guerre	demande	agent
mine	service	chambre	ministre
loi	militaire	voix	fer
compagnie	officier	scrutin	chemin
ouvrier	armée	président	compagnie

4.2. Categorisation of topics into classes

These topics can easily be divided into two main categories. The first broad category includes topics related to the functioning of the Chamber: speaking tours, organisation of the sitting, votes, etc. The conduct of a sitting (even if it is sometimes disrupted) is highly codified. Even if the aim of the stenographers is to reproduce the naturalness of the exchanges and speeches, the transcription of the debates must accurately record each stage of the parliamentary sittings, from the bill's introduction to its final vote, but also all the elements relating to the functioning of the assembly (announcement of leave, composition of committees, questions to the government, etc.).

The second category includes topics that are semantically more significant for our study. The latter captures the different issues that dominated the parliamentary debates. Naturally, there are also some useless topics - for instance Topic 12 is nothing but the list of all French departments. This is because the names of the departments often appear in debates: at the time of the verification of election results (each deputy represents a department), during debates which frequently concern local life, or at the time of the vote on bills because the voters are identified by their name and the department they represent.

Also some topics are very similar to each other, especially those dealing with how the Chamber works. Hence we categorised the 50 topics in 16 classes with the following labels (Table 2):

Table 2

List of labeled topics and contributions in the corpus. Some very similar topics are put together in a single class for the sake of readability.

Label	Topics	Examples of words	Weight in the corpus
Names of MPs	0,5,10,14,18, 23,35,37,39	Duval, Sigismond, Jules, Martin	0.101
government/parliament	1,6,9,13,17, 19,22,36,38, 41,45,46,49	tribune, projet, adoption, majorité	0.284
economy	2,4,16	agriculture, commerce, patente, betterave	0.069
working class	7,8,31,34	travailler, salaire, usine, mutuelle	0.070
army	11,48	général, régiment, contrôle, militaire	0.041
department	12	Calais, Alpes, Saône, Charente	0.006
trains/communications	15,44	télégraphe, ingénieur, train, travaux	0.069
local politics	20, 33	ville, arrondissement, local, département	0.030
law enforcement	21,40	police, préfet, tribunal, délit	0.055
school	24	lycée, faculté, classe, enfant	0.023
alcohol	25	bouilleur, degré, raisin, octroi	0.019
budget	26,29,30,43	chiffre, budget, dépense, exercice	0.097
colonies	28	métropole, juif, algérien, tonkin	0.018
navy	32	marin, flotte, mer, bâtiment	0.021
building works	27, 42	construction, théâtre, hectare, terrain	0.024
foreign affairs	47	puissance, Madagascar, Angleterre, traité	0.034

In Table 2, we also calculated the contribution of each of these classes in the corpus (Cf.

column "Weight in the corpus"). This allows us to better understand whether a class of topics was more or less frequently addressed in the corpus.

If we disregard the first two classes, which bring together topics concerning the functioning of the Chamber of Deputies, we can see that "budget", "working class", "economy" and "trains/communications" are the four classes of topics that appear most often in the corpus. "Budget" is naturally the most important class of topics, because the key role of the Chamber of Deputies is to discuss the state budget and to allocate the funds needed to enforce government policy. The growth of the working class and the rise of socialism are also well reflected in the debates: MPs address social struggles in their speeches ; we also see the (timid) beginning of social legislation in the 1890s. "Economy" is one the class of topics most often dealt with by the Chamber of Deputies, as it is frequently the subject of legislation (particularly with the question of taxation). This class of topics is also frequently present, as it relates to many sectors (agriculture, trade agreements, industrialisation, etc.). "Train/communications" reflects the significant investment in the development of communications infrastructures, and the creation of the French railway network - one of the most developed in Europe at the beginning of the twentieth century. More generally, an examination of this figures confirms the coherence of the topics we have identified: they are quite consistent with the major themes that marked political life during the early Third Republic [20].

Beyond this simple comparison between their weight in the corpus, we see (Figure 1) that the various classes have unequal variances. Let us consider the topic "army". While the quartiles are not extremely far from the median, there are some strong outliers. They can go up to 0.2, and 6 of them are greater than 0.1. It seems that when "army" is the main topic, it tends to become hegemonic. On the two extremes, the topic "budget" has a very high variance while "school" never gets to be really prominent, the maximum never goes higher than 0.07.

## 4.3. Distribution of the topic "army" over time

Figure 2 shows, for every month, the percentage of the vocabulary that belongs to a given class of topics. As expected, this signal is quite noisy, since many topics are discussed over small periods of time. However, it is possible to notice some patterns in these graphs.

The topic "army" was very popular at four different times: first in 1881, then around 1884, then in 1887-1888, and finally around 1895. These four peaks can be explained perfectly well in terms of military history. The conquest of Tunisia began in 1881, with the intervention of the French army in Kroumirie, located in northwestern Tunisia. The year 1884 was marked by two events concerning the army: the Tonkin campaign, and discussions on reducing military service to three years. In 1887 and 1888, discussions resumed on the reform of military service. The year 1887 was also marked by renewed tensions with Germany (Schnaebelé affair). In 1895, the difficult conquest of Madagascar gave rise to new debates concerning the French colonial armies.

For the sake of the comparison, we present in Figure 2 the evolution of other topics over time, namely "law enforcement", "school" and "colonies". Observe that "law enforcement" remains a significant proportion of the corpus, while "school" and "colonies" are almost non-existent over large periods. We can clearly see that these three graphs show different patterns, confirming that each topic captures distinct phenomena. For instance, "law enforcement" and "school" are

**Figure 1:** Comparative variances of several topic classes
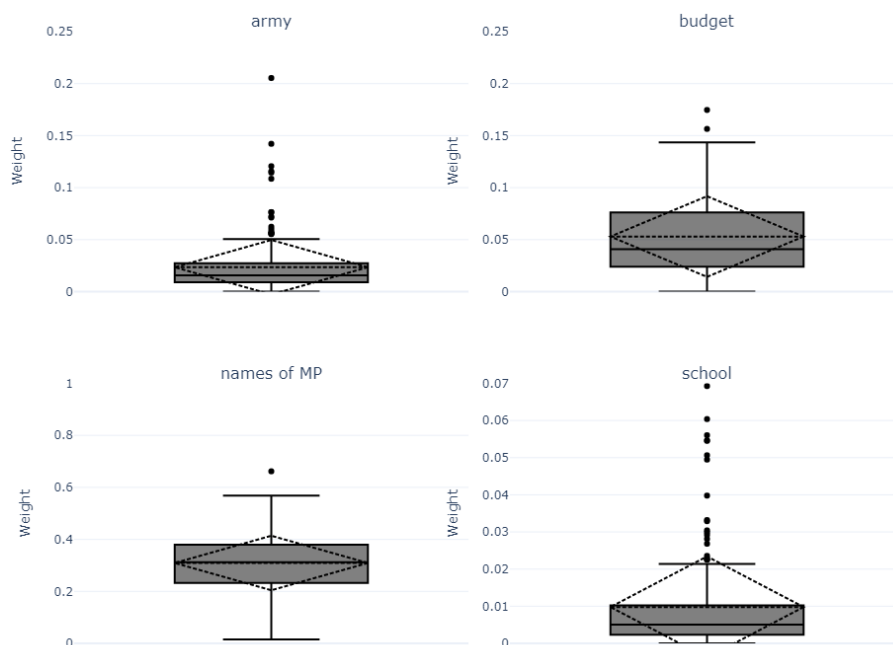
more represented in the first half of the period, while "colonies" circulates quietly in the corpus, and gains in importance in the second half of the period with strong peaks in 1886, 1895 and at the very end of the 1890s.

Figures 3 and 4 show distribution of the topic "army", respectively for all the years considered, and for the year 1884 during which the topic is particularly present. In Figure 3, several peaks can be seen that were not visible in the previous figure. These peaks can be explained by the colonial policy conducted by France, by the military reforms that took place during the period, and by the Dreyfus affair. In 1888 and 1889, the law reducing the length of military service was discussed and voted. The years 1892 and 1893 were marked by the continuation of colonial conquests (Comoros, Tunisia, Sudan, Dahomey, Ivory Coast, Siam). The Dreyfus affair began in 1894 and continued until the end of the period studied. In 1897, the borders of the French colonial empire are stabilised with the last conquests (Indochina and Madagascar), and the Franco-Russian military alliance was affirmed in case of war.

Figure 4 shows that 1884 was a year in which there were several intense discussions about the army in the Chamber of Deputies. Two issues occupied the MPs. On the one hand, they deliberated on a bill to reduce the length of military service between April and June (see peaks in April and June). On the other hand, they also had to discuss the military operations carried out by France in Tonkin. The way in which the government conducted this war of conquest can be seen in the shape of the graph: (1) the government asked for new credits in February, which led to heated debates; (2) the government sought to increase the number of colonial troops to satisfy its ambitions and proposed a project to this effect in June; (3) the Chamber discussed the budget in December, and in particular the credits allocated to colonial troops.

Figure 2: Distribution of four different topics over time
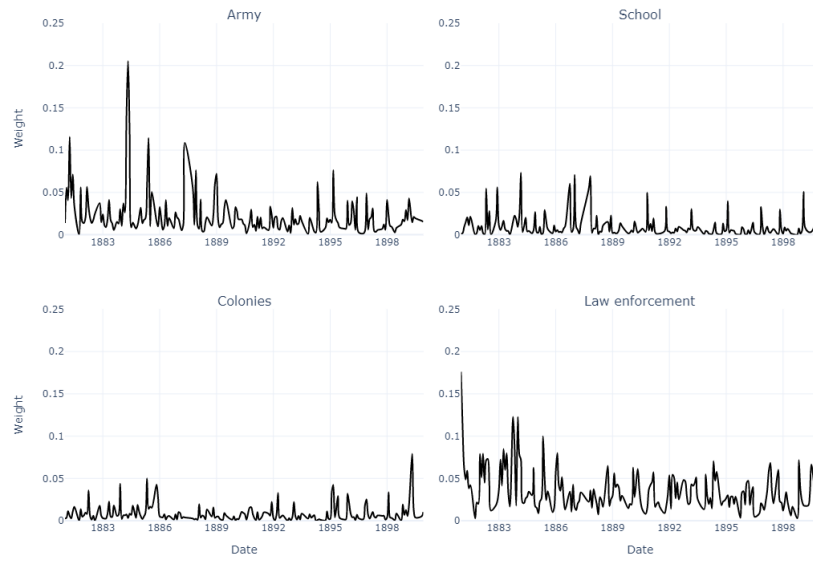
Figure 3: Distribution of the topic "army" over time, per day (all years)
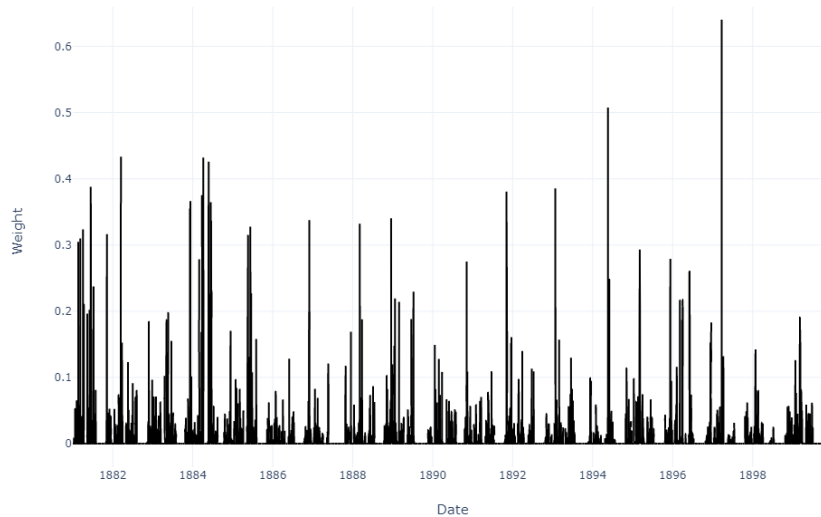

5. Cross study of the topics' prevalence

5.1. Time-based correlation between topics

We are now looking for high co-intensity topics, i.e. topics that tend to be frequently associated with each other in specific parts of the corpus. Correlation is based on the number of text units that contain a significant percentage of both subjects. These text units are not defined semantically, but on the basis of a fixed length window of 6000 characters.

We create a first indicator by dividing our corpus according to the date of production of the

**Figure 4:** Distribution of the topic "army" over time, per day (1884)
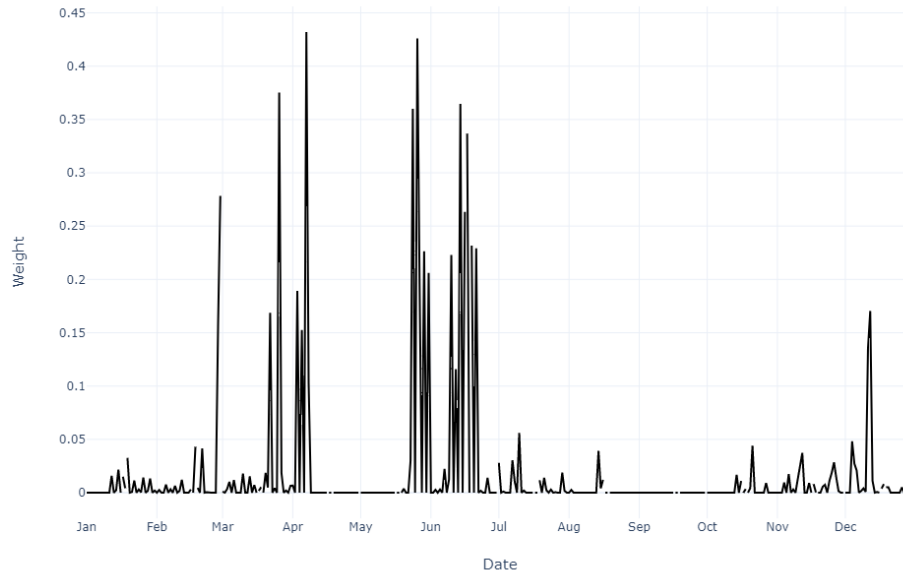
texts composing it. This allows us to divide the corpus into smaller segments (by month or by year), and then determine whether a high proportion of the topic "army" is correlated with a high or low proportion of other topics in the same period. We take the average weight of each topic per month and calculate the Pearson correlation coefficient between the topic "army" and all other topics. We then observe that the intensity of the topic "army" over the course of a month (Figure 5) is positively correlated with the following topics: "colonies", "navy", "foreign affairs". Perhaps most surprising is the strong correlation with the topic "school".

However this correlation is rather weak in absolute terms. In the course of a given period (even a single parliamentary sitting, i.e. a single day), many different issues are addressed by MPs. Information therefore tends to be spread across most subjects. In particular, some topics such as the names of MPs, the departments they come from, and the vocabulary describing the functioning of parliament, are evenly spread over the period. We therefore decided to look for a correlation at the lowest level. We are in fact looking to answer the following question: what proportion of the blocks that address with high intensity the topic "army" (more than 15% of the vocabulary) also deals with high intensity with another given topic?

We find a very strong correlation between army and navy (15.5% of documents with a high proportion of the topic "navy" also have a high proportion of the topic "army"), followed by "colonies", "school", "law enforcement" and "budget" (see Table 3 and Figure 6). Since the topic "government/parliament" is fairly evenly distributed throughout the corpus, other topics such as "working class", "alcohol" or "local politics" are almost completely disconnected from the topic "army". The case of the names of the MPs and the departments is specific, as these two

Figure 5: Correlation between the topic "army" and the other identified topics (by month)
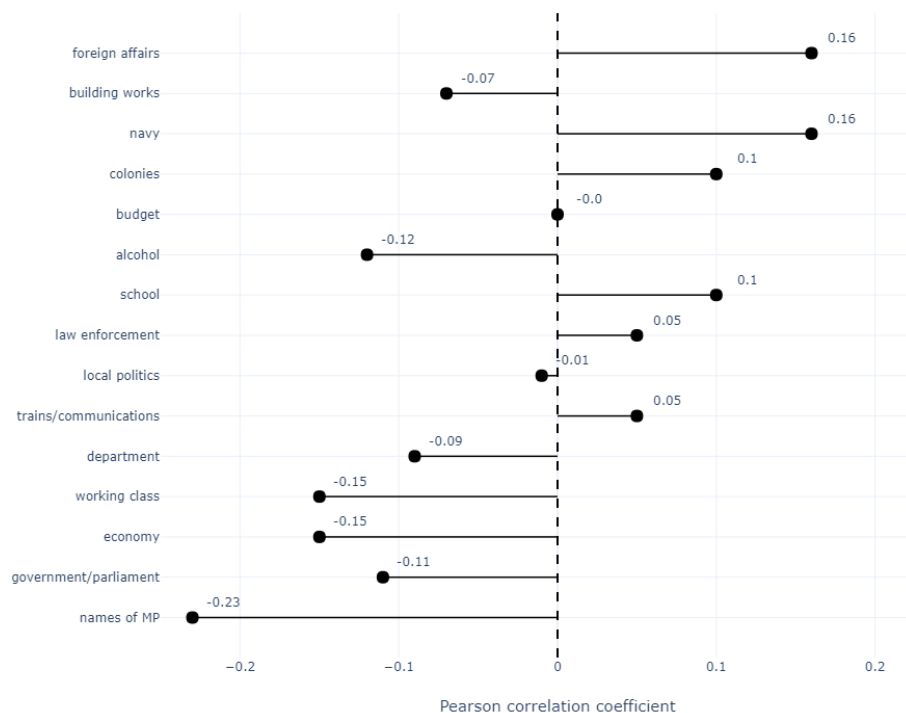
topics are mainly present in specific sections, i.e. the vote count.

Table 3

Number of topics with more than 15% of vocabulary from "army" and more than 15% from another topic

topic name	army	MPs	gov./parl.	economy	work. class	depart	trains	local pol.
topic + army	1737	5	1009	74	48	1	85	31
topic only	1737	11651	17731	2489	2554	534	2559	3678
proportion	100%	0.04%	5.69%	2.97%	1.88%	0.19%	3.32%	0.84%

topic name	law inft	school	alcohol	budget	colonies	navy	building	foreign affairs
topic + army	202	69	1	310	62	124	33	65
topic only	2968	923	587	4678	723	800	962	1404
proportion	6.81%	7.48%	0.17%	6.63%	8.58%	15.50%	3.43%	4.63%

5.2. Study of topics with a strong correlation with the topic "army"

We aim here to focus on the strong correlations we have just identified. We will examine the nature of these correlations by "close reading" [10] the texts we are studying. We will also check the validity of these results in the light of current historical knowledge.

Figure 7 shows that there is a strong correlation between "army" and "school"; this is mainly related to the debates on the reform of military service. The law of 1872 had established a

**Figure 6:** Graph of topics with high correlation
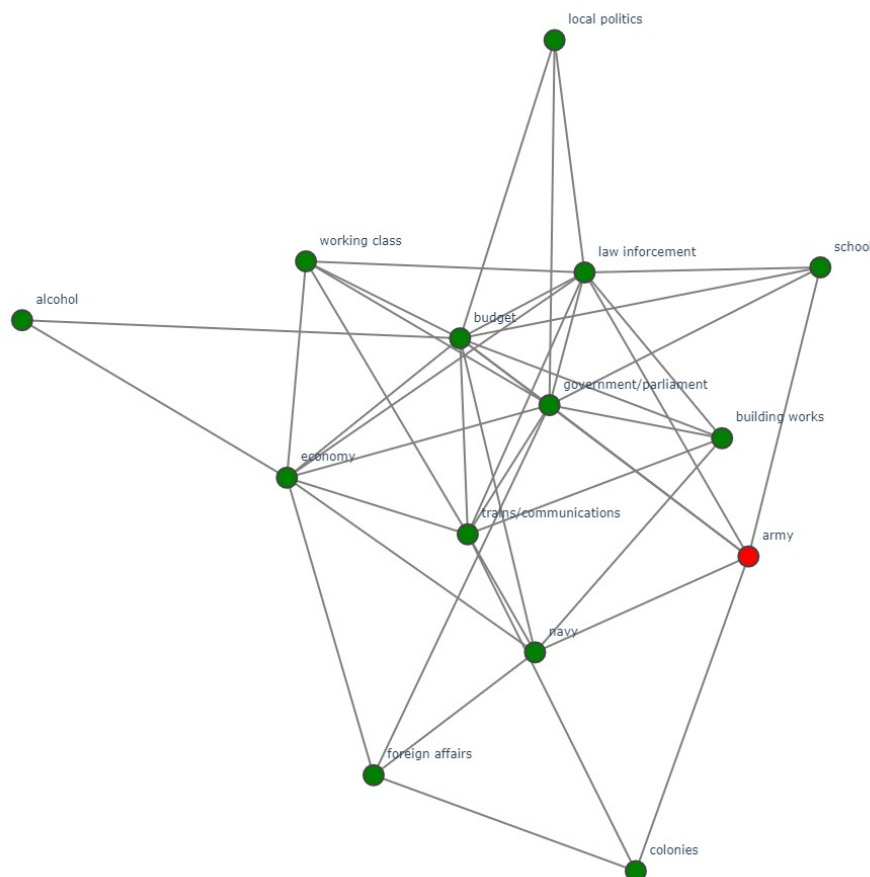
military service that could last up to five years, but with a certain number of exemptions (teachers, students of *grandes écoles*[10] and seminarians [16]). The association of "army" with "school" refers to the exemptions granted to teachers and students of *grandes écoles*, which the Republican MPs wanted to put an end to.

This correlation was most intense in 1887, although the law removing these exemptions and reforming the military service was passed in 1889. This was because this law was in the making from the early 1880s [16]. Between 1876 and 1889, there were twelve bills related to this issue [21]. But it was really in 1887, with the renewed tensions between Germany and France, that the Ministry of War, in agreement with the Chamber of Deputies, decided to transform the law of 1872 [16]. The report on this project was proposed and discussed between June and July 1887. After passing through the Senate, the bill was presented to the MPs again in December 1888 and voted on in January 1889 [21].

The topic "law enforcement" includes vocabulary related to the creation of the law, as well as references to punishments and means of control inside the military. Figure 7 reflects the intense

---

[10]For more information on the French system of *grandes écoles*, please see this Wikipedia article.

Figure 7: Cross-topics between "army", "school" and "law enforcement"
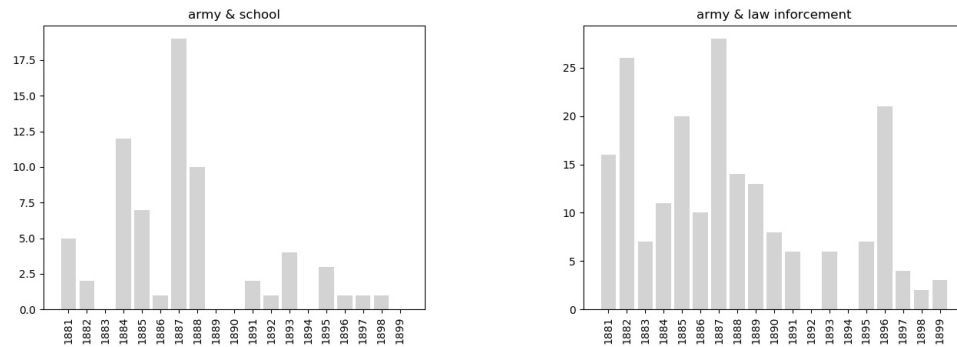
legislative activity relating to the army during this period. Discussions about military service laws explain its strong correlation with "army" between 1881 and 1889; they also cause a strong peak in 1887 for the same reason.

These recruitment issues were also linked to the international context. The increase in the intensity of the association of "army" with "foreign affairs" in 1894-1895 (see Figure 8) can be explained by the introduction of a law in November 1894 extending the duration of incorporation to two years, in order to increase the army's strength [16]. This was a reaction to the changes that were taking place in the German military, whose growing power frightened the MPs. Between 1893 and 1894, the number of German soldiers increased following the introduction of the two-year service. Our model captures this trend well by clearly associating the topic "army" with the topic "foreign affairs".

Figure 8: Cross-topics between "army", "colonies" and "foreign affairs"
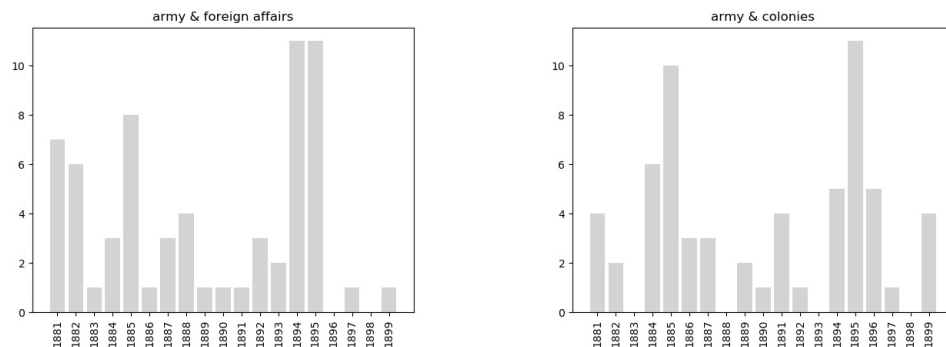
The peak in 1895 can be explained by the return of the project to the Chamber in June 1895, as the German strength had just increased by 70000 men [16]. The association between "army" and "foreign affairs" also reveals the competition with Great Britain in colonial affairs. In 1884, France took control of Annam, while Great Britain extended its influence over Burma. Both imperialisms were in contact with South China, which led to tensions, especially in 1885 over Siam. In 1885, there were also strong tensions with the British in the face of growing French

appetite for Madagascar. These tensions were also high at the time of the second Madagascar expedition in 1895 [22] (see peaks in 1885 and 1895 in Figure 8); The peak in 1881 can also be explained by tensions with another European competitor for the conquest of new territory: a Franco-Italian crisis broke out in June following the Treaty of Bardo, which placed Tunisia under the French protectorate.

We also note the correlation of the topic "army" with the topic "colonies". This correlation refers to the crucial role of the army in the acquisition and defence of colonies. This association follows a pattern quite similar to the previous association (see Figure 8): the topic is present throughout the period, but with a strong intensity in the early years (1884-1885), and a second peak in the mid-1890s. Our model succeeds in capturing the way in which the executive power imposes its colonialist policy on Parliament. After the defeat of 1870, the colonial enterprises were blamed for the domestic defeat, as they were said to have taken away the men and funds needed for national defence. Public opinion - and the MPs with it - was at best indifferent, at worst hostile, to new conquests. The arrival in power of the opportunist Republicans in 1879 nevertheless saw the renewal of colonial expansion, which resumed in 1880 and continued intensively until 1885. This policy of conquest was carried out in parallel on several fronts: notably in Tunisia (1880-1881), Annam and Tokin (1883-1885), not to mention Sudan, Congo and Madagascar [23]. The government therefore had to "trick" public opinion and Parliament, and act on the sly to conceal the extent of its ambitions. Then, as the difficulties accumulated, it insensibly obtained an increase in credits, the sending of increasingly large reenforcements, and irresistibly dragged the MPs into the spiral of conquest [22].

The 1884-1885 peak in Figure 8 is explained by the launching of the Tonkin expedition, for which the government asked the Chamber for new credits and troops in 1884 and early 1885. Debates were particularly intense on this subject in 1885, as the difficulties encountered by the French army in April (Retreat from Lạng Sơn at the end of March 1885) led to an outcry in the Chamber and the fall of the government [16]. The Chamber elected in 1885 was more anti-colonialist than the previous one and avoided any colonial adventure of the importance of Tonkin; but from 1890 onwards, the opposition began to diminish until it disappeared. The very principle of colonisation was progressively accepted and increasingly supported by the MPs [23], even if this did not avoid stormy debates in the assembly. The intensity of the correlation between "army" and "colonies" from 1894 to 1896 is mainly explained by the second expedition led by the French army in Madagascar. In November 1894, the government submitted a request for credits to send an expeditionary corps to the island. This expedition was partly a failure; and in March 1895, the government was interpellated by the Chamber about the pitiful state of the troops. In July 1895, the conquest of Madagascar was resumed but it was stalled. A text is then presented to the Chamber to reform the recruitment of the colonial armies [16].

Let us examine the year 1896 in particular. We can see that the correlation between "army" and "colonies" is quite strong. In March and July 1896, a bill on colonial armies was discussed in the Chamber. It is interesting to note that, in its second version, the bill proposed to entrust the entire management of colonial units to the Navy [16]. The text shows the birth of a new trend in the Chamber in favour of this branch of the armed forces. On 27 October 1896, the government proposed a new bill on the colonial army, which the Navy would be responsible for,

as it was the only one capable of ensuring the continuity of transport and logistics[11] [16]. This explains the strong correlation between the topics "army" and "navy" in 1896 see (Figure 9). The topics "army" and "navy" are frequently associated, whether for cooperation - (the Navy transports colonial troops) - or competition between the two branches of the military.

Figure 9: Cross-topics between "army" and "navy".
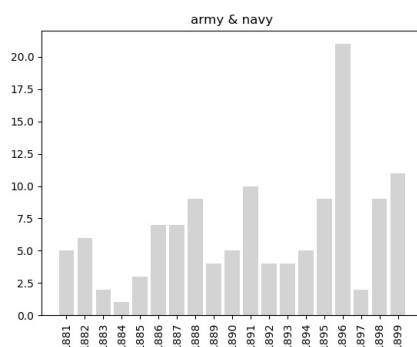
This association was rather weak during the 1880s but reached a peak in 1896. Until 1895, the Navy had been relatively indifferent to colonial troops. In June 1895, however, the Minister of the Navy claimed responsibility for the management of colonial units from the Ministry of War. This request was the consequence of the rivalry between the two armies over Madagascar, as the Navy could not bear the idea that the Army had taken charge of the expedition [16]. The financial competition between the two armies was becoming tougher, especially as the Navy needed new investments to modernise the fleet and train staff [24]. This is why in 1896 a great wave of legislative reforms was launched concerning the organisation of the Navy, notably the creation of a naval school [24].

6. Conclusion

The results of our study show the validity of topic modelling for the analysis of parliamentary debates. This confirms the interest of using such a method to facilitate the analysis of this major historical source. This study also allows us to draw a number of interesting insights on parliamentary debates, which we wish to explore further.

We see that the weight of the descriptive vocabulary of parliamentary activity itself is very important in the corpus; but this problem is rather well solved thanks to the topic model. We then observe that parliamentary debates follow their own rhythm. This rhythm is in fact imposed by the legislative process, which requires long debates before a law is finally voted. This means that subjects can be dealt with by the Chamber of Deputies long before they become newsworthy. Conversely, issues that make the news are rarely discussed during parliamentary sittings; they are usually dealt with long after they have made the headlines. Topic modelling therefore seems to us to be a method that makes it easier to identify underlying political trends.

---

[11]The project was finally rejected in December.

We also note the reactive nature of parliamentary work: this means that a major legislative effort can take place a few months or even several years after the triggering events (as shown in [25] for instance). Finally, there is another consequence of the way legislative work is carried out, namely the weight that discussions on sensitive subjects can take on, without leading to a vote or the production of a law. The Chamber can indeed seize on a subject to interpellate the government - this is for instance the case of Tonkin after the Retreat from Lạng Sơn in 1885.

While encouraging, these results are still preliminary. We are working in two directions To further complement and improve them. In order to obtain information on more specific and hopefully unexpected correlations (e.g. the role of the church in the army, or the influence of the executive branch), we will use additional tools, such as word embedding, to further divide the corpus into a few hundred groups, some of them very specific, and to study their life cycle in relation to the army. To improve our model, we are planning to enlarge the period studied, and to work on a less faulty corpus. Within the framework of the AGODA project, we are thus evaluating the solutions available to us to improve the results of the OCR, hoping to further enhance these first results.

Acknowledgments

References

[1] H. Coniez, L'invention du compte rendu intégral des débats en france (1789-1848), Parlement[s], Revue d'histoire politique 2 (2010) 146–159. doi:10.3917/parl.014.0146.

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

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

[4] C. Lermercier, Le vocabulaire des débats sur la loi de 1841 sur le travail des enfants : Premiers résultats sur la chambre des pairs, 4-10 mars 1840, 2006. URL: https://halshs.archives-ouvertes.fr/halshs-0010745.

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

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

[7] H. Bonin, From antagonist to protagonist: 'democracy' and 'people' in british parliamentary debates, 1775–1885, Digital Scholarship in the Humanities 35 (2010) 759–775. doi:10.1093/llc/fqz082.

[8] S. Mollin, The hansard hazard: gauging the accuracy of british parliamentary transcripts, Corporas 2 (2008) 187–201. doi:10.3366/cor.2007.2.2.187.

[9] F. Clavert, Vers de nouveaux modes de lecture des sources, in: O. L. Deuff (Ed.), Le temps des humanités digitales, FYP EDITIONS, Roubaix, 2014.

[10] F. Moretti, Distant Reading, Verso, London, 2013.

[11] M. Puren, P. Vernus, Agoda : Analyse sémantique et graphes relationnels pour l'ouverture et l'étude des débats à l'assemblée nationale, 2021. URL: https://hal.archives-ouvertes.fr/hal-03382765.

[12] G. Shawn, I. Milligan, S. Weingart, Exploring big historical data: the historian's macroscope, Imperial College Press, London, 2016.

[13] G. Lavenir, N. Bourgeois, Old people, video games and french press: a topic model approach on a study about discipline, entertainment and self-improvement., MedieKultur: Journal of media and communication research (2017).

[14] L. Violla, J. Verheul, Mining ethnicity: Discourse-driven topic modelling of immigrant discourses in the usa, 1898–1920, Digital Scholarship in the Humanities 35 (2020) 921–943. doi:10.1093/llc/fqz068.

[15] O. Forcade, Éric Duhamel, P. Vial (Eds.), Militaires en République, 1870-1962, Éditions de la Sorbonne, Paris, 1999. doi:10.4000/books.psorbonne.61562.

[16] J.-C. Jauffret, Parlement, gouvernement, commandement : l'armée de métier sous la 3è république 1871-1914, Ph.D. thesis, Université de Paris I Panthéon Sorbonne, Paris, 1987.

[17] D. Blei, A. Ng, M. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

[18] D. Blei, Topic modeling and digital humanities, Journal of Digital Humanities 2 (2012).

[19] X. Wang, A. McCallum, Topics over time: a non-markov continuous-time model of topical trends, KDD'06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (2006) 424–433.

[20] J.-M. Mayeur, Les débuts de la IIIe République 1871-1898, Editions du Seuil, Paris, 1973.

[21] A. Crépin, Défendre la France : Les Français, la guerre et le service militaire, de la guerre de Sept Ans à Verdun, Presses universitaires de Rennes, Rennes, 2005.

[22] M. Battesti, La Marine au XIXe siècle. Interventions extérieures et colonies, Du May, Paris, 1993.

[23] D. Bouche, Histoire de la colonisation française. Flux et reflux : 1815-1962, Le Grand livre du mois, Paris, 2004.

[24] R. Monaque, Une histoire de la marine de guerre française, Perrin, Paris, 2016.

[25] J. Alerini, M. Olteanu, J. Ridgway, Markov and the duchy of savoy: Segmenting a century with regime-switching models, Journal de la Société Française de Statistique (2017).

A. Online Resources

The data and source code are available via GitHub.