Explorer les débats parlementaires français de la Troisième République par leurs sujets

Marie Puren, Aurélien Pellet

Explorer les débats parlementaires français de la Troisième République par leurs sujets : une approche méthodologique pour étudier de grands corpus de textes historiques

Marie Puren
LRE, EPITA
CJM, École nationale des chartes
marie.puren@epita.fr

Aurélien Pellet
Epitech
aurelien.pellet@epitech.eu

Résumé

Cet article compare trois méthodes pour explorer de grands corpus de documents historiques par leurs sujets. Nous travaillons ici sur les débats parlementaires français de la Troisième République, qui se prêtent particulièrement bien à ce type d'analyse. Après avoir présenté le contexte de cette étude, nous exposons les résultats obtenus avec trois méthodes issues du traitement automatique des langues et appliquées sur des textes publiés entre 1876 et 1914 : l'allocation de Dirichlet latente, les plongements de mots et le *Transfer Learning*.

1 Introduction

Les comptes rendus des débats parlementaires naissent spontanément en France en 1789. À part quelques rares éclipses au XIX^e^ siècle, ils assurent jusqu'à aujourd'hui la publicisation des discussions entre les parlementaires (Coniez, 2010). Si les compte-rendus les plus contemporains sont disponibles sous la forme de données « nativement numériques » [1], les débats parlementaires historiques ont été numérisés et mis en ligne sur *Gallica* entre 2008 et 2016, dans le cadre d'un vaste programme de numérisation des sources pour les sciences juridiques [2] (Alix, 2008).

Le projet AGODA, soutenu par le DataLab de la Bibliothèque nationale de France (Puren et Vernus, 2021), a pour objectif d'exploiter ces sources numérisées pour en faciliter la consultation et l'analyse. L'un des principaux objectifs d'AGODA consiste à créer une plateforme de consultation de ces débats anciens, en produisant également un corpus de textes encodés en XML-TEI. Il s'agit non seulement de publier un texte édité et encodé selon les normes de la *Text Encoding Inititative*, mais aussi d'offrir de « nouveaux modes de lecture » de ces sources (Clavert, 2014), notamment grâce aux méthodes offertes par le traitement automatique des langues. Dans le cadre d'AGODA, nous travaillons sur les débats parlementaires qui se sont tenus à la Chambre des députés durant la Troisième République (1870-1940).

Dans cet article, nous nous intéressons plus particulièrement aux méthodes computationnelles à notre disposition pour valoriser ces sources. La modélisation de sujets nous paraît être une excellente manière d'approcher ce corpus massif et de développer un nouveau « mode de lecture » de ces documents. Nous présenterons d'abord le corpus sur lequel nous travaillons, avant de nous intéresser aux enjeux entourant l'analyse de cet important corpus de sources historiques. Nous exposerons ensuite les résultats que nous obtenons avec les trois méthodes que nous avons employées pour extraire les sujets de notre corpus. Enfin, nous terminerons avec une proposition pour faire de ces sujets des annotations linguistiques, qui pourront être utilisées pour explorer les documents.

2 Les comptes rendus des débats parlementaires durant la Troisième République

2.1 Une source précieuse pour l'histoire

Proclamée le 4 septembre 1870, la Troisième République est fondée constitutionnellement en 1875 comme solution provisoire. Les élections législatives de 1876 permettent d'élire la Chambre des députés, qui succède ainsi à l'Assemblée nationale élue en 1871 dans des conditions particulièrement difficiles après la défaite française. Mais ce n'est qu'entre 1876 et 1879 que la Troisième République est devenue pleinement républicaine, grâce à la conquête de la Chambre des députés et du Sénat par les républicains. Les républicains ont alors établi un régime parlementaire dans lequel la Chambre des députés occupait une place centrale. Le rôle

1. Par exemple : https://www.assemblee-nationale.fr/dyn/15/comptes-rendus/seance.
2. Cf. La Mission de recherche Droit et Justice et le programme national de numérisation concertée en sciences juridiques

considérable joué par la Chambre des députés fait des débats à la Chambre, une source cruciale pour l'histoire politique (Ouellet et Roussel-Beaulieu, 2003), mais aussi pour d'autres recherches historiques, puisqu'ils permettent de suivre les grandes étapes de l'élaboration du cadre législatif de divers champs d'activité sociaux, économiques, religieux ou culturels (Lemercier, 2021; Marnot, 2000).

2.2 Un corpus massif

C'est à partir de 1881 que les débats parlementaires à la Chambre sont publiés dans le *Journal officiel de la République française. Débats parlementaires. Chambre des députés : compte rendu in-extenso* [3]. La forme que prend alors les compte-rendus va perdurer jusqu'à nos jours, avec une organisation très similaire à celle des comptes rendus contemporains. Dans le cadre d'AGODA, nous avons adopté une approche « proof of concept » en travaillant sur un sous-corpus test, à savoir les débats du cycle parlementaire 1889-1893. Cette approche nous a permis de concevoir le projet comme un cas d'utilisation.

Toutefois nous envisageons d'agrandir le corpus pour couvrir tous les débats qui se sont tenus à la Chambre des députés durant la Troisième République. Par exemple, la période 1876-1879 est particulièrement intéressante d'un point de vue de l'historien car c'est le moment où la Chambre devient véritablement républicaine. Nous avons ainsi décidé de réaliser nos analyses en traitement automatique des langues à un corpus beaucoup plus important que celui sur lequel nous travaillons dans le cadre du projet AGODA. Ces analyses à plus grande échelle ont également été envisagées comme une nouvelle étape destinée à ouvrir la voie au traitement de l'ensemble du corpus. Nous procédons donc par étapes successives, en élargissant progressivement la période étudiée. Par ailleurs, les méthodes que nous avons utilisées sont particulièrement gourmandes en données : afin d'assurer la robustesse de nos résultats, nous avons préféré travailler sur les débats publiés entre 1876 et le tournant de la Grande Guerre. Nous avons ainsi inclus dans le corpus les débats extraits des *Annales du Sénat et de la Chambre des députés* [4] qui enregistre les débats pour la période 1876 à 1881.

Sur la période étudiée, nous avons travaillé sur 4644 débats. Téléchargeables au format TXT [5], ces fichiers ont été obtenus grâce au logiciel d'OCR *ABBY FineReader*, et les textes générés ont été mis en ligne sans post-correction poussée. La mise à disposition de ces documents en version texte offre de précieuses fonctionnalités pour la consultation de ces derniers (notamment la recherche plein-texte), mais il faut être conscients que les textes océrisés contiennent des erreurs. Si les OCR sont de qualité variable, nous considérons toutefois que la qualité est en moyenne correcte et que la grande taille de notre corpus nous assure une quantité suffisante de texte « propre » pour effectuer nos analyses.

Les figures suivantes donnent une bonne idée de l'importance de ce corpus, qui ne cesse de grandir au cours du temps. La figure 1 montrent la tendance à l'augmentation de la taille des débats parlementaires : si la médiane du nombre de tokens est à un peu plus de 20.000 en 1876, celle-ci atteint plus de 60.000 à la fin de la période. Sur la figure 2, on remarque également une tendance à l'augmentation du nombre de débats, même si celle-ci est beaucoup moins régulière.

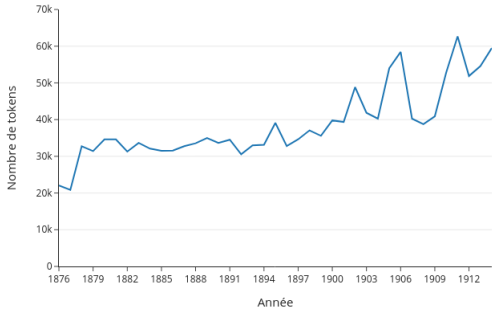


FIGURE 1 – Médiane du nombre de tokens par années

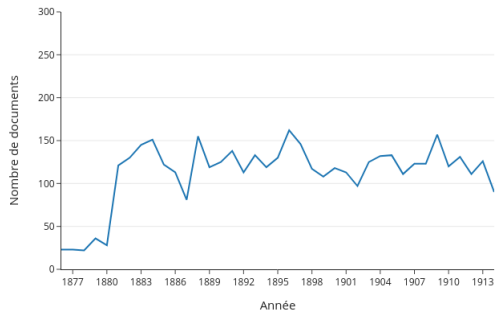


FIGURE 2 – Évolution du nombre de documents par années

3. Disponible sur Gallica : https://gallica.bnf.fr/ark:/12148/cb328020951/.
4. Disponible sur Gallica : https://gallica.bnf.fr/ark:/12148/cb32694473t/
5. Nous avons utilisé l'API Document de *Gallica* pour constituer le corpus : https://api.bnf.fr/fr/api-document-de-gallica.

3 « Lire à distance » ces sources historiques

La taille importante de ce corpus rend difficile l'utilisation de méthodes de lecture « classiques ». Surtout lorsqu'il s'agit d'identifier les débats menés dans le temps, dont l'étude permet d'observer l'évolution des questions politiques et sociales traitées par l'assemblée (Baker et al., 2017). Les débats parlementaires constituent en effet un « corpus remarquable » pour l'histoire politique et intellectuelle (Bonin, 2020). Face à une telle quantité de documents, nous proposons de nous appuyer sur la « lecture distante » prônée par Franco Moretti (Moretti, 2013). La distance est ici une « condition de la connaissance » (Moretti, 2000) car elle permet d'identifier plus facilement des anomalies ou des régularités qu'une lecture proche n'aurait pas détectées. Les débats parlementaires se prêtent particulièrement bien aux « analyses assistées par ordinateur », dans le sens où l'ordinateur facilite les changements d'échelle (Ihalainen, 2020) avec le passage d'une lecture distante à une lecture proche et vice-versa. AGODA vise ainsi à promouvoir une lecture « mixte » (Stulpe et Lemke, 2016) des débats parlementaires, c'est-à-dire l'ensemble des lectures historiques entre les approches macro et micro (Annales, 2015).

Pour analyser les travaux parlementaires, l'utilisation de l'ordinateur est un prérequis quasi indispensable (Bonin, 2020; Ihalainen, 2020; Blaxill, 2022), même si son utilisation en sciences humaines a diminué depuis les années 1970, notamment en histoire (Lemercier et Zalc, 2019; Kemman, 2021; Salmi, 2021). Il est donc nécessaire de doter les historiens d'outils et de méthodes adaptés à l'exploration de ces grands ensembles de données (Salmi, 2021; Shawn et al., 2016). Si, comme les auteurs d'*Exploring Big Historical Data*, nous pensons que « Big Data analysis skills are on the verge of no longer being a 'nice to have' for historians but nearly a necessity » (Shawn et al., 2016), nous pensons également qu'il est essentiel de fournir des outils clés en main pour faciliter la lecture de ces grands corpus par le grand public, ainsi que par les chercheurs dont les besoins se limitent à la consultation des documents. Une piste consiste à développer des interfaces permettant d'interroger facilement ces corpus (Ihalainen et al., 2022). Ce type de plateforme est indispensable pour que les historiens prennent conscience du potentiel de recherche offert par cette source numérisée. C'est ce que nous proposons de développer dans le cadre du projet AGODA (Puren et Vernus, 2021). Il nous semble essentiel de proposer une telle plateforme dans le contexte français, où l'histoire numérique souffre d'une désaffection pour les méthodes quantitatives (Lemercier et Zalc, 2019), et d'un manque de formation dans ce domaine (Ruiz, 2022). Cette plateforme offrira ainsi deux niveaux de lecture : d'une part en créant un corpus interrogeable et lisible par des humains (close reading) ; d'autre part en offrant de nouvelles portes d'entrée dans ce corpus au moyen des méthodes computationnelles (*distant reading*).

Un écueil des méthodes de « lecture distante » est l'imposition de catégories inappropriées (Karila-Cohen et al., 2018). Plutôt que d'imposer des catégories prédéfinies, nous avons choisi d'adopter une approche inductive basée sur les résultats de la modélisation des sujets. Cette méthode permet de faire émerger les sujets à partir des textes eux-mêmes, sans intervention humaine (Lemercier et Zalc, 2019; Shawn et al., 2016). Lauren Klein et ses coauteurs proposent ainsi d'utiliser les sujets générés avec la modélisation de sujets pour proposer une « nouvelle façon de lire » les collections de journaux anciens (Klein et al., 2015). De nombreuses études ont en effet montré l'intérêt de la méthode pour étudier les corpus de presse ancienne (par exemple (Violla et Verheul, 2020)). Comme la presse, les débats parlementaires constituent un corpus volumineux, sériel et traversé par de nombreux sujets différents évoluant dans le temps ; ils se prêtent donc bien à une exploration par leurs sujets comme le montrent des études menées sur les corpus contemporains - par exemple (Abercrombie et Batista-Navarro, 2020). Dans une première étude (Bourgeois et al., 2022), nous avons également constaté que l'allocation de Dirichlet latente (*Latent Dirichtet Allocation* ou LDA) (Blei et al., 2003) donne de bons résultats sur notre corpus de débats anciens. Il nous semble donc pertinent de créer une interface qui permettrait aux utilisateurs de parcourir les documents en fonction des thèmes qui les traversent.

4 Explorer les débats avec la modélisation de sujets

Les analyses ont été menées sur un corpus lemmatisé avec la bibliothèque Spacy. Les *stop words*

ou mots vides ont également été supprimés [6]. Nous avons aussi divisé les débats en plus petites unités : en effet, il est très peu probable qu'un sujet circule dans tout un compte-rendu ; il est beaucoup plus probable que certaines parties vont aborder des sujets bien définis, et que ces sujets, pourtant abordés dans un même débat, vont avoir très peu de rapports entre eux [7]. On s'assure ainsi de « capter » de manière plus fine les sujets. Par ailleurs, certains modèles de langue n'acceptent qu'un nombre limité de tokens. Ces derniers ont une complexité d'entraînement quadratique par rapport au nombre de mots à cause des couches d'attention ; on ajoute une baisse en qualité du modèle au-delà de cette limite. Nous avons ainsi divisé chaque débat sur la base d'une expression régulière qui identifie les suites de caractères en majuscules. Ces derniers sont les marqueurs d'une nouvelle phase dans les délibérations, et donc d'un nouveau sujet évoqué. Nous conservons uniquement les blocs d'au moins 2000 caractères et nous obtenons ainsi 75.775 blocs de textes.

4.1 Résultats

Nous avons décidé de comparer trois méthodes traditionnellement employés pour la modélisation de sujets : l'allocation Dirichlet latente ou LDA, les plongements de mots ou *word embeddings*, et le *Transfer learning*. Chacune de ces méthodes repose sur des bases différentes et propose des avantages et des inconvénients propres à leur fonctionnement et leur application sur le corpus. Nous détaillons dans la section suivante ces trois méthodes et les détails de notre implémentation. On évoquera aussi les résultats obtenus avec ces dernières ainsi que les limites de l'analyse.

4.1.1 L'allocation de Dirichlet latente ou LDA

LDA (Blei et al., 2003) repose sur l'hypothèse qu'un corpus de textes se compose d'un nombre N [8] de sujets ou *topics*. Chaque corpus est (selon le modèle) créé en choisissant un sous-ensemble de ces *topics* selon une probabilité de distribution, et en « piochant » des mots parmi chacun de ces *topics*. Le rôle de LDA est d'inverser ce processus de génération afin de retrouver les sujets originaux, en espérant que leur cohérence statistique reflète une certaine homogénéité sémantique. Ainsi LDA cherche à reconstituer ces probabilités de distribution, c'est-à-dire les probabilités de distributions des *topics* par documents, et des mots par *topics*. En sortie, LDA produit donc une répartition des *topics* pour chacun des documents et une répartition des mots pour chacun des *topics*. On note aussi que LDA est « non contextuel », ce qui signifie que l'ordre des mots n'a pas d'importance parmi les documents. Le nombre de sujets est à fixer à l'avance. Nous avons ainsi testé LDA avec un nombre de sujets fixé à 50, 75, 100,200 et 300. Une première validation manuelle montre qu'au-delà de 50, les nouveaux *topics* sont répétitifs. Nous n'avons pas réalisé de mesure de cohérence pour valider ce paramètre, parce que nous avons utilisé la bibliothèque scikit-learn qui n'implémente pas une telle mesure. C'est une faiblesse de notre analyse, dont nous sommes conscients et que nous souhaitons corriger dans le futur.

L'étude de notre corpus avec LDA a fait l'objet d'un travail détaillé dans (Bourgeois et al., 2022) mais sur une période plus restreinte (1881-1899). Comme le montre la figure 3, l'utilisation de LDA sur les années 1876-1914 donne des résultats tout à fait cohérents avec notre étude précédente, confirmant l'intérêt d'employer cette méthode pour extraire les sujets de notre corpus. Le tableau 1 montre quelques-uns des *topics* trouvés avec nos nouvelles analyses. Là aussi on remarque la similarité de ces derniers avec ceux analysés dans (Bourgeois et al., 2022).

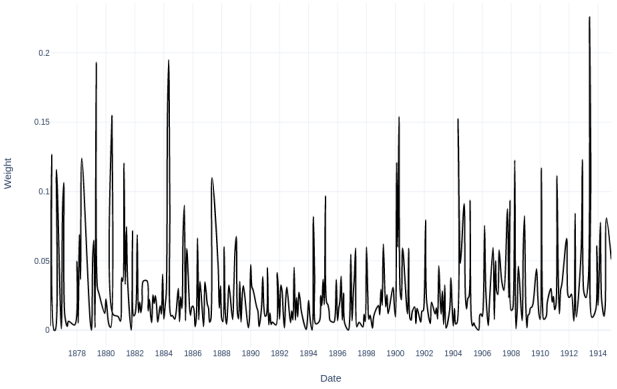


FIGURE 3 – Varation du poids des *topics* liés à l'armée, moyenne par mois

La sélection a-priori du nombre de *topics* assure une certaine cohérence des sujets au cours du temps. En fixant ce nombre à 50, on remarque que ces derniers restent relativement similaires à

6. Nous utilisons la liste des *stop words* en français de NLTK.

7. Lors d'une séance, les députés vont en effet discuter de questions très diverses (agriculture, santé, agitation politique, infrastructures, etc.), qui ne vont pas avoir de cohérence thématique.

8. N est un paramètre choisi par l'utilisateur.

Topic 9	Topic 8	Topic 48	Topic 43
armée	eau	retraite	école
guerre	ville	caisse	primaire
homme	paris	pension	supérieur
militaire	canal	assurance	lycée

TABLEAU 1 – Quatre sujets ou « *topics* »parmi les 50 définis avec LDA : armée (9), gestion des eaux dans la capitale (8), retraite et pensions (48) et éducation (43).

ceux obtenus sur la période 1881-1899. Cela nous permet donc de pouvoir généraliser nos analyses sur une période plus longue. Toutefois, définir un nombre de *topics* a priori restreint inévitablement le nombre de sujets identifiés. Par définition, LDA construit un nombre limité de grandes unités sémantiques et offre peu de contrôle sur le processus. Potentiellement, on rate donc des nouveaux sujets intéressants. Les méthodes de plongements de mots permettent de palier à ce problème.

4.1.2 Les plongements de mots

Cette méthode a montré son intérêt dans l'étude des débats parlementaires contemporains (Rheault et Cochrane, 2020) ; il nous a semblé donc tout à fait pertinent de l'utiliser pour analyser des débats parlementaires plus anciens. Nous avons utilisé l'algorithme Word2Vec (Mikolov et al., 2013), qui se base sur un apprentissage par réseau de neurones pour entraîner des vecteurs sur des mots. Le principe est que des mots qui apparaissent dans un contexte similaire se voient accorder des vecteurs similaires (proche en termes de similarité cosinus).

Deux méthodes sont proposées : Skip-Gram (SG) et *Continuous Bag Of Words* (CBOW). Cette dernière méthode cherche à prédire un vecteur associé à un mot en se basant sur une fenêtre de mot autour de lui. Skip-Gram fait l'inverse en prédisant les mots du contexte à partir d'un mot cible. Le contexte correspond à une fenêtre de mots à gauche et à droite du mot cible. Nous utilisons l'implémentation *Skip-Gram* de la bibliothèque Python Gensim avec une fenêtre de 15 mots. En complément de cette méthode, l'algorithme Doc2Vec a été proposé (Le et Mikolov, 2014) pour associer chaque document à un vecteur le représentant.

C'est avec ces méthodes que Top2Vec (Angelov, 2020) cherche à extraire des *topics*. Une fois qu'on a associé à chaque élément (mots/documents) de notre corpus un vecteur, on recherche des zones denses de documents dont on va extraire les *topics* comme les centroïdes de ces zones. Les mots composant chacun des *topics* sont choisis selon la proximité de leur vecteur avec celui du *topic* trouvé. Ainsi chacun des mots, documents et *topics* sont projetés dans le même espace permettant des analyses comparées.

Nous utilisons la méthode de *clustering* HDBSCAN (*Hierarchichal DBSCAN*), connue pour son efficacité à regrouper en grande dimension. De plus, cette méthode non-supervisée identifie elle-même le nombre de *clusters*, c'est à dire le nombre de *topics*. On note que contrairement à LDA, l'implémentation de base associe un unique *topic* à chaque document. On utilise l'implémentation Python de l'auteur avec le modèle le plus précis (*deep-learn*). S'il n'est pas nécessaire d'enlever les mots vides ou bien de lemmatiser, nous avons décidé de le faire de façon à travailler avec le même corpus que LDA, et de réduire un peu sa taille.

L'algorithme une fois entraîné identifie un nombre total de 375 *topics* pour nos 75.775 documents. Nous obtenons donc un grand nombre de *topics*, ce qui contraste fortement avec l'implémation de LDA pour laquelle nous en avions défini une cinquantaine seulement.

Topic 37	Topic 40	Topic 42	Topic 54
institutrice	concordat	artiste	lycée
instituteur	pape	beaux-arts	enseignement
paris	primaire	musée	universitaire
canal	élémentaire	louvre	internat
enseignement	catholique	décoratif	bachelier

TABLEAU 2 – Quatre *topics* obtenus avec Top2Vec

Le tableau 2 illustre ici quelques-uns des *topics* très pertinents que nous avons trouvés. On remarque à la fois la cohérence dans les mots trouvés et leurs spécificités. Prenant par exemple le cas des *topics* 37 et 54. Bien qu'appartenant à une même thématique, celle de l'enseignement (ce mot est présent dans les deux *topics*), le premier se concentre sur l'enseignement primaire, quand l'autre traite du secondaire et du supérieur. Cette diversité et spécificité des *topics* se retrouve dans une grande partie des 375 identifiés.

Dans le graphique 4, on illustre le principe de

projection des *topics* dans l'espace et l'identification des mots associés. En rouge sont projetés

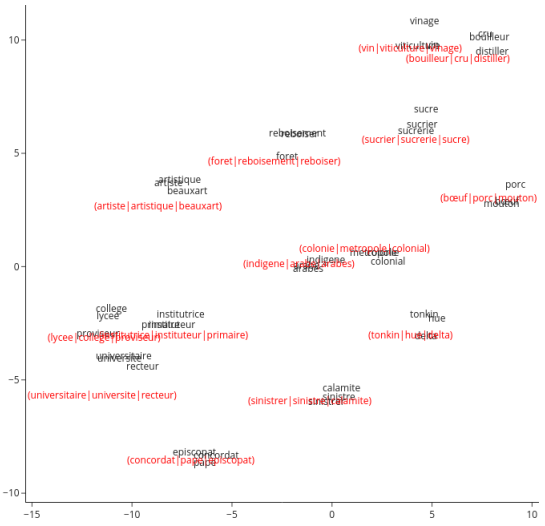


FIGURE 4 – Projection des mots par *topics* avec t-SNE

les *topics* auxquels on va associer les mots les plus proches dans l'espace [9]. Chacun des éléments étant originellement dans un espace de 300 dimensions, on utilise la méthode t-SNE (van der Maaten et Hinton, 2008) pour projeter les mots dans le plan. Si une projection dans un plan entraîne nécessairement une perte d'informations, on remarque toutefois que certains *topics* sont regroupés. Par exemple, le *topic* (vin | viticulture | vinage) est proche de (bouilleur | cru | distiller) ; et les *topics* (universitaire | université | recteur), (lycée | collège | proviseur) et (instituteur | institutrice | primaires) sont dans la même zone en bas à gauche. Cette représentation nous confirme la bonne qualité de projection de nos documents et de la représentation de nos *topics*.

Topic 34	Topic 45	Topic 55
Vin	Cheminot	Sinistré
Viticulture	Syndicaliste	Calamité
Alcoolisation	Confédération	Cyclone
Coupage	Militant	Gelée

TABLEAU 3 – Trois *topics* extraits de Top2Vec et les quatre mots les plus représentatifs

Le tableau 3 et la figure 5 permettent d'étudier certains *topics* plus en détails. On note un pic intéressant concernant le *topic* (cheminot | syndicaliste

9. Chaque *topic* est représenté par les trois mots les plus proches séparés par des *pipes*.

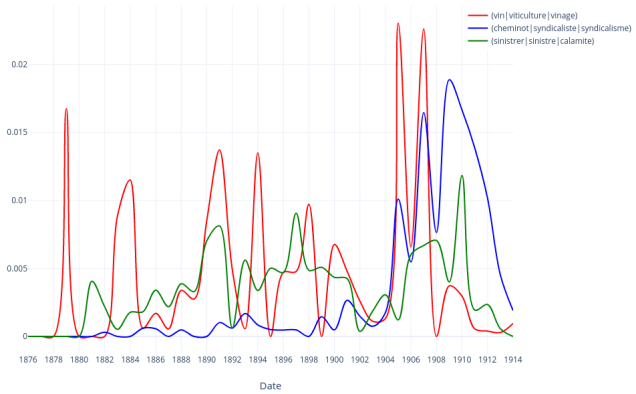
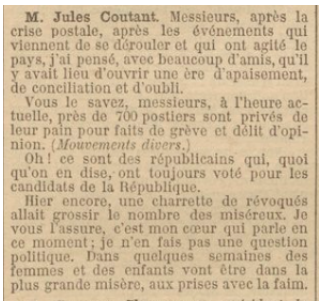


FIGURE 5 – Proportion des documents associés à chacun des trois *topics*, agrégation par années. Chacun des *topics* est identifié par ses trois mots les plus proches.

| syndicalisme) pendant l'année 1909. En utilisant la proximité entre *topics* et documents qui ont été projetés dans le même espace, on peut notamment retrouver des exemples de documents : la figure 6 est un extrait du *Journal Officiel* du 29 mai 1909 où il est question des suites à donner à la grève des PTT (Postes, télégraphes et téléphones). L'année 1909 voit en effet une importante grève des PTT qui affecte inévitablement la distribution du courrier, en grande partie acheminé par le train.



M. Jules Coutant. Messieurs, après la crise postale, après les événements qui viennent de se dérouler et qui ont agité le pays, j'ai pensé, avec beaucoup d'amis, qu'il y avait lieu d'ouvrir une ère d'apaisement, de conciliation et d'oubli.

Vous le savez, messieurs, à l'heure actuelle, près de 700 postiers sont privés de leur pain pour faits de grève et délit d'opinion. *(Mouvements divers.)*

Oh ! ce sont des républicains qui, quoi qu'on en dise, ont toujours voté pour les candidats de la République.

Hier encore, une charrette de révoqués allait grossir le nombre des miséreux. Je vous l'assure, c'est mon cœur qui parle en ce moment ; je n'en fais pas une question politique. Dans quelques semaines des femmes et des enfants vont être dans la plus grande misère, aux prises avec la faim.

FIGURE 6 – Extrait d'un document ayant une forte similarité avec le *topic* de la grève, au cours de l'année 1909

On peut aussi faire une analyse temporelle plus fine en faisant une agrégation par mois (figure 7). On note un gros pic en octobre 1910 : en revenant vers le corpus, on comprend que ce pic fait référence à la grève des cheminots. La figure 8 présente un extrait d'un débat portant sur ce sujet.

Plutôt que d'étudier des *topics* individuellement on peut décider d'étudier un groupe de *topics* autour d'une même thématique. Par exemple la figure 9 montre la proportion de documents dont le *topic* associé appartient au champ sémantique de la grève.

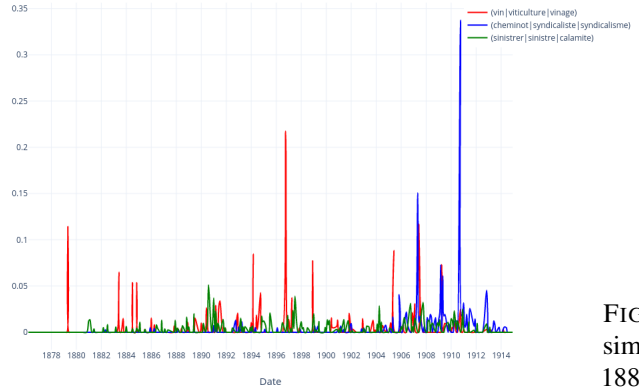
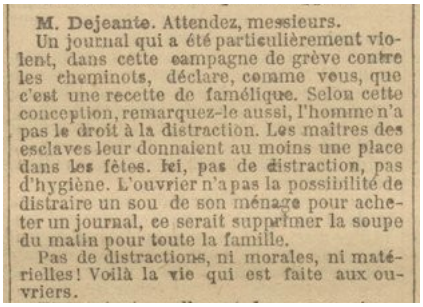


FIGURE 7 – Proportion des documents associés à chacun des trois *topics*, agrégation par mois. Chacun des *topics* est identifié par ses trois mots les plus proches.



M. Dejeante. Attendez, messieurs.

Un journal qui a été particulièrement violent, dans cette campagne de grève contre les cheminots, déclare, comme vous, que c'est une recette de famélique. Selon cette conception, remarquez-le aussi, l'homme n'a pas le droit à la distraction. Les maîtres des esclaves leur donnaient au moins une place dans les fêtes. Ici, pas de distraction, pas d'hygiène. L'ouvrier n'a pas la possibilité de distraire un sou de son ménage pour acheter un journal, ce serait supprimer la soupe du matin pour toute la famille.

Pas de distractions, ni morales, ni matérielles! Voilà la vie qui est faite aux ouvriers.

FIGURE 8 – Extrait d'un document ayant une forte similarité avec le topic de la grève pendant la période d'octobre 1910

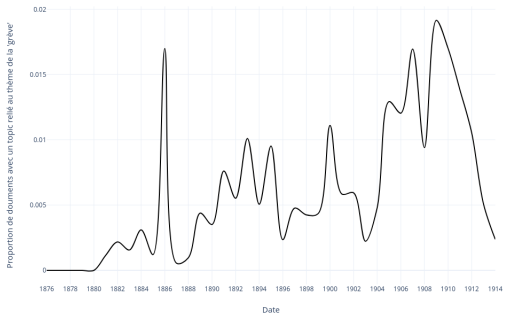
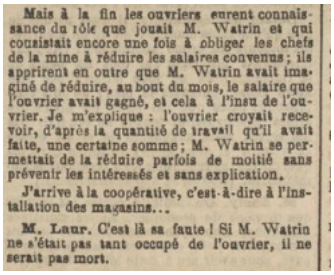


FIGURE 9 – Proportion des documents associés aux *topics* sémantiquement similaires au mot "grève", agrégation par années

Dans la figure 9, en plus des pics constatés précédemment autour des années 1910, on remarque un autre pic important pour l'année 1886. Il correspond aux discussions suite à la grève tragique à Decazeville survenue le 26 janvier 1886. La figure 10 est un exemple de prise de parole survenue pendant le débat du 11 février 1886.



Mais à la fin les ouvriers eurent connaissance du rôle que jouait M. Watrin et qui consistait encore une fois à obliger les chefs de la mine à réduire les salaires convenus ; ils apprirent en outre que M. Watrin avait imaginé de réduire, au bout du mois, le salaire que l'ouvrier avait gagné, et cela à l'insu de l'ouvrier. Je m'explique : l'ouvrier croyait recevoir, d'après la quantité de travail qu'il avait faite, une certaine somme ; M. Watrin se permettait de la réduire parfois de moitié sans prévenir les intéressés et sans explication.

J'arrive à la coopérative, c'est-à-dire à l'installation des magasins...

M. Laur. C'est là sa faute ! Si M. Watrin ne s'était pas tant occupé de l'ouvrier, il ne serait pas mort.

FIGURE 10 – Extrait d'un document ayant une forte similarité avec le « *topic* »de la grève pendant l'année 1886

Ces analyses montrent que le grand nombre de *topics* ne semble pas réduire leur pertinence ; au contraire, on constate une plus grande précision des *topics* identifiés.

4.1.3 Modèles de langue et BERTopic

De nouvelles méthodes de traitement automatique de la langue se sont popularisées ces dernières années, plus particulièrement les modèles de langue basés sur de l'apprentissage profond. Par exemple, l'architecture *Transformers* (Vaswani et al., 2017) propose un apprentissage basé sur des mécanismes d'attention. Dans la foulée, se sont développés des modèles plus avancés comme BERT (Devlin et al., 2019). Ces modèles pré-entrainés permettant de réaliser un grand nombre de tâches comme la reconnaissance d'entités nommées ou encore l'analyse de sentiment.

Le mécanisme d'attention est un aspect spécifique des modèles de langue. Avec Word2Vec, une fois la fenêtre contextuelle choisie, les mots dans la fenêtre ont tous la même importance pour déterminer le mot cible. À l'inverse, le mécanisme d'attention peut choisir spécifiquement les tokens auxquels il veut accorder de l'importance. L'autre conséquence est que l'*embedding* d'un mot pourra varier selon son contexte. En effet la limite des algorithmes comme CBOW ou Skip-Gram venait du fait qu'un unique mot (Devlin et al., 2019) a un unique vecteur. Ce problème est résolu par les *Transformers*. La conséquence principale est l'augmentation du coût en temps de calcul, car la complexité est quadratique par rapport aux nombres de mots du contexte.

Afin de faire une première évaluation de cette méthode, nous avons décidé d'utiliser CamemBERT (Martin et al., 2020), un modèle pré-entraîné

sur des documents en français [10], et d'évaluer les résultats obtenus. Sur cette base d'*embeddings*, on utilise le modèle BERTopic pour générer les plongements de mots (Grootendorst, 2022). Cette implémentation très modulable génère des *topics* sur les bases d'*embeddings* de vecteur.

De manière similaire à Top2Vec, ces plongements vont être utilisés pour extraire les *topics* de nos documents. Contrairement à l'implémentation Doc2Vec, les modèles pré-entraînés ont une limite pour le nombre de tokens. Si les documents donnés en entrée sont trop longs, on gardera uniquement la première partie des tokens correspondant à la limite fixée par le modèle, et le reste sera ignoré. Pour des modèles comme CamemBERT, la limite est de 512 tokens.

Contrairement à l'approche précédente où nous utilisions un modèle agnostique qui n'avait aucune connaissance préalable et qui apprenait tout à partir de nos documents, nous utilisons cette fois-ci un modèle pré-entraîné sur des documents en français. Les principaux avantages de ces méthodes résident dans la complexité accrue des modèles de langage et leur pré-entraînement massif. Cependant, malgré ce pré-entraînement, l'encodage contextuel de nos documents reste très long en raison de leur taille, et donc du contexte à prendre en compte. Alors que Word2Vec utilise une fenêtre de 15 mots, ici, tout le contexte dans la limite théorique de 512 tokens de CamemBERT est utilisé avec la fenêtre d'attention.

Cette complexité accrue nous a conduit à prendre deux décisions : réduire la taille des blocs à 256 tokens, en créant des « *chunks* » de taille plus petite, et ne pas effectuer de phase d'entraînement spécifique sur nos données. Aux deux atouts majeurs de cette méthode s'ajoutent donc, dans notre contexte, deux inconvénients : le redécoupage des blocs et donc un changement de la structure, et l'absence d'apprentissage spécifique sur nos données. Une fois que les *chunks* ont été extraits et que le nettoyage a été effectué, nous nous retrouvons avec environ 800.000 d'entre eux sur lesquels nous lançons directement CamemBERT afin d'extraire les *embeddings* contextuels.

10. Un grand nombre de modèles de langues différents peuvent en effet être intégrés.

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**
de	de	suffrages	voté
la	discussion	exprimés	déclare
le	du	de	porté
et	loi	inscrits	ayant

TABLEAU 4 – Quatre *topics* obtenus avec BERTopic (CamemBERT)

Comme on peut le voir, les quelques *topics* présentés dans le tableau 4 sont très décevants. C'est d'ailleurs le cas de la totalité des *topics* obtenus avec cette implémentation. On note que si nos *topics* contiennent des mots vides, c'est parce qu'ils n'ont pas été enlevés de notre corpus. Les modèle de langue basant leur compréhension sur la totalité de la phrase, il est préférable d'en garder la structure et donc les mots vides.

Plusieurs raisons expliquent un tel résultat. D'abord, notre modèle ne s'est pas entraîné spécifiquement sur nos données. Ces dernières ont en effet deux spécificités très particulières :

— Le cadre : dans les débats parlementaires, le langage et les thématiques spécifiques à ce domaine ne se retrouve pas nécessairement dans les données de pré-entrainement de CamemBERT.

— Le contexte : nos documents sont des OCR et contiennent donc des erreurs produisant du vocabulaire et des phrases que CamemBERT n'a probablement jamais rencontré, et une grammaire inconnue qu'il n'a pas pu apprendre.

La deuxième distinction concerne le choix du modèle de langue. En effet, CamemBERT a été entraîné pour accomplir une tâche très spécifique : prédire le prochain token ou le token manquant dans une phrase donnée. Bien que cet entraînement lui confère une compréhension précise de la grammaire, et constitue une base essentielle pour nos analyses, il lui manque l'aspect principal : la compréhension du concept de similarité entre les documents. Dans le cadre de la modélisation de sujet, notre intérêt réside précisément dans l'identification des documents similaires, afin d'en extraire les sujets ou thématiques qui les caractérisent, ainsi que les mots-clés associés à ces sujets. Malheureusement, CamemBERT ne possède pas cette capacité.

Heureusement, il existe des modèles qui répondent à ces besoins spécifiques. Ces modèles,

en plus de leur capacité à prédire les tokens manquants, proposent également des *embeddings* de documents basés sur des scores de similarité. Ces scores permettent au modèle de générer des vecteurs similaires pour des documents identifiés comme étant similaires (dans une base d'entraînement préalable). Dans notre cas, nous avons décidé d'utiliser le modèle multilingue « paraphrase-multilingual-MiniLM-L12-v2 [11] », en association avec BERTopic, pour extraire nos *topics*. Ce modèle offre des fonctionnalités avancées pour capturer la similarité entre les documents en se basant sur les mots et leurs contextes.

L'utilisation de ce modèle multilingue nous permet d'aborder des documents dans différentes langues et d'obtenir des résultats plus pertinents et cohérents. En combinant les capacités de « paraphrase-multilingual-MiniLM-L12-v2 » et BERTopic, nous sommes en mesure d'extraire plus efficacement les sujets des documents analysés.

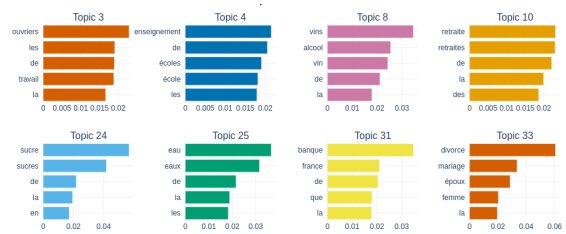


FIGURE 11 – Huit *topics* obtenus avec BERTopic, et les cinq mots les plus représentatifs

Les *topics* obtenus dans la figure 11, toujours imparfaits, sont cependant plus intéressants. On note un *topic* spécifique au sujet de la banque qu'on n'obtenait pas via la méthode Top2Vec. S'il est possible de pousser l'analyse comme avec Top2Vec, on constate que les *topics* sont moins cohérents. En effet, si la présence de mots vides s'explique car ils n'ont pas été retirés (pour les raisons expliquées plus haut), ils ne devraient pas apparaître dans un modèle performant, car ils ne sont pas caractéristiques du sujet (par exemple le token « en » dans le topic 24). BERTopic propose une implémentation pour les retirer à posteriori, mais le constat ne change pas. Par ailleurs, une particularité de cette méthode est qu'elle s'autorise de ne pas classifier certains documents, ou plutôt d'en associer à la classe « *outliers* »c'est à dire des documents présents dans une zone non suffisamment dense pour en extraire un sujet. Ces *outliers* représentent près de la moitié de nos documents : le risque d'avoir mis de côté un trop grand nombre de sujets significatifs est donc très élevé.

FIGURE 12 – Projection des huit *topics* précédents avec t-SNE

La figure 12 représente dans le plan les *topics* obtenus ainsi que les documents. On remarque en gris les documents *outliers* qui sont très nombreux. Si certaines projections sont intéressantes, comme la proximité des *topics* (suffrages | exprimés | inscrits) et (dépouillement | scrutins | votants), la représentation semble moins efficace qu'avec Top2Vec.

4.2 Comparaison des trois méthodes

LDA nous fournit des sujets plus génériques et constants dans le temps, alors que Top2Vec fait apparaître des sujets plus précis mais également beaucoup plus nombreux. C'est parce que LDA voit chaque document comme un mélange de sujets, tandis que Top2Vec associe un sujet à un document [12] - le nombre de *topics* est donc proportionnel au nombre de documents. Top2Vec risque donc de créer de nombreux *topics* très spécifiques. LDA offre l'avantage de donner une vision plus générale du corpus. Les méthodes de plongements de mots proposent eux une analyse plus fine de nos documents, notamment avec la prise en compte du contexte, la structure de la phrase permettant d'avoir une vision plus détaillée de nos thématiques. Lorsque, par exemple, on s'intéresse à un sujet en particulier et à son évolution, il peut donc être particulièrement fructueux d'associer les deux méthodes.

Malgré les promesses offertes par les modèles d'apprentissage profond, nous avons constaté la limite des techniques basées sur du simple *Transfer Learning*. Utiliser un modèle, même adapté à notre

11. Disponible en ligne : https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

12. On se rappelle qu'un document dans ce contexte est un bloc de mots.

objectif (ici, la recherche de similarité entre documents) reste limité si ce modèle n'a pas aussi été entraîné sur les documents analysés. Les méthodes comme Word2Vec ou Doc2Vec, sont plus simples dans leur structure mais apprennent spécifiquement sur nos documents ; et pour notre cas d'étude, elles semblent plus efficaces. Cependant les premiers résultats obtenus avec BERTopic restent prometteurs ; cela nous pousse à approfondir les méthodes de modélisation de sujets par modèles de langue. Nous envisageons deux pistes de développement : utiliser FlauBERT [13] un autre modèle de langue en français, entraîné à reconnaître la similarité entre documents ; si les résultats obtenus avec FlauBERT s'avère décevant, implémenter une phase d'apprentissage spécifique à nos données.

5 Annoter les débats avec leurs sujets

Ces résultats nous encouragent également à utiliser les sujets générés par la modélisation de sujets, pour offrir aux utilisateurs la possibilité de naviguer entre les sujets et pas seulement entre les différents numéros du *Journal officiel*. Les sujets extraits avec LDA nous semblent les plus appropriés : leur « petit » nombre rend leur gestion plus aisée, et leur dimension générique permet d'embrasser plus facilement l'ensemble du corpus.

Pour intégrer ces résultats à la plateforme que nous sommes en train de développer, il nous a semblé naturel d'exploiter les possibilités offertes par la TEI pour l'annotation linguistique, et de faire des sujets, un nouveau type d'annotations. En utilisant l'élément `<standOff>`, nous pouvons stocker ces annotations directement dans les fichiers XML concernés. Le nom du *topic* est choisi à la main ; pour associer cette annotation sémantique à la liste de mots correspondante, nous proposons d'utiliser l'élément `<span>` qui permet d'attacher une note analytique à des parties sépcifiques du texte. La figure 13 montre que chaque mot du texte est annoté par un élément `<w>` accompagné d'un identifiant unique composé de l'identifiant du document (formé par le préfixe `ps` pour séance parlementaire, suivi de la date de la séance) et d'un numéro correspondant à la place du mot dans le texte. Un attribut `ref` associe ensuite le sujet aux mots correspondants. Sur la figure 14, on voit que nous choisi de regrouper les annotations sémantiques dans l'élément `<standOff>` afin de faciliter leur gestion. Ces balises `<span>` sont également re-

13. https://github.com/getalp/Flaubert

```
<body>
   <!-- [...] -->
   <u>
      <!-- [...] -->
      <!-- "some of the war material in Madagascar" -->
      <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>.
      <!-- [...] -->
   </u>
   <!-- [...] -->
</body>
```

FIGURE 13 – Annotations dans le corps de texte

groupées dans un élément `<spanGrp>` associé à un attribut `type`.

```
<standOff>
   <spanGrp type="topic">
      <span target="#ps1895022_119">
         army</span>
      <span target=" #ps1895022_123">
         colonization</span>
   </spanGrp>
</standOff>
```

FIGURE 14 – Utilisation de l'élément `<standOff>`

Comme nous l'avons souligné, les intitulés des *topics* sont choisis à la main. Nous proposons d'utiliser un vocabulaire contrôlé pour identifier ces sujets. Une des options envisagées consisterait à utiliser le langage d'indexation matière RAMEAU [14], qui fournit des vedettes matière réutilisables dans le contexte de notre projet. Par exemple, RAMEAU fournit une vedette matière « Colonisation » [15]. L'utilisation de ce langage d'indexation permettrait de trouver des intitulés homogènes pour les sujets identifiés et donc de faciliter l'annotation à grande échelle de ces textes.

6 Conclusion

Grâce à cet article, nous avons pu évaluer différentes méthodes pour extraire les sujets des débats parlementaires de la Troisième République. Au vu de nos résultats, LDA et Top2vec proposent des approches complémentaires. Une bonne manière de procéder consiste à réaliser une première analyse avec LDA, avant d'utiliser les plongements de mots pour affiner certains résultats. On a vu que les méthodes basées sur l'apprentissage profond présentent l'inconvénient de ne pas être parfaitement adaptées à notre corpus d'étude ; il faut donc se demander si l'entraînement d'un modèle de langue sur nos données nous permettrait d'obte-

14. https://rameau.bnf.fr/
15. https://catalogue.bnf.fr/ark:/12148/cb119596539

nir des résultats significativement meilleurs ou non. Un tel entraînement est en effet coûteux en termes de temps ; mais il pourrait être envisagé de créer un modèle de langue en français adapté au langage politique du XIX^{e} et du début du XX^{e} siècle - voire jusqu'à nos jours.

Pour affiner ce travail, nous prévoyons de relancer les mêmes analyses sur des textes réocérisés et donc moins fautifs (Puren et al., 2022), et d'évaluer les nouveaux résultats ainsi obtenus. Une prochaine étape consistera également à annoter les débats avec les sujets ainsi extraits. Il sera alors nécessaire de réfléchir à une stratégie d'annotation automatique, permettant d'associer les mots et les sujets.

Bibliographie

Gavin Abercrombie et Riza Batista-Navarro. 2020. Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. *Journal of Computational Social Science*, 3(1) :245–270.

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

Dimo Angelov. 2020. Top2vec: Distributed representations of topics.

Annales. 2015. La longue durée en débat. *Annales. Histoire, Sciences Sociales*, 70e année(2) :285–287.

Helen Baker, Vaclav Brezina, et Tony McEnery. 2017. Ireland in British parliamentary debates 1803–2005:Plotting changes in discourse in a large volume of time-series corpus data. In *Exploring Future Paths for Historical Sociolinguistics. Advances in Historical Sociolinguistics*, pages 83–107. John Benjamins.

Luke Blaxill. 2022. Parliamentary Corpora and Research in Political Science and Political History. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 33–34, Marseille, France. European Language Resources Association.

David Blei, Andrew Ng, et Michael Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3 :993–1022.

Hugo Bonin. 2020. From antagonist to protagonist: 'democracy' and 'people' in british parliamentary debates, 1775–1885. *Digital Scholarship in the Humanities*, 35(4) :759–775.

Nicolas Bourgeois, Aurélien Pellet, et Marie Puren. 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*. CEUR Workshop.

Frédéric Clavert. 2014. Vers de nouveaux modes de lecture des sources. In Olivier Le Deuff, éditeur, *Le temps des humanités digitales*. FYP EDITIONS, Roubaix.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, et Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Maarten Grootendorst. 2022. Bertopic : Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv :2203.05794*.

Pasi Ihalainen. 2020. European History as a Nationalist and Post-Nationalist Project.

Pasi Ihalainen, Berit Janssen, Jani Marjanen, et Ville Vaara. 2022. Building and Testing a Comparative Interface on Northwest European Historical Parliamentary Debates : Relative Term Frequency Analysis of British Representative Democracy. In *DHNB 2022 – Digital Humanities in Action - Workshop "Digital Parliamentary Data in Action*. CEUR Workshop.

Karine Karila-Cohen, Claire Lemercier, Isabelle Rosé, et Claire Zalc. 2018. Nouvelles cuisines de l'histoire quantitative. *Annales. Histoire, Sciences Sociales*, 73(4) :771–783. Bibliographie_available : 0 Cairndomain : www.cairn.info Cite Par_available : 0 Publisher : Éditions de l'EHESS.

Max Kemman. 2021. *Trading zones of digital history*. Numéro Volume 1 in Studies in digital history and hermeneutics. De Gruyter Oldenbourg, Berlin. Country : DE Diagramme, Pläne. 24 cm. Bibliogr. p. 161-178.

Lauren F. Klein, Jacob Eisenstein, et Iris Sun. 2015. Exploratory Thematic Analysis for Digitized Archival Collections. *Digital Scholarship in the Humanities*, 30(suppl_1) :i130–i141.

Quoc V. Le et Tomas Mikolov. 2014. Distributed representations of sentences and documents.

Claire Lemercier. 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.

Claire Lemercier et Claire Zalc. 2019. *Quantitative methods in the humanities : an introduction*. University of Virginia Press, Charlottesville. OCLC : 1099931813.

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

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, et Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, et Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *ICLR*.

Franco Moretti. 2000. Conjectures on World Literature. *New Left Review*, 238(1) :54–68.

Franco Moretti. 2013. *Distant reading*. Verso, London.

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

Marie Puren, Aurélien Pellet, Nicolas Bourgeois, Pierre Vernus, et Fanny Lebreton. 2022. Between History and Natural Language Processing: Study, Enrichment and Online Publication of French Parliamentary Debates of the Early Third Republic (1881-1899). In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 33–34. European Language Resources Association.

Marie Puren et Pierre Vernus. 2021. Agoda : Analyse sémantique et graphes relationnels pour l'ouverture et l'étude des débats à l'assemblée nationale.

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

Emilien Ruiz. 2022. Former « au numérique » en sciences humaines et sociales ? Propositions d'un historien. In Clarisse Bardiot, Esther Dehoux, et Emilien Ruiz, éditeurs, *La fabrique numérique des corpus en sciences humaines et sociales*, Humanités numériques et science ouverte. Presses universitaires du Septentrion, Lille.

Hannu Salmi. 2021. *What is digital history ?* Polity Press, Cambridge, UK.

Graham Shawn, Ian Milligan, et Scott Weingart. 2016. *Exploring big historical data : the historian's macroscope*. Imperial College Press, London.

Alexander Stulpe et Matthias Lemke. 2016. Blended Reading. In Matthias Lemke et Gregor Wiedemann, éditeurs, *Text Mining in den Sozialwissenschaften : Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse*, pages 17–61. Springer Fachmedien, Wiesbaden.

Laurens van der Maaten et Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86) :2579–2605.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, et Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lorella Violla et Japp Verheul. 2020. Mining ethnicity: Discourse-driven topic modelling of immigrant discourses in the usa, 1898–1920. *Digital Scholarship in the Humanities*, 35(4) :921–943.