EXTRAIRE DU TEXTE DES IMPRIMÉS ANCIENS : QUELS DÉFIS, QUELLES PROMESSES ?

Illustration par le projet AGODA

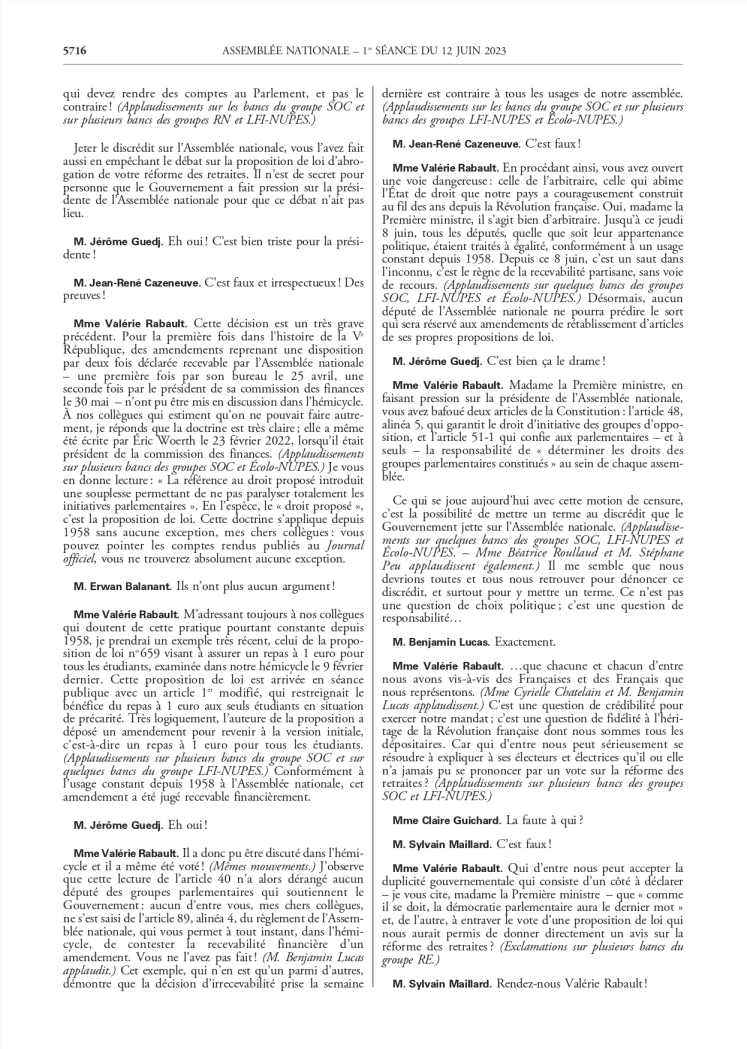
Marie Puren[1 2] (et Aurélien Pellet et Pierre Vernus)

8 février 2024 - Séminaire de l'Axe de recherche en histoire numérique du LARHRA

[1]LRE, EPITA [2]CJM, Ecole nationale des chartes

LES COMPTES-RENDUS DES DÉBATS PARLEMENTAIRES FRANÇAIS

- Naissance spontanée des comptes-rendus des débats en 1789
- Publicisation des discussions entre parlementaires jusqu'à nos jours
- A partir de 1881 : diffusion des débats à la Chambre des députés et au Sénat via des publications dédiées
    - Forme très similaire à celles des débats publiés aujourd'hui



5716 ASSEMBLÉE NATIONALE – 1re SÉANCE DU 12 JUIN 2023

qui devez rendre des comptes au Parlement, et pas le contraire ! *(Applaudissements sur les bancs du groupe SOC et sur plusieurs bancs des groupes RN et LFI-NUPES.)*

Jeter le discrédit sur l'Assemblée nationale, vous l'avez fait aussi en empêchant le débat sur la proposition de loi d'abrogation de votre réforme des retraites. Il n'est de secret pour personne que le Gouvernement a fait pression sur la présidente de l'Assemblée nationale pour que ce débat n'ait pas lieu.

M. Jérôme Guedj. Eh oui ! C'est bien triste pour la présidente !

M. Jean-René Cazeneuve. C'est faux et irrespectueux ! Des preuves !

Mme Valérie Rabault. Cette décision est un très grave précédent. Pour la première fois dans l'histoire de la Ve République, des amendements reprenant une disposition par deux fois déclarée recevable par l'Assemblée nationale – une première fois par son bureau le 25 avril, une seconde fois par le président de sa commission des finances le 30 mai – n'ont pu être mis en discussion dans l'hémicycle. À nos collègues qui estiment qu'on ne pouvait faire autrement, je réponds que la doctrine est très claire ; elle a même été écrite par Éric Woerth le 23 février 2022, lorsqu'il était président de la commission des finances. *(Applaudissements sur plusieurs bancs des groupes SOC et Écolo-NUPES.)* Je vous en donne lecture : « La référence au droit proposé introduit une souplesse permettant de ne pas paralyser totalement les initiatives parlementaires ». En l'espèce, le « droit proposé », c'est la proposition de loi. Cette doctrine s'applique depuis 1958 sans aucune exception, mes chers collègues : vous pouvez pointer les comptes rendus publiés au *Journal officiel*, vous ne trouverez absolument aucune exception.

M. Erwan Balanant. Ils n'ont plus aucun argument !

Mme Valérie Rabault. M'adressant toujours à nos collègues qui doutent de cette pratique pourtant constante depuis 1958, je prendrai un exemple très récent, celui de la proposition de loi n°659 visant à assurer un repas à 1 euro pour tous les étudiants, examinée dans notre hémicycle le 9 février dernier. Cette proposition de loi est arrivée en séance publique avec un article 1er modifié, qui restreignait le bénéfice du repas à 1 euro aux seuls étudiants en situation de précarité. Très logiquement, l'auteure de la proposition a déposé un amendement pour revenir à la version initiale, c'est-à-dire un repas à 1 euro pour tous les étudiants. *(Applaudissements sur plusieurs bancs du groupe SOC et sur quelques bancs du groupe LFI-NUPES.)* Conformément à l'usage constant depuis 1958 à l'Assemblée nationale, cet amendement a été jugé recevable financièrement.

M. Jérôme Guedj. Eh oui !

Mme Valérie Rabault. Il a donc pu être discuté dans l'hémicycle et il a même été voté ! *(Mêmes mouvements.)* J'observe que cette lecture de l'article 40 n'a alors dérangé aucun député des groupes parlementaires qui soutiennent le Gouvernement : aucun d'entre vous, mes chers collègues, ne s'est saisi de l'article 89, alinéa 4, du règlement de l'Assemblée nationale, qui vous permet à tout instant, dans l'hémicycle, de contester la recevabilité financière d'un amendement. Vous ne l'avez pas fait ! *(M. Benjamin Lucas applaudit.)* Cet exemple, qui n'en est qu'un parmi d'autres, démontre que la décision d'irrecevabilité prise la semaine dernière est contraire à tous les usages de notre assemblée. *(Applaudissements sur les bancs du groupe SOC et sur plusieurs bancs des groupes LFI-NUPES et Écolo-NUPES.)*

M. Jean-René Cazeneuve. C'est faux !

Mme Valérie Rabault. En procédant ainsi, vous avez ouvert une voie dangereuse : celle de l'arbitraire, celle qui abîme l'État de droit que notre pays a courageusement construit au fil des ans depuis la Révolution française. Oui, madame la Première ministre, il s'agit bien d'arbitraire. Jusqu'à ce jeudi 8 juin, tous les députés, quelle que soit leur appartenance politique, étaient traités à égalité, conformément à un usage constant depuis 1958. Depuis ce 8 juin, c'est un saut dans l'inconnu, c'est le règne de la recevabilité partisane, sans voie de recours. *(Applaudissements sur quelques bancs des groupes SOC, LFI-NUPES et Écolo-NUPES.)* Désormais, aucun député de l'Assemblée nationale ne pourra prédire le sort qui sera réservé aux amendements de rétablissement d'articles de ses propres propositions de loi.

M. Jérôme Guedj. C'est bien ça le drame !

Mme Valérie Rabault. Madame la Première ministre, en faisant pression sur la présidente de l'Assemblée nationale, vous avez bafoué deux articles de la Constitution : l'article 48, alinéa 5, qui garantit le droit d'initiative des groupes d'opposition, et l'article 51-1 qui confie aux parlementaires – et à seuls – la responsabilité de « déterminer les droits des groupes parlementaires constitués » au sein de chaque assemblée.

Ce qui se joue aujourd'hui avec cette motion de censure, c'est la possibilité de mettre un terme au discrédit que le Gouvernement jette sur l'Assemblée nationale. *(Applaudissements sur quelques bancs des groupes SOC, LFI-NUPES et Écolo-NUPES. – Mme Béatrice Roullaud et M. Stéphane Peu applaudissent également.)* Il me semble que nous devrions toutes et tous nous retrouver pour dénoncer ce discrédit, et surtout pour y mettre un terme. Ce n'est pas une question de choix politique ; c'est une question de responsabilité…

M. Benjamin Lucas. Exactement.

Mme Valérie Rabault. …que chacune et chacun d'entre nous avons vis-à-vis des Françaises et des Français que nous représentons. *(Mme Cyrielle Chatelain et M. Benjamin Lucas applaudissent.)* C'est une question de crédibilité pour exercer notre mandat ; c'est une question de fidélité à l'héritage de la Révolution française dont nous sommes tous les dépositaires. Car qui d'entre nous peut sérieusement se résoudre à expliquer à ses électeurs et électrices qu'il ou elle n'a jamais pu se prononcer par un vote sur la réforme des retraites ? *(Applaudissements sur plusieurs bancs des groupes SOC et LFI-NUPES.)*

Mme Claire Guichard. La faute à qui ?

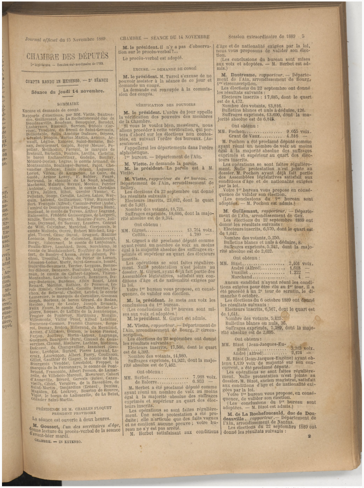
M. Sylvain Maillard. C'est faux !

Mme Valérie Rabault. Qui d'entre nous peut accepter la duplicité gouvernementale qui consiste d'un côté à déclarer – je vous cite, madame la Première ministre – que « comme il se doit, la démocratie parlementaire aura le dernier mot » et, de l'autre, à entraver le vote d'une proposition de loi qui nous aurait permis de donner directement un avis sur la réforme des retraites ? *(Exclamations sur plusieurs bancs du groupe RE.)*

M. Sylvain Maillard. Rendez-nous Valérie Rabault !

FIGURE – Séance parlementaire du 12 juin 2023

LES DÉBATS PARLEMENTAIRES DURANT LA IIIE RÉPUBLIQUE


Journal officiel du 15 Novembre 1889

CHAMBRE — SÉANCE DU 14 NOVEMBRE

Session extraordinaire de 1889 5

CHAMBRE DES DÉPUTÉS

COMPTE RENDU IN EXTENSO. — 2e SÉANCE

Séance du jeudi 14 novembre.

SOMMAIRE

M. le président. Il n'y a pas d'observation sur le procès-verbal ?...

Le procès-verbal est adopté.

EXCUSE. — DEMANDE DE CONGÉ

VÉRIFICATION DES POUVOIRS

PRÉSIDENCE DE M. CHARLES FLOQUET
PRÉSIDENT PROVISOIRE

La séance est ouverte à deux heures.

FIGURE – Séance parlementaire du 14 novembre 1889

- Projet AGODA : **A**nalyse sémantique et **G**raphes relationnels pour l'**O**uverture et l'étude des **D**ébats à l'**A**ssemblée nationale
- Débats à la Chambre des députés (chambre basse du parlement) transcrits :
  - entre 1876 et 1880, dans les Annales du Sénat et de la Chambre des députés
  - à partir de 1881, dans le Journal officiel de la République française. Débats parlementaires (1881-1940)
- Disponible en ligne via Gallica (bibliothèque numérique de la Bibliothèque nationale de France)

Des sources pour l'histoire

- Permettent d'observer l'évolution des questions politiques et sociales traitées à l'assemblée
- « Corpus remarquable » pour l'histoire politique et intellectuelle
- Sources précieuses pour l'histoire sociale, économique, culturelle et l'histoire du droit

OBJECTIFS

- Créer une plateforme de consultation
- Produire des données textuelles structurées et sémantiquement enrichies à partir de ces débats numérisés
- Contribuer à la conception d'un workflow adapté à la préparation, à la publication et à l'analyse de grands corpus de documents historiques

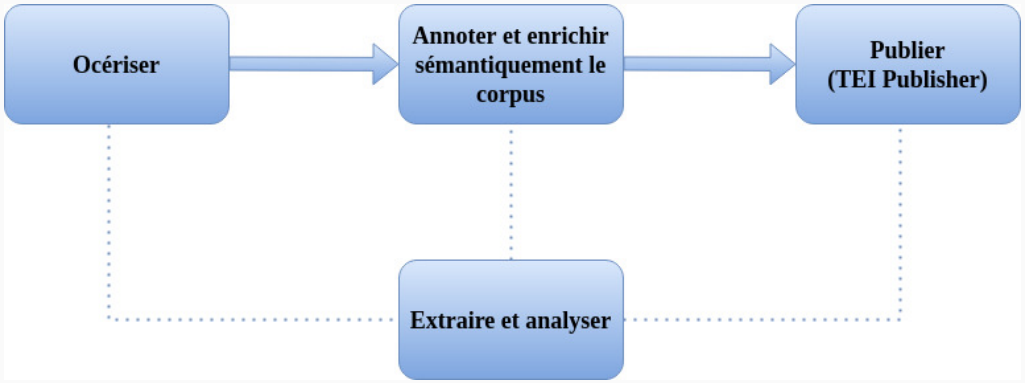
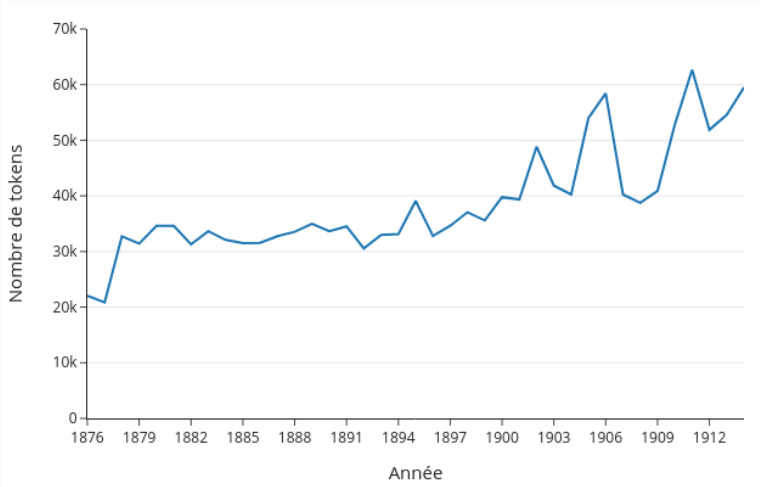


Figure – Les étapes de la chaîne de traitement

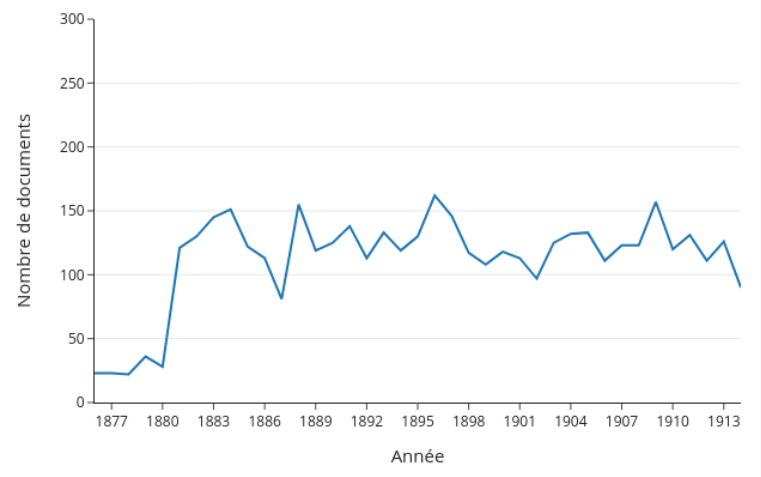
UNE PREUVE DE CONCEPT

- Projet « preuve de concept »
- Ne pas traiter l'ensemble des débats parlementaires depuis 1789 mais travailler sur des sous-corpus test
  - Débats à la Chambre des députés
  - Plateforme de publication : période 1889-1893
  - Analyses en TAL menées sur un corpus allant de 1876 à 1914

Un corpus « massif ». Exemple : 15 législatures entre 1881 et 1940 à la Chambre des députés / 10-12.000 images par législature



(a) Médiane du nombre de tokens par années (1876-1914)



(b) Évolution du nombre de documents par années 1876-1914

FIGURE – Un corpus qui ne cesse de grandir au cours du temps

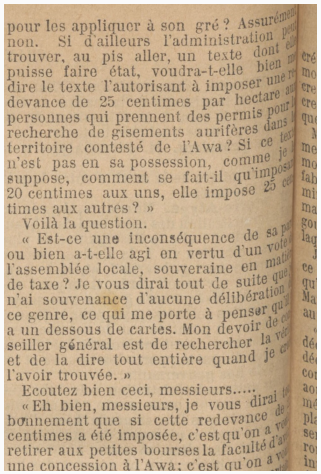
OCÉRISER LES DÉBATS

LE CAS DES DÉBATS PARLEMENTAIRES

- Récupération des textes océrisés via l'API Document de Gallica
  => qualité inégale de l'OCR
- Erreurs dues à :
  - qualité du document : tâches et surimpression
  - la **courbure de la page** au niveau de la reliure

FIGURE – Evaluation de la qualité de l'OCR fourni par Gallica

EFFET DE LA COURBURE SUR LES RÉSULTATS DE L'OCR



pour les appliquer à son gré ? Assure.

non. Si d'ailleurs l'administration fj, trouver, au pis aller, un texte (1011t eu puisse faire état, voudra-t-elle dire le texte l'autorisant à imposer u devance de 25 centimes par hectar , personnes qui prennent des permis Pour1 recherche de gisements aurifères a teSI territoire contesté de l'Awa ? Si ce tes

n'est pas en sa possession, cornas le suppose, comment se fait-il qulIWjs3j 20 centimes aux uns, elle impose 20 cet times aux autres ? » , Voilà la question. é

« Est-ce une inconséquence de sa te ou bien a-t-elle agi en vertu d'un v l'assemblée locale, souveraine en OlqallC de taxe? Je vous dirai tout de suite Il n'ai souvenance d'aucune délibératl°ll'V ce genre, ce qui me porte à penser a un dessous de cartes. Mon devoir seiller général est de rechercher la cJ'!" et de la dire tout entière quand Je l'avoir trouvée. »

Ecoutez bien ceci, messieurs. l «Eh bien, messieurs, je vous dIra de bonnement que si cette redevance J centimes a été imposée, c'est qu'on a, retirer aux petites bourses la faculté" une concession à l'Awa; c'est qu'on a r¡;:i\

FIGURE – 20 octobre 1890 (p.1718)

FIGURE – Résultat de l'OCR

Décision de ré-océriser le corpus

TRAITER LES IMAGES

Améliorer la qualité de l'image avec une méthode de « dewarping »
=> résultats peu probants

- Gérer la courbure des pages avec le dewarping ?
- Utiliser des outils plus avancés ?

(a) Image d'origine

(b) Image « dewarpée »

FIGURE – Dewarping : pas adapté à nos documents

NETTOYAGE DE L'IMAGE

Année 1881 DÉBATS ET DOCUMENTS PARLEMENTAIRES 25 Février [CHAMBRE] 345

CHAMBRE DES DÉPUTÉS
Session ordinaire de 1881

COMPTE RENDU IN EXTENSO. — 23e SÉANCE

Séance du jeudi 24 février 1881

SOMMAIRE

Procès-verbal : M. Devès.
Excuse.
Dépôt, par M. Marquiset, du rapport de la commission chargée d'examiner le projet de loi ayant pour objet de faire déclarer d'utilité publique la construction d'un canal de navigation dans la vallée de la Chiers, entre Longwy et la Meuse, à Mouzon.
Dépôt, par M. Morel, du rapport de la 2e commission d'intérêt local, sur le projet de loi tendant à autoriser le département de la Haute-Marne à contracter un emprunt pour l'amélioration de l'asile des aliénés de Saint-Dizier.
Dépôt, par M. Sarrien, du rapport de la commission chargée d'examiner le projet de loi ayant pour objet de faire déclarer d'utilité publique l'amélioration du canal du centre.
Suite de la 1re délibération sur le projet de loi, adopté par le Sénat, concernant l'administration de l'armée : M. le comte de Roys, rapporteur. — Ajournement à samedi.
Adoption en 2e délibération du projet de loi ayant pour objet l'établissement d'un canal de Dombasle à Saint-Dié.
Discussion des conclusions du rapport de la 2e commission des pétitions sur la pétition du sieur Jean Dupuis, citoyen français, demeurant à Han-Kow (Chine) : M. Emile Bouchet, rapporteur. Adoption.
1re délibération sur les propositions de loi : 1° de M. Germain Casse et de plusieurs de ses collègues, tendant à régler les rapports entre les compagnies de chemins de fer et leurs agents commissionnés ; 2° de M. le baron de Janzé, relative à la réglementation des compagnies de chemins de fer : MM. René Brice, le baron de Janzé.
Règlement de l'ordre du jour : MM. Latrade, Sarrette, de Gasté, Bardoux.
Reprise de la discussion des propositions tendant à régler les rapports entre les compagnies de chemins de fer et leurs agents commissionnés : M. Trarieux.
Présentation, par M. le ministre de l'intérieur et des cultes, d'un projet de loi tendant à autoriser le département des Côtes-du-Nord à contracter deux emprunts pour les travaux des chemins vicinaux ordinaires.
Demande, par M. le ministre de la marine, de discussion immédiate du projet de loi, adopté par la Chambre des députés, adopté avec modifications par le Sénat, ayant pour objet d'ouvrir au ministre de la marine et des colonies un crédit de 8,352,751 fr., sur l'exercice 1881, pour les chemins de fer du Haut-Sénégal. Adoption. = Article 1er : M. Des Rotours. Adoption. = Adoption de l'article 2. = Adoption, au scrutin, de l'ensemble du projet de loi.
Règlement de l'ordre du jour : MM. le président, de Gasté.
Présentation par M. le sous-secrétaire d'Etat des travaux publics :
1° D'un projet de loi ayant pour objet : 1° L'incorporation dans le réseau d'intérêt général du chemin de fer d'intérêt local de Mézidon à Dives ; 2° l'approbation d'une convention passée avec la compagnie des chemins de fer de la vallée d'Auge pour le rachat de ladite ligne ;
2° D'un projet de loi ayant pour objet : 1° De déclarer d'utilité publique l'établissement, dans le département des Landes, de plusieurs chemins de fer d'intérêt local et d'en autoriser l'exécution ; 2° d'approuver un traité passé entre la compagnie du Midi et les concessionnaires de ces chemins.
Dépôt par M. Cavalié d'un rapport de la 2e commission d'intérêt local sur le projet de loi tendant à autoriser le département de la Haute-Saône à contracter un emprunt pour les travaux des chemins vicinaux.

PRÉSIDENCE DE M. GAMBETTA

La séance est ouverte à deux heures.
M. Frémiet, *l'un des secrétaires*, donne lecture du procès-verbal de la séance de mardi dernier.
M. le président. La parole est à M. Devès sur le procès-verbal.
M. Paul Devès. Messieurs, MM. Tassin, de Sonnier et Deniau auraient voté contre la proposition d'enquête faite par M. Lorgé s'ils eussent assisté à la séance de mardi. Je le déclare en leur nom.
M. le président. Il n'y a pas d'autre observation ?...
Le procès-verbal est adopté.

EXCUSE

M. le président. M. Dremel s'excuse de ne pouvoir assister à la séance de ce jour.

DÉPÔT DE RAPPORTS

M. le président. Je donne la parole à ceux de MM. les députés qui ont des rapports à déposer.
M. Sarrien. Au nom de la commission des voies navigables, j'ai l'honneur de déposer sur le bureau de la Chambre un rapport sur le projet de loi ayant pour objet de faire déclarer d'utilité publique l'amélioration du canal du Centre.
M. Marquiset. J'ai l'honneur de déposer sur le bureau de la Chambre, au nom de la commission des voies navigables, un rapport sur le projet de loi ayant pour objet de faire déclarer d'utilité publique la construction d'un canal de navigation dans la vallée de la Chiers, entre Longwy et la Meuse, à Mouzon.
M. Morel. Au nom de la 2e commission d'intérêt local, j'ai l'honneur de déposer sur le bureau de la Chambre un rapport sur le projet de loi tendant à autoriser le département de la Haute-Marne à contracter un emprunt pour l'amélioration de l'asile des aliénés de Saint-Dizier.
M. le président. Ces rapports seront imprimés et distribués.

AJOURNEMENT DE LA SUITE DE LA DISCUSSION DU PROJET DE LOI SUR L'ADMINISTRATION DE L'ARMÉE.

M. le président. L'ordre du jour appelle la suite de la première délibération sur le projet de loi, adopté par le Sénat, concernant l'administration de l'armée.
M. le rapporteur a la parole.
M. le comte de Roys, *rapporteur*. Messieurs, M. le ministre de la guerre étant appelé aujourd'hui au Sénat par une interpellation, la commission vous prie de vouloir bien ajourner à la séance de samedi la suite de la discussion sur l'administration de l'armée, en lui maintenant son rang en tête de l'ordre du jour de cette séance.
M. le président. Il n'y a pas d'opposition ?...
L'ajournement à samedi est prononcé.

DISCUSSION DU PROJET DE LOI CONCERNANT LE CANAL DE DOMBASLE A SAINT DIÉ

M. le président. L'ordre du jour appelle la deuxième délibération sur le projet de loi ayant pour objet l'établissement d'un canal de Dombasle à Saint-Dié.
Art. 1er. — Sont déclarés d'utilité publique les travaux à faire pour la construction du canal latéral à la Meurthe, de Dombasle à Saint-Dié, et son raccordement par voie ferrée avec la gare du chemin de fer de l'Est à Saint-Dié, conformément aux dispositions de l'avant-projet adopté par le conseil général des ponts et chaussées dans ses délibérations des 7 novembre 1878 et 29 janvier 1880. »
(L'article 1er est mis aux voix et adopté.)
« Art. 2. — Un péage dont le produit est garanti, comme il est dit à l'article 3, par les départements des Vosges et de Meurthe-et-Moselle, sera établi sur la voie navigable, au fur et à mesure de la mise en exploitation de ses diverses sections.
« Ce péage se composera, par tonne, d'une constante de 0,30 et d'un élément variable de 0,0005 par kilomètre.
« Il prendra fin vingt ans après l'ouverture du canal sur toute sa longueur. » — (Adopté.)
« Art. 3. — Le ministre des travaux publics est autorisé à accepter, au nom de l'Etat :
« 1° Les subventions souscrites par les industriels intéressés à titre de libération du péage ;
« 2° L'offre faite par les deux départements de Meurthe-et-Moselle et des Vosges, ainsi qu'il résulte des délibérations des conseils généraux en date des 22 et 23 août 1879, de garantir solidairement et à parts égales les subventions précitées ainsi qu'un produit annuel de 140,000 francs pour le péage à partir de la mise en exploitation du canal sur toute sa longueur, sauf déduction, au taux de 4 1/2 p. 100, de l'annuité d'amortissement en vingt ans de ces subventions ;
« 3° Les offres de concours des villes de Lunéville, Baccarat, Raon l'Etape, Saint-Dié et Fraize montant ensemble à la somme de 510,000 francs. » — (Adopté.)
« Art. 4. — La somme de 20 millions de francs à laquelle sont évalués les travaux sera imputée sur les ressources extraordinaires inscrites au budget de chaque année.
« Viendra en déduction de la dépense à la charge du Trésor le montant des subventions mentionnées à l'article précédent. » — (Adopté.)
(L'ensemble du projet de loi est mis aux voix et adopté.)

DISCUSSION D'UN RAPPORT DE LA 2e COMMISSION DES PÉTITIONS

M. le président. L'ordre du jour appelle la discussion des conclusions du rapport fait au nom de la 2e commission des pétitions sur la pétition du sieur Jean Dupuis, citoyen français, demeurant à Han-Kow (Chine).
La parole est à M. le rapporteur.
M. Emile Bouchet, *rapporteur*. Messieurs, sur cette question qui aurait comporté de longs développements, je n'aurai que très-peu de mots à prononcer : je demande simplement à la Chambre de vouloir bien accueillir les conclusions de la commission, tendant au renvoi de la pétition au ministre de la marine, conclusions acceptées par M. le ministre de la marine et par le Gouvernement.
M. de Mahy. Très-bien ! très-bien !
M. La Vieille. Quelles sont ces conclusions ?
M. le président. La commission conclut au renvoi de la pétition au ministre de la marine ; le renvoi est accepté par le Gouvernement.
M. le rapporteur. Les conclusions sont celles de toutes les commissions de pétitions qui, après avoir examiné les faits, concluent, quand elles les trouvent dignes d'une attention particulière, au renvoi de la pétition au ministre compétent. C'est ce qu'a fait la deuxième commission des pétitions.
M. de Larochefoucauld, duc de Bisaccia. Quelle est cette pétition ?
M. le président. Le rapport a été imprimé et distribué, et la discussion est inscrite au feuilleton depuis deux mois.
M. de Larochefoucauld, duc de Bisaccia. Avant de voter, je demande des explications.
M. le président. La commission conclut au renvoi à M. le ministre de la marine.

(a) Image d'origine

Année 1881 DÉBATS ET DOCUMENTS PARLEMENTAIRES 25 Février [CHAMBRE] 345

CHAMBRE DES DÉPUTÉS
Session ordinaire de 1881

COMPTE RENDU IN EXTENSO. — 23e SÉANCE

Séance du jeudi 24 février 1881

SOMMAIRE

Procès-verbal : M. Devès.
Excuse.
Dépôt, par M. Marquiset, du rapport de la commission chargée d'examiner le projet de loi ayant pour objet de faire déclarer d'utilité publique la construction d'un canal de navigation dans la vallée de la Chiers, entre Longwy et la Meuse, à Mouzon.
Dépôt, par M. Morel, du rapport de la 2e commission d'intérêt local, sur le projet de loi tendant à autoriser le département de la Haute-Marne à contracter un emprunt pour l'amélioration de l'asile des aliénés de Saint-Dizier.
Dépôt, par M. Sarrien, du rapport de la commission chargée d'examiner le projet de loi ayant pour objet de faire déclarer d'utilité publique l'amélioration du canal du centre.
Suite de la 1re délibération sur le projet de loi, adopté par le Sénat, concernant l'administration de l'armée : M. le comte de Roys, rapporteur. — Ajournement à samedi.
Adoption en 2e délibération du projet de loi ayant pour objet l'établissement d'un canal de Dombasle à Saint-Dié.
Discussion des conclusions du rapport de la 2e commission des pétitions sur la pétition du sieur Jean Dupuis, citoyen français, demeurant à Han-Kow (Chine) : M. Emile Bouchet, rapporteur. Adoption.
1re délibération sur les propositions de loi : 1° de M. Germain Casse et de plusieurs de ses collègues, tendant à régler les rapports entre les compagnies de chemins de fer et leurs agents commissionnés ; 2° de M. le baron de Janzé, relative à la réglementation des compagnies de chemins de fer : MM. René Brice, le baron de Janzé.
Règlement de l'ordre du jour : MM. Latrade, Sarrette, de Gasté, Bardoux.
Reprise de la discussion des propositions tendant à régler les rapports entre les compagnies de chemins de fer et leurs agents commissionnés : M. Trarieux.
Présentation, par M. le ministre de l'intérieur et des cultes, d'un projet de loi tendant à autoriser le département des Côtes-du-Nord à contracter deux emprunts pour les travaux des chemins vicinaux ordinaires.
Demande, par M. le ministre de la marine, de discussion immédiate du projet de loi, adopté par la Chambre des députés, adopté avec modifications par le Sénat, ayant pour objet d'ouvrir au ministre de la marine et des colonies un crédit de 8,352,751 fr., sur l'exercice 1881, pour les chemins de fer du Haut-Sénégal. Adoption. = Article 1er : M. Des Rotours. Adoption. = Adoption de l'article 2. = Adoption, au scrutin, de l'ensemble du projet de loi.
Règlement de l'ordre du jour : MM. le président, de Gasté.
Présentation par M. le sous-secrétaire d'Etat des travaux publics :
1° D'un projet de loi ayant pour objet : 1° L'incorporation dans le réseau d'intérêt général du chemin de fer d'intérêt local de Mézidon à Dives ; 2° l'approbation d'une convention passée avec la compagnie des chemins de fer de la vallée d'Auge pour le rachat de ladite ligne ;
2° D'un projet de loi ayant pour objet : 1° De déclarer d'utilité publique l'établissement, dans le département des Landes, de plusieurs chemins de fer d'intérêt local et d'en autoriser l'exécution ; 2° d'approuver un traité passé entre la compagnie du Midi et les concessionnaires de ces chemins.
Dépôt par M. Cavalié d'un rapport de la 2e commission d'intérêt local sur le projet de loi tendant à autoriser le département de la Haute-Saône à contracter un emprunt pour les travaux des chemins vicinaux.

PRÉSIDENCE DE M. GAMBETTA

La séance est ouverte à deux heures.
M. Frémiet, *l'un des secrétaires*, donne lecture du procès-verbal de la séance de mardi dernier.
M. le président. La parole est à M. Devès sur le procès-verbal.
M. Paul Devès. Messieurs, MM. Tassin, de Sonnier et Deniau auraient voté contre la proposition d'enquête faite par M. Lorgé s'ils eussent assisté à la séance de mardi. Je le déclare en leur nom.
M. le président. Il n'y a pas d'autre observation ?...
Le procès-verbal est adopté.

EXCUSE

M. le président. M. Dremel s'excuse de ne pouvoir assister à la séance de ce jour.

DÉPÔT DE RAPPORTS

M. le président. Je donne la parole à ceux de MM. les députés qui ont des rapports à déposer.
M. Sarrien. Au nom de la commission des voies navigables, j'ai l'honneur de déposer sur le bureau de la Chambre un rapport sur le projet de loi ayant pour objet de faire déclarer d'utilité publique l'amélioration du canal du Centre.
M. Marquiset. J'ai l'honneur de déposer sur le bureau de la Chambre, au nom de la commission des voies navigables, un rapport sur le projet de loi ayant pour objet de faire déclarer d'utilité publique la construction d'un canal de navigation dans la vallée de la Chiers, entre Longwy et la Meuse, à Mouzon.
M. Morel. Au nom de la 2e commission d'intérêt local, j'ai l'honneur de déposer sur le bureau de la Chambre un rapport sur le projet de loi tendant à autoriser le département de la Haute-Marne à contracter un emprunt pour l'amélioration de l'asile des aliénés de Saint-Dizier.
M. le président. Ces rapports seront imprimés et distribués.

AJOURNEMENT DE LA SUITE DE LA DISCUSSION DU PROJET DE LOI SUR L'ADMINISTRATION DE L'ARMÉE.

M. le président. L'ordre du jour appelle la suite de la première délibération sur le projet de loi, adopté par le Sénat, concernant l'administration de l'armée.
M. le rapporteur a la parole.
M. le comte de Roys, *rapporteur*. Messieurs, M. le ministre de la guerre étant appelé aujourd'hui au Sénat par une interpellation, la commission vous prie de vouloir bien ajourner à la séance de samedi la suite de la discussion sur l'administration de l'armée, en lui maintenant son rang en tête de l'ordre du jour de cette séance.
M. le président. Il n'y a pas d'opposition ?...
L'ajournement à samedi est prononcé.

DISCUSSION DU PROJET DE LOI CONCERNANT LE CANAL DE DOMBASLE A SAINT DIÉ

M. le président. L'ordre du jour appelle la deuxième délibération sur le projet de loi ayant pour objet l'établissement d'un canal de Dombasle à Saint-Dié.
Art. 1er. — Sont déclarés d'utilité publique les travaux à faire pour la construction du canal latéral à la Meurthe, de Dombasle à Saint-Dié, et son raccordement par voie ferrée avec la gare du chemin de fer de l'Est à Saint-Dié, conformément aux dispositions de l'avant-projet adopté par le conseil général des ponts et chaussées dans ses délibérations des 7 novembre 1878 et 29 janvier 1880. »
(L'article 1er est mis aux voix et adopté.)
« Art. 2. — Un péage dont le produit est garanti, comme il est dit à l'article 3, par les départements des Vosges et de Meurthe-et-Moselle, sera établi sur la voie navigable, au fur et à mesure de la mise en exploitation de ses diverses sections.
« Ce péage se composera, par tonne, d'une constante de 0,30 et d'un élément variable de 0,0005 par kilomètre.
« Il prendra fin vingt ans après l'ouverture du canal sur toute sa longueur. » — (Adopté.)
« Art. 3. — Le ministre des travaux publics est autorisé à accepter, au nom de l'Etat :
« 1° Les subventions souscrites par les industriels intéressés à titre de libération du péage ;
« 2° L'offre faite par les deux départements de Meurthe-et-Moselle et des Vosges, ainsi qu'il résulte des délibérations des conseils généraux en date des 22 et 23 août 1879, de garantir solidairement et à parts égales les subventions précitées ainsi qu'un produit annuel de 140,000 francs pour le péage à partir de la mise en exploitation du canal sur toute sa longueur, sauf déduction, au taux de 4 1/2 p. 100, de l'annuité d'amortissement en vingt ans de ces subventions ;
« 3° Les offres de concours des villes de Lunéville, Baccarat, Raon l'Etape, Saint-Dié et Fraize montant ensemble à la somme de 510,000 francs. » — (Adopté.)
« Art. 4. — La somme de 20 millions de francs à laquelle sont évalués les travaux sera imputée sur les ressources extraordinaires inscrites au budget de chaque année.
« Viendra en déduction de la dépense à la charge du Trésor le montant des subventions mentionnées à l'article précédent. » — (Adopté.)
(L'ensemble du projet de loi est mis aux voix et adopté.)

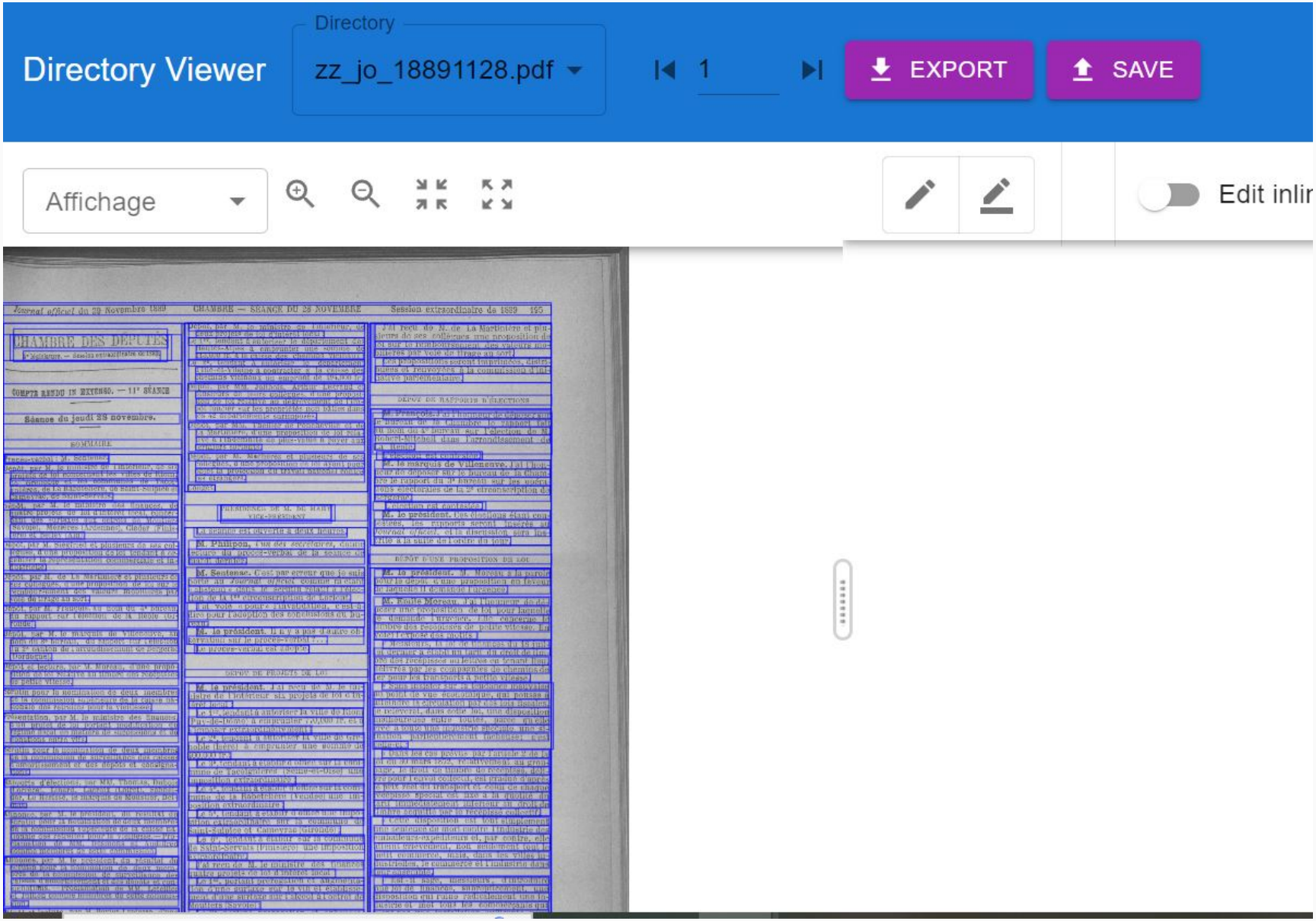
DISCUSSION D'UN RAPPORT DE LA 2e COMMISSION DES PÉTITIONS

M. le président. L'ordre du jour appelle la discussion des conclusions du rapport fait au nom de la 2e commission des pétitions sur la pétition du sieur Jean Dupuis, citoyen français, demeurant à Han-Kow (Chine).
La parole est à M. le rapporteur.
M. Emile Bouchet, *rapporteur*. Messieurs, sur cette question qui aurait comporté de longs développements, je n'aurai que très-peu de mots à prononcer : je demande simplement à la Chambre de vouloir bien accueillir les conclusions de la commission, tendant au renvoi de la pétition au ministre de la marine, conclusions acceptées par M. le ministre de la marine et par le Gouvernement.
M. de Mahy. Très-bien ! très-bien !
M. La Vieille. Quelles sont ces conclusions ?
M. le président. La commission conclut au renvoi de la pétition au ministre de la marine ; le renvoi est accepté par le Gouvernement.
M. le rapporteur. Les conclusions sont celles de toutes les commissions de pétitions qui, après avoir examiné les faits, concluent, quand elles les trouvent dignes d'une attention particulière, au renvoi de la pétition au ministre compétent. C'est ce qu'a fait la deuxième commission des pétitions.
M. de Larochefoucauld, duc de Bisaccia. Quelle est cette pétition ?
M. le président. Le rapport a été imprimé et distribué, et la discussion est inscrite au feuilleton depuis deux mois.
M. de Larochefoucauld, duc de Bisaccia. Avant de voter, je demande des explications.
M. le président. La commission conclut au renvoi à M. le ministre de la marine.

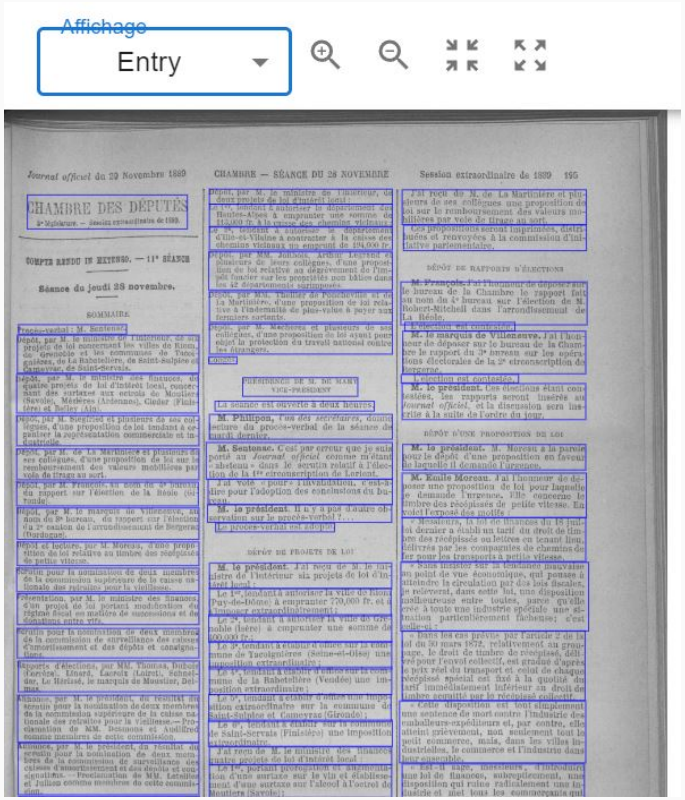
(b) Image nettoyée

FIGURE – Exemple d'image nettoyée

ANALYSE DE LA MISE EN PAGE



ANALYSE DE LA MISE EN PAGE



(a) Zones ”entrées”

II.JPG II.bb

(b) Zones ”Titre II”

FIGURE – Distinction entre différentes ”zones”

RÉSULTATS

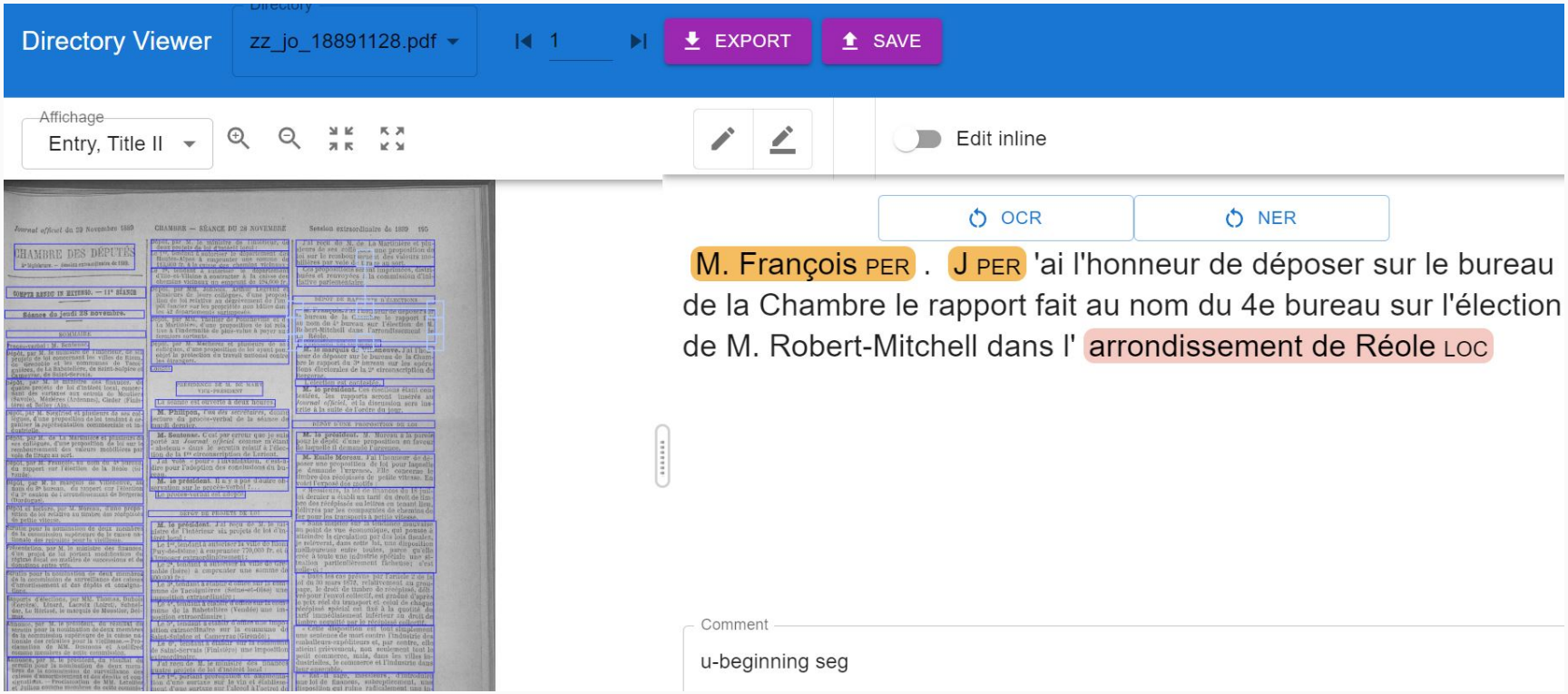


FIGURE – Démonstration de l'outil sur une page de débat parlementaire

PERSPECTIVES

1. Évaluation, analyse comparative
2. Outils prêts pour la production, vraiment en libre accès
3. D'autres méthodes que le NER pour l'extraction de données structurées (entryseg + structured output + reading order + …)
4. Projet Mezanno (financé dans le cadre du plan quadriennal de la BnF) : DH sans informaticiens ?

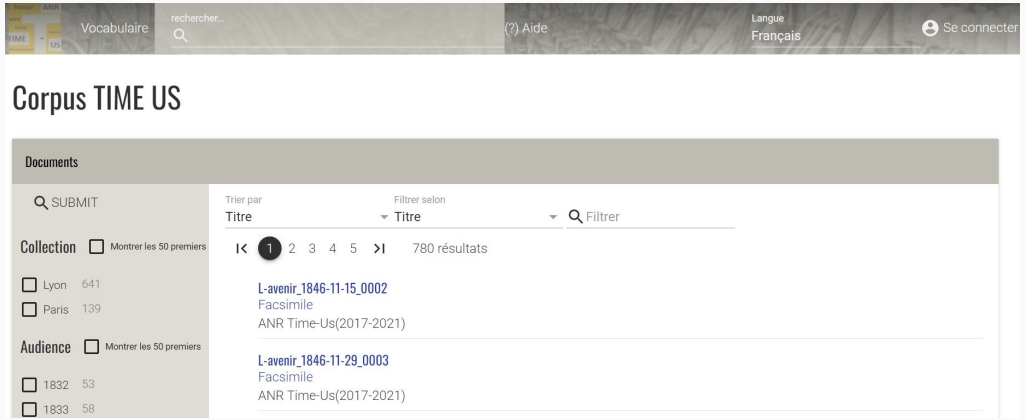
MEZANNO

- Page web sans installation ni serveur
- Interface de constitution de corpus à partir de ressources IIIF publiques
  - Gallica
  - Images sur Nakala, Zenodo…
- Interface de création et modification d'annotations
  - Adaptée à la transcription
  - Stockage local ou serveur d'annotation collaboratif
- Possibilité d'appeler des API distantes standardisées
  - Layout detection
  - OCR
- Facilité d'export des données
  - CSV…

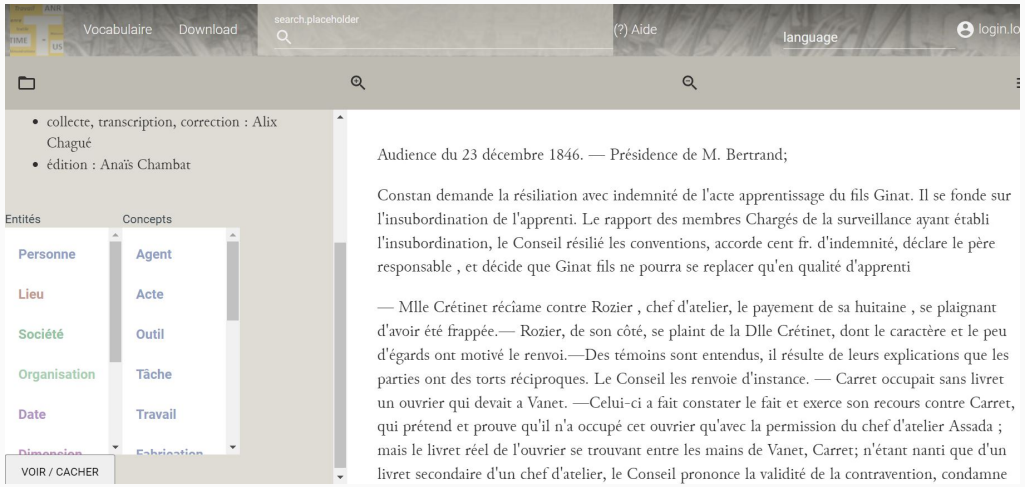
Réutiliser tous les outils existants, et maximiser l'indépendance chercheurs/ses en SHS

Annoter les débats en XML-TEI

PUBLIER UN CORPUS ÉDITORIALISÉ



(a) Interface de recherche



(b) Affichage d'un document

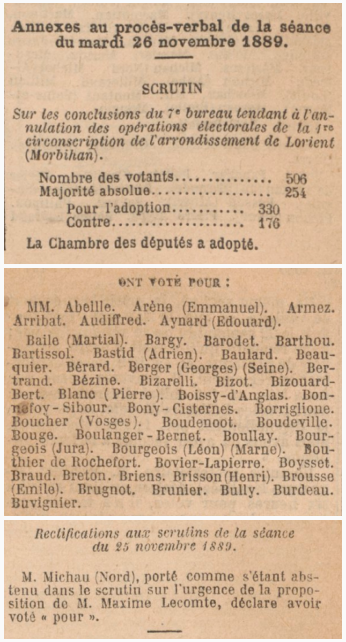
FIGURE – Corpus du projet ANR TIME-US publié avec TEI Publisher

Corps d'un débat encodé en XML-TEI

```
        <div type="part" corresp="#rapportelections">
           <head>SUITE DE LA VÉRIFICATION DES POUVOIRS</head>
           <u who="#pers_ID" xml:id="CR_1889-11-26_u23" ana="#chair">
              <seg xml:id="CR_1889-11-26_u23.1"><persName ref="#pers_ID">M. le <roleName ref="#pers_ID">président</roleName></persName>. L'ordre du jour
appelle la suite de la vérification des pouvoirs.</seg>
              <seg xml:id="CR_1889-11-26_u23.2"><persName ref="#pers_ID">M. Reybert</persName> a la parole pour donner lecture d'un rapport sur une
élection non contestée.</seg>
           </u>

           <u who="#pers_ID" xml:id="CR_1889-11-26_u24" ana="#rapporteur">
              <!-- Lecture d'un rapport -->
              <quote>
                 <seg xml:id="CR_1889-11-26_u24.1"><persName ref="#pers_ID">M. Reybert, <roleName ref="#pers_ID">rapporteur</roleName></persName>.
—<placeName ref="#lieu_ID">Département de la Corrèze, arrondissement de Tulle, <num>1re</num> circonscription</placeName>.</seg>
                 <seg xml:id="CR_1889-11-26_u24.2">Les élections du <date when="1889-09-22">22 septembre</date> ont donné les résultats suivants :</seg>
                 <seg xml:id="CR_1889-11-26_u24.3">Electeurs inscrits, <num>17,950</num>, dont le quart est de <num>4,263</num>.</seg>
                 <seg xml:id="CR_1889-11-26_u24.4">Nombre des votants, <num>12,322</num>.</seg>
                 <seg xml:id="CR_1889-11-26_u24.5">Bulletins blancs et nuls, à déduire, <num>133</num>.</seg>
                 <seg xml:id="CR_1889-11-26_u24.6">Suffrages exprimés, <num>12,189</num>.</seg>
                 <seg xml:id="CR_1889-11-26_u24.7">Ont obtenu : MM. <persName ref="#pers_ID">Borie, <roleName ref="#pers_ID">député
sortant</roleName></persName>... <num>7.508</num> voix</seg>
                 <seg xml:id="CR_1889-11-26_u24.8"><persName ref="#pers_ID">Vachal, <roleName ref="#pers_ID">ancien député</roleName></persName>...
<num>4.748</num></seg>
                 <seg xml:id="CR_1889-11-26_u24.9"><persName ref="#pers_ID">M. Borie</persName> a été proclamé député comme ayant réuni un nombre de voix
au moins égal à la majorité absolue des suffrages</seg>
              </quote>
           </u>

           <floatingText><body><div><pb n="176"/></div></body></floatingText>

           <!-- [...] -->
        </div>
```

Figure – Encodage - Séance parlementaire du 26 novembre 1889 (extrait)

ANNEXES D'UN DÉBAT ENCODÉES EN XML-TEI



Annexes au procès-verbal de la séance du mardi 26 novembre 1889.

SCRUTIN

Sur les conclusions du 7e bureau tendant à l'annulation des opérations électorales de la 1re circonscription de l'arrondissement de Lorient (Morbihan).

Nombre des votants..... 506
Majorité absolue..... 254
Pour l'adoption..... 330
Contre..... 176

La Chambre des députés a adopté.

ONT VOTÉ POUR :

MM. Abeille. Arène (Emmanuel). Armez. Arribat. Audifred. Aynard (Edouard).

Baile (Martial). Bargy. Barodet. Barthou. Bartissol. Bastid (Adrien). Baulard. Beauquier. Bérard. Berger (Georges) (Seine). Bertrand. Bézine. Bizarelli. Bizot. Bizouard-Bert. Blanc (Pierre). Boissy-d'Anglas. Bonnefoy-Sibour. Bony-Cisternes. Borriglione. Boucher (Vosges). Boudenoot. Boudeville. Bouge. Boulanger-Bernet. Boullay. Bourgeois (Jura). Bourgeois (Léon) (Marne). Bouthier de Rochefort. Bovier-Lapierre. Boysset. Braud. Breton. Briens. Brisson (Henri). Brousse (Emile). Brugnot. Brunier. Bully. Burdeau. Buvignier.

Rectifications aux scrutins de la séance du 25 novembre 1889.

M. Michau (Nord), porté comme s'étant abstenu dans le scrutin sur l'urgence de la proposition de M. Maxime Lecomte, déclare avoir voté « pour ».

(a) Source numérisée

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

   <div xml:id="vot18891126">
      <!-- VOTE 1 -->
      <div xml:id="vot18891126_vot1" type="voting" corresp="#discussion7ebureau">
         <head>
            <label>SCRUTIN</label>
            <note><seg>Sur les conclusions du <num>7e</num> bureau tendant à l'annulation des opérations électorales de la
<placeName ref="#lieu_ID"><num>1re</num> circonscription de l'arrondissement de Lorient (Morbihan)</placeName>.</seg></note>
         </head>
         <!-- Détail du vote -->
         <desc>
            <measure type="nbvotants" quantity="506">Nombre des votants <num>506</num></measure>
            <measure type="maj" quantity="254">Majorité absolue <num>254</num></measure>
            <measure type="ayes" quantity="330">Pour l'adoption <num>330</num></measure>
            <measure type="noes" quantity="176">Contre <num>176</num></measure>
         </desc>

         <note type="result"><seg>La <orgName ref="#org_ID">Chambre des députés</orgName> a adopté.</seg></note>

         <floatingText><body><div><pb n="192"/></div></body></floatingText>

         <!-- Liste des votants -->
         <note type="voterslist">
            <desc>Ont voté pour :</desc>
            <seg> MM.<persName ref="#pers_ID">Abeille</persName>. <persName ref="#pers_ID">Arène (Emmanuel)</persName>. <persName
ref="#pers_ID">Armez</persName>. <persName ref="#pers_ID">Arribat</persName>. <persName ref="#pers_ID">Audifred</persName>. <persName ref
="#pers_ID">Aynard (Edouard)</persName>.</seg>
            <!-- [...] -->
         </note>
      </div>

      <!-- RECTIFICATIONS -->
      <div corresp="#vot18891125" type="rectification">
         <head>Rectifications aux scrutins de la séance du <date>25 novembre 1889</date>.</head>
         <note corresp="#vot18891125_vot1"><seg><persName ref="#pers_ID">M. Michau</persName> <placeName
ref="#lieu_ID">(Nord)</placeName>, porté comme s'étant abstenu dans le scrutin sur l'urgence de la proposition de <persName
ref="#pers_ID">M. Maxime Lecomte</persName>, déclare avoir voté « pour ».</seg></note>
         <!-- [...] -->
      </div>
   </div>
</back>
```

(b) Modèle d'encodage

FIGURE – Séance parlementaire du 26 novembre 1889 - votes, liste des votants, rectifications (extrait annexes)

STRUCTURE GÉNÉRALE DES FICHIERS XML TEI

```xml
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0" xml:id="FR_3R_5L" xml:lang="fr">

   <!-- Métadonnées du corpus (non développées) -->
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title></title>
         </titleStmt>
         <publicationStmt>
            <p></p>
         </publicationStmt>
         <sourceDesc>
            <p></p>
         </sourceDesc>
      </fileDesc>
   </teiHeader>

   <!-- Données liées et informations contextuelles -->
   <standOff>
      <listPerson>
         <person></person>
      </listPerson>

      <listOrg>
         <org></org>
      </listOrg>

      <listPlace>
         <place></place>
      </listPlace>
   </standOff>

   <!-- Stockage du composant correspondant à la séance du 26 novembre 1889 -->
   <xi:include xmlns:xi="http://www.w3.org/2001/XInclude" href="FR_3R_5L_1889-11-26.xml"/>


   <!-- Stockage des autres composants du corpus de façon identique -->
   <!-- ... -->

</teiCorpus>
```

(a) Structure générale d'un fichier corpus

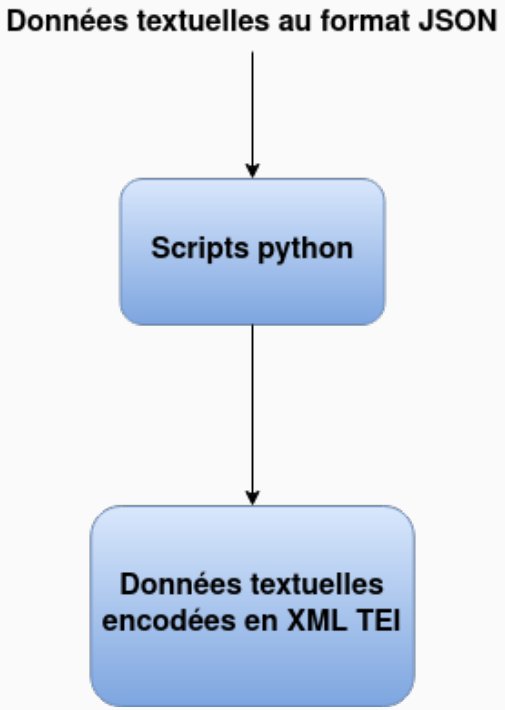
```xml
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="FR_3R_5L_1889-11-26" xml:lang="fr">

   <!-- Métadonnées du composant (non développées) -->
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title></title>
         </titleStmt>
         <publicationStmt>
            <p></p>
         </publicationStmt>
         <sourceDesc>
            <p></p>
         </sourceDesc>
      </fileDesc>
   </teiHeader>

   <text>

      <!-- Transcription du compte rendu -->
      <body>
         <div></div>
      </body>

      <!-- Annexes du compte rendu -->
      <back></back>

   </text>

</TEI>
```

(b) Structure générale d'un fichier composant

FIGURE – Encodages - Structure générale des fichiers corpus et composant

APPLIQUER L'ENCODAGE : AUTOMATISATION


```
19:
  activities:        []
  addresses:         []
  box:
    0:               220.12998972250773
    1:               282.1299897225077
    2:               561.8527187504491
    3:               80.4418437486525
  checked:           true
  comment:           "seg"
  id:                276
  ner_xml:           "<PER>M. Borie</PER> a <ACT>déjà fait partie des Assemblées\u2029législatives et satisfait aux conditions d</ACT>'<ACT>âge\u2029et de
                     nationalité</ACT> exigées par la loi.\u2029L'élection de la 2e circonscription de"
  origin:            "computer"
  parent:            261
  persons:
    0:               "M. Borie"
  text_ocr:          "M. Borie a déjà fait partie des Assemblées\nlégislatives et satisfait aux conditions d'âge\net de nationalité exigées par la loi."
  type:              "ENTRY"
```

DE JSON VERS XML-TEI

- Données textuelles dans les fichiers JSON = valeurs des clés « text_ocr »
- Clés contenues dans des objets qui correspondent à la segmentation au niveau paragraphe
- Appliquer des règles de transformation sur les valeurs des clés « text_ocr »

A LA RECHERCHE DE « FEATURES »

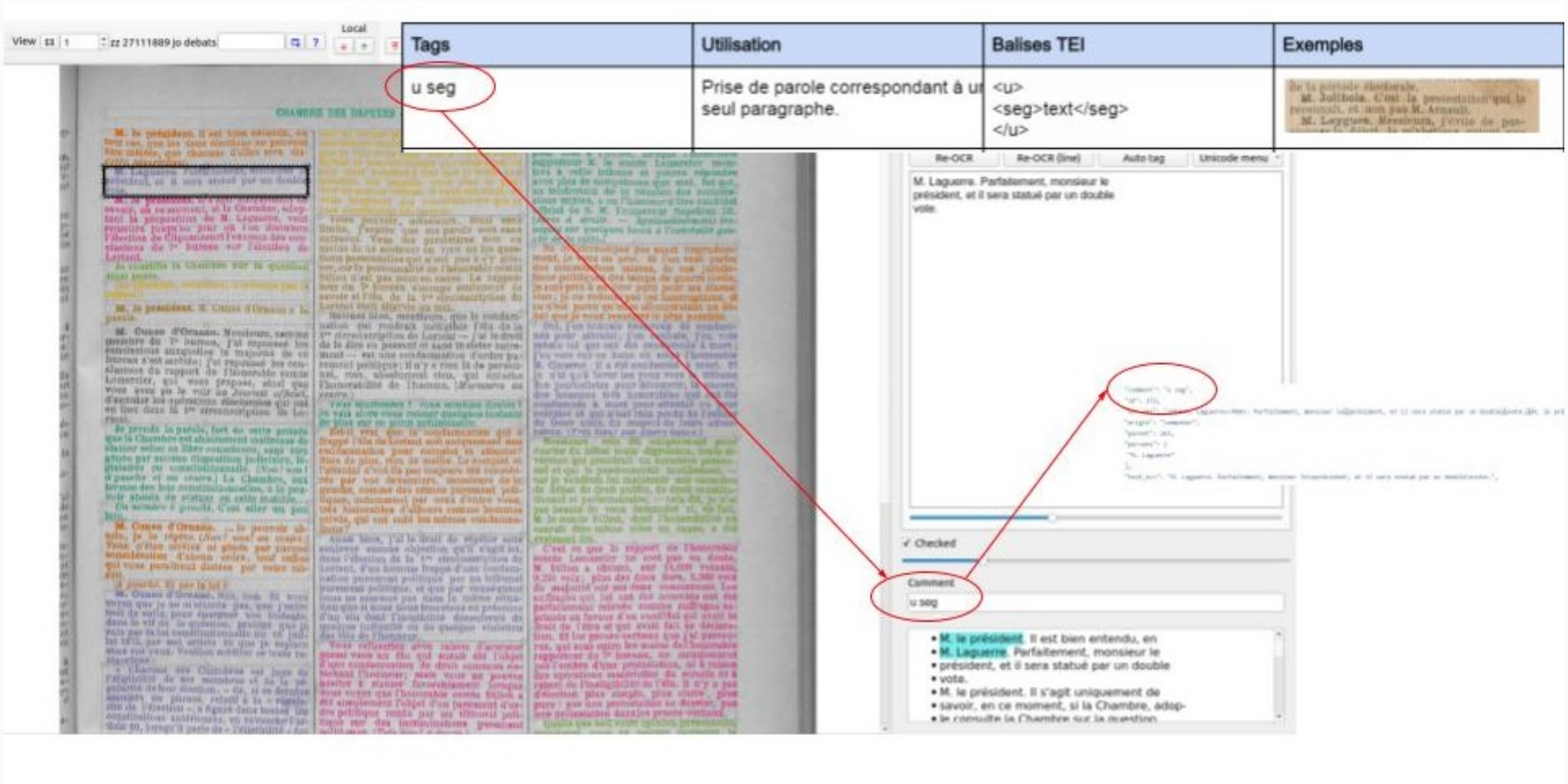
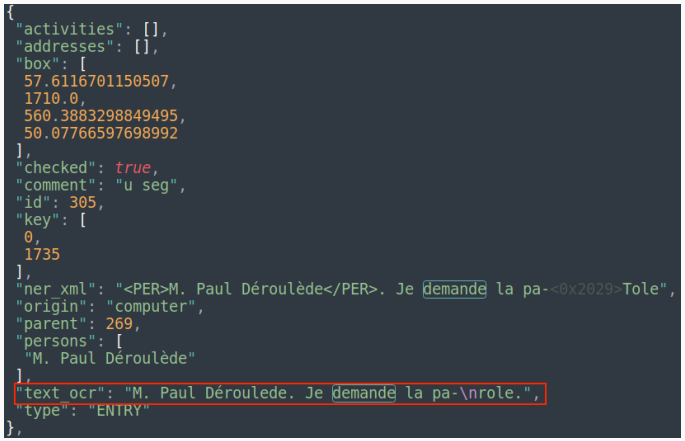


Figure – Du guide d'annotation à l'export JSON

RÉSULTATS



(a) Fichier JSON avec une prise de parole

```python
def add_utterance(data):
    """
    Ajout de l'élément TEI "u" pour chaque boxe étiquetée "u" ou "u-beginning" et "u-end"
    :param data: dictionnaire contenant l'ensemble des données issues des JSON
    """
    for i in range(len(data)):
        if "comment" in data[i]:
            if re.search(r"\bu(?!-)\b", data[i]["comment"]):
                data[i]['text_ocr'] = "".join(['<u>', data[i]['text_ocr'], '</u>'])
            elif re.search(r"u-beginning", data[i]["comment"]):
                data[i]['text_ocr'] = "".join(['<u>', data[i]['text_ocr']])
            elif re.search(r"u-end", data[i]["comment"]):
                data[i]['text_ocr'] = "".join([data[i]['text_ocr'], '</u>'])
            else:
                pass
    return data
```

(b) Fonction add_utterance

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

(c) Résultat

FIGURE – Exemple d'application de la fonction add_utterance

CRÉATION DES FICHIERS COMPOSANTS

1. Récupération des données JSON stockées dans une variable « data » sous forme d'un dictionnaire.
2. Gestion de l'encodage des changements de page.
3. Intégration des scripts contenant les règles de transformation dans la chaîne de traitement. Permettent d'encoder les données contenues dans la variable « data ».
4. Création de l'élément <teiHeader> contenant les métadonnées :
    - Rédaction d'une partie du <teiHeader> à la main contenant les métadonnées « fixes »
    - Construction de règles permettant de rechercher les métadonnées propres à chaque compte rendu, présentes dans la variable « data », par la suite intégrées au <teiHeader> préalablement établi
5. Création des fichiers XML valides.
6. Nettoyage des fichiers XML (gestion des césures).

Création du fichier corpus

1. Génération du fichier corpus.
2. Intégration des métadonnées.
3. Gestion des <xi :include>.

Annotation semi-automatique et export en XML-TEI

1. Annotation manuelle très consommatrice de temps et non exempte d'erreurs
    - Ajout de manière semi-automatique des étiquettes
    - Permettre aux utilisateurs de réutiliser les étiquettes et de créer leurs propres étiquettes (vocabulaire contrôlé ?)
2. Utiliser ces scripts pour proposer un export XML-TEI (outil SoDUCo)
    - Ajouter les entités nommées (<persName>, <orgName>, etc.)

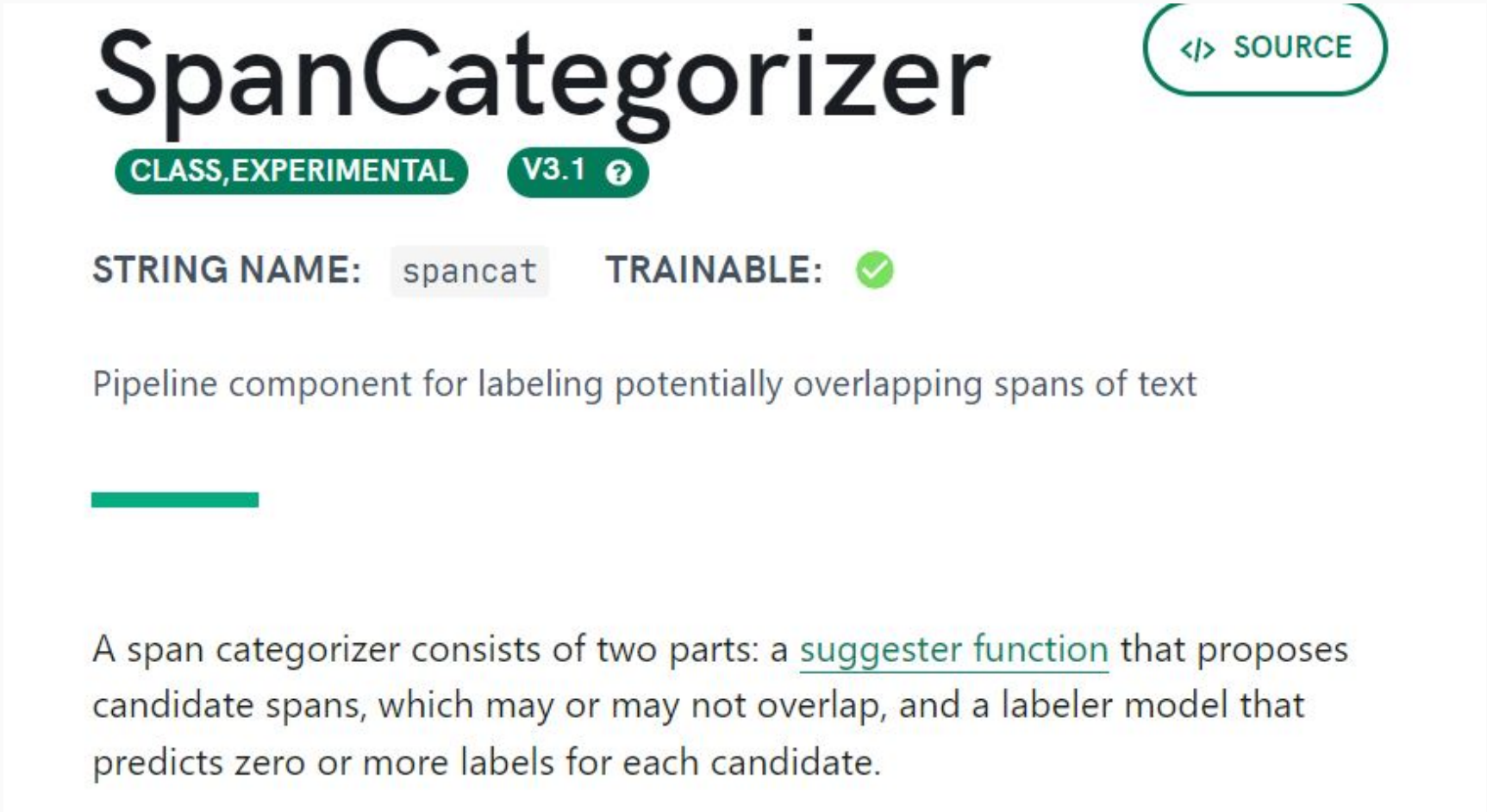
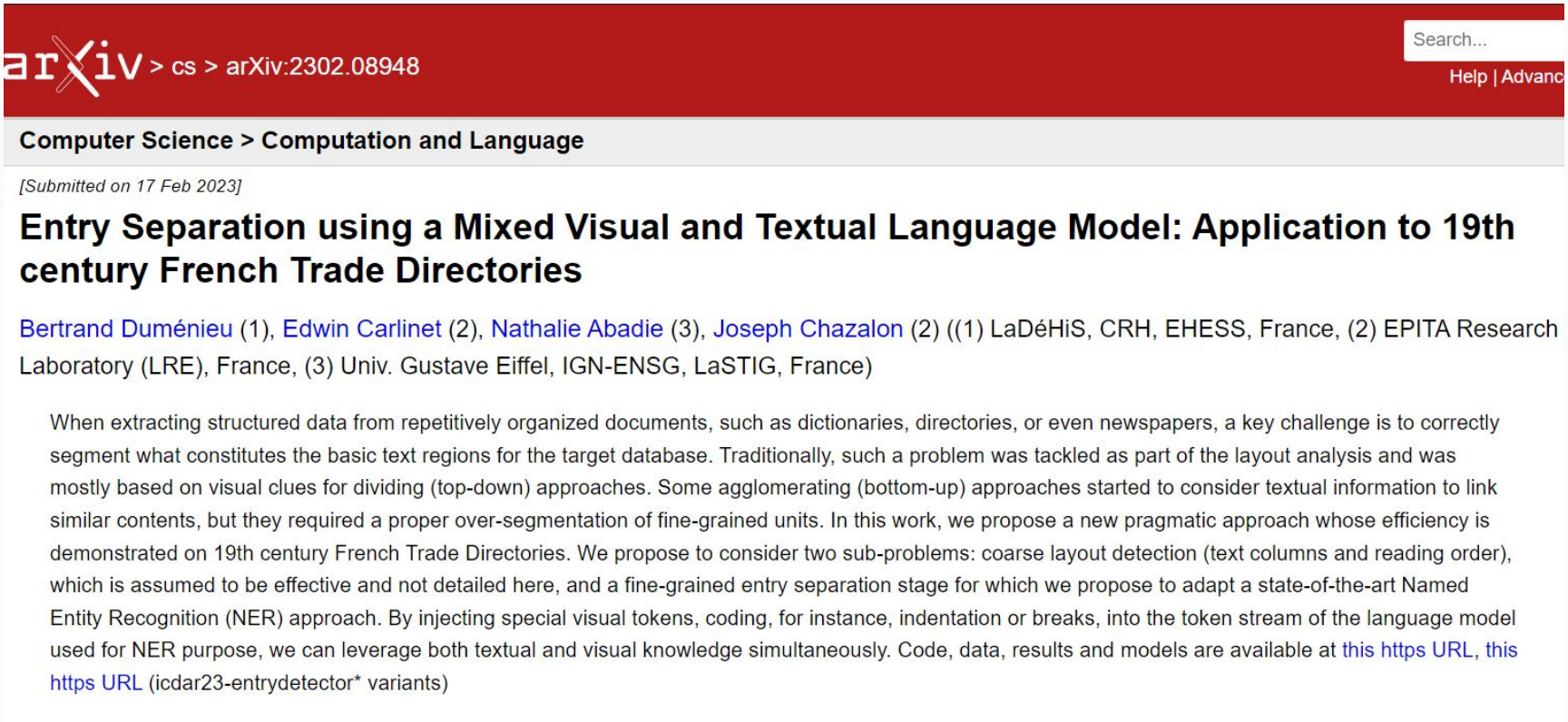


FIGURE – SpanCategorizer : https://spacy.io/api/spancategorizer

AUTOMATISER L'AJOUT DES TAGS



arXiv > cs > arXiv:2302.08948

Search...

Help | Advanc

Computer Science > Computation and Language

[Submitted on 17 Feb 2023]

Entry Separation using a Mixed Visual and Textual Language Model: Application to 19th century French Trade Directories

Bertrand Duménieu (1), Edwin Carlinet (2), Nathalie Abadie (3), Joseph Chazalon (2) ((1) LaDéHiS, CRH, EHESS, France, (2) EPITA Research Laboratory (LRE), France, (3) Univ. Gustave Eiffel, IGN-ENSG, LaSTIG, France)

When extracting structured data from repetitively organized documents, such as dictionaries, directories, or even newspapers, a key challenge is to correctly segment what constitutes the basic text regions for the target database. Traditionally, such a problem was tackled as part of the layout analysis and was mostly based on visual clues for dividing (top-down) approaches. Some agglomerating (bottom-up) approaches started to consider textual information to link similar contents, but they required a proper over-segmentation of fine-grained units. In this work, we propose a new pragmatic approach whose efficiency is demonstrated on 19th century French Trade Directories. We propose to consider two sub-problems: coarse layout detection (text columns and reading order), which is assumed to be effective and not detailed here, and a fine-grained entry separation stage for which we propose to adapt a state-of-the-art Named Entity Recognition (NER) approach. By injecting special visual tokens, coding, for instance, indentation or breaks, into the token stream of the language model used for NER purpose, we can leverage both textual and visual knowledge simultaneously. Code, data, results and models are available at this https URL, this https URL (icdar23-entrydetector* variants)

FIGURE – Modèle de langue basé sur les Transformeurs (BERT) : https ://arxiv.org/abs/2302.08948

TOPIC MODELING ET WORD EMBEDDING

« LIRE À DISTANCE »

- Corpus massif : long à traiter, impossible à lire dans le détail
- Difficile de repérer les grandes tendances

« Lecture distante » (Franco Moretti) comme condition de la connaissance : permet plus facilement de repérer des régularités ou des anomalies qu'une « lecture proche » n'aurait pas détectées.

De nouvelles méthodes d'exploration

- Méthodes d'analyse permettant de « lire à distance » le corpus
  - « Topic modeling » ou modélisation de sujets
  - « Word embedding » ou plongement de mots ou plongement sémantique
- Appliquées au corpus océrisé extrait de Gallica (sans post-correction)

QUELLE MÉTHODE ?

- Écueil de la lecture distante : imposer des catégories de lecture inappropriées
- Choix d'une méthode inductive : modélisation de sujets qui fait émerger les sujets à partir des textes eux-mêmes, sans intervention humaine

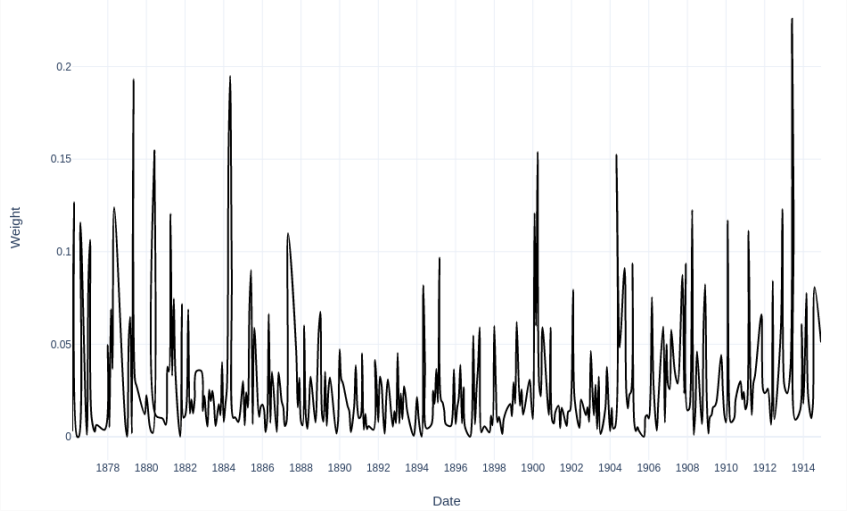
POURQUOI LA MODÉLISATION DE SUJETS ?

Méthode bien adaptée à l'étude de grands corpus historiques

- Particulièrement bien adaptée aux corpus de type « presse » : grand volume, sériels, traversés par de nombreux sujets différents évoluant dans le temps
- Identifier les sujets dans les débats :
  - Nouveau point d'entrée
  - Nouveau « mode de lecture »
  - Meilleure compréhension de l'évolution des idées/ sujets/ débats au cours du temps

MODÉLISATION DE SUJETS

- Méthode d'apprentissage non supervisée permettant de découvrir des sujets abstraits, dits « topics », issus d'un corpus à l'aide d'un modèle probabiliste LDA (Latent Dirichlet Allocation)
- Sujets = champs sémantiques préexistants à l'écriture du texte
- Textes construits à partir de ces champs sémantiques
- Principe de LDA : inverser le processus de génération textuelle, en partant du texte pour extraire les sujets



(a) Variation du poids des sujets liés à l'armée : moyenne par mois

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

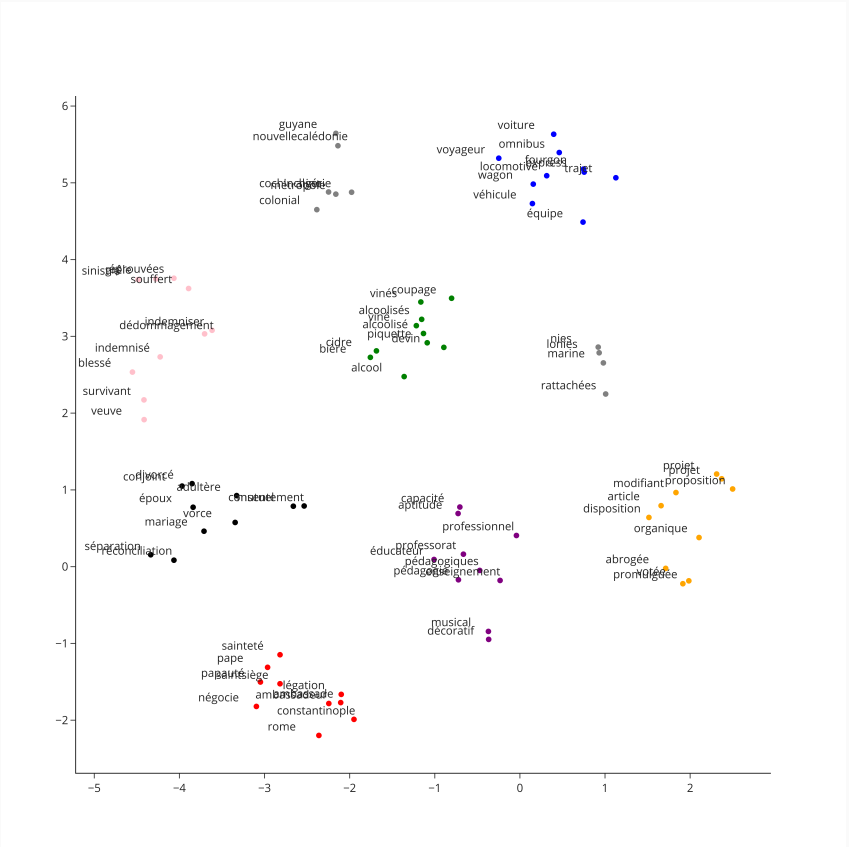
(b) 4 sujets ou « topics »parmi les 50 définis avec LDA

Figure – Résultats obtenus avec LDA

LE PLONGEMENT DE MOTS

- Fournit une représentation des mots dans un espace vectoriel en fonction de leur similarité sémantique
- Repose sur l'hypothèse qu'un mot est caractérisé par son contexte, autrement dit par les mots qui l'entourent
    - Si les mots partagent des contextes similaires, ils partagent alors aussi des significations similaires

LES PLONGEMENTS DE MOTS : WORD2VEC ET TOP2VEC



(a) Projection t-SNE des centroïdes des vecteurs (word2vec)

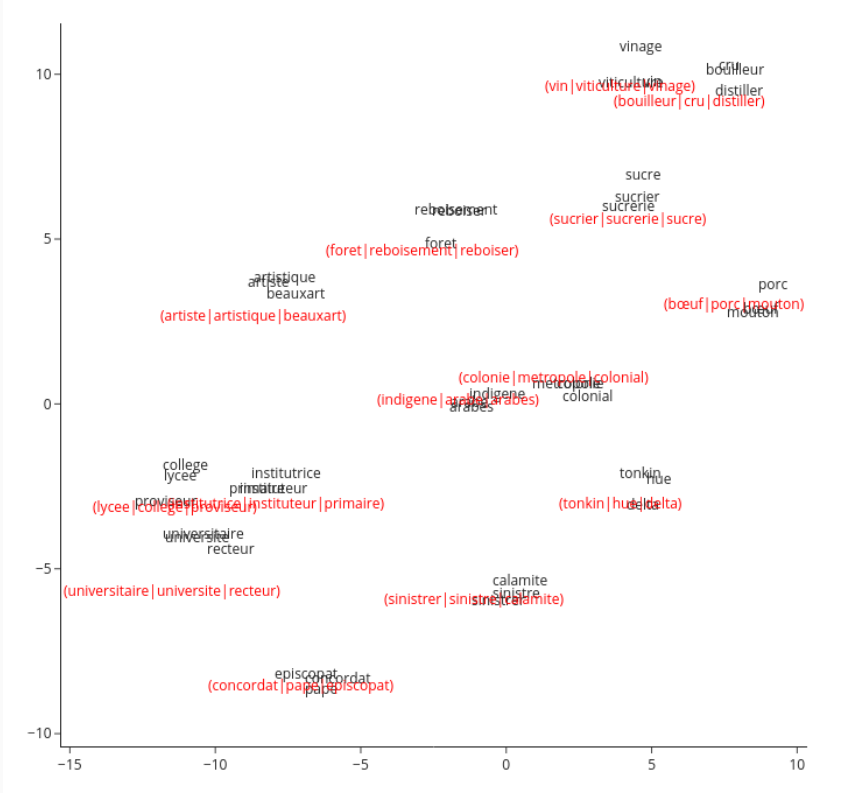
Cluster 55	Cluster 68	Cluster 70
victimes	divorce	enveloppes
inondations	epoux	timbres
secourir	mariage	poste
eprouvees	conjugal	postale
orages	divorces	timbre
sinistres	adultere	recepisses
grele	conjugale	postes
secours	remarier	postaux
venir	separation	telegraphes
infortunes	indissolubilite	colis
ravages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

(b) 3 clusters parmi les 113 obtenus avec top2vec : tempêtes (55), divorce (68) et poste (70)

FIGURE – Résultats obtenus avec word2vec et top2vec

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

(a) 4 sujets obtenus avec Top2vec



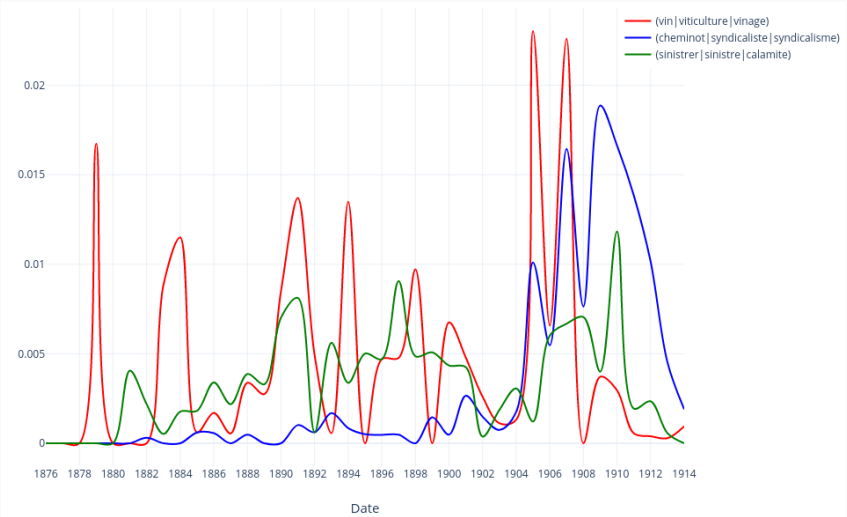
(b) Projection des mots par sujets avec la méthode t-SNE

FIGURE – Résultats obtenus avec Top2vec

Méthode « basée » sur Word2vec et Doc2vec. Nombre de sujets obtenus : 375.

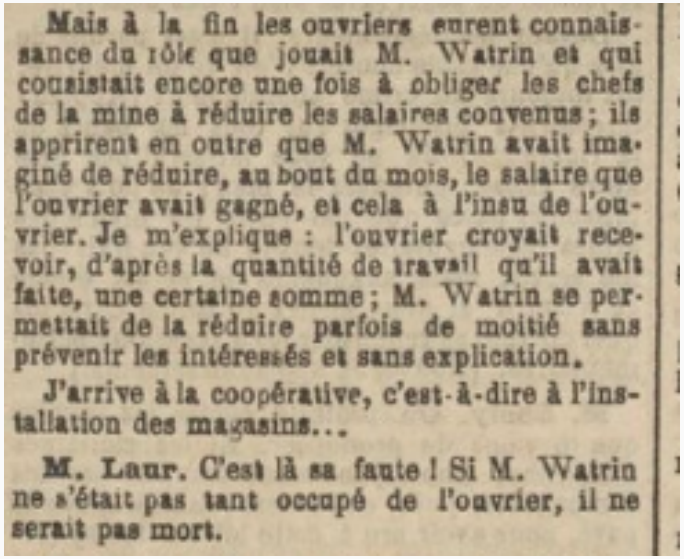
Topic 34	Topic 45	Topic 55
Vin	Cheminot	Sinistré
Viticulture	Syndicaliste	Calamité
Alcoolisation	Confédération	Cyclone
Coupage	Militant	Gelée

(a) 3 sujets et leurs 4 mots les plus représentatifs



(b) Proportion des documents associés à chacun des trois sujets, agrégation par années

FIGURE – Evolution de sujets au cours du temps

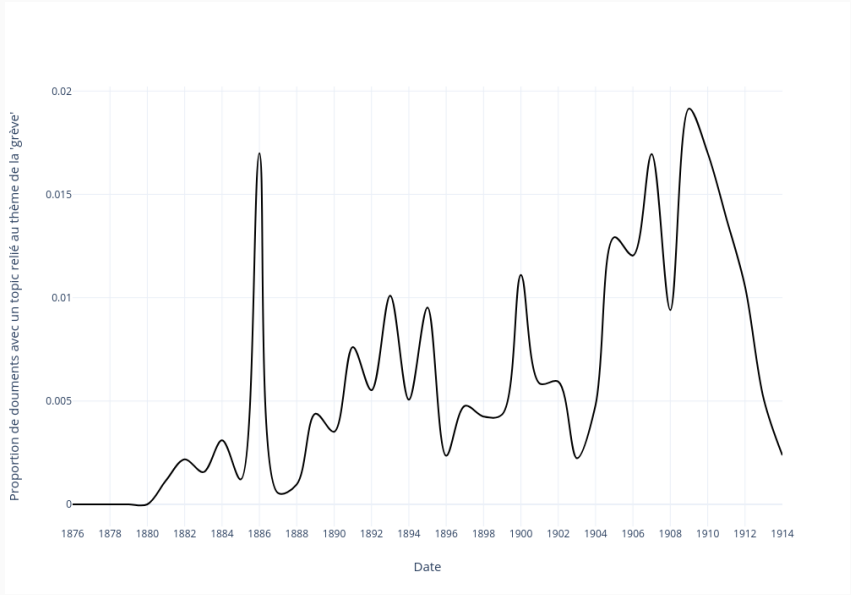


Mais à la fin les ouvriers eurent connaissance du rôle que jouait M. Watrin et qui consistait encore une fois à obliger les chefs de la mine à réduire les salaires convenus ; ils apprirent en outre que M. Watrin avait imaginé de réduire, au bout du mois, le salaire que l'ouvrier avait gagné, et cela 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.

(a) Extrait d'un document ayant une forte similarité avec le sujet de la grève pendant l'année 1886



(b) Proportion des documents associés aux topics sémantiquement similaires au mot "grève", agrégation par années

Figure – Evolution de sujets liés au champ sémantique de la grève dans le temps

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

Figure – Les 4 sujets obtenus avec CamemBERT

MODÈLES DE LANGUE ET BERTOPIC

(a) Huit sujets obtenus avec BERTopic, et les cinq mots les plus représentatifs

(b) Projection des huit sujets avec t-SNE

FIGURE – 2ème itération avec `paraphrase-multilingual-MiniLM-L12-v2`

COMPARAISON DES TROIS MÉTHODES

- Associer LDA et Top2vec :
  - LDA donne une vision plus générale du corpus ;
  - Top2vec permet d'approfondir certains sujets.
- BERTopic : résultats assez décevants :
  - Travailler sur un corpus plus « propre » (ré-océrisé) ;
  - Utiliser un autre modèle de langue développé pour le français comme FlauBERT.

Intégrer ces résultats aux fichiers TEI

ANNOTER LES DÉBATS AVEC LEURS SUJETS

```
<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>
```

(a) Annotations dans le corps de texte

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

(b) Utilisation de l'élément `<standOff>`

Figure – Traitement des sujets comme une annotation linguistique dans le fichier XML-TEI

PERSPECTIVES

- Travailler sur des textes moins fautifs
- Extraire les sujets
  - LDA et Top2vec : complémentaires
  - Nouveau modèle de langue : FlauBERT ou spécialement développé pour le language politique XIXe-XXe siècle
- Annoter automatiquement les textes avec leurs sujets ? Quelle stratégie ?

CONCLUSION

- Travailler sur des textes moins fautifs : nouvel OCR
- Extraire les sujets
  - LDA et Top2vec : complémentaires
  - Nouveau modèle de langue : FlauBERT ou spécialement développé pour le language politique XIXe-XXe siècle
- Annoter automatiquement les textes avec leurs sujets ? Quelle stratégie ?

MERCI POUR VOTRE ATTENTION !

Marie Puren : marie.puren@epita.fr