ANALYSIS OF THE FRENCH PARLIAMENTARY DEBATES OF THE THIRD REPUBLIC WITH TOPIC MODELLING AND WORD EMBEDDING

Methodological Challenges and First Results

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

19-20 mai 2022

Journées MASHS-19-20 mai

PLAN

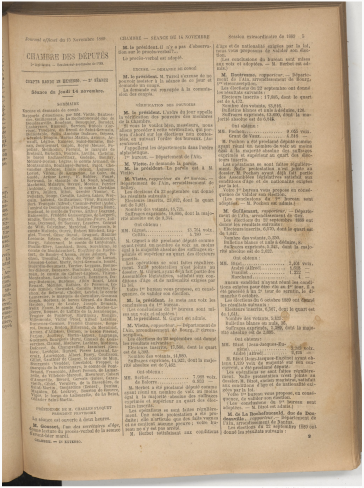
1. Context : the AGODA project
2. Ocerisation of digitised documents
3. Topic modelling with LDA
4. Word, Document and Topic embedding

CONTEXT : THE AGODA PROJECT

THE AGODA PROJECT

- AGODA : **A**nalyse sémantique et **G**raphes relationnels pour l'**O**uverture et l'étude des **D**ébats à l'**A**ssemblée nationale
- Project funded by the Bibliothèque nationale de France for one year
- One of the five pilot projects supported by the DataLab of the National Library of France
- Collaboration beetween Epitech (MNSHS), Inria (ALMAnaCH) and l'Université Lumière Lyon 2 (LARHRA).

FRENCH PARLIAMENTARY DEBATES OF THE THIRD REPUBLIC



Source gallica.bnf.fr / Bibliothèque et Archives de l'Assemblée nationale

Figure – Session of November 14, 1889

- The debates in the lower house of the French Parliament were published in the Journal officiel de la République française. Débats parlementaires (1881-1940)
- The issues of the Journal Officiel are available on Gallica (bibliothèque numérique de la Bibliothèque nationale de France)
- Difficult to work on this corpus, yet interesting for various discipline (history, sociology, political science, linguistics)

GOALS

- Facilitate access to French parliamentary debates of the Third Republic
- Facilitate research in this corpus
- Allowing the constitution of sub-corpora
- Providing new ways of viewing the documents

RESULTS EXPECTED

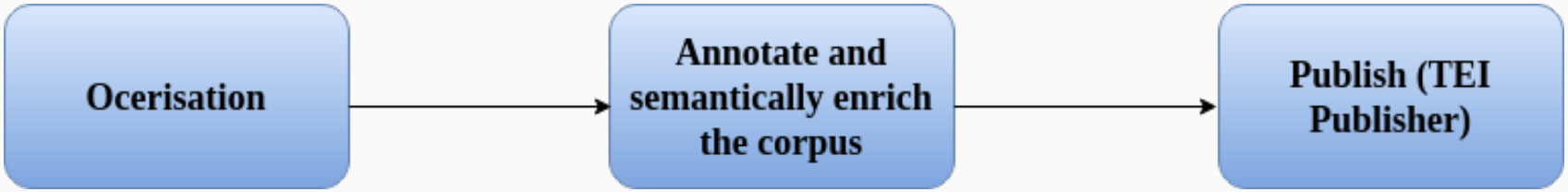
- Creating an online platform for consulting and exploring parliamentary debates
- Producing structured and semantically enriched textual data
- Contribute to the design of a workflow adapted to the analysis of high number of historical documents

PROOF OF CONCEPT

Processing a subpart of the corpus : legislature 1889-1893 : **10418 images to process**

- Partial renewal of the political staff (boulangisme and Panama scandal)
- First signs of Catholic's rallying to the Republic
- Turnaround in customs policy (Méline laws)
- Rise of socialism and trade unionism (Fourmies)
- First anarchist terrorist attacks

WORKFLOW


Ocerisation
Annotate and semantically enrich the corpus
Publish (TEI Publisher)

OCERISATION OF DIGITISED DOCUMENTS

OCERISATION

Mass digitization of texts : how to access/ process and analyze their content ?

- OCR : Optical Character Recognition
- Processing of an image (digitized text) by an OCR engine
- Use of AI : « translation » of the image into text
  - Pages are digitised, then transformed into letters and « discreet »words » (countable : discontinuous, separate, distinct)
  - Possible to explore these pages and « differentiate »them by period, genre, language, year of publication, etc.

FRENCH PARLIAMENTARY DEBATES

- Retrieval of ocerised texts via an API of Gallica, a certain number of erros are identified
- Erros tue to the curvature of the page at the binding + stains, shadows...etc

Figure – 1st evaluation of the Gallica's OCR quality

SOME SOLUTIONS

Improving OCR quality with « dewarping »methods => Inconclusive results

- Dewarping of pages ?
- More advanced cleaning Tools ?
  - Tesseract documentation provides a lot of tools to clean the images
  - Tools developed by other teams
- Bench marking different OCR engines

SODUCO CLEANING TOOL

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 25e 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, Secrétan, 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 25e 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. Drumel 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 la construction 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 25e 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) Original

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 25e 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, Secrétan, 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 25e 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. Drumel 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 la construction 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 25e 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

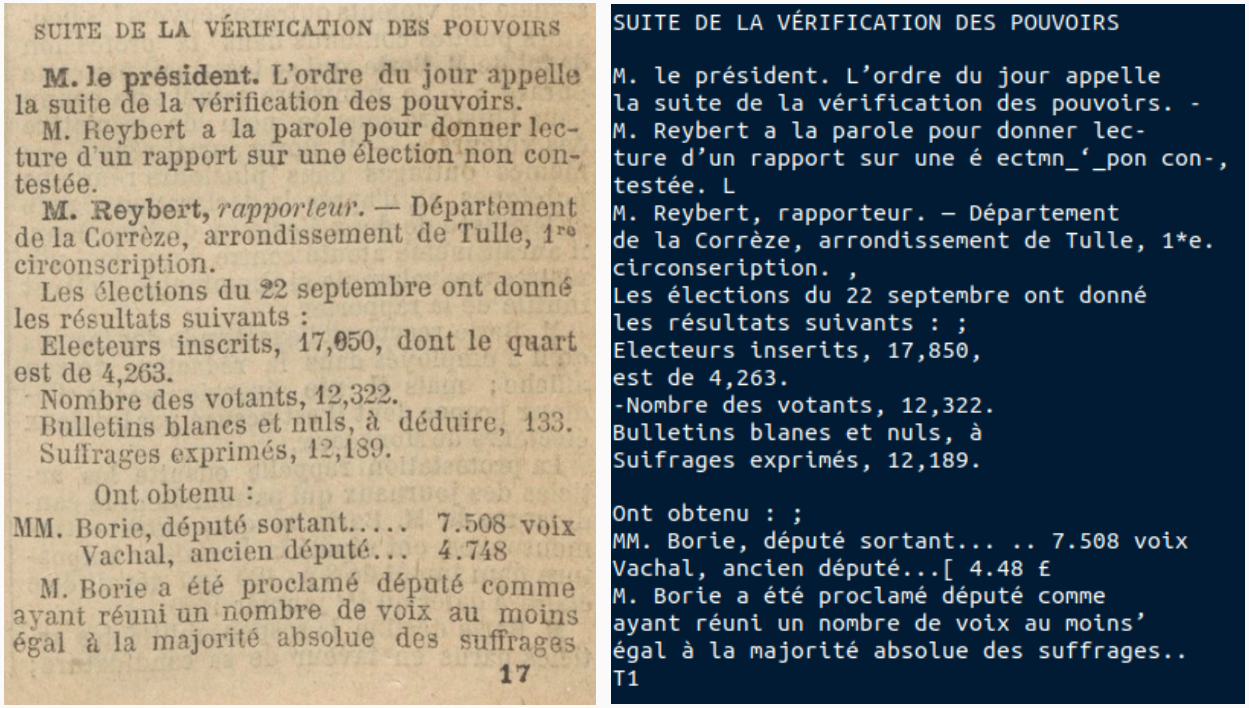
M. le président. L'ordre du jour appelle la discussion des conclusions du rapport fait au nom de la 2e commission des pétitions sur la pétition du sieur Jean Dupuis, citoyen français, demeurant à Han-Kow (Chine).
La parole est à M. le rapporteur.
M. Emile Bouchet, *rapporteur*. Messieurs, sur cette question 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) (pre)Processed

Figure – SODUCO tool applied to one page of debate

Selecting and comparing multiple OCR engines :

- Tesseract (ocr-tesseract or pytesseract)
- ABBYY FineReader



SUITE DE LA VÉRIFICATION DES POUVOIRS

M. le président. L'ordre du jour appelle la suite de la vérification des pouvoirs.

M. Reybert a la parole pour donner lecture d'un rapport sur une élection non contestée.

M. Reybert, *rapporteur.* — Département de la Corrèze, arrondissement de Tulle, 1re circonscription.

Les élections du 22 septembre ont donné les résultats suivants :

Electeurs inscrits, 17,050, dont le quart est de 4,263.

Nombre des votants, 12,322.

Bulletins blancs et nuls, à déduire, 133.

Suffrages exprimés, 12,189.

Ont obtenu :

MM. Borie, député sortant..... 7.508 voix
Vachal, ancien député... 4.748

M. Borie a été proclamé député comme ayant réuni un nombre de voix au moins égal à la majorité absolue des suffrages

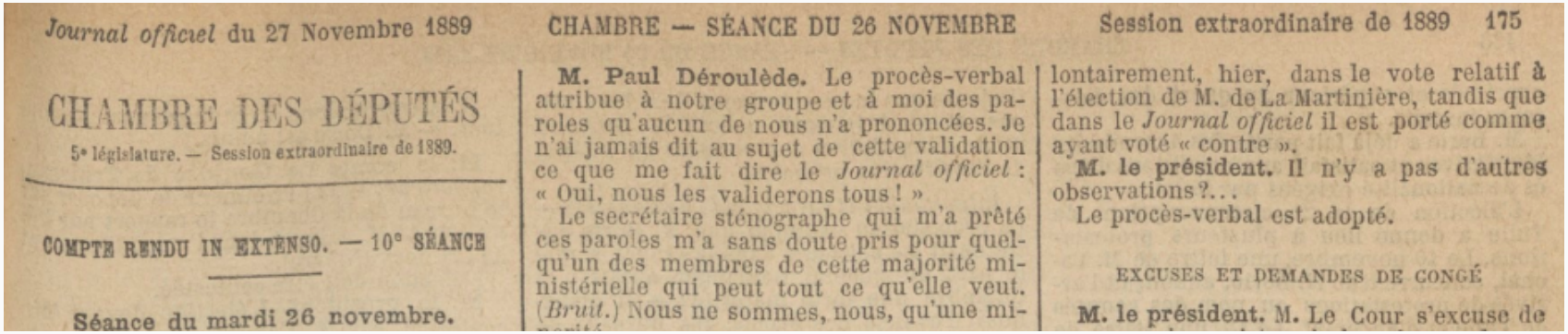
17

```
SUITE DE LA VÉRIFICATION DES POUVOIRS

M. le président. L'ordre du jour appelle
la suite de la vérification des pouvoirs. -
M. Reybert a la parole pour donner lec-
ture d'un rapport sur une é ectmn_'_pon con-,
testée. L
M. Reybert, rapporteur. — Département
de la Corrèze, arrondissement de Tulle, 1*e.
circonseription. ,
Les élections du 22 septembre ont donné
les résultats suivants : ;
Electeurs inserits, 17,850,
est de 4,263.
-Nombre des votants, 12,322.
Bulletins blanes et nuls, à
Suifrages exprimés, 12,189.

Ont obtenu : ;
MM. Borie, député sortant... .. 7.508 voix
Vachal, ancien député...[ 4.48 £
M. Borie a été proclamé député comme
ayant réuni un nombre de voix au moins'
égal à la majorité absolue des suffrages..
T1
```

(a) original (zoom) **(b)** OCR (zoom)

Figure – Zoom in on a text block + OCR (tesseract)

SOME SOLUTIONS



Journal officiel du 27 Novembre 1889

CHAMBRE DES DÉPUTÉS

5e législature. — Session extraordinaire de 1889.

COMPTE RENDU IN EXTENSO. — 10e SÉANCE

Séance du mardi 26 novembre.

CHAMBRE — SÉANCE DU 26 NOVEMBRE

M. Paul Déroulède. Le procès-verbal attribue à notre groupe et à moi des paroles qu'aucun de nous n'a prononcées. Je n'ai jamais dit au sujet de cette validation ce que me fait dire le *Journal officiel* : « Oui, nous les validerons tous ! »

Le secrétaire sténographe qui m'a prêté ces paroles m'a sans doute pris pour quelqu'un des membres de cette majorité ministérielle qui peut tout ce qu'elle veut. (*Bruit.*) Nous ne sommes, nous, qu'une mi-

Session extraordinaire de 1889 175

lontairement, hier, dans le vote relatif à l'élection de M. de La Martinière, tandis que dans le *Journal officiel* il est porté comme ayant voté « contre ».

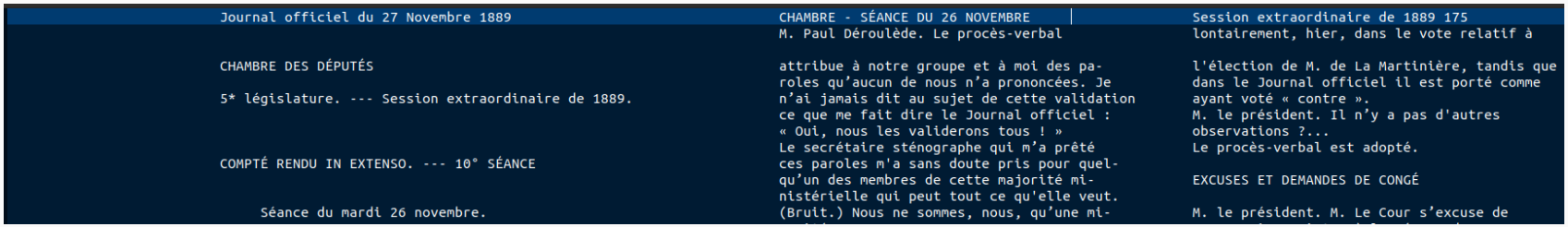
M. le président. Il n'y a pas d'autres observations ?...

Le procès-verbal est adopté.

EXCUSES ET DEMANDES DE CONGÉ

M. le président. M. Le Cour s'excuse de

(a) original (zoom)



Journal officiel du 27 Novembre 1889

CHAMBRE DES DÉPUTÉS

5* législature. --- Session extraordinaire de 1889.

COMPTÉ RENDU IN EXTENSO. --- 10° SÉANCE

Séance du mardi 26 novembre.

CHAMBRE - SÉANCE DU 26 NOVEMBRE

M. Paul Déroulède. Le procès-verbal attribue à notre groupe et à moi des paroles qu'aucun de nous n'a prononcées. Je n'ai jamais dit au sujet de cette validation ce que me fait dire le Journal officiel : « Oui, nous les validerons tous ! »

Le secrétaire sténographe qui m'a prêté ces paroles m'a sans doute pris pour quelqu'un des membres de cette majorité ministérielle qui peut tout ce qu'elle veut. (Bruit.) Nous ne sommes, nous, qu'une mi-

Session extraordinaire de 1889 175

lontairement, hier, dans le vote relatif à l'élection de M. de La Martinière, tandis que dans le Journal officiel il est porté comme ayant voté « contre ».

M. le président. Il n'y a pas d'autres observations ?...

Le procès-verbal est adopté.

EXCUSES ET DEMANDES DE CONGÉ

M. le président. M. Le Cour s'excuse de

(b) OCR (zoom)

Figure – Zoom in on a text block + OCR (ABBYY)

TOOL DEVELOPED BY LRDE (BASED THE OCR ENGINE : PERO OCR)

PERO OCR : very efficient of historical sources

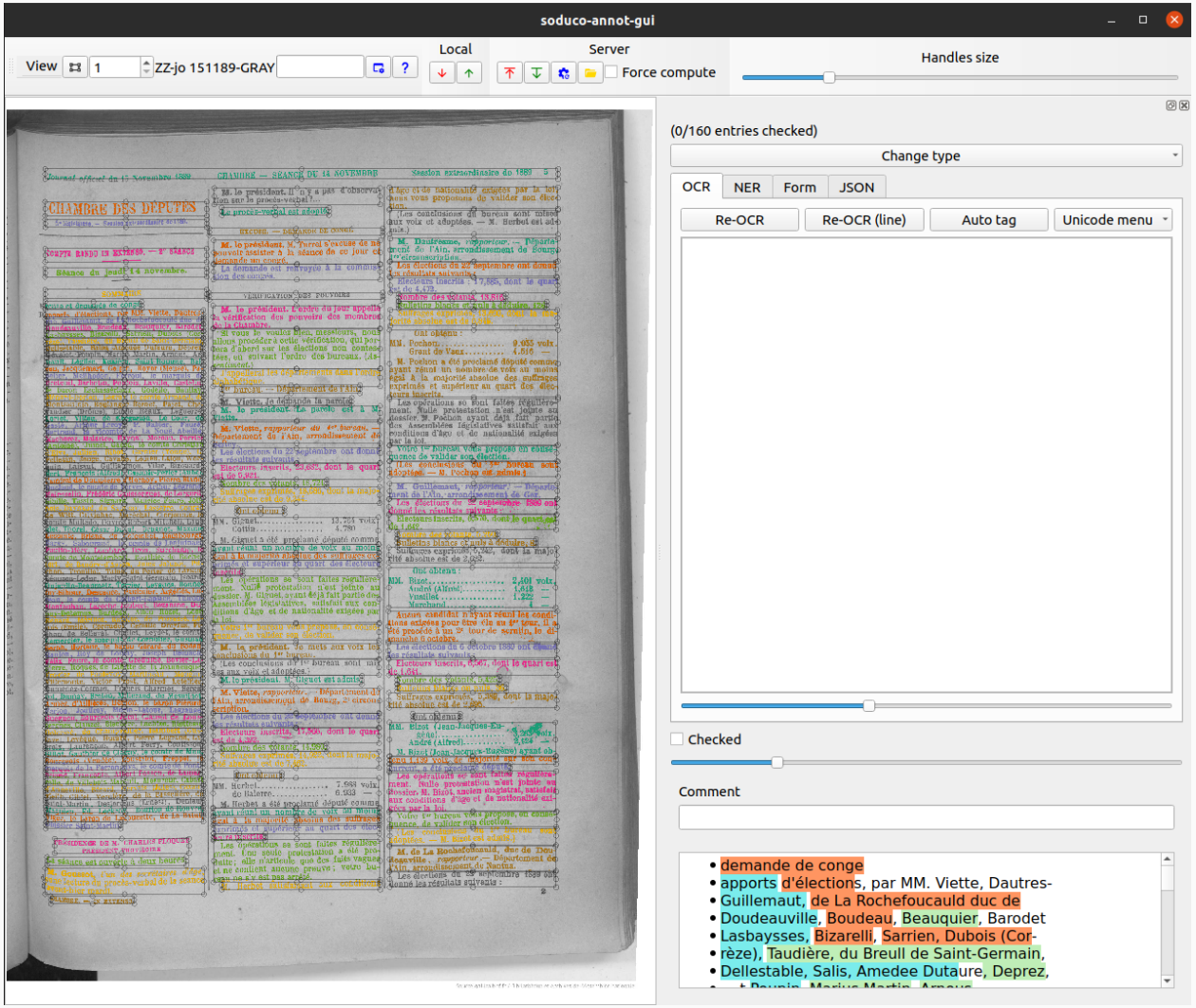
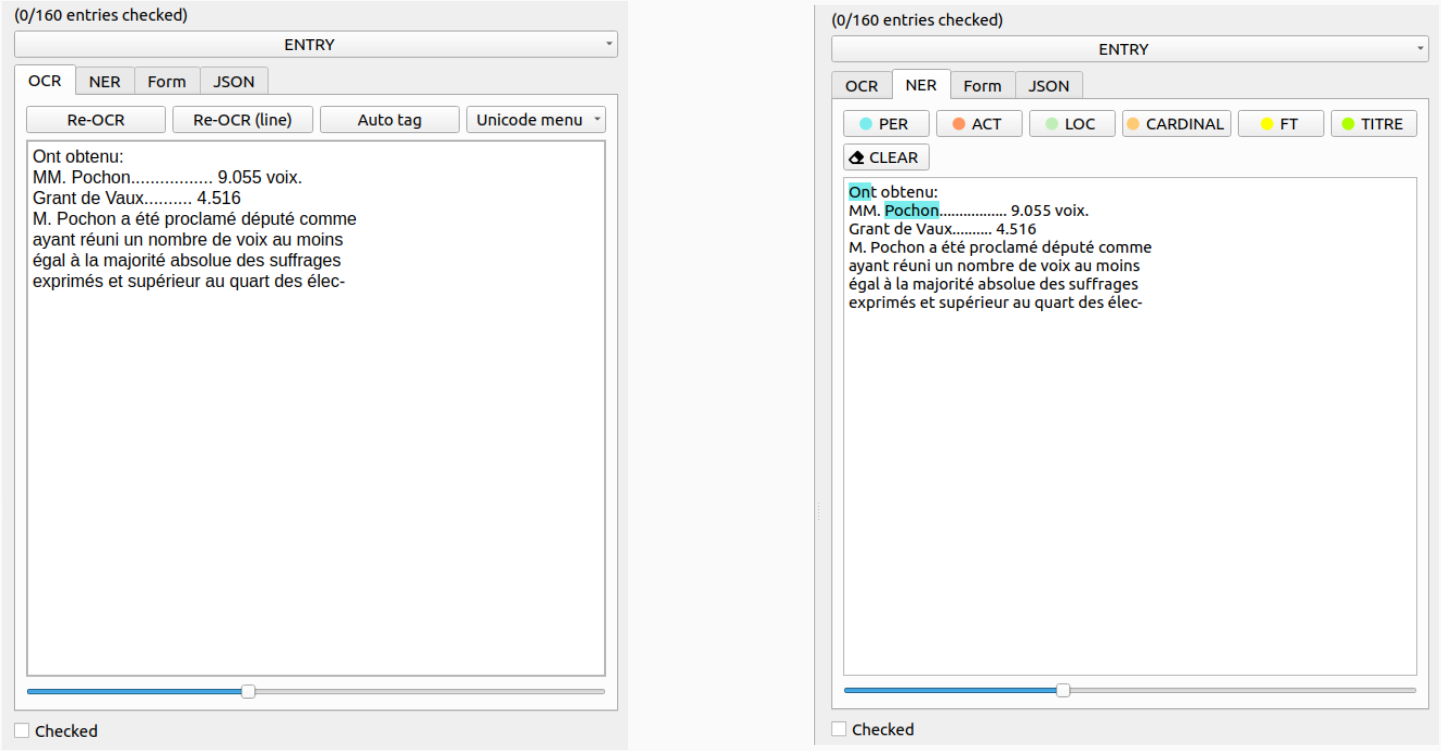


Figure – LRDE's Tool

TOOL DEVELOPED BY LRDE (BASED THE OCR ENGINE : PERO OCR)



(a) OCR

(b) NER

Figure – OCR and NER areas

OCR QUALITY EVALUATION

2 kind of metrics :

- Unsupervised methods : Dictionary...
- Supervised methods
  - Bag Of Words : inherits from classical metrics, has limitations
  - Levenshtein's distance

Levenshtein's distance : Minimum number of insertions, deletions and substitutions on single characters to transform a text into another. Depending on the implementation, spaces can be taken into account or ignored.

OCR (string|token)	Ground Truth	$LEV_{1,1,1}$
« des ates sont accomplis »	« Des actes sont accomplis »	3

Table – An example

METRIC : CHARACTER ACCURACY

$$\text{CharacterErrorRate(CER)} = \frac{\text{LEV}(\text{text}_{gt}, \text{text}_{o}\text{cr})}{\text{len}(\text{text}_{gt})}$$

$$\text{CharacterAccuracy} = \max(0, 1 - \text{CER})$$

OCR	Ground Truth	LEV	CharacterAccuracy
« des ates sont accomplis »	« Des actes sont accomplis »	3	0.88
« as aboue s0 belo »	« as above so below »	3	0.82

Table – Accuracy on 2 examples

LIBRARIES

Examples of libraries on python and in command line

- Jiwer : `https://pypi.org/project/jiwer`
- Fastwer : `https://pypi.org/project/fastwer`
- PRImA's tool : `https://www.primaresearch.org/tools/PerformanceEvaluation`

Different implementations :

1. Penalizing multiple white space
2. Pre-processing needed before the calculation of the metric
3. Possibility of not penalizing certain character
4. ...

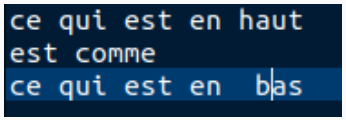
RESULTS ON 1 PAGE OF DEBATE

OCR	Document	CharAccuracy
ocr-tesseract (eng,psm3)	26 Novembre 1889	0.93
ocr-tesseract (fr,psm3)	26 Novembre 1889	0.95
ABBYY	26 Novembre 1889	0.22

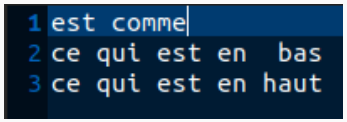
- We see the main limitation of Levenshtein's distance : it's highly sensible of the detection order and layout. ABBYY is overly penalized by the metric.
- Moreover, other problems aroused such as bad table detection
- We are interested in a metric more independent from the reading order

Solution : Flexible Character Accuracy (Clausner et al, 2020)

METRIC : FLEXIBLE CHARACTER ACCURACY

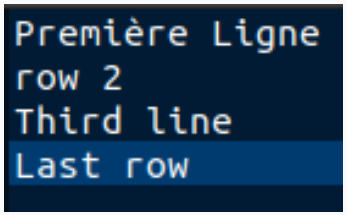


(a) Ground Truth

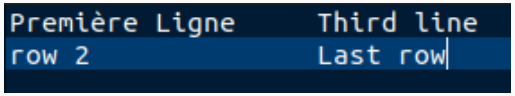


(b) OCR

Figure – Example A



(a) Ground Truth



(b) OCR

Figure – Example B

Exemple	CharAccuracy	FlexCharAccuracy
A	0.45	1
B	0.39	0.95

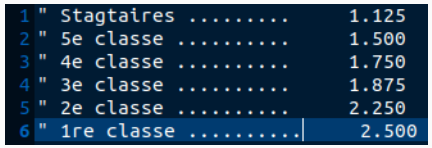
Table – Character Accuracy vs Flexible Character Accuracy

Flexible character accuracy (Clausner, 2020)

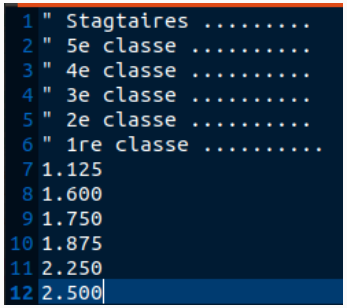
A metric, less sensible to reading order, over and under segmentation :

1. Split the two input text into line
2. Sort the ground truth text lines by length (in descending order)
3. For the first ground truth line, find the best matching OCR result line segment (by minimising a penalty that is partly based on string edit distance)
4. If full match :
   - Mark as done and remove line from list
   - Else subdivide and add to respective list of text lines ; resort
5. If any more lines available repeat step 3
6. Count non-matched lines / strings as insertions or deletions (depending on origin : ground truth or result)
7. Sum up all partial edit distances and calculate overall character accuracy

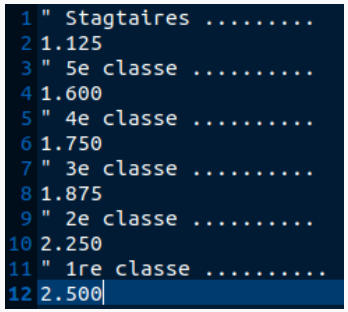
ANOTHER EXAMPLE



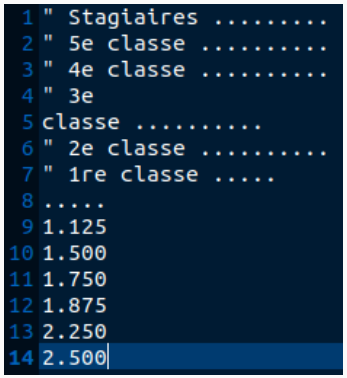
(a) Ground Truth



(b) Example A



(c) Example B



(d) Example C

Figure – Our ground truth with 3 different OCR

Example	CharAccuracy	FlexCharAccuracy
A	0.59	0.96
B	0.84	0.96
C	0.58	0.96-

FLEXIBLE CHARACTER ACCURACY FOR 1 PAGE OF DEBATE

OCR	Document	CharAccuracy	FlexCharAccuracy
ocr-tesseract	1889/11/26	0.95	0.95
ABBY	1889/11/26	0.22	0.97

.

- Results are similar to character accuracy when reading order is good
- Measurement of correct character detection somehow independent from the reading order
- Time consuming : according to the original paper it took around 17 s for one pages of 900 characters. Our debates have a average of 10000 characters per pages

POST-CORRECTION

- Post correction dictionary
- Using Regex
- Endogenous corrections

CONCLUSION

Results are to be improved

- OCR quality is highly impacted by the digitization process
- Post correction needed
- Defining a well suited metric

TOPIC MODELLING WITH LDA

LDA

LDA : Latent Dirichlet Allocation

- A corpus is written based on a finite number of topics. The number is a parameter of the model
- A Topic can be seen as a semantic field, a list of words linked by their meaning
- We want to find Topic as distributions over words and documents as a mixture of topics, conditional on the observed words
- Each document is then produced by selecting words from a subset of topics given a certain probability distribution that we aim to find

Bag of Words method, order of the word doesn't impact; OCR quality does.

Pre-processing

1. Removing stop word
2. 50 Topics
3. Whole corpus is divided in similar size chunks, we hope to find a better way

	topic_0	topic_1	topic_2	topic_3	topic_4	topic_5	t
0	and	and	the	you	of	no	
1	was	the	of	to	had	gutenberg	
2	the	of	and	it	she	project	
3	miss	they	her	not	the	of	
4	in	to	to	be	his	the	
5	all	in	was	is	her	or	
6	of	as	his	but	to	is	
7	from	his	in	have	been	you	
8	elizabeth	their	she	will	in	indeed	
9	he	he	elizabeth	for	that	tm	
10	but	was	their	that	and	this	
11	to	were	by	if	was	work	
12	his	it	with	do	he	and	
13	cried	mr	were	the	as	works	
14	at	she	for	of	which	any	
15	an	had	on	we	not	terms	
16	am	all	which	me	could	foundation	
17	bingley	could	be	can	with	are	
18	it	that	from	so	mr	electronic	
19	bennet	not	as	must	have	all	

Figure – LDA over Pride and Prejudice, without removing stop words

4 STRAIGHTFORWARD TOPICS

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

HAND LABELLING OF TOPICS

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

Figure – topic-classes

Topics a quite consistent with the major themes that marked political life during the early third republic

TOPIC EVOLUTION OVER TIME (BY MONTH)

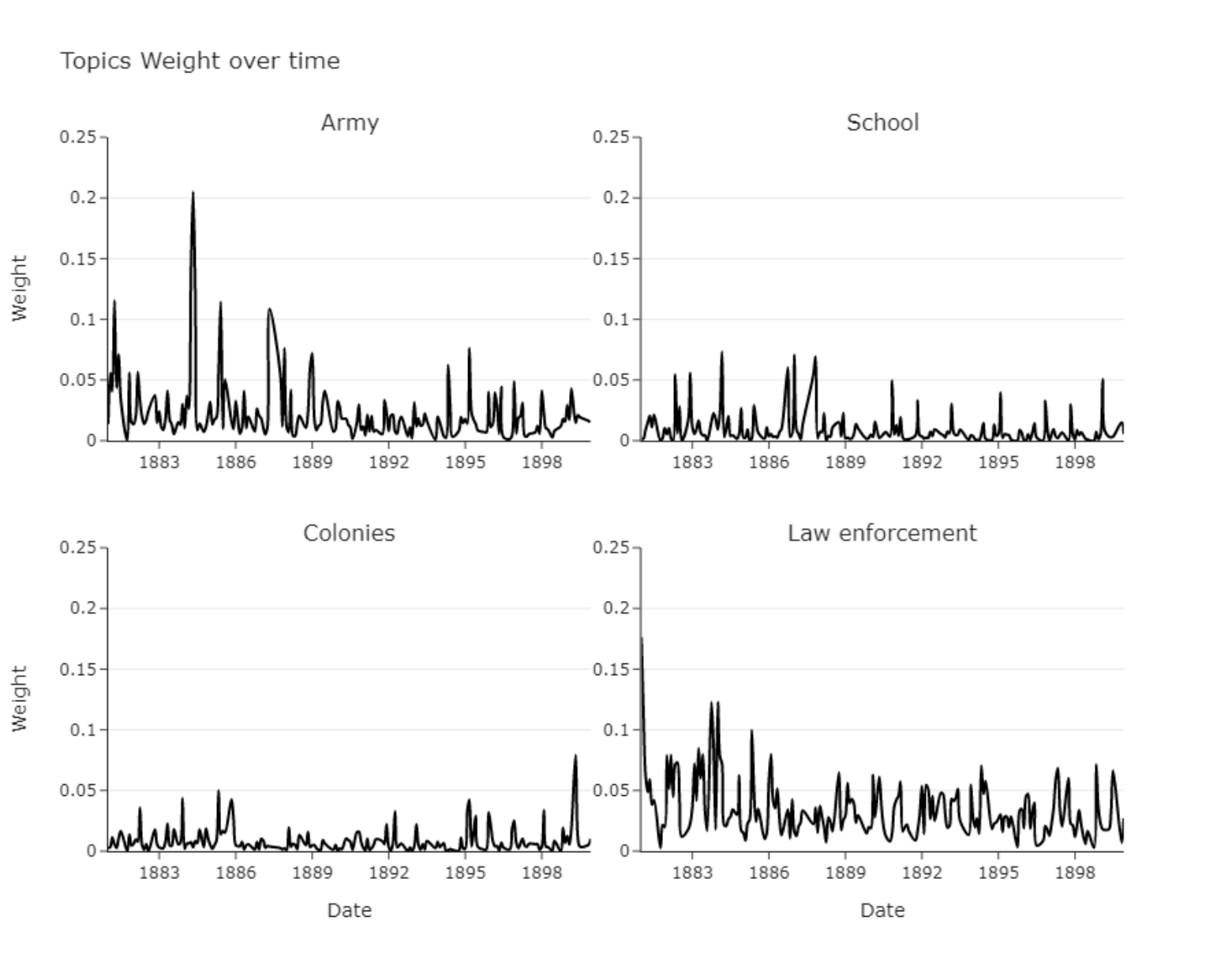


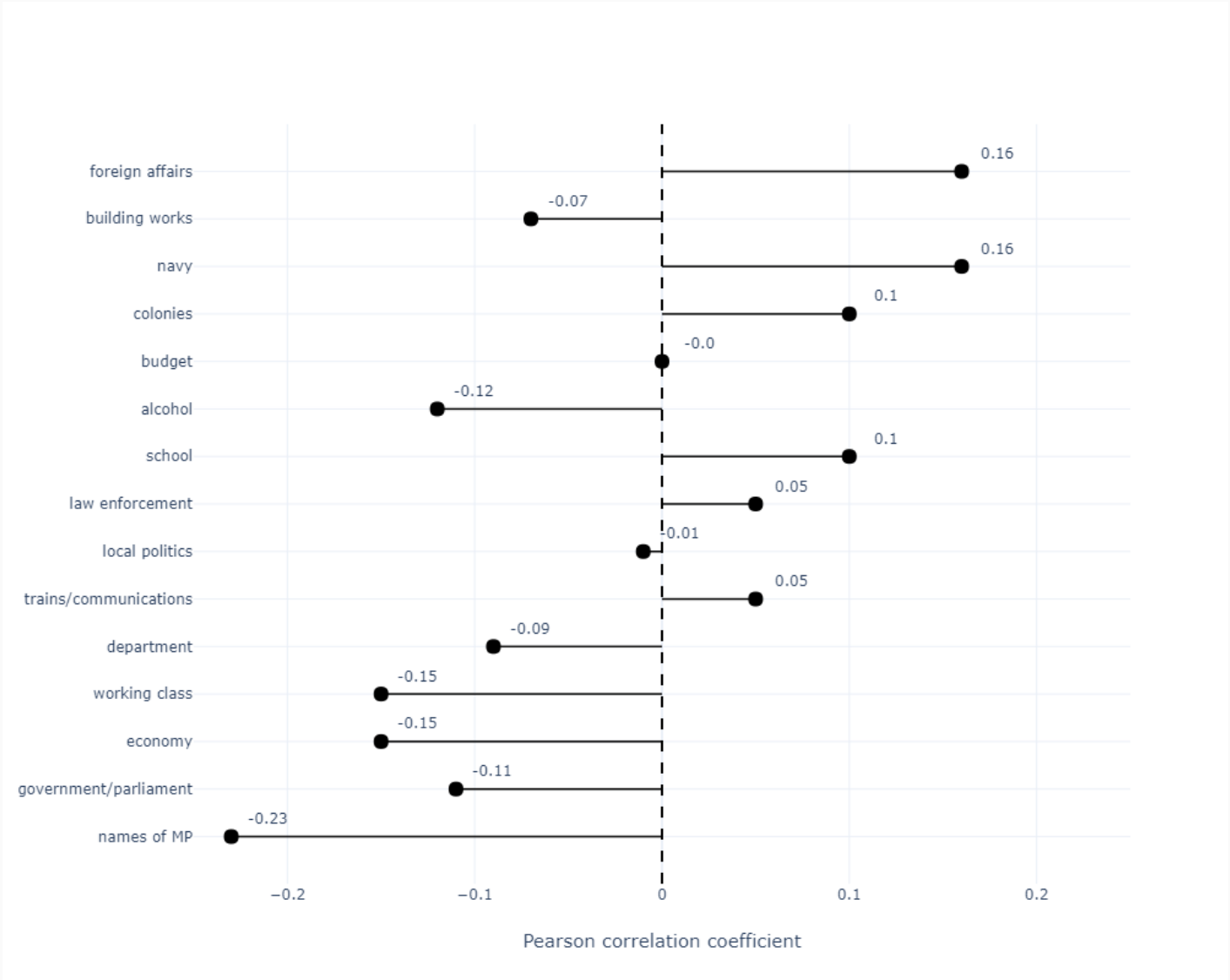
Figure – 4 topic and the evolution of their weight in the corpus

TOPIC ARMY

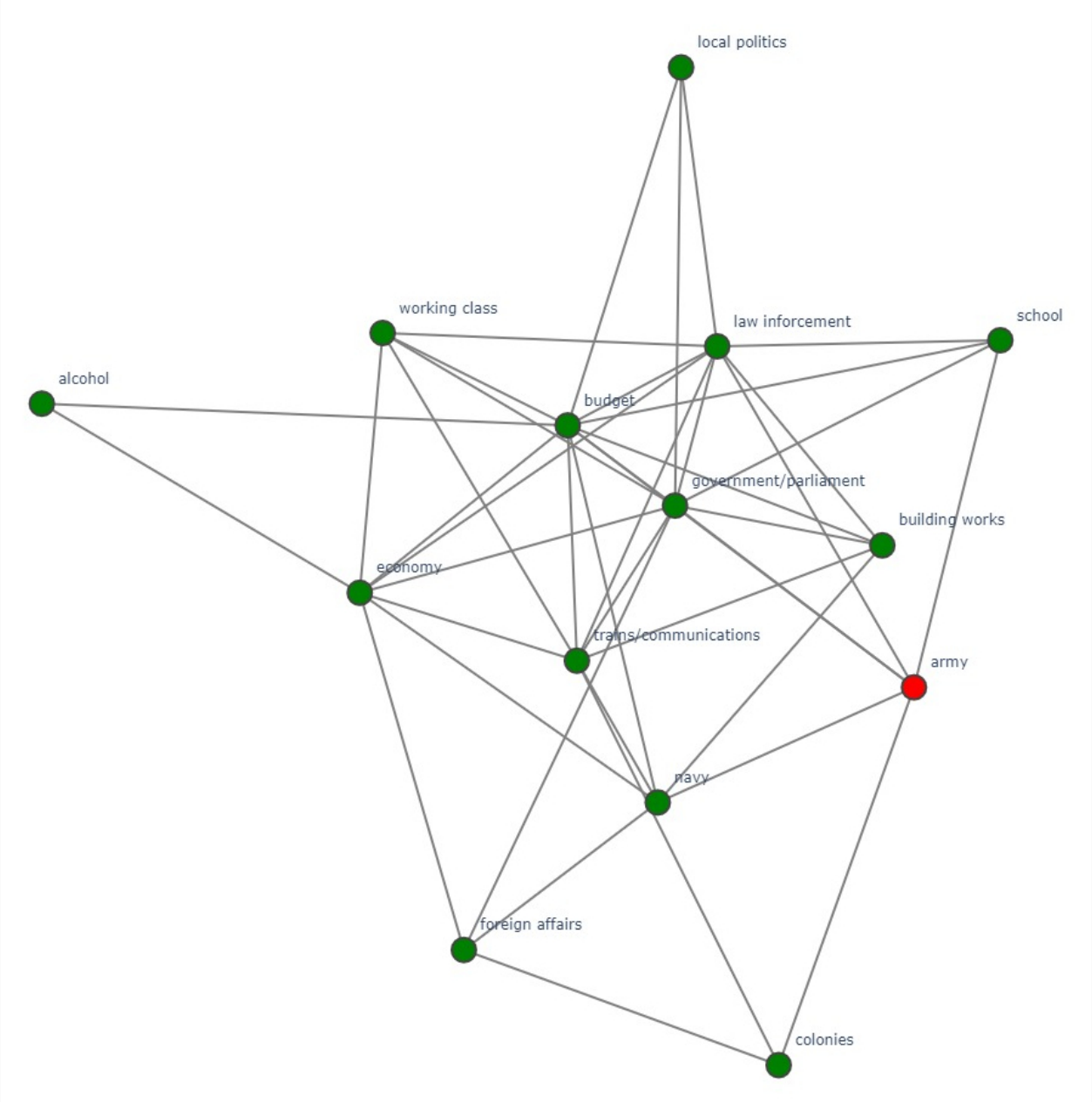
Popularity of the topic army at 4 four different times :

- In 1881 began the conquest of Tunisia
- The year 1884 was marked by the Tonkin campaign and discussion on reducing military service to 3 years
- In 1887 and 1888 : new discussions on the reform of military service.
- In 1895 the difficult conquest of Madagascar gave rise to new debates

CORRELATION WITH THE TOPIC ARMY



STRONGER CORRELATION



WORD EMBEDDING

A BETTER UNDERSTANDING OF OUR CORPUS

A non bag of words method

LDA doesn't specify a notion of similarity between words and also between topics

Stop words have a huge impact, even if a Tf-idf is possible, we look for other possibilities

We look for a word encoding that capture a certain form of semantic similarity

- One-Hot Encoding of vectors (too simple)
- Singular Value Decomposition (too complex)

WORD2VEC

- Neural Network architecture with an hidden layer
- 2 methods : Continuous Bag of Words and Skip-Gram
- CBOW : predict the target word based on the context words
- Skip-Gram : predict the context words based on a target word

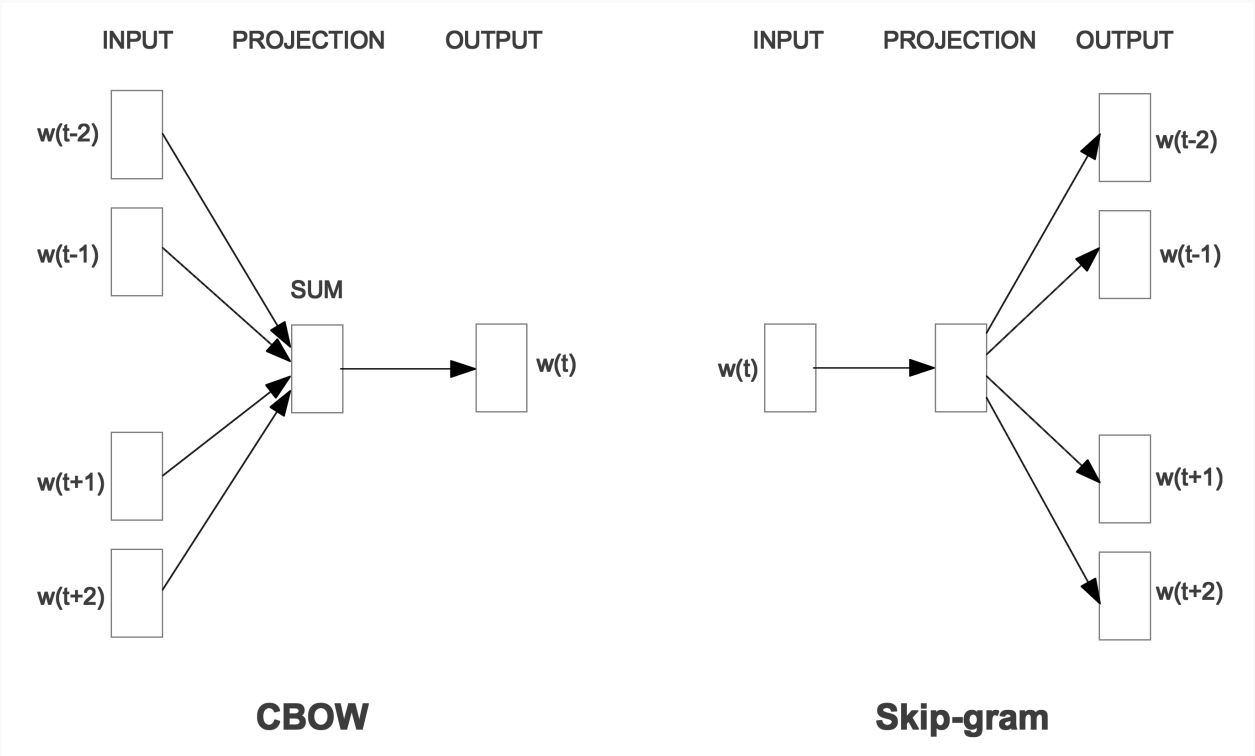


Figure – Word2Vec : CBOW et Skip-gram

église	train	alcool	colonie
presbytère	express	vin	colonial
cimetière	wagon	vinaigre	métropole
temple	locomotive	cidre	nies
cathédrale	omnibus	alcoolisés	marine
geneviève	véhicule	viné	lonies
basilique	voyageur	bière	cochinchine
chantre	fourgon	hectolitre	guyane
chapelain	trajet	poirés	algérie
sacristie	équipe	hydromel	rattachées
rabbin	voiture	hecto	nouvellecalédonie

Table – 4 words, with the top10 most similar words (cosine similarity)

« DIMENSION REDUCTION »USING T-SNE

Once again we selected 8 words, we looked for their top10 most similar words. Then we apply t-SNE for dimension reduction in order to visualize them (vatican, train, vin, divorce, victime, loi, colonie, pédagogique)

SOME COMMENTS - 1

- A word is representend with a unique vector (size 300 typically)
- Vector coordinates do not have an intuitive interpretation
- A word (vector) is defined (learned) by the context
- The size of the context, i.e. the **Window**, is a parameter of the model

Boire (Window=1)	Boire (Window=15)
chercher	indécis
rire	vodka
trouver	allumer
changer	fuma
garder	commanda
entendre	pastis
marquer	coktail
choisir	bière
souffir	comptoire
accepter	champagne

Table – Closest words to « Boire »from a 1 and 15 window size model, French author corpus

SOME COMMENTS - 2

- Vectors can encode more subtle semantic relationships : small is to big what smallest is to biggest (Mikilov et al, 2013)
- It's not systematic, both large corpus and dimensions are needed
- Analogy reported may be biased (Nissim et al, 2019) :
  - **Man** is to **King** what **Woman** is to **Queen** ? In fact King is the first answer returned, Queen comes in second place
- It does not work as well on our corpus
- Both dimension and corpus size need to be improved to keep increasing the quality of our vectors
- Skip-Gram has been shown to give better result with respect to the semantic analogies

DOC2VEC

We want to assign vectors to document

- Paragraph Vector : Distributed Memory model
- Paragraph Vector : Distributed Bag of Words

TOP2VEC (ANGELOV, 2020)

Creating a jointly embedded topic, documents and word vectors

The main assumption is that similar document indicate a topic. A Topic can be seen as an area of close documents in the projected words/documents space

1. We start by finding vectors for words and documents (ex : Doc2Vec)
2. Similar documents will be closed to each other and close to similar words
3. Find dense areas of documents, i.e. a topic
4. Topic vector is then computed as the centroid of the documents, the closest words form the vocabulary of the topic

We obtained a large number of Topics : 113. They are quite coherent and precise. We can study them in relation to each other, but also in relation with words and documents

SOME TOPICS OBTAINED

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

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

SIMILARITY WORD/TOPIC

Topic 80	Topic 32	Topic 64	Topic 81
pedagogique	lycées	polytechnique	séminaristes
enseignement	enseignement	ecole	ecclesiatiques
aptitude	collèges	élèves	vocations
licencie	secondaire	cyr	prêtes
secondaire	proviseur	examens	clergé
université	internants	jeunes	paroissial
diplome	universités	destinent	caserne
bachelier	boursiers	caserne	sacerdoce
baccalauréat	aggregation	forestieres	dispense
infortunes	indissolubilite	études	religion
ravages	conjoints	bachelier	vouer

Table – 5 closest Topics to the word « Education »

Cluster 57	Cluster 97	Cluster 54
paquebots	trains	strategiques
postal	freins	loulan
escale	wagons	ligne
messageries	mecaniciens	chemins
antilles	signaux	timbre
transatlantiques	train	tronçons
west	voyageurs	kilomètres

Table – A cluster made a three topics

LIMITATIONS

By definition, each document is assign to at most 1 Topic. Therefore we loose a property that LDA had to get a mixture of Topics for each documents. Yet with the possibility to look for similarity beetween topics/words, topics/topics, document/topics we think we can retrieve some of the temporal information that we had with LDA

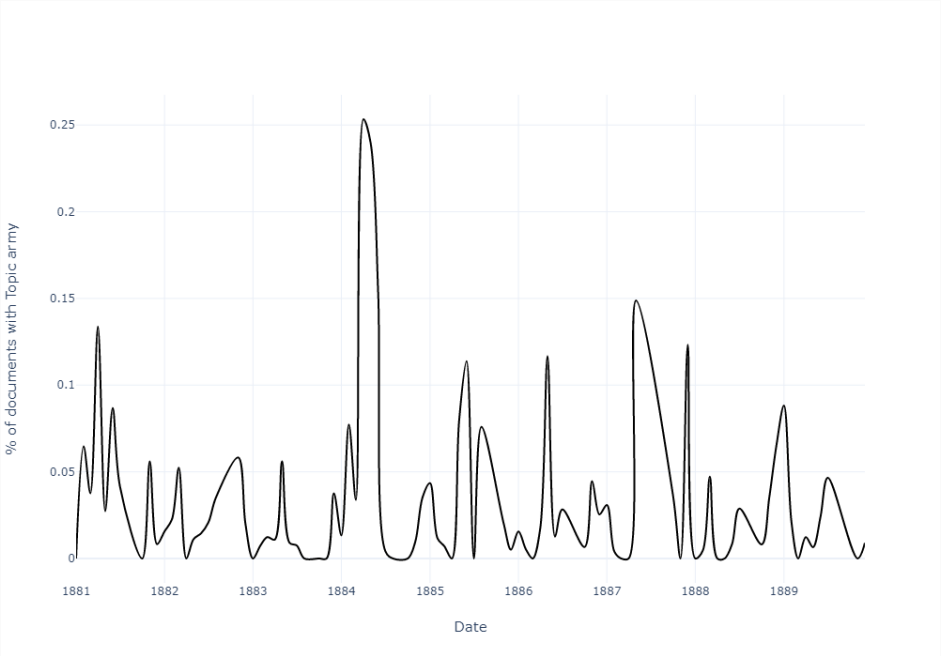
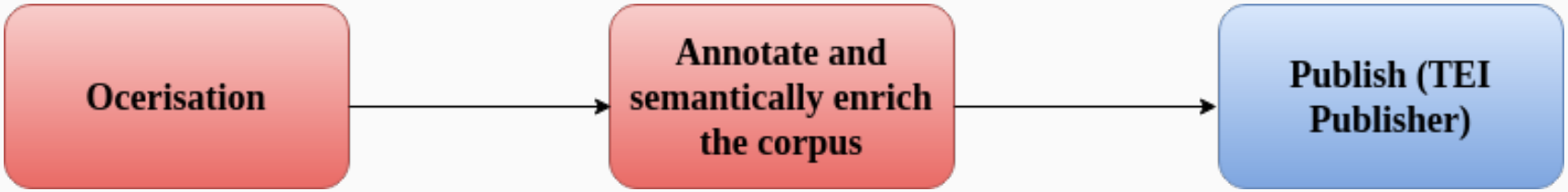


Figure – Evolution of the « Topic Army »We can find similarities with the previous graph

CONCLUSION - 1

- First results obtained with LDA on our « raw »corpus
- Word, document and topic embedding can lead to more interpretations
- The analysis is limited by the quality of the OCR : we must improve that part

CONCLUSION - 2



MERCI POUR VOTRE ATTENTION !

Aurélien Pellet : aurelien.pellet@epitech.eu

Nicolas Bourgeois : nicolas.bourgeois@epitech.eu

Fanny Lebreton : fanny.lebreton@chartes.psl.eu

Marie Puren : marie.puren@epitech.eu