

## LE PROJET AGODA

Annoter et publier les débats parlementaires français de la fin du XIXe siècle : défis et solutions

---

Marie Puren, Pierre Vernus, Aurélien Pellet, Nicolas Bourgeois, Fanny Lebreton  
21 mai 2022

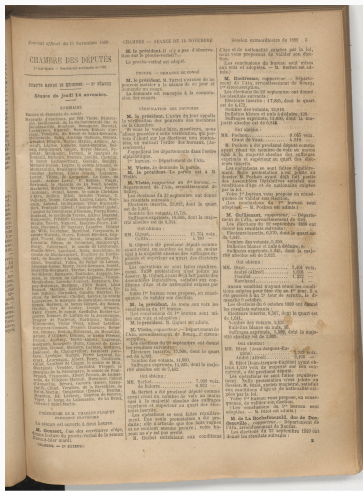
Colloque Humanistica 2022, Université de Montréal

# LE PROJET AGODA

---

- AGODA : **A**nalyse sémantique et **G**raphes relationnels pour l'**O**uverture et l'étude des **D**ébats à l'**A**ssemblée nationale
- Projet financé par la Bibliothèque nationale de France pour une durée d'un an
- L'un des 5 projets-pilote soutenu par le **DataLab**
- Collaboration entre Epitech (MNSHS), Inria (ALMAnaCH) et l'Université Lumière Lyon 2 (LARHRA).

# LES DÉBATS PARLEMENTAIRES DURANT LA TROISIÈME RÉPUBLIQUE



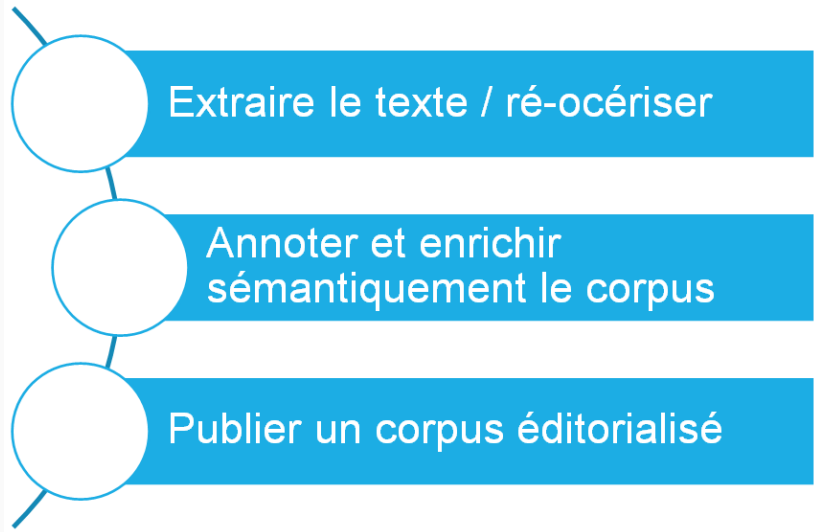
- Débats à la Chambre des députés (chambre basse du parlement) transcrits en détail dans le **Journal officiel de la République française. Débats parlementaires** (1881-1940)
- Disponible en ligne via **Gallica** (bibliothèque numérique de la Bibliothèque nationale de France)
- Difficile de travailler sur ce corpus, pourtant intéressant pour diverses disciplines (histoire, sociologie, science politique, linguistique)

Figure – Séance parlementaire du 14 novembre 1889

- Donner plus facilement accès aux retranscriptions anciennes des débats parlementaires
- Faciliter la recherche dans ce corpus
- Permettre la constitution de sous-corpus
- Offrir de nouveaux modes de visualisation des documents

Traitement d'une sous-partie du corpus : législature 1889-1893 soit **10418 images à traiter**

- Renouvellement partiel du personnel politique (boulangisme et scandale de Panama)
- Premières manifestations du Ralliement des catholiques à la République
- Tournant de la politique douanière (lois Méline)
- Essor du socialisme et du syndicalisme (Fourmies)
- Premiers attentats anarchistes



- 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 l'analyse de gros corpus de documents historiques



# OCÉRISER LES DÉBATS

---

- Récupération des textes océrisés via **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

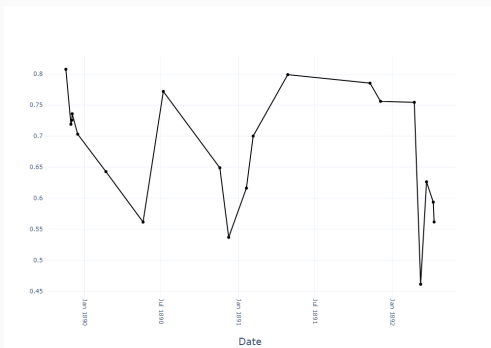


Figure – Evaluation de la qualité de l'OCR fourni par Gallica

# EFFET DE LA COURBURE SUR LES RÉSULTATS DE L'OCR

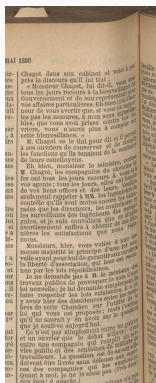


Figure – 14  
mai 1890  
(p.786)

Chagot dans son cabinet et voi il près le discours qu'il lui tint : « Monsieur Chagot, lui dit-il, uS eut tous les jours recours à labienveLu Gouvernement et de son représsu c t vos affaires particulières. Eh bien! rapof neut de vous avertir que, si vous tez pas les mesures, à mon sens i ^oS ow' bles, que vous avez prises contre l'ur alle vriers, vous n'aurez pas à courtoier S'il cette bienveillance. » l per;ll M. Chagot se le tint pour dit et il etelit à ses ouvriers de conserver et de poolplir les fonctions qu'ils tenaient de n de leurs concitoyens. , cow! Eh bien, monsieur le

M. Chagot, les compagnies de r fer ont tous les jours recours à vou90m vos agents ; tous les jours, elles onV'euiUde de vos bons offices et des leurso.nvelliiez seulement rappeler à MM. les libe c-contrôle qu'ils sont moins encorrôleurs e rades que les directeurs, les con coll11P les surveillants des ingénieurs dee coH11P les surveillants des ingénieurs c0 der, gnies, et je suis convaincu que ces def avertissement suffira à obtenir nières les satisfactions que no r mons. une Messieurs, hier, vous vouez à loi ou j0j no mense majorité le principe ou velle ayant pour but de garantir x oUvro la liberté d association, qui leur est recoe, me par les lois républicaines. inist8, de Je ne demande pas à M. ie III iigtre, travaux publics de provoquer \* j ^eiii loi nouvelle ; je lui demande Simple gleut P faire respecter les lois exis" leS ro9% y avoir hier des dissidences en bres de cette Chambre sur ais j'es l'OH loi qui vous est propose; Illa quest qu'il ne saurait y en avoir sur 011 que je soulevé aujourd'hui. un t Ce n'est pas simplement en 'élève, C'ef et un ouvrier que le débat s lit ull S de entre une compagnie q lui ren'Ptger.

vice public et des centaines de jugi gaVOir j, travailleurs. La question est de anlc Jlt.

doivent être livrés sans défense au\* JieV ces des compagnies qui les ifres 0'10 et, Quant à moi, je ne le crois p \*

Figure – Résultat de l'OCR

# LES SOLUTIONS À NOTRE DISPOSITION

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

# OUTIL DE NETTOYAGE DÉVELOPPÉ PAR L'ANR SODUCO



(a) Image d'origine



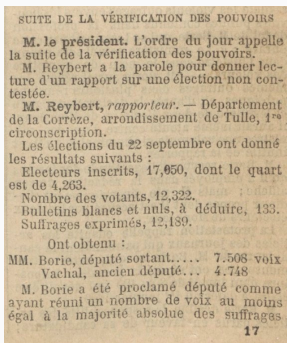
(b) Image nettoyée

Figure – Démonstration de l'outil SODUCO sur une page de débat

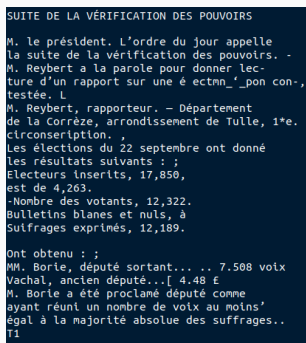
# LES SOLUTIONS À NOTRE DISPOSITION

Sélectionner et comparer plusieurs moteurs OCR :

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



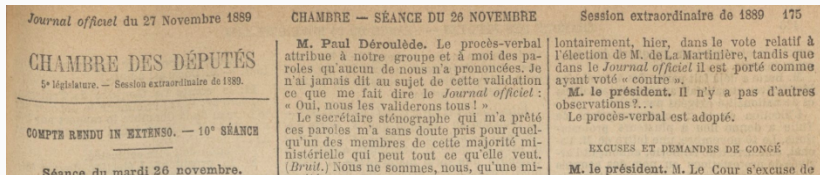
(a) Image d'origine



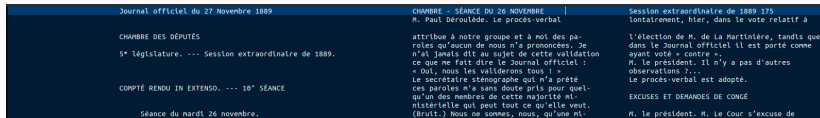
(b) OCR

Figure – Zoom sur un bloc de texte + OCR (tesseract)

# LES SOLUTIONS À NOTRE DISPOSITION



(a) Image d'origine



(b) OCR

Figure – Zoom sur un bloc de texte + OCR (ABBY)

# OUTIL DÉVELOPPÉ PAR SODUCO - 1

Outil basé sur le moteur d'OCR **PERO OCR** : très efficace sur les textes historiques

Développé dans le cadre de l'ANR **SODUCO**

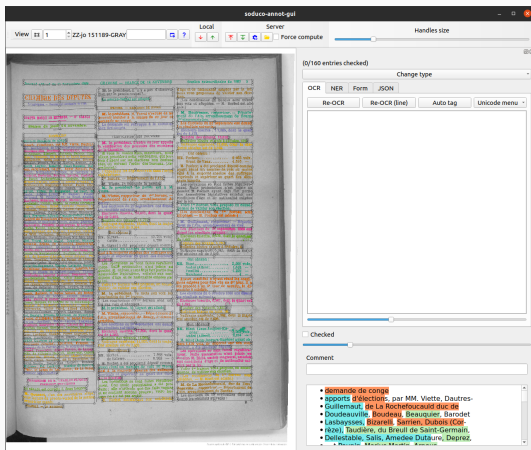
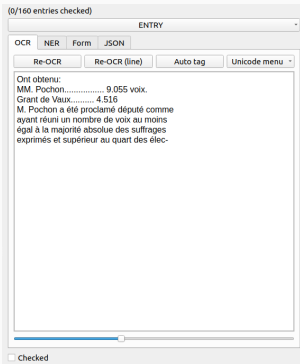


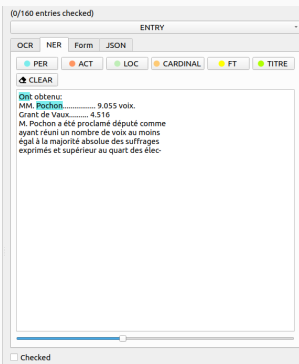
Figure – Outil d'OCR développé par l'ANR SODUCO



# OUTIL DÉVELOPPÉ PAR SODUCO - 2



(a) OCR



(b) NER

Figure – Zones d'OCR et de NER

```
50
},
"id": 302,
"ner_xml": "<PER>Suirrages</PER> exprim\u00e9s, 9,90<CARDINAL>8</CARDINAL>, dont la majo-\u2029rit\u00e9 absolue est de <CARDINAL>4,455</CARDINAL>.",
"origin": "computer",
"parent": 263,
"persons": ["Suirrages"],
"text_ocr": "Suirrages exprim\u00e9s, 9,908, dont la majo-\nrit\u00e9 absolue est de 4,455.",
"type": "ENTRY",
"activities": [],
"comment": "",
"checked": false
```

Figure – Sortie Json (extrait)

- Dictionnaire de post-correction (pyspellchecker)
- Utilisation d'expressions régulières : par exemple gérer les espaces multiples, passage à la ligne, supprimer les « - »
- Corrections endogènes

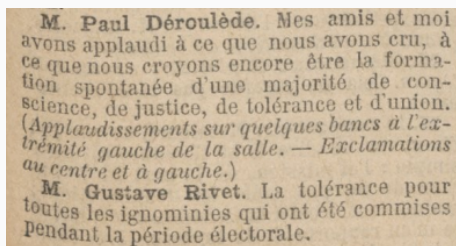
# ANNOTER LES DÉBATS EN XML-TEI

---

Encodage pensé selon 4 principes :

- Les différentes exploitations des textes
- Les particularités de la source
- Les projets similaires : [ParlaClarín](#) et [ParlaMint](#)
- Le processus de balisage automatique

Principes permettant d'orienter, de déterminer, d'influencer, et de contraindre nos choix.



**M. Paul Déroulède.** Mes amis et moi avons applaudi à ce que nous avons cru, à ce que nous croyons encore être la formation spontanée d'une majorité de conscience, de justice, de tolérance et d'union. (*Applaudissements sur quelques bancs à l'extrémité gauche de la salle. — Exclamations au centre et à gauche.*)

**M. Gustave Rivet.** La tolérance pour toutes les ignominies qui ont été commises pendant la période électorale.

**Figure** – Source numérisée - Séance parlementaire du 26 novembre 1889 (extrait)

# CONSERVER LA MISE EN PAGE? - 2

```
<lb/><u who="#pers_ID" xml:id="CR_1889-11-26_u5" ana="#speaker">
  <seg xml:id="CR_1889-11-26_u5.1">
    <persName ref="#pers_ID">M. Paul Déroulède</persName>. Mes amis et moi
    <lb/>avons applaudi à ce que nous avons cru, à
    <lb/>ce que nous croyons encore être la forma-
    <lb/>tion spontanée d'une majorité de con-
    <lb/>science, de justice, de tolérance et d'union.
    <lb/><incident><desc>(Applaudissements sur quelques bancs à l'ex-
    <!-- Pas de lb possible dans incident --> trémité gauche de La salle. – Exclamations
    <!-- Pas de lb possible dans incident --> au centre et à gauche.)</desc></incident>
  </seg>
</u>

<lb/><u who="pers_ID" xml:id="CR_1889-11-26_u6" ana="#speaker">
  <seg xml:id="CR_1889-11-26_u6.1">
    <persName ref="#pers_ID">Gustave Rivet</persName>. La tolérance pour
    <lb/>toutes Les ignominies qui ont été commises
    <lb/>Pendant la période électorale.
  </seg>
</u>
```

## (a) Modèle d'encodage 1 : sémantique et mise en page

```
<u who="#pers_ID" xml:id="CR_1889-11-26_u5" ana="#speaker">
  <seg xml:id="CR_1889-11-26_u5.1"><persName ref="#pers_ID">M. Paul Déroulède</persName>. Mes amis et moi
avons applaudi à ce que nous avons cru, à ce que nous croyons encore être la formation spontanée d'une majorité de
conscience, de justice, de tolérance et d'union. <incident><desc>(Applaudissements sur quelques bancs à l'extrémité
gauche de la salle. –Exclamations au centre et à gauche.)</desc></incident></seg>
</u>

<u who="#pers_ID" xml:id="CR_1889-11-26_u6" ana="#speaker">
  <seg xml:id="CR_1889-11-26_u6.1"><persName ref="#pers_ID">Gustave Rivet</persName>. La tolérance pour
toutes les ignominies qui ont été commises pendant la période électorale.</seg>
</u>
```

## (b) Modèle d'encodage 2 : sémantique

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

# ENCODER LES ANNEXES

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

---

SCRUTIN

Sur les conclusions de l<sup>e</sup> bureau tendant à l'annulation des opérations électorales de la 1<sup>re</sup> 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, Adon (Emanuel), Armez, Arribat, Audifred, Aynard (Eduard), Baile (Marial), Berg, Barodet, Barthou, Bataillon, Bédit (Edron), Benard, Beauquier, Bérard, Berger (Georges) (Séclé), Bertrand, Bézine, Bisarull, Bisol, Bissonard-Bert, Bono (Pierre), Boulay-Castelnau, Bondy-Ly-Sibour, Bony-Casténo, Boriglione, Bouchet (Vierge), Bouchonnet, Boudetille, Bouge, Boulanger-Benet, Boullay, Bourgeois (Jules), Bourgeois (Léon) (Marie), Boulière de Bouchéger, Boyer-Lapierre, Buvard, Brand, Breton, Briens, Brisson (Henri), Broussé (Emile), Bugeon, Buisson, Bully, Boudou, Davignier.

Rectifications aux scrutins de la séance du 22 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 numérisisée

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

<div xsl:id="vot18891126">
<!-- VOTE 1 -->
<div xsl:id="vot18891126_vot1" type="voting" corresp="#discussion7ebureau">
<head>
<label>SCRUTIN</label>
<note>seg<!-- Sur les conclusions du <num>7</num> bureau tendant à l'annulation des opérations électorales de la
<placeName ref="#lieu_ID"><num>1</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="yes" 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 num="192"/></div></floatingText>
<!-- Liste des votants -->
<note type="voterslist">
<desc>Ont voté pour :</desc>
<seg>MM. <persName ref="#pers_ID">Abeille</persName>, <persName ref="#pers_ID">Adon (Emanuel)</persName>, <persName
ref="#pers_ID">Armez</persName>, <persName ref="#pers_ID">Arribat</persName>, <persName ref="#pers_ID">Audifred</persName>, <persName ref="#pers_ID">Aynard (Eduard)</persName>, <persName
ref="#pers_ID">Baile (Marial)</persName>, <persName ref="#pers_ID">Berg</persName>, <persName ref="#pers_ID">Barodet</persName>, <persName ref="#pers_ID">Barthou</persName>, <persName
ref="#pers_ID">Bataillon</persName>, <persName ref="#pers_ID">Bédit (Edron)</persName>, <persName ref="#pers_ID">Benard</persName>, <persName ref="#pers_ID">Beauquier</persName>, <persName
ref="#pers_ID">Bérard</persName>, <persName ref="#pers_ID">Berger (Georges) (Séclé)</persName>, <persName ref="#pers_ID">Bertrand</persName>, <persName ref="#pers_ID">Bézin</persName>, <persName
ref="#pers_ID">Bisarull</persName>, <persName ref="#pers_ID">Bisol</persName>, <persName ref="#pers_ID">Bissonard-Bert</persName>, <persName ref="#pers_ID">Bono (Pierre)</persName>, <persName
ref="#pers_ID">Boulay-Castelnau</persName>, <persName ref="#pers_ID">Bondy-Ly-Sibour</persName>, <persName ref="#pers_ID">Bony-Casténo</persName>, <persName
ref="#pers_ID">Boriglione</persName>, <persName ref="#pers_ID">Bouchet (Vierge)</persName>, <persName ref="#pers_ID">Bouchonnet</persName>, <persName ref="#pers_ID">Boudetille</persName>, <persName
ref="#pers_ID">Bouge</persName>, <persName ref="#pers_ID">Boulanger-Benet</persName>, <persName ref="#pers_ID">Boullay</persName>, <persName ref="#pers_ID">Bourgeois (Jules)</persName>, <persName
ref="#pers_ID">Bourgeois (Léon) (Marie)</persName>, <persName ref="#pers_ID">Boulière de Bouchéger</persName>, <persName ref="#pers_ID">Boyer-Lapierre</persName>, <persName
ref="#pers_ID">Buvard</persName>, <persName ref="#pers_ID">Brand</persName>, <persName ref="#pers_ID">Breton</persName>, <persName ref="#pers_ID">Briens</persName>, <persName ref="#pers_ID">Brisson (Henri)</persName>, <persName
ref="#pers_ID">Broussé (Emile)</persName>, <persName ref="#pers_ID">Bugeon</persName>, <persName ref="#pers_ID">Buisson</persName>, <persName ref="#pers_ID">Bully</persName>, <persName
ref="#pers_ID">Boudou</persName>, <persName ref="#pers_ID">Davignier</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<!-- M. Michau <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)



# VERS LES LINKED DATA

---

# VERS LES LINKED DATA

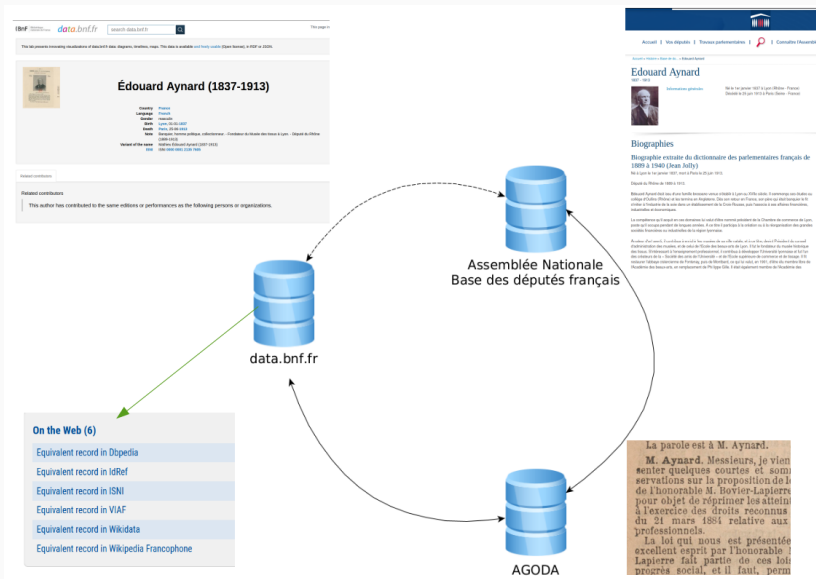


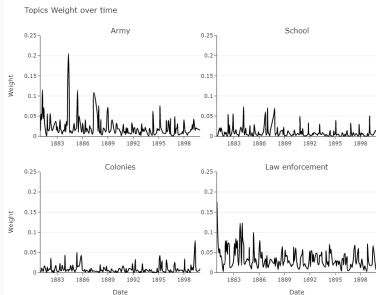
Figure – Linked data

# TOPIC MODELING ET WORD EMBEDDING

---

# EXPLORER LES DÉBATS AVEC LA MODÉLISATION DE SUJETS

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



- (a) 3 sujets parmi 40 : classe ouvrière (8), armée (11) et infrastructures (15)
- (b) Evolution de quatre sujets au cours du temps

Figure – Résultats de la modélisation de sujets

# 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
eprouvées	conjugal	postale
orages	divorces	timbre
sinistres	adultère	recepissés
grele	conjugale	postes
secours	remarié	postaux
venir	separation	telegraphes
infortunes	indissolubilité	colis
ravages	conjoints	fixe
misères	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

Nicolas Bourgeois, Aurélien Pellet, Marie Puren. "Using Topic Generation Model to explore the French Parliamentary Debates during the early Third Republic (1881-1899)". (hal-03526254v2)

Marie Puren, Nicolas Bourgeois, Aurélien Pellet, Pierre Vernus, Fanny Lebreton. "Between History and Natural Language Processing : Study, Enrichment and Online Publication of French Parliamentary Debates of the Early Third Republic (1881-1899)". ParlaCLARIN III at LREC2022 - Workshop on Creating, Enriching and Using Parliamentary Corpora, Jun 2022, Marseille, France. (hal-03623351)



Marie Puren : `marie.puren@epitech.eu`

Pierre Vernus : `pierre.vernus@msh-lse.fr`

Aurélien Pellet : `aurelien.pellet@epitech.eu`

Nicolas Bourgeois : `nicolas.bourgeois@epitech.eu`

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