



HAL
open science

Le projet AGODA. Annoter et publier les débats parlementaires français de la fin du XIX e siècle : défis et solutions

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

► To cite this version:

Nicolas Bourgeois, Fanny Lebreton, Aurélien Pellet, Marie Puren, Pierre Vernus. Le projet AGODA. Annoter et publier les débats parlementaires français de la fin du XIX e siècle : défis et solutions. Présentation des projets AGODA et Gallicorpora, Bibliothèque nationale de France, Jun 2022, Paris, France. hal-03762957

HAL Id: hal-03762957

<https://hal.science/hal-03762957>

Submitted on 29 Aug 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

LE PROJET AGODA

Annoter et publier les débats parlementaires français de la fin du XIX^e siècle : défis et solutions

Nicolas Bourgeois¹, Fanny Lebreton², Aurélien Pellet¹, Marie Puren^{1 3}, Pierre Vernus^{4 5}

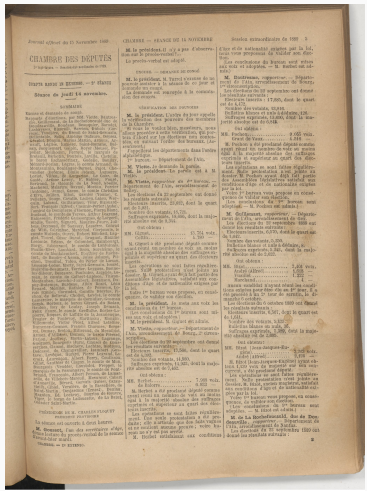
17 juin 2022 - Présentation des projets AGODA et Gallicorpora, Bibliothèque nationale de France (Paris)

¹MNSHS, Epitech ²Ecole nationale des chartes ³CJM, Ecole nationale des chartes ⁴LARHRA ⁵Université Lyon 2

PRÉSENTATION DU PROJET

- AGODA : **A**nalyse sémantique et **G**raphes relationnels pour l'**O**uverture et l'étude des **D**ébats à l'**A**ssemblée nationale
- Collaboration entre Epitech (MNSHS), l'Université Lumière Lyon 2 (LARHRA) et Inria (ALMA_{na}CH)

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

- 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 gros corpus de documents historiques

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

- Renouveau 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



OCÉRISER LES DÉBATS

LE CAS DES DÉBATS PARLEMENTAIRES

- 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

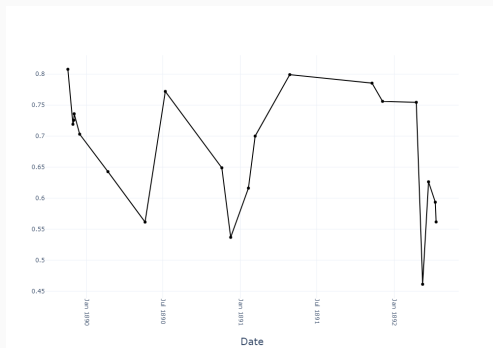


Figure – Evaluation de la qualité de l'OCR fourni par Gallica

EFFET DE LA COURBURE SUR LES RÉSULTATS DE L'OCR

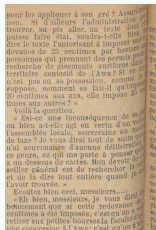


Figure – 20
octobre
1890
(p.1718)

pour les appliquer à son gré ? Assure.

non. Si d'ailleurs l'administration f, trouver, au pis aller, un texte (10111 en 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 Pour l recherche de gisements aurifères a teSI territoire contesté de l'Awá ? Si ce tes

n'est pas en sa possession, cornas le suppose, comment se fait-il qu'ilWjsj} 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 souveraineté d'aucune délibération l'IV ce genre, ce qui me porte à penser à un dessous de cartes. Mon devoir seiller général est de rechercher la cJ? » et de la dire tout entière quand Je l'avoir trouvée. »

Écoutez bien ceci, messieurs. l «Eh bien, messieurs, je vous dirá 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'Awá; c'est qu'on a rj;:j}

Figure – Résultat de l'OCR

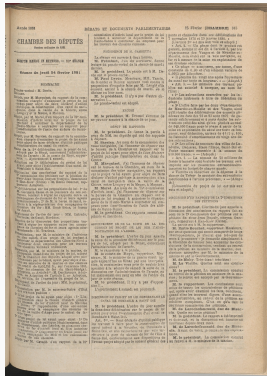
Décision de ré-océreriser le corpus

1. Segmentation incorrecte
2. OCR moins performant
3. Texte très courbé et même coupé : peu lisible par un être humain, encore moins lisible par une machine
4. Annotation automatique en XML-TEI plus difficile car liée à la segmentation
5. Texte trop fautif : impossible à publier tel quel en ligne
6. Influence négative des erreurs de l'OCR sur la performance des analyses TAL (NER, topic modeling, word embedding) (cf. [Assessing the Impact of OCR Quality on Downstream NLP Tasks](#))

QUELLES SOLUTIONS?

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

- Redresser la courbure => redresser le texte tout entier et peut recourber les parties du texte non courbées.
- Solution envisagée : redresser uniquement la partie de l'image concernée (colonne en bord de reliure)

Quand le texte n'est pas lisible par l'humain et donc par la machine :

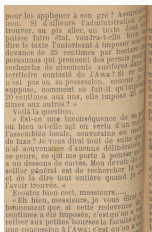


Figure – Mots coupés par la forte courbure

- Utiliser un modèle de langue pour « prédire » le ou les mot(s) manquants
- Modèle de langue pour le français **Camembert**
 - ... Mais entraîné sur du français (trop?) contemporain
 - Peut-être nécessaire de réentraîner sur du français de l'époque (utiliser la vérité de terrain, des corpus de presse corrigé...)

Deux problèmes récurrents :

- Mots coupés à cause de la courbure
- Mots océrés mais apparaissant dans le désordre :
conséquence de la mauvaise segmentation due à la courbure
de la page

Utiliser une IA pour « restaurer » le texte, comme le fait Ithaca de Deep Mind (cf. [Restoring and attributing ancient texts using deep neural networks](#))

DES solutions

Meilleure approche = combiner plusieurs approches

OCÉRISER ET POST-CORRIGER

Dans le cadre du financement, le traitement de l'image est impossible.

Décision de travailler d'abord sur les numéros les moins affectées :

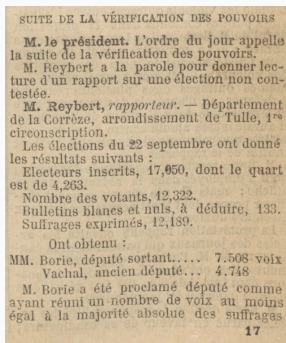
- Océriser ces documents en priorité
- Corriger les textes extraits (cf. post-correction)

Utiliser le taux d'erreur dans l'OCR pour repérer les documents les plus problématiques

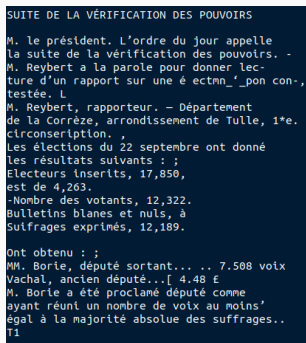
LES SOLUTIONS À NOTRE DISPOSITION

Comparer les performances des moteurs OCR offerts par Huma-Num (Sharedocs) :

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



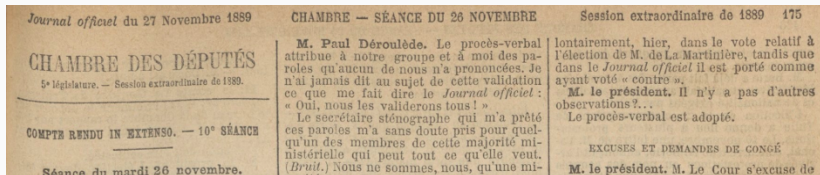
(a) Image d'origine



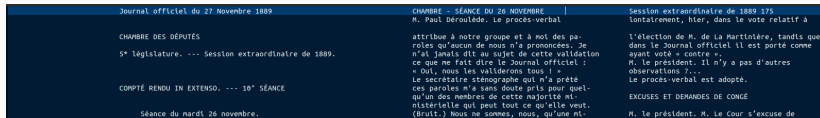
(b) OCR

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 LRDE (EPITA) - 1

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

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

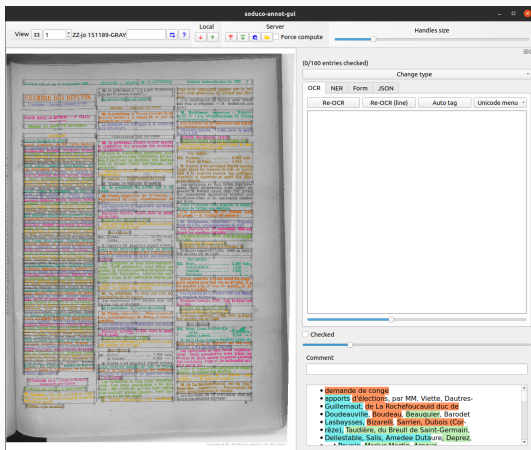


Figure – Outil LRDE (EPITA)

OUTIL DÉVELOPPÉ PAR LRDE (EPITA) - 2

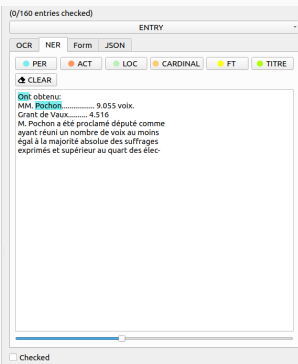
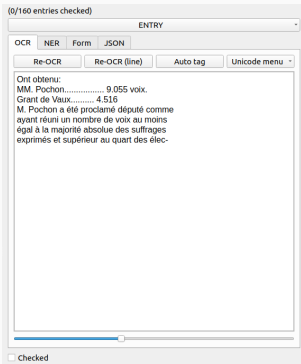


Figure – Zones d'OCR et de NER

```
"box": [
  827.8018264183805,
  2374.55023569486,
  589.3638270234404,
  207.5897599646105
],
"checked": true,
"comment": "u-par",
"id": 352,
"ner_xml": "<PER>M. Francis Laur</PER>. Je n'ai pas l'intention\nd'apporter \u00e0 la Chambre d'autr\u2029 de",
"origin": "computer",
"parent": 274,
"persons": ["M. Francis Laur"],
"text_ocr": "M. Francis Laur. Je n'ai pas l'intention\nd'apporter \u00e0 la Chambre d'autres documents\nque ceux qui me sont fournis par la lettre\nde mission de M. Lev\u00e9aque, adress\u00e9 \u00e0 M. le ministre des finances. Je demanderai\nla Chambre si elle connait cette lettre ou\nsi elle veut que j'en donne lecture.",
"type": "ENTRY"
},
{
  "activities": [],
  "addresses": [],
```

Figure – Sortie Json (extrait)

2 Groupes de métriques :

- Méthodes non-supervisées : Dictionnaire...
- Méthodes supervisées
 - Bag Of Words : hérite de métrique classique, des limites
 - Distance de Levenshtein

Distance de Levenshtein : Nombre minimal d'insertion, délétions et substitutions sur caractères seuls pour transformer un texte en un autre. Selon les implémentations, les espaces peuvent être pris en compte ou ignorés

OCR (string token)	Ground Truth Dict	LEV _{1,1,1}
« des ates sont accomplis »	« Des actes sont accomplis »	3
« tensor »	[tenseur, trois, cube]	2

Table – Exemples de distances de Levenshtein

$$\text{CharacterErrorRate(CER)} = \frac{\text{Lev}(\text{text}_{\text{gt}}, \text{text}_{\text{ocr}})}{\text{len}(\text{text}_{\text{gt}})}$$

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

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 – Calcul de l'Accuracy

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

Table – Résultats sur un de nos document

On remarque la limite de la distance de Lenvenshtein, elle est trop sensible à l'ordre de détection.

Une métrique moins sensible à l'ordre de détection du texte :

1. Sépare les deux textes en sacs de lignes (GT + OCR)
2. Ordonne les lignes de la GT de la plus longue à la plus petite
3. Pour la première ligne, chercher le meilleur match dans les lignes de l'OCR
4.
 - Si match complet on sort la ligne
 - Sinon on sous-divise et on rajoute à la pile de lignes
5. On recommence l'étape précédente
6. On compte le nombre de modifications nécessaires pour les lignes qu'on ne peut pas matcher
7. Calcul de l'accuracy

MÉTRIQUE : FLEXIBLE CHARACTER ACCURACY

```
ce qui est en haut  
est comme  
ce qui est en bas
```

(a) Ground Truth

```
1 est comme  
2 ce qui est en bas  
3 ce qui est en haut
```

(b) OCR

Figure – Exemple A

```
Première Ligne  
row 2  
Third line  
Last row
```

(a) Ground Truth

```
Première Ligne   Third line  
row 2           Last row
```

(b) OCR

Figure – Exemple B

Exemple	CharAccuracy	FlexCharAccuracy
A	0.45	1
B	0.39	0.95

Table – Character Accuracy vs Flexible Character Accuracy

OCR	Document	CharAccuracy	FlexCharAccuracy
ocr-tesseract (fr,psm3)	26 Novembre 1889	0.95	0.95
ABBY	26 Novembre 1889	0.22	0.97

Table – Résultats sur un de nos documents

- La flexible character accuracy est moins sensible à l'ordre de détection
- Résultats sensiblement équivalents quand l'ordre est le bon
- Comparaison de la qualité de détection des caractères toutes choses égales par ailleurs
- Utile si on n'utilise des algorithmes qui ne s'occupent pas de l'ordre des mots (Clustering via frequency matrix, topic modelling : LDA...)

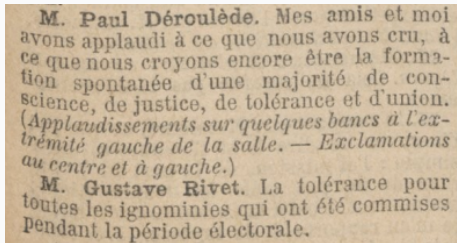
- Dictionnaire de post-correction
- Utilisation d'expressions régulières : gérer les espaces multiples, passage à la ligne, les « - »
- Corrections endogènes
- Modèle de langue CamemBERT pour corriger les mots fautifs (cf. ci-dessus)

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)

PROBLÉMATIQUE : 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 : logique, sémantique et formel

```
<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 : logique et sémantique

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

PROBLÉMATIQUE : CONSERVER LA MISE EN PAGE ? - 3

```
<div type="part" corresp="#rapportelections">
  <head-SUITE DE LA VERIFICATION 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>
```

(a) Modèle d'encodage 3 : solution médiane

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

PROBLÉMATIQUE : ENCODER LES ANNEXES

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

SCRUTIN

Sur les conclusions de l^e bureau tendant à l'annulation des opérations électorales de la 1^{re} 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.

NOT VOTÉ POUR :

MM. Abeille, Adon (Emanuel), Armez, Arribat, Audifred, Aynard (Eduard).

Balle (Marial), Bergy, Barodet, Barthou, Bataillon, Bédit (Edron), Benard, Beauquier, Bérard, Berger (Georges) (Séclé), Bertrand, Bézina, Bisarull, Bisol, Bissonard-Bert, Bono (Pierre), Boulay-Castelnau, Bondeyroy-Sibour, Bony-Casténo, Boriglione, Bouchet (Vierge), Bouchonnet, Boudetille, Bouga, Boulanger-Benet, Boullay, Bourgeois (Juri), Bourgeois (Léon) (Marie), Boulière de Bouchéger, Boyer-Lapierre, Buvard, Brand, Breton, Briens, Brisson (Henri), Broasat (Henri), Brognon, Buvard, Bully, Boudou, Davignier.

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="mdiscussion7ebureau">
<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>pb <num>192</num></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 (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>.</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)

STRUCTURE GÉNÉRALE DES FICHIERS XML TEI

```
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0" xsl:id="FR_3R_5L" xsl:lang="fr">
  <!-- Métadonnées du corpus (non développées) -->
  <teiHeader>
    <fileDesc>
      <titleStnt>
        <title></title>
      </titleStnt>
      <publicationStnt>
        <sp></sp>
      </publicationStnt>
      <sourceDesc>
        <sp></sp>
      </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 -->
  <xli:include xmlns:xli="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

```
<TEI xmlns="http://www.tei-c.org/ns/1.0" xsl:id="FR_3R_5L_1889-11-26" xsl:lang="fr">
  <!-- Métadonnées du composant (non développées) -->
  <teiHeader>
    <fileDesc>
      <titleStnt>
        <title></title>
      </titleStnt>
      <publicationStnt>
        <sp></sp>
      </publicationStnt>
      <sourceDesc>
        <sp></sp>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <!-- Transcription du compte rendu -->
    <body>
      <div></div>
    </body>
    <!-- Annexes du compte rendu -->
    <back></back>
  </text>
</TEI>
```

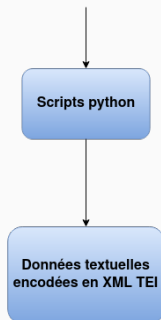
(b) Structure générale d'un fichier composant

Figure – Encodages - Structure générale des fichiers corpus et composant

APPLIQUER L'ENCODAGE : AUTOMATISATION

```
▼ 19:
activities:      []
addresses:      []
▼ box:
  0:             220.12998972250773
  1:             202.1299897225077
  2:             561.8527187504491
  3:             00.4418437486525
checked:        true
comment:        "seg"
id:             276
▼ ner_xml:
  origin:        "<PER>M. Borie</PER> a «ACT>déjà fait partie des Assemblées\u2029législatives et satisfait aux conditions d«ACT>«ACT>âge\u2029et de
  parent:        "computer"
  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"
```

Données textuelles au format JSON



- Publication avec **TEI Publisher** : outil de publication de corpus en TEI
- Basé sur la technologie Web Components : permet de développer de nouvelles fonctionnalités
- Développer de nouveaux modes de visualisation : par exemple, explorer les débats en fonction des sujets qu'ils évoquent , et pas seulement par des mots-clés

VERS LES LINKED DATA

VERS LES LINKED DATA

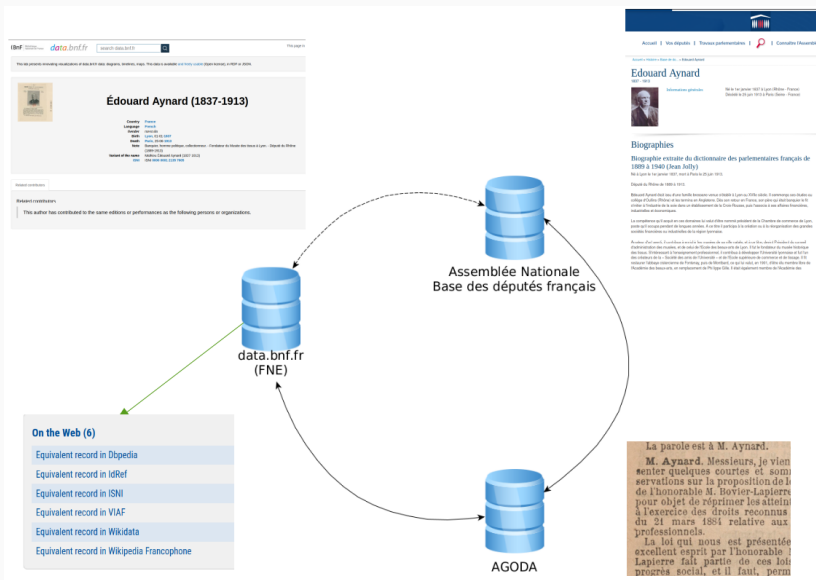
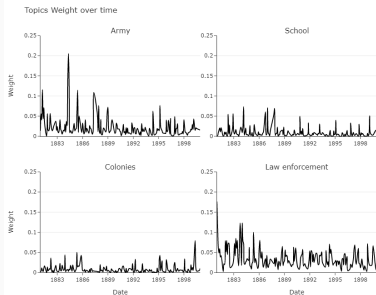


Figure – Linked data

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	adultere	recepissés
grele	conjugale	postes
secours	remarier	postaux
venir	separation	telegraphes
infortunes	indissolubilité	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

- 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)



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

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

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

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

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

Répertoire Github du projet : <https://github.com/mpuren/agoda>