

**Création et annotation d'un corpus de presse
historique à destination du TAL: avancées du projet
EMONTAL**

Nicolas Gutehrlé

C.R.I.T., Université de Franche-Comté

Séminaire du Centre Tesnière, 28/01/2022

Sommaire

1. Contextualisation

- 1.1 Introduction
- 1.2 Problématiques
- 1.3 Histoire numérique et TAL
- 1.4 Projet EMONTAL

2. Création du corpus

- 2.1 Collection des données
- 2.2 Suppression des césures
- 2.3 Correction de l'OCR
- 2.4 Analyse de la structure logique des documents
- 2.5 Segmentation en articles
- 2.6 Lisibilité du texte
- 2.7 Conversion au format Docbook

3. Conclusion et ouverture

Introduction

Introduction

- Les documents d'archives (presse, livres, parchemins, ...) représentent une source d'information inestimable pour l'étude de l'histoire
- Les campagnes de numérisation menées par les archives et bibliothèques ont eu un impact conséquent sur l'accès à ces documents :
 - A permis de préserver ces documents de manière plus sécurisée
 - A ouvert l'accès au grand public aux documents d'archives
 - A permis l'émergence de nouveaux domaines d'études, tel que l'histoire numérique

Histoire numérique

Apparue en 1998, l'histoire numérique (HN) se distingue de l'histoire traditionnelle de part son utilisation des documents numériques et des liens hypertextes :

- Elle permet une exploration non-linéaire et interactive de l'histoire
- Les projets de HN peuvent être mis à jour régulièrement avec de nouveaux documents ou fonctionnalités
- Elle demande cependant un investissement plus important de la part de l'utilisateur, qui doit apprendre à consulter les documents et à poser des questions de recherches

Histoire numérique

Initié par Edward L. Ayers en 1998, [The Valley of the Shadows](#)¹ est considéré comme le tout premier projet d'histoire numérique :

- Le projet traite de l'esclavage avant, pendant et après la guerre civile américaine (1861 - 1865) dans les comtés voisins d'Augusta (Virginie) et de Franklin (Pennsylvanie)
- S'est rapidement inscrit dans une démarche d'enseignement de l'histoire, en poussant les chercheurs à développer leur propre interprétation de l'histoire
- S'est complété au fil des années avec de nouveaux documents (via crowdsourcing) et nouvelles fonctionnalités (SIG)

Aujourd'hui, le site permet de consulter plus de 12 000 documents[12]

The Valley of the Shadows

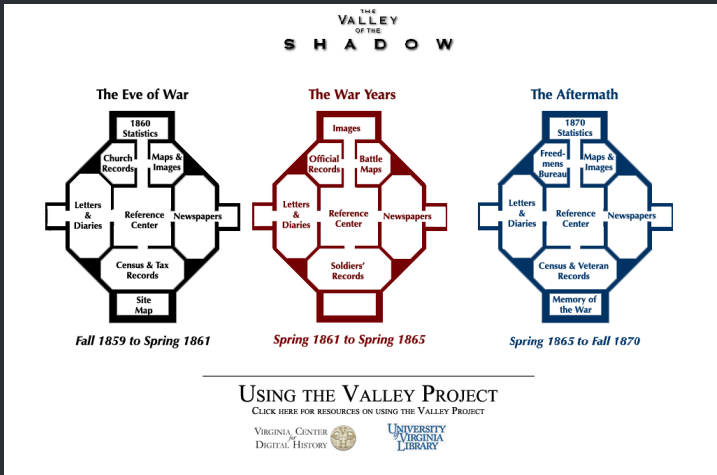


Figure – Interface de *The Valley of the Shadows*

Problématiques

Problématiques

La quantité massive de documents historiques désormais disponible soulève deux problèmes importants :

- Le contenu textuel des scans n'est pas directement accessible : il faut donc "transcrire" ces documents. La transcription manuelle est cependant exclue, au vu de la quantité de documents
- Les méthodes de recherches manuelles ne sont plus adaptées à cette masse d'information

OCR

Pour extraire le contenu textuel d'images, on peut employer les outils d'**Optical Character Recognition (OCR)** :

- La plupart des logiciels de scan ou de lecture de PDF tel qu'Adobe proposent aujourd'hui une fonction d'OCR
- D'autres logiciels ou programmes dédiés à l'OCR existent tels que ABBYFineReader ou Tesseract

OCR

Océriser des documents anciens représente plusieurs difficultés :

- En général, plus le document est ancien et/ou usé, plus les performances d'OCR sont basses
- La reconnaissance de textes manuscrits nécessite des outils dédiés : on parle alors de **Handwritten Text Recognition (HTR)**, avec des logiciels tels que **Transkribus** ou **eScriptorium**

La qualité de l'OCR est un critère essentiel pour permettre des traitements textuels ultérieurs de qualité

Big Data of the Past

La surabondance de documents historiques correspond à ce que Kaplan appelle le **Big Data of the Past**[7] :

- le terme Big Data désigne généralement la masse de données produites ces dernières années par l'ensemble des objets connectés
- ceci a poussé à développer de nouvelles méthodes pour traiter ces données, notamment par l'emploi d'algorithmes issus du Machine et Deep Learning

Big Data of the Past

Pourtant, l'histoire contient de nombreuses périodes où des quantités massives de données ont été produites :

- Les archives vénitiennes contiennent près de 80km de documents produits sous la République de Venise, connue pour son système administratif
- La cour criminelle de Londres (Old Bailey) a fait retranscrire de manière structurée les 197 000 procès qui y ont eu lieu entre 1674 et 1913, ce qui représente 127 millions de mots

Big Data of the Past

Ainsi comme pour le Big Data contemporain, il est nécessaire aujourd'hui de créer de nouveaux outils et protocoles pour structurer le flux de données issu du passé. On peut notamment citer les solutions apportées par les projets suivants :

- Trading Consequences
- impresso : Media Monitoring of the Past
- NewsEye

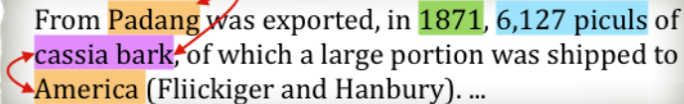
Trading Consequences

Trading Consequences

- Le projet **Trading Consequences (TC)** est une collaboration multi-institutionnelle et internationale entre des historiens de l'environnement au Canada et des informaticiens au Royaume-Uni.
- Il applique les méthodes de text-mining pour explorer des milliers de pages de documents historiques liés au commerce international de marchandises dans l'Empire britannique au cours du 19^e siècle, et son impact sur l'économie et l'environnement.

Trading Consequences

Trading Consequences s'intéresse particulièrement aux entités nommées de type Produits, Lieux et Date, et les relations entre elles :



From Padang was exported, in 1871, 6,127 piculs of cassia bark, of which a large portion was shipped to America (Fliückiger and Hanbury). ...

The image shows a text excerpt with several entities highlighted in different colors: 'Padang' (orange), '1871' (green), '6,127 piculs' (blue), 'cassia bark' (purple), and 'America' (orange). Red arrows indicate relationships: one arrow points from 'Padang' to 'cassia bark', and another points from 'cassia bark' to 'America'.

Figure 7: Excerpt from "Spices" (Ridley, 1912). The text mined information is highlighted in colour and relations are visualised using arrows.

Figure – Entités et les relations entre elles

Trading Consequences

La Reconnaissance d'Entités Nommées (REN) est effectuée à l'aide d'un lexique contenant 20 476 mentions de produits échangés durant la période traitée :

- Un système de règles permet d'identifier chaque entrée du lexique dans le corpus, ainsi que de résoudre les éventuelles ambiguïtés (ex : Markham le botaniste contre Markham, Ontario)
- Ce lexique a été construit sous la forme d'une ontologie afin de s'adapter aux multiples noms des produits (*rubber, caoutchouc, caou-chouc, Indian rubber, ...*), et de regrouper les entités par catégorie (Fruit, Textile, Pêche, ...)

Trading Consequences

Deux types de relations entre entités sont considérées :

- La relation produit-lieu (point de départ, d'arrivée, de transit)
- La relation produit-date (si celle-ci est disponible, sinon la date du document est utilisée)

La relation entre entité est détectée selon une règle simple : il y a relation entre ces entités si elles sont présentes dans la même phrase. Les relations entre entités de plusieurs phrases ne sont pas détectées.

Trading Consequences

La qualité de l'OCR a été une problématique :

- Les auteurs n'ont pas pu utiliser d'analyseur syntaxique pour l'extraction de relations à cause de la mauvaise qualité de l'OCR
- Beaucoup d'erreurs sont dues à la confusion entre le s long et le f ainsi que la césure qui divise un mot en deux
- Les autres erreurs sont dues à la mauvaise qualité du papier, de l'utilisation d'anciennes polices inconnues du logiciel d'océrisation, et de la mise en page (tableaux, entêtes, pieds de pages, etc.)

Impresso & NewsEye

Impresso & NewsEye

impresso : Media Monitoring of the Past. Mining 200 years of historical newspapers est un projet financé par le Fonds national suisse de la recherche scientifique qui s'est déroulé de 2017 à 2020 :

- Soutenu par un consortium interdisciplinaire, il vise à fournir un cadre pour l'extraction, le traitement, la mise en relation et l'exploration de données provenant d'archives de médias imprimés.
- La plateforme donne accès aux journaux publiés en Suisse et au Luxembourg au XXème siècle, rédigés en français, allemand et anglais²

Impresso & NewsEye

NewsEye est un projet Horizon 2020 qui s'est déroulé de 2018 à 2021. Il est coordonné par l'université de La Rochelle et est également soutenu par un consortium interdisciplinaire de chercheurs en sciences sociales et informatiques international

- De même, ce projet vise à créer de nouveaux outils et méthodes pour "changer la façon dont les données du patrimoine numérique européen sont (re)recherchées, accessibles, utilisées et analysées"[1].
- La plateforme donne accès aux journaux publiés en France et en Allemagne entre 1850 et 1950³

Impresso & NewsEye

Les deux projets visent à développer de nouvelles normes pour traiter, accéder et étudier des corpus historiques massifs. Pour ce faire, les deux projets appliquent des techniques issues du Machine et Deep Learning, qui sont habituellement utilisées pour traiter les Big Data contemporains :

- Reconnaissance d'Entités Nommées
- Topic Modelling
- Word embeddings (unilangue, multilingue, diachroniques)

Impresso & NewsEye

The screenshot displays the Impresso search interface. On the left, there are filters for 'SEARCH ARTICLES', 'SEARCH IMAGES', and 'NGRAMS'. The search terms 'titanic' and 'Belfast' are entered. Below this, there are options for 'Frontpage' and 'FIND SIMILAR WORDS'. A 'PUBLICATION DATE' filter shows a bar chart with a 'SUM' button. Below that is a 'FILTER BY CONTENT LENGTH' bar chart. Further down are filters for 'LANGUAGE OF ARTICLES' (French: 57 results, German: 8 results) and 'NEWSPAPER TITLES' (L'Express: 15 results, L'Impartial: 15 results). The main search results area shows '65 articles found containing **titanic** mentioning BELFAST'. The first result is from 'Luxemburger Wort' dated Wednesday, July 31, 1912, with the title 'Großbritannien.' The second result is from 'L'Express' dated Friday, May 23, 2014, with the title 'AU BERCEAU DU «TITANIC»'. Each result includes a thumbnail image and a 'VIEW' button.

Figure – Recherche de l'entité "Titanic" dans la plateforme *impresso*

Impresso & NewsEye

Afin d'évaluer leurs plateformes respectives, chaque projet a défini des cas d'études :

- l'expression de la résistance à l'idée européenne de la fin du 19ème siècle à 1950 dans les journaux francophones, germanophones et anglophones (impresso)
- le statut de la femme au XIXe siècle français tel qu'il est dépeint dans la presse (NewsEye)
- l'évolution du vocabulaire lié aux questions de nationalisme et de révolution (NewsEye)
- l'émergence de la profession de journaliste (NewsEye)

Projet EMONTAL

EMONTAL

EMONTAL

Extraction et Modélisation ONTologique des Acteurs et Lieux pour la valorisation du patrimoine de Bourgogne Franche-Comté

Le projet EMONTAL :

- S'inscrit dans une démarche similaire à Trading Consequences, impresso et NewsEye
- vise l'exploitation et la valorisation des documents issus du patrimoine de Bourgogne Franche-Comté
- est financé par la région Bourgogne Franche-Comté sur la période 2020 - 2023
- est dirigé par Dr. Iana Atanassova

EMONTAL

Ce projet a pour objectifs de développer :

- des méthodologies pour l'extraction automatique d'informations dans les documents d'archives (journaux, comptes rendus, documents administratifs, ...)
- des interfaces pour assister l'utilisateur dans l'exploration de ces documents

Une première étape importante est de constituer un corpus de documents annotés afin de pouvoir y appliquer des traitements issus du TAL

Collection des données

Corpus

Notre corpus est constitué de documents de presses et de périodiques publiés au XXème siècle appartenant aux fonds régionaux Franche-Comté et Bourgogne ⁴ de Gallica (BnF) ⁵



Figure – Extrait de la première page du second numéro du journal communiste *Le Semeur*, publié le 23 avril 1932

Corpus

Chaque document hébergé dans Gallica possède un identifiant nommé **ark**, commun à tous les services de la BnF. Il existe deux types d'identifiant ark :

bptXXXXXXXXXX : un ark préfixé par **bpt** indique un document

cbXXXXXXXXXX : un ark préfixé par **cb** représente une collection de documents

L'OCR est disponible pour une majeure partie des documents des deux fonds régionaux ayant le préfixe **bpt**.

Metadonnées

Les métadonnées des documents collectés sur Gallica sont stockées au format **Dublin Core** :

dc :identifiant : l'url vers le document sur la plateforme Gallica

dc :date : la date de publication du document

dc :title : le titre du document

dc :creator : le créateur du document

...

XML ALTO

Quand il est disponible, l'OCR des documents est enregistré au format **XML ALTO**. La mise en page et le contenu textuel des documents est présenté comme suit :

- Les lignes de texte sont contenues dans des tags **TextLine**, qui eux-mêmes contiennent des tags **String** pour les mots et des tags **SP** pour les espaces
- La valeur textuelle d'un tag String est contenue dans son attribut **content**
- Les tags Textline sont regroupés dans des tags **TextBlock**

XML ALTO

Les tags TextBlock et TextLine ont les attributs suivants :

Id l'identifiant du tag

Height, Width la hauteur et largeur du texte

Vpos la position verticale du texte sur la page. Plus la valeur est haute, plus le texte est bas sur la page

Hpos la position horizontale du texte sur la page. Plus la valeur est haute, plus le texte est situé sur la droite de la page

Language la langue du texte (uniquement pour les tags Textblock)

XML ALTO description

```
<textblock height="139" hpos="727" id="PAG_00000001_TB000003" language="fr" vpos="1064" width="531">
  <shape>
    <polygon points="738,1122 1268,1122 1268,1256 738,1256 738,1122">
    </polygon>
  </shape>
  <textline height="55" hpos="743" id="PAG_00000001_TL000005" vpos="1076" width="498">
    <string content="Directeur" height="37" hpos="743" id="PAG_00000001_ST000014" vpos="1077" wc="0.5811111331" width="232">
    </string>
    <sp hpos="976" id="PAG_00000001_SP000010" vpos="1090" width="18">
    </sp>
    <string content="politique" height="51" hpos="995" id="PAG_00000001_ST000015" vpos="1079" wc="0.7722222209" width="212">
    </string>
    <sp hpos="1208" id="PAG_00000001_SP000011" vpos="1093" width="21">
    </sp>
    <string content=";" height="25" hpos="1230" id="PAG_00000001_ST000016" vpos="1094" wc="0.2899999917" width="11">
    </string>
  </textline>
  <textline height="50" hpos="794" id="PAG_00000001_TL000006" vpos="1148" width="392">
    <string content="Henri" height="44" hpos="794" id="PAG_00000001_ST000017" vpos="1148" wc="0.6859999895" width="153">
    </string>
    <sp hpos="947" id="PAG_00000001_SP000012" vpos="1150" width="26">
    </sp>
    <string content="JACOB" height="48" hpos="973" id="PAG_00000001_ST000018" vpos="1150" wc="0.8199999928" width="213">
    </string>
  </textline>
</textblock>
```

Figure – Extrait de la transcription en XML ALTO du *Semeur* publié le 23 avril 1932

Statistiques

Tags / Corpus	Fond Franche-Comté	Fond Bourgogne	Total
Collections	46	118	164
Publications	2650	6120	8770
TextBlocks	3 735 851	8 770 516	12 506 367
TextLines	11 381 384	26 584 643	37 966 027
Strings	83 063 089	211 601 426	294 664 515
Pages	255 908	709 250	965 158

Table – Statistiques des données collectées depuis Gallica

Suppression des césures

Suppression des césures

En typographie, le tiret peut servir à combiner ou à diviser un mot, notamment en fin de ligne. Le format XML ALTO utilise le tag **HYP** pour représenter le tiret séparateur. Ce tag ne possède que quatre attributs :

content : le symbole du tiret ("-")

hpos : position horizontale dans le document

vpos : position verticale dans le document

width : largeur du texte

Suppression des césures

L'attribut **content** contient la valeur textuelle d'un tag. Cependant, dans le contexte d'un tag HYP, cet attribut ne contient que la portion de texte qui précède ou suit le tiret, et pas le mot complet.

La première étape de pre-traitements consiste donc à :

- mettre à jour les tags Strings environnants un tag HYP afin que leur valeur représente le mot complet
- supprimer les tags HYP des documents

Suppression des césures

```
<string content="fau-" height="33" hpos="4080" id="PAG_00000003_ST002258" stylerefs="TXT_1"
subs_content="faudrait" subs_type="HypPart1" vpos="5128" wc="1" width="79">
</string>
<hyp content="-" hpos="4159" vpos="5161" width="30">
</hyp>
</textline>
<textline height="43" hpos="2967" id="PAG_00000003_TL000285" stylerefs="TXT_1" vpos="5167" width="1193">
<string content="drait" height="33" hpos="2967" id="PAG_00000003_ST002259" stylerefs="TXT_1"
subs_content="faudrait" subs_type="HypPart2" vpos="5167" wc="1" width="93">
</string>
```

Figure – Exemple d'utilisation du tag HYP dans un fichier XML ALTO

Correction de l'OCR

Correction de l'OCR

La correction de l'OCR est une étape nécessaire, puisqu'elle permet d'améliorer les performances de toutes les tâches suivantes (REN, Topic Modelling, ...). De simples corrections peuvent corriger une grande partie des erreurs :

- [10] précise que 81.49% des erreurs d'OCR dans leur corpus peuvent être corrigés avec une distance d'édition de 1 ou 2
- 72% de ces erreurs d'OCR de TC ont pu être traitées à l'aide de dictionnaires, ce qui a augmenté la qualité générale de l'OCR de 12%

Correction de l'OCR

Nous avons conçu un système de règles pour corriger l'OCR. Ces règles déterminent si l'une des trois opérations suivantes doit être appliquées aux tags **Strings** :

Conserver : Opération par défaut. Le contenu du tag String est gardé tel quel

Supprimer : le tag String est supprimé du fichier XML

Substituer : le contenu du tag String est modifié

Nous utilisons l'algorithme **SymSpell (Symmetric Delete spelling correction algorithm)[5]**⁶ pour identifier des corrections possibles de transcriptions erronées.

Correction de l'OCR

Notre système emploie plusieurs ressources :

- une liste de mots vides en français⁷
- Le dictionnaire de base d'unigrammes en français de SymSpell
- Un dictionnaire d'unigrammes en français du 19ème et 20ème siècles, généré à l'aide de SymSpell à partir du corpus ICDAR2017[4]

Correction de l'OCR

La première étape du système consiste à extraire les propriétés du contenu textuel de chaque tag String d'un document. Ces propriétés sont enregistrées dans une matrice :

Propriété	Description
word_length	nombre de caractères dans le mot
stw_capital	Vrai si le mot commence par une lettre capitale, sinon Faux
...	...
operation	opération à appliquer à ce mot. Par défaut, la valeur est Conserver
correction	le text corrigé. Uniquement utilisé si l'opération est Substituer

Table – Extraits des propriétés nécessaires aux post-traitement de l'OCR

Correction de l'OCR

Les règles suivantes sont ensuite appliquées à la matrice pour définir les opérations à appliquer. Les règles 1 et 2 identifient les candidats pour la **Suppression**, tandis que la règle 3 identifie les candidats pour la **Substitution** :

Règle	Objectif	Conditions	Opération
1	Supprimer mots non-alphanumériques	W.non_alpha_prop > 75 et W.is_punct est Faux	Supprimer
2	Supprimer les mots longs d'un seul caractère	W.is_oneletter_word est Faux et W.word_length = 1 et W.is_punct est Faux et W.is_digit est Faux	Supprimer
3	Proposer une correction	W.stw_capital est Faux et W.ends_punct est Faux et W.is_punct est Faux et W.is_digit est Faux et W.stw_elision est Faux et W.operation n'est pas Supprimer	Substituer

Table – Règles de post-traitement de l'OCR

Frame Title

Les mots identifiés par la règle 3 sont traités de la manière suivante :

- une expression régulière corrige les caractères qui se répètent plus de deux fois (ex : **mercrediiii** => **mercredi**)
- SymSpell est ensuite employé pour proposer une correction, avec une distance d'édition maximale de 1. Le mot est retourné à l'identique si aucune correction n'est proposée

Les opérations sont ensuite appliquées à chaque tag String du fichier XML ALTO d'origine.

Analyse de la structure logique des documents

Analyse de la structure logique des documents

L'extraction du contenu textuel d'une image se fait au moins en trois étapes :

Optical Character Recognition (OCR) pour extraire le texte des images

Physical Layout Analysis (PLA) pour identifier

- les régions physiques du texte et leurs limites
- l'ordre de lecture du document

Logical Layout Analysis (LLA) pour identifier la structure logique des documents (titres, en-têtes, notes de bas de page, paragraphes, ...)

- Ces étiquettes peuvent intégrer une ou plusieurs régions physiques extraites par l'étape PLA

LLA appliquée aux documents historiques

Les systèmes existants de LLA font appel à diverses méthodes qui vont :

- des systèmes heuristiques (par exemple, [9], [11])
- des systèmes hybrides (par exemple [8])
- aux architectures plus récentes utilisant des réseaux de neurones (par exemple, [2], [13])

LLA appliquée aux documents historiques

Cependant, les approches les plus courantes de LLA ne sont pas adaptées aux documents historiques, car la présentation du document change au fil du temps :

- la mise en page et la structure d'une publicité dans un même journal peuvent présenter des changements importants sur plusieurs années.

Les systèmes de LLA appliqués aux documents historiques doivent alors tenir compte de l'aspect diachronique de leur mise en page et s'adapter aux changements, comme dans [3].

Methodologie

Notre système à base de règles attribue des **catégories logiques** aux tags TextBlock et TextLine des documents au format XML ALTO. Pour cette tâche, nous définissons les catégories d'annotation suivantes :

catégorie pour TextBlock : Text, Title, Header, Other

catégorie pour TextLine : Text, Firstline, Title, Header, Other

Methodologie

Important

- L'étiquette "Firstline" doit être comprise comme "première ligne du paragraphe". Ainsi, toute balise TextLine étiquetée Firstline indiquera le début d'un paragraphe.
- Certains éléments, comme les tableaux ou les publicités, ne sont pas pertinents pour notre étude. Ces éléments sont étiquetés comme "Autres" et sont ignorés pour l'évaluation.
- Le système attribue l'étiquette "Autre" à un bloc de texte ou à une balise de ligne de texte uniquement si aucune autre étiquette ne lui a déjà été attribuée.

Dataset

Nous avons constitué un corpus d'étude à partir de documents du "Fond régional : Franche-Comté". Ces documents ont été regroupés en trois catégories de mise en page (MeP), puis divisé en un train et test sets :

- 1c** MeP sur une colonne, comme les livres
- 2c** MeP sur deux colonnes, comme dans certains magazines
- 3c+** MeP sur au moins trois colonnes, comme dans les journaux

Dataset / (MeP)	1c	2c	3c+	Total
Train	18	5	25	48
Test	2	2	2	6

Table – Distribution des documents selon la mise en page dans le train et test set

Catégories d'annotation

	Label	Train		Test	
		Count	Percentage	Count	Percentage
TextBlock	Text	2 064	45.724	1 102	80.203
	Title	429	9.503	90	6.550
	Header	333	7.377	53	3.857
	Other	1 686	37.35	128	9.314
TextLine	Text	36 272	70.138	6 648	75.881
	Firstline	9 785	18.921	1 563	17.840
	Title	1 820	3.519	234	2.670
	Header	740	1.430	115	1.312
	Other	3 098	5.989	201	2.293

Table – Distribution des tags TextBlock et TextLine dans le train et test set

XML ALTO description

Certaines tags TextBlock possèdent également un attribut **Type**. Cet attribut est utile car il contient les catégories logiques des lignes dans le bloc. Malheureusement, les TextBlock avec un attribut Type sont rares dans notre jeu de données :

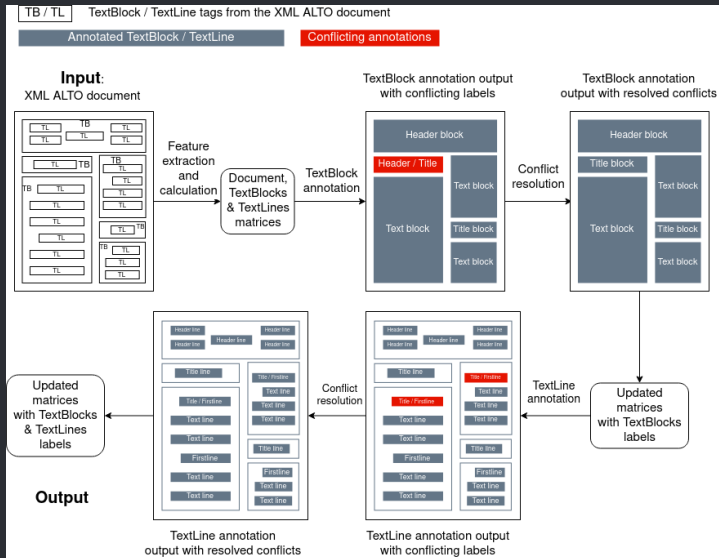
Attribut Type	Train		Test	
	Count	Perc.	Count	Perc.
No attribute	4 514	97.96	1 423	98.48
illegible	79	1.71	15	1.04
titre1	15	0.33	0	0
advertisement	0	0	4	0.28
table	0	0	2	0.14
textStamped	0	0	1	0.07

Table – Distribution de l'attribut Type des tags TextBlock dans le train et le test set

Description de l'algorithme

- Les règles sont appliquées aux documents indépendamment de la catégorie de mise en page à laquelle ils appartiennent.
- Le système attribue l'étiquette Other à un tag TextBlock ou TextLine uniquement si aucune autre étiquette ne lui a déjà été attribuée.
- L'annotation de TextBlock est une étape intermédiaire nécessaire dans l'algorithme qui remplit les attributs de type manquants.

Description de l'algorithme



Extraction de caractéristiques du document

Feature	Description	TextLine	TextBlock	Document
<i>page</i>	page number of the page containing the element	X	X	
<i>blockType</i>	type of the block	X	X	
<i>wordCount</i>	number of words	X	X	
<i>precedingSpace, followingSpace</i>	spaces between the element and those before and after it	X	X	
<i>capitalPro, digitProp</i>	proportion of capital letters and digits	X	X	
<i>height, width</i>	height and width values of the line	X		
<i>hpos, vpos</i>	coordinates of the line on the page, i.e. its horizontal and vertical position	X		
<i>diffHpos</i>	the difference between <i>hpos</i> and the median <i>hpos</i> value in the block	X		
<i>stwCapital, stwDigit</i>	True if the line starts either by a capital letter or a number, False otherwise	X		
<i>headerMark1</i>	True if the element contains the word "Page" or a dash sign. False otherwise.	X	X	
<i>headerMark2</i>	True if the element contains a date, a currency, an address. False otherwise.	X	X	
<i>simTitle</i>	similarity of the line with the title of the document, calculated by the Levenshtein distance	X		
<i>simHeaderSet</i>	highest similarity of the line with the words contained in the header words set, calculated by the Levenshtein distance	X		
<i>firsthpos, firstvpos</i>	coordinates of the first line of the block		X	
<i>lasthpos, lastvpos</i>	coordinates of the last line of the block		X	
<i>linecount</i>	number of lines		X	
<i>medHeight, medWidth</i>	median line height and line width		X	X
<i>medHpos, medVpos</i>	median <i>hpos</i> and <i>vpos</i> values in the block		X	
<i>medWordCount, med-LineSpace</i>	median number of words by line and the median space between lines in the block		X	X
<i>wordRatio</i>	number of words by line		X	

Figure – Exemples de caractéristiques extraites pour les tags TextLine, TextBlock et le document

Exemple de règles d'annotation des tags TextBlocks

Règle	Condition	Catégorie
1	$(B.\text{linecount} > D.\text{medLineCount})$ ou $(B.\text{wordCount} > D.\text{medWordCount}/3)$	Text
2	Previous et next TextBlocks est Text et $(B.\text{linecount} < D.\text{medLineCount})$ and $(B.\text{medHeight} < D.\text{medBlockHeight})$	Text
3	Previous et next TextBlocks est Text et B n'est pas Text et $(B.\text{linecount} < 4)$ et $(B.\text{precedingSpace} > D.\text{medBlockSpace})$ ou $(B.\text{followingSpace} > D.\text{medBlockSpace})$	Title

Table – Exemple de règles d'annotation des tags TextBlocks

Résolution de conflit pour les tags TextBlock

Règle	Condition	Catégorie
1	Annotation conflictuelle : Header et (Text ou Title) : (<i>B.linecount</i> < 15) et (<i>B.wordCount</i> < 50) Otherwise	Header Text / Title
2	Annotation conflictuelle : Text et Title : <i>B.medHeight</i> > <i>D.medBlockHeight</i> / 2 Otherwise	Title Text

Table – Règles pour résoudre les conflit d'annotation des tags TextBlock

Exemple de règles d'annotation des tags TextLine

Règle	Condition	Catégorie
1	L.precedingSpace = 0 et L.followingSpace > D.medLineSpace et L.simTitle < 60 et L.simHeaderSet < 60 et L.stwCapital	Title
5	L.hpos > B.medHpos et L.diffHpos < 105 et (L.stwCapital ou L.stwDigit)	Firstline
10	Aucune des règle de 1-9 n'est Vraie	Text

Table – Exemple de règles d'annotation des tags TextLine

Résolution de conflit pour les tags TextLine

Règle	Condition	Catégorie
1	Annotation conflictuelle : Header et autre catégorie : Previous TextLine est Header et next TextLine is Header	Header
2	Annotation conflictuelle : Title et FirstLine : L.followingSpace < B.medLineSpace et L.capitalProp < 15	Title

Table – Règles pour résoudre les conflit d'annotation des tags TextLine

Evaluation de l'annotation des tags TextBlock

Cat	Text			Title			Header		
	P	R	F1	P	R	F1	P	R	F1
1c	0.947	0.938	0.942	0.312	0.357	0.333	0.679	0.373	0.476
2c	0.973	0.989	0.981	0.899	1.000	0.947	1.000	0.271	0.411
3c+	0.958	0.973	0.965	0.589	0.560	0.551	0.500	0.250	0.333
Mean	0.959	0.966	0.962	0.600	0.639	0.610	0.726	0.298	0.406

Table – Précision, Rappel et score F1 pour l'annotation des tags TextBlock

Evaluation de l'annotation des tags TextBlock

- Les règles d'annotation des tags TextBlock fonctionnent le mieux pour la mise en page 2c
- L'annotation des Title obtient un score F1 de 0.61 en moyenne mais atteint un score de 0.94 pour la catégorie 2c
- L'annotation des Header obtient une bonne précision (0.72) mais un mauvais rappel (0.29)

Evaluation de l'annotation des tags TextLine

Cat	Text			Title		
	P	R	F1	P	R	F1
1c	0.979	0.986	0.983	0.354	0.720	0.473
2c	0.961	0.995	0.978	0.746	0.765	0.747
3c+	0.975	0.992	0.983	0.703	0.702	0.702
Mean	0.969	0.991	0.979	0.595	0.733	0.639

Cat	Firstline			Header		
	P	R	F1	P	R	F1
1c	0.943	0.854	0.895	0.909	0.598	0.721
2c	0.955	0.859	0.902	1.000	0.118	0.197
3c+	0.952	0.877	0.913	0.500	0.400	0.444
Mean	0.949	0.861	0.902	0.803	0.348	0.435

Table – Précision, Rappel et score F1 pour l'annotation des tags TextLine

Evaluation de l'annotation des tags TextLine

- Comme pour l'annotation des tags TextBlock, l'annotation des tags TextLine fonctionne le mieux pour la mise en page 2c
- L'annotation des Title fonctionne le moins bien pour la mise en page 1c et obtient un score F1 moyen de 0.63 pour toute les mises en page
- L'annotation de Firstline obtient un score F1 supérieur à 0.9 en moyenne
- De même, l'annotation des Header obtient un bon score de précision (0.80), mais un mauvais rappel (0.34), ce qui suggère un manque de règle

Evaluation

Deux erreurs récurrentes se démarquent :

Mauvaise annotation des TextBlock : comme toute ligne d'un bloc Title ou Header hérite de cette annotation, la précision de l'annotation des TextBlock est un facteur important pour la performance globale de l'algorithme

Confusion Titre-Première ligne : la plupart des Title mal étiquetés comme Firstline sont des titres courts de sous-section. En tant que tels, ils sont similaires à d'autres lignes de texte en termes de typographie, et sont difficiles à détecter avec les caractéristiques que nous utilisons. Cette confusion se produit principalement dans les documents des catégories 2c et 3c+.

Ouverture

- À notre connaissance, il n'existe pas de jeux de données annotés suffisamment grands pour entraîner des modèles d'apprentissage profond pour la LLA. Pour cette raison, l'algorithme à base de règles que nous proposons vise principalement à produire des jeux de données annotés suffisamment grands pour envisager de telles méthodes.
- Le jeu de données créé pour cette étude est disponible sur Zenodo[6]
- Cette étude a fait l'objet d'un article à paraître. Il a été présenté dans le cadre du workshop NLP4DH en décembre 2021

Ouverture

- La comparaison entre les performances de ces règles et les résultats des récentes architectures d'apprentissage profond fera l'objet de futurs travaux.
- Afin de créer des ensembles de règles qui traitent l'aspect diachronique de la LLA, nous prévoyons dans de futurs travaux d'appliquer des algorithmes d'apprentissage de règles pour généraliser la création de règles.

Segmentation en articles

Segmentation en articles

La segmentation en article est réalisée à partir des résultats de la LLA :

- un article est constitué par un titre ou groupe de titres, suivi de son texte jusqu'à rencontrer un autre titre
- un identifiant est associé à chaque article, permettant de suivre l'article sur plusieurs pages

Division des paragraphes en phrases

Division des paragraphes en phrases

Par défaut, l'algorithme divise un paragraphe en phrases en se basant sur les ponctuations `.!?`. Des règles permettent de spécifier des exceptions, comme :

- N.B., Nota. Bene
- P.S. p.s., p. s.
- U.R.S.S., C.G.T.

Lisibilité du texte

Lisibilité du texte

Malgré la post-correction de l'OCR, beaucoup de textes dans le corpus restent peu lisibles. Il nous faut donc définir une mesure pour indiquer le degré de lisibilité d'un texte. Pour cela, nous calculons **la proportion de syllabes connues dans le texte**, une syllabe étant connue **si elle est présente dans un dictionnaire de syllabes**.

Lisibilité du texte

Nous avons créé un dictionnaire de syllabes françaises contenant 33 000 entrées à partir du dataset ICDAR 2017 à l'aide de la librairie `pyphen`⁸ :

- Pour un mot donné, `pyphen` indique toutes les positions possibles pour insérer un tiret, indiquant ainsi toutes les syllabes de ce mot
- `pyphen` utilise des dictionnaires Hunspell pour la division de mots qui sont notamment employés par LibreOffice, Google Chrome ou Mozilla Firefox⁹

Lisibilité du texte

Le calcul de la lisibilité d'un texte est réalisé comme suit :

1. Une première étape consiste à normaliser le texte en le passant en minuscule, en supprimant les élisions et en le tokenisant
2. Chaque token est ensuite divisé en syllabes à l'aide de `pyphen`
3. Enfin, le programme calcule la proportion de syllabes dans le texte qui sont aussi présentes dans le dictionnaire.

Le système retourne deux valeurs :

- un score de lisibilité entre 0 et 1, qui correspond à la proportion de syllabes connues dans le texte
- une note de lisibilité, qui assigne le score à une certaine catégorie

Lisibilité du texte

Score	Note	Description
0 - 20	E	Très mauvais : le texte est (presque) entièrement illisible
21 - 40	D	Mauvais : une majeure partie du texte est illisible
41 - 60	C	Moyen : une portion du texte est lisible seulement
61 - 80	B	Bon : une majeure partie du texte lisible
81 - 100	A	Très bon : le texte est (presque) entièrement lisible

Table – Correspondance entre la note et le score de lisibilité

Conversion au format Docbook

Conversion au format Docbook

Après application de ces traitements, les documents du corpus sont convertis au format XML Docbook. Ils ont la structure suivante :

metadata contient les métadonnées relatives au document

content contient le contenu textuel du document

Conversion au format Docbook

Le tag **metadata** contient les éléments suivants :

- ark** : l'identifiant ark du document, similaire à celui de Gallica
- identifrier** : l'url vers le document sur Gallica
- date** : la date de publication d'origine du document
- title** : le titre du document
- publisher** : l'éditeur original du document
- creator** : le créateur du document
- source** : l'endroit d'où le document est originaire
- typedoc** : le type de document
- dewey** : la catégorie du document selon la classification décimale de Dewey
- image_url** : l'url vers le document numérisé en haute définition

Conversion au format Docbook

```
<document>
  <metadata>
    <ark>
      bpt6k881591v
    </ark>
    <identifiant>
      https://gallica.bnf.fr/ark:/12148/bpt6k881591v
    </identifiant>
    <date>
      1944-08
    </date>
    <title>
      La Haute-Saône libre : organe départemental du Front national
    </title>
    <contributor/>
    <publisher>
      [s.n.]
    </publisher>
    <langue/>
    <creator>
      Front national de lutte pour la libération et l'indépendance de la France. Auteur du texte
    </creator>
    <source>
      Bibliothèque nationale de France, département Réserve des livres rares, RES-G-1470 (626)
    </source>
    <typedoc>
      fascicule
    </typedoc>
    <nqamoyen>
      86.99
    </nqamoyen>
    <dewey/>
    <image_url>
      https://gallica.bnf.fr/ark:/12148/bpt6k881591v/highres
    </image_url>
  </metadata>
```

Figure – Extrait de la section metadata dans un fichier docbook (*La Haute-Saône Libre*, août 1944)

Conversion au format Docbook

Le tag **content** contient les éléments suivants :

page : le contenu d'une page du document

header : l'entête de cette page

articles : contient tous les articles qui débutent dans cette page

article : le contenu d'un article. Peut courir sur plusieurs pages

title : le titre de cet article

text : le contenu textuel de cet article

para : un paragraphe

sent : une phrase dans ce paragraphe

Conversion au format Docbook

Les tags <header>, <article>, <title>, <text>, <para> and <sent> ont les attributs suivants :

id : l'identifiant du tag

readability : la note de lisibilité du texte (de A à E)

readability_score : le score de lisibilité du texte (entre 0 et 1)

Les tags <para> ont également un attribut **block_id**, qui permet de les associer au tag TextBlock dans le fichier XML ALTO d'origine

Conversion au format Docbook

```
<content>
<page id="1">
  <header id="header_1" readability="E" readability_score="0"/>
  <articles>
    <article id="article_01" readability="A" readability_score="0.83">
      <title id="title_01" readability="D" readability_score="0.27">
        M lj.lt fui l.'jü lloi):»;v 'l;0 *: , ! ! 11
      </title>
      <text id="text_01" readability="A" readability_score="0.85">
        <para block_id="PAG_00000001_TB000001" id="para_1" readability="A" readability_score="0.85">
          <sent id="sent_1" readability="A" readability_score="0.88">
            Tous debout et au combat !
          </sent>
          <sent id="sent_2" readability="A" readability_score="0.82">
            Haute-Saône Libre pas morte ; après un long sileuce, elle réparait plus vivante que jamais ;
            compatriotes, d'en faire une réalité.
          </sent>
          <sent id="sent_3" readability="A" readability_score="0.9">
            Les armées alliées sont à nos portes : suivons le mot d'ordre du général de Gualle.
          </sent>
        </para>
      </text>
    </article>
  </articles>
</page>
</content>
```

Figure – Extrait de la section content d'un fichier docbook (*La Haute-Saône Libre*, août 1944)

Statistiques du corpus

Elements / Corpus	Fond Franche-Comté	Fond Bourgogne	Total
Title	607 705	1 450 634	2 058 339
Header	246 349	683 835	930 184
Article	607 705	1 450 634	2 058 339
Para	2 028 416	4 411 038	6 439 454
Sent	4 398 498	10 596 716	14 995 214
Other	92 286	218 894	311 180
Lisibilité	0.77%	0.75%	0.76%

Table – Statistiques du corpus final

Conclusion

Conclusion

Conclusion

- Nous avons présenté les différentes étapes nécessaires pour la constitution d'un corpus de presse historique
- Ce travail a été réalisé dans le cadre du projet EMONTAL, qui vise à développer des outils et interfaces pour la valorisation du patrimoine de Bourgogne Franche-Comté
- Ce projet s'inscrit à la suite de grands projets tels que Trading Consequences, impresso ou NewsEye, qui ont montré l'intérêt d'appliquer les techniques issues du TAL pour structurer l'immense flux de données historiques disponible aujourd'hui

Prochaines étapes

- Intégration des données à un serveur Solr
- Appliquer la Reconnaissance d'Entités Nommées au corpus
- Etablir la méthodologie pour l'extraction d'information liées aux acteurs et lieux :
 - Conception d'une ontologie
 - Open Information Extraction

Merci pour votre attention !

Bibliographie

- [1] *A propos*. 2020. URL : <https://www.newseye.eu/fr/a-propos/>.
- [2] Hanna Abi Akl, Anubhav Gupta et Dominique Mariko. "FinTOC-2019 Shared Task : Finding Title in Text Blocks". In : *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)*. Turku, Finland : Linköping University Electronic Press, sept. 2019, p. 58-62. url : <https://www.aclweb.org/anthology/W19-6408>.
- [3] Raphaël Barman et al. "Combining Visual and Textual Features for Semantic Segmentation of Historical Newspapers". In : *ArXiv abs/2002.06144* (2020).
- [4] Guillaume Chiron et al. "ICDAR2017 Competition on Post-OCR Text Correction". In : *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Kyoto, France : IEEE, nov. 2017, p. 1423-1428. doi : [10.1109/icdar.2017.232](https://hal.archives-ouvertes.fr/hal-03025499). url : <https://hal.archives-ouvertes.fr/hal-03025499>.
- [5] Wolf Garbe. *T1000x Faster Spelling Correction algorithm*. url : <https://seekstorm.com/blog/1000x-spelling-correction/>.
- [6] Nicolas Gutehrlé et Iana Atanassova. *Dataset for Logical-layout analysis on French historical newspapers*. Version 1.0. Zenodo, oct. 2021. doi : [10.5281/zenodo.5752440](https://doi.org/10.5281/zenodo.5752440). url : <https://doi.org/10.5281/zenodo.5752440>.
- [7] Frédéric Kaplan et Isabella di Lenardo. "Big Data of the Past". In : *Frontiers in Digital Humanities* 4 (2017), p. 12. issn : 2297-2668. doi : [10.3389/fdigh.2017.00012](https://www.frontiersin.org/article/10.3389/fdigh.2017.00012). url : <https://www.frontiersin.org/article/10.3389/fdigh.2017.00012>.
- [8] S. Klampfl et Roman Kern. "An Unsupervised Machine Learning Approach to Body Text and Table of Contents Extraction from Digital Scientific Articles". In : *TPDL*. 2013.
- [9] Anoop Namboodiri et Anil Jain. "Document Structure and Layout Analysis". In : mars 2007, p. 29-48. isbn : 978-1-84628-501-1. doi : [10.1007/978-1-84628-726-8_2](https://doi.org/10.1007/978-1-84628-726-8_2).
- [10] Thi-Tuyet-Hai Nguyen et al. "Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing". In : *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Champaign, France : IEEE, juin 2019, p. 29-38. doi : [10.1109/jcdl.2019.00015](https://hal.archives-ouvertes.fr/hal-02519302). url : <https://hal.archives-ouvertes.fr/hal-02519302>.