Information extraction from unstructured documents for the valorisation of historical periodicals : application to the heritage of the Bourgogne Franche-Comté Region in France

Nicolas Gutehrlé

UNIVERSITÉ DE FRANCHE-COMTÉ

THESE DE DOCTORAT DE L'ETABLISSEMENT UNIVERSITE BOURGOGNE FRANCHE-COMTE

PREPAREE A UNIVERSITÉ DE FRANCHE-COMTÉ

Ecole doctorale n°592

ÉCOLE DOCTORALE LECLA

Doctorat de Sciences du Langage, mention Traitement Automatique des Langues

Par

Nicolas GUTEHRLÉ

Sous la direction de Iana ATANASSOVA

Information extraction from unstructured documents for the valorisation of historical periodicals. Application to the heritage of the Bourgogne Franche-Comté Region in France

Extraction d'informations appliquée aux documents non-structurés pour la valorisation de périodiques historiques : application au patrimoine de la région Bourgogne Franche-Comté en France

Thèse présentée et soutenue à Besançon, le 21/06/2024

Composition du Jury :

Professeur Doucet, Antoine	L3i, La Rochelle Université (France)	Président
Professeur Bachimont, Bruno	COSTECH, Université de technologie de Compiègne (France)	Rapporteur
Docteur Pecina, Pavel	ÚFAL, Charles University (République Tchèque)	Rapporteur
Professeur Boughanem, Mohand	IRIS, Université de Toulouse 3 (France)	Examinateur
Professeur Jatowt, Adam	Data Science Group, Universität Innsbruck (Autriche)	Examinateur
Docteur Lamirel, Jean-Charles	SYNALP, LORIA, Université de Strasbourg (France)	Examinateur
Docteur Atanassova, Iana	CRIT, Université de Franche-Comté ; IUF (France)	Directrice de thèse

LECLA
ÉCOLE DOCTORALE

Titre : Extraction d'informations appliquée aux documents non-structurés pour la valorisation de périodiques historiques : application au patrimoine de la région Bourgogne Franche-Comté en France

Mots clés : Extraction d'information, Gestion des connaissances, Annotation sémantique, Interfaces de recherche, Exploitation et exploration des documents, Humanités Numériques

Résumé : Ces dernières années, les bibliothèques et archives ont entrepris de nombreuses campagnes de numérisation afin d'élargir l'accès du public à leurs collections d'archives. Cependant, le défi de promouvoir le contenu des collections et de rendre ces ressources accessibles reste entier. La numérisation produit souvent un contenu non structuré dans lequel il est difficile de naviguer, tandis que les interfaces qui s'appuient sur des requêtes basées sur des mots clés pour accéder aux documents d'archives peuvent fournir aux utilisateurs des résultats non pertinents. Afin d'exploiter le potentiel des « Big Data of the Past », notion introduite par Kaplan et di Lenardo en 2017, il est essentiel de développer des méthodes et des cadres pour structurer le contenu textuel des documents, dans le but d'en améliorer l'exploration et l'exploitation. Dans ce contexte, la présente thèse de doctorat aborde le problème du traitement des documents historiques numérisés, en se concentrant sur l'extraction des Entités Nommées et des Relations afin de créer des interfaces pour l'exploitation efficace des données textuelles historiques. Premièrement, nous proposons une nouvelle méthode pour déterminer la structure logique des journaux historiques en utilisant une approche à base de règles. Deuxièmement, nous présentons une méthode pour extraire les entités et les relations concernant les personnes et les lieux mentionnés dans les textes. Notre approche s'intitule Extensible, Lightweight and Interpretable Joint Extraction of Relations and Entities (ELIJERE). Elle est basée sur des ressources linguistiques obtenues par supervision distante. Enfin, nous proposons un cadre général pour l'étude de l'expression d'informations spatiales dans les documents, et un autre cadre pour l'application des méthodes de TimeLine Summarisation à des collections de documents. Nous montrons comment ces méthodes peuvent être appliquées pour produire des interfaces sémantiquement riches, telles que des frises chronologiques et des cartes, qui permettent au grand public une lecture proche ou distante de ces collections.

Title : Information extraction from unstructured documents for the valorisation of historical periodicals: application to the heritage of the Bourgogne Franche-Comté Region in France

Keywords : Information Extraction, Knowledge management, Semantic Annotation, Search interfaces, Exploration and exploitation of documents, Digital Humanities

Abstract : In recent years, libraries and archives have undertaken numerous digitisation campaigns to widen public access to their archival collections. However, the challenge of promoting the content of collections and making these resources discoverable remains. Digitisation often produces unstructured content that is difficult to navigate, while interfaces that rely on keyword-based queries to access archival materials can provide users with irrelevant results. In order to exploit the potential of the "Big Data of the Past", notion introduced by Kaplan and di Lenardo in 2017, it is essential to develop methods and frameworks for structuring the textual content of documents, with the aim of improving their exploration and discoverability. In this context, the present Ph.D. thesis addresses the problem of processing digitised historical documents, focusing on the extraction of Named Entities and Relations in order to create interfaces for the efficient exploitation of historical textual data. Firstly, we propose a new method for determining the logical structure of historical newspapers using a rule-based approach. Secondly, we present a method for extracting entities and relations about people and places mentioned in texts. Our approach is called Extensible, Lightweight and Interpretable Joint Extraction of Relations and Entities (ELIJERE). It is based on linguistic resources obtained through distant supervision. Finally, we propose a general framework for studying the expression of spatial information in documents, and another framework for applying TimeLine Summarisation methods to document collections. We show how these methods can be applied to produce semantically rich interfaces, such as timelines and maps, that allow the general public a close and a distant reading of these collections.

UBFC
UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besançon

Acknowledgement

First of all, I would like to express all my gratitude to my supervisor, Dr. Iana Atanassova. This work would not have been possible without her guidance, her advice, her support, her encouragements, her help in the difficult moments, her trust and all the interesting conversations we've had throughout the years.

I would also like to thank the *Bourgogne Franche-Comté* region for funding this Ph.D thesis.

I would like to express my gratitude towards Pr. Bruno Bachimont, Pr. Mohand Boughanem, Pr. Antoine Doucet, Pr. Adam Jatowt, Dr. Jean-Charles Lamirel and Dr. Pavel Pecina, for accepting to be members of my jury, and in particular to Pr. Bruno Bachimont, Dr. Jean-Charles Lamirel and Dr. Pavel Pecina, who accepted to be the reviewers of this work.

I would also like to express my special thanks to Pr. Adam Jatowt and Pr. Antoine Doucet, who welcomed me at the Digital Science Center (DiSC) of the University of Innsbruck (Austria) between April and July 2022 for a three-months research stay. This research stay, as well as their advice, their supervision and the conversations we've had made me discover other research methods, extended my knowledge of the Natural Language Processing field, and made me grow as a researcher.

I would also like to thank the members of my *Comités de Suivi de Thèse* (CST), Dr. Virginie Lethier, Dr. Marc Bertin and Dr. Pierre Verschueren for their advice and their suggestions, and for helping me consider this work from other perspectives than that of the Natural Language Processing field.

I would also like to express my special thanks to Dr. Izabella Thomas, who was my first teacher at the university, who made me discover the fields of linguistics and Natural Language Processing, and pushed me to pursue my studies into these fields.

I would like to thank the staff of the *Maison des Sciences de l'Homme et de l'Environnement* (MSHE) in Besançon, and in particular Marion, Yuji, Mattieu, Etienne, Marie, Florence, Julien, Manu, Quentin, Arnaud, with whom I worked during the ORTEP project. This project was my first contact with the task of processing historical documents, and gave me the desire to pursue such work with a Ph.D.

I would also like to thank the *Sciences, arts et culture* team of the University of Franche-Comté

for organising events for the popularisation of science such as *La Fête de la Science*, *La Nuit Européenne des Chercheur.e.s* or *OVNI*, which gave me the opportunity to discuss my work with the public.

I would like to express my thanks to all my students, and in particular to Bruno, Laura, Florine, Marine, Maya, Lilian, Jacob, Charline and Veronika, for their involvement and their curiosity. You are the best students a teacher can dream of.

I would also like to express my gratitude to the member of the DiSC laboratory, Adam, Carina, Lorenzo, Alexander, Maija, Sébastien, Reto and Matthias, who welcomed me in Innsbruck, made me discover the city and the Austrian culture, and integrated me into the life of the laboratory.

I could not conclude these acknowledgements without expressing all my gratitude to my closest friends Nathan, Jean-Baptiste, Charline, Céline, Thomas, Jean-Baptiste, Jim, Dani, Quentin, Inès, Julien, Cecil, Marc-Antoine, Emel and Matthieu, to my siblings, Thomas and Sarah, to my sister-in-law Elise and to my friends from the Centre de Recherches Interdisciplinaires et Transculturelles (C.R.I.T.) laboratory, Yağmur, Aurélie, Ning, François-Claude, Salah, Laure, Cécile, and Manolo. Thank you for all the good moments we've had throughout the years, for letting me discuss extensively of this work with you, and for supporting me in the difficult moments as well as during the good moments.

And of course, last but definitively not the least, I want to express all my gratitude to my partner Nicolas. Thank you for your support, for your encouragements and understanding, for making me laugh and trying to distract me during the difficult moments, for giving me the space to focus and work when I needed it, but overall, thank you for always being there for me.

Remerciements

Tout d'abord, je tiens à exprimer toute ma gratitude à ma directrice de thèse, Dr. Iana Atanassova. Ce travail n'aurait pas été possible sans ses conseils, son soutien, ses encouragements, son aide dans les moments difficiles, sa confiance et toutes les conversations intéressantes que nous avons eues au fil des ans.

Je tiens également à remercier la région Bourgogne Franche-Comté pour le financement de cette thèse de doctorat.

Je tiens à exprimer ma gratitude envers Pr. Bruno Bachimont, Pr. Mohand Boughanem, Pr. Antoine Doucet, Pr. Adam Jatowt, Dr. Jean-Charles Lamirel et Dr. Pavel Pecina, d'avoir accepté d'être membres de mon jury, et en particulier au Pr. Bruno Bachimont, Dr. Jean-Charles Lamirel et Dr. Pavel Pecina, qui ont accepté d'être les relecteurs de ce travail.

Je tiens également à remercier tout particulièrement Pr. Adam Jatowt et Pr. Antoine Doucet, qui m'ont accueilli au *Digital Science Center* (DiSC) de l'Université d'Innsbruck (Autriche) entre avril et juillet 2022 pour un séjour de recherche de trois mois. Ce séjour de recherche, ainsi que leurs conseils, leur encadrement et les conversations que nous avons eues m'ont fait découvrir d'autres méthodes de recherche, ont élargi mes connaissances dans le domaine du Traitement Automatique des Langues, et m'ont fait grandir en tant que chercheur.

Je tiens également à remercier les membres de mon Comités de Suivi de Thèse (CST), Dr. Virginie Lethier, Dr. Marc Bertin et Dr. Pierre Verschueren pour leurs conseils et leurs suggestions, et pour m'avoir aidé à considérer ce travail sous d'autres perspectives que celle du domaine du Traitement Automatique des Langues.

Je tiens également à remercier tout particulièrement Dr. Izabella Thomas, qui a été mon premier professeur à l'université, qui m'a fait découvrir les domaines de la linguistique et du Traitement Automatique des Langues, et qui m'a poussé à poursuivre mes études dans ce domaine.

Je tiens à remercier le personnel de la Maison des Sciences de l'Homme et de l'Environnement (MSHE) de Besançon, et en particulier Marion, Yuji, Mattieu, Etienne, Marie, Florence, Julien, Manu, Quentin, Arnaud, avec qui j'ai travaillé dans le cadre du projet ORTEP. Ce projet a été mon premier contact avec le travail de traitement des documents historiques, et m'a donné l'envie de poursuivre ce travail avec un doctorat.

Je remercie également l'équipe Sciences, arts et culture de l'Université de Franche-Comté pour l'organisation d'événements de vulgarisation scientifique tels que La Fête de la Science, La Nuit Européenne des Chercheur.e.s ou OVNI, qui m'ont donné l'occasion de discuter de mes travaux avec le public.

Je tiens à remercier tous mes étudiants, et en particulier Bruno, Laura, Florine, Marine, Maya, Lilian, Jacob, Charline et Veronika, pour leur implication et leur curiosité. Vous êtes les meilleurs élèves dont un professeur puisse rêver.

Je tiens également à exprimer ma gratitude aux membres du laboratoire DiSC, Adam, Carina, Lorenzo, Alexander, Maija, Sébastien, Reto et Matthias, qui m'ont accueilli à Innsbruck, m'ont fait découvrir la ville et la culture autrichienne, et m'ont intégré à la vie du laboratoire.

Je ne saurais terminer ces remerciements sans exprimer toute ma gratitude à mes amis les plus proches, Nathan, Jean-Baptiste, Charline, Céline, Thomas, Jean-Baptiste, Jim, Dani, Quentin, Inès, Julien, Cecil, Marc-Antoine, Emel et Matthieu, à mes frères et sœurs, Thomas et Sarah, à ma belle-sœur Elise et à mes amis du laboratoire du Centre de Recherches Interdisciplinaires et Transculturelles (C.R.I.T.), Yağmur, Aurélie, Ning, François-Claude, Salah, Laure, Cécile et Manolo. Merci pour tous les bons moments que nous avons passés au fil des années, pour m'avoir permis de discuter largement de ce travail avec vous, et pour m'avoir soutenu dans les moments difficiles comme dans les bons moments.

Et bien sûr, en dernier lieu, mais certainement pas le moins important, je veux exprimer toute ma gratitude à mon partenaire Nicolas. Merci pour ton soutien, pour tes encouragements et ta compréhension, pour m'avoir fait rire et avoir essayé de me distraire dans les moments difficiles, pour m'avoir donné l'espace nécessaire pour me concentrer et travailler quand j'en avais besoin, mais surtout, merci d'avoir toujours été là pour moi.

Table of contents

PARTIE I

Extraction d'informations appliquée aux documents non-structurés pour la valorisation de périodiques historiques : application au patrimoine de la région Bourgogne Franche-Comté en France

Résumé substantiel de la thèse en français

Introduction

L'exploitation à grande échelle des documents historiques contenus dans les archives, tels que les journaux, les registres, les cartes ou les certificats, pour n'en citer que quelques-uns, est un problème majeur dans le domaine des humanités numériques, et reste un défi à bien des égards. Ces dernières années, les campagnes de numérisation menées par les archives et les bibliothèques ont permis de préserver le contenu des documents historiques dans un format numérisé et de les rendre accessibles au grand public. De plus, grâce à des méthodes automatiques telles que la reconnaissance optique de caractères (*Optical Character Recognition*, OCR), ces documents numérisés peuvent être convertis en formats lisibles par la machine et exploitables par les moteurs de recherche, ce qui permet à un public plus large d'accéder à ces collections et d'y effectuer des recherches.

Cependant, le contenu textuel des documents historiques extraits par des méthodes automatiques telles que l'OCR est généralement peu structuré. Ainsi, si l'extraction de ces contenus permet d'effectuer des recherches par mots-clés dans les documents, cela ne permet pas en revanche de développer des interfaces plus avancées pour la visualisation et l'exploitation des données. Lorsqu'il s'agit de grandes collections, les utilisateurs effectuant des recherches par mot-clé peuvent rapidement être submergés par une multitude de résultats qui ne sont pas toujours pertinents par rapport à leur recherche.

Par conséquent, un problème important subsiste : les documents historiques contiennent de grandes quantités d'informations qui peuvent être potentiellement utiles à la fois au grand public et aux chercheurs en sciences humaines. Cependant, l'extraction d'informations spécifiques à partir de grandes collections, bien que possible, nécessite des efforts considérables. Il est donc difficile d'envisager toutes les applications possibles dans ce domaine et d'explorer tout le potentiel de l'exploitation des documents historiques. Dans ce contexte, Kaplan and di Lenardo (2017) définissent la notion des "Big Data du passé" (*Big Data of the Past*), en référence au phénomène de production intensive et rapide de données historiques résultant de la numérisation de masse des collections historiques, et à la nécessité de développer de nouvelles méthodes et de nouveaux outils pour traiter ces données.

La structuration du contenu textuel des documents d'archives constitue la pierre angulaire de la création d'interfaces permettant d'explorer efficacement les ensembles de données historiques

et d'améliorer les interfaces de recherche. La structure logique des documents, c'est-à-dire la division des éléments de texte en titres, en-têtes, paragraphes, etc., transmet une partie du sens, permet une indexation différenciée des différents types d'éléments et donc la création de moteurs de recherche plus efficaces. Le domaine du Traitement Automatique des Langues (TAL) offre à cet égard diverses méthodes et outils, qui doivent être améliorés et adaptés aux types de documents présents dans les collections de documents historiques. En outre, les outils de TAL peuvent être utilisés pour annoter le contenu textuel des documents avec des catégories sémantiques telles que les sentiments (par exemple, positifs, négatifs), les types d'entités (par exemple, personne, lieu, expressions temporelles) ou les relations (par exemple, lieu de naissance, lieu d'études), pour n'en citer que quelques-unes. Ces annotations sémantiques peuvent être intégrées dans les index des interfaces de recherche pour aider à explorer les collections de documents au-delà des possibilités permises par la recherche par mot-clé. Ces annotations permettent aussi de construire des interfaces de recherche augmentées (Atanassova, 2012), telles que des cartes ou des frises chronologiques, qui permettraient une lecture distante (Moretti, 2013) des collections.

Cependant, l'annotation sémantique des documents historiques reste un défi pour plusieurs raisons. Les documents en format lisible par machine obtenus par des méthodes automatiques telles que l'OCR contiennent généralement des erreurs qui affectent les résultats obtenus par les moteurs de recherche et les méthodes de TAL (Boros, Nguyen, et al., 2022; M. Ehrmann et al., 2021a; Linhares Pontes et al., 2019a). Par conséquent, des méthodes de post-traitement sont généralement nécessaires pour éliminer ces erreurs. En outre, la plupart des outils TAL disponibles aujourd'hui sont conçus pour traiter des documents rédigés dans l'état actuel de la langue. Ils sont de plus généralement dédiés à des styles d'écriture spécifiques. Par conséquent, ces outils ne sont généralement pas adaptés au traitement des documents historiques, qui peuvent être rédigés dans des styles différents et dans un état antérieur de la langue. Bien que des ressources pour le traitement des documents historiques soient apparues ces dernières années, telles que des jeux de données annotées (M. Ehrmann, Romanello, Flückiger, & Clematide, 2020; M. Ehrmann et al., 2022; Gutehrlé & Atanassova, 2021a; Hamdi et al., 2021), de telles ressources restent rares.

Dans cette thèse de doctorat, nous abordons la question de l'exploration et de l'exploitation des collections de documents historiques. Notre contribution majeure dans ce domaine consiste en trois parties :

- proposer une nouvelle méthode pour déterminer la structure logique des journaux historiques en utilisant une approche basée sur des règles
- proposer une méthode pour extraire les entités et les relations concernant les personnes et les lieux mentionnés dans les textes. Notre approche s'appelle *Extensible, Lightweight and Interpretable Joint Extraction of Relations and Entities* (ELIJERE). Elle s'appuie sur des

ressources linguistiques obtenues par supervision distante

- proposer deux cadres généraux pour la conception d'interfaces de recherche augmentées, basées respectivement sur des cartes et des frises chronologiques générées automatiquement, qui permettent de passer facilement d'une lecture proche à une lecture distante des documents

Pour ce travail, nous avons construit un corpus de périodiques imprimés d'origines diverses publiés aux 19ème et 20ème siècles dans les régions de Bourgogne et de Franche-Comté en France. Nous avons choisi de nous concentrer sur les périodiques en raison de leur importance en tant que sources primaires d'information pour le grand public reflétant les événements locaux et internationaux (Tibbo, 2007). Les documents de notre corpus traitent de sujets variés tels que la guerre, la religion ou la science, et sont rédigés dans des styles différents. Nous avons collecté le corpus à partir de Gallica, l'archive numérique de la Bibliothèque Nationale de France[1]. Gallica fournit le contenu du document au format XML ALTO, qui fournit la transcription OCR du document, son ordre de lecture et son organisation physique. Nous nous concentrons sur le traitement du contenu textuel de ces documents plutôt que sur d'autres types de contenu tels que les images, les publicités ou les tableaux, puisque la plupart des informations de notre corpus sont véhiculées au travers de moyens textuels. Ce travail fait partie du projet EMONTAL (Extraction et Modélisation ONTologique des Acteurs et Lieux pour la valorisation du patrimoine de Bourgogne Franche-Comté)[2] et est financé par la région Bourgogne Franche-Comté pour la période 2020-2023.

Le reste de ce travail est structuré comme suit : dans le chapitre 1, nous abordons la question de l'application des méthodes de TAL pour structurer le contenu textuel des documents historiques. Nous donnons d'abord un aperçu des différentes méthodes de TAL qui peuvent être appliquées, ainsi qu'un aperçu d'autres projets similaires au nôtre. Nous décrivons ensuites la tâche d'annotation sémantique, en nous concentrant en particulier sur les tâches de reconnaissance d'entités nommées et d'extraction de relations. Dans le chapitre 2, nous présentons les documents du corpus EMONTAL et le thématiques qui y sont traitées. Nous décrivons ensuite notre approche pour nettoyer les erreurs produites par le processus d'OCR, ainsi que notre approche pour déterminer la structure logique des documents.

Nous présentons l'approche ELIJERE dans le chapitre 3. Nous décrivons d'abord le jeu de données DARES, qui consiste en des phrases faiblement annotées avec des relations et des entités collectées à partir de Wikipedia et Wikidata. Ensuite, nous décrivons notre méthode pour construire les ressources linguistiques à partir de l'ensemble de données DARES, avant de présenter comment nous exploitons ces ressources pour la tâche d'extraction conjointe de relations et d'entités. Dans

[1] https://gallica.bnf.fr/
[2] http://tesniere.univ-fcomte.fr/projet-emontal/

le chapitre 4, nous évaluons l'approche ELIJERE. Nous évaluons une implémentation de base de cette approche, qui s'appuie sur les deux ressources linguistiques pour extraire et catégoriser les relations candidates. De plus, nous évaluons une implémentation hybride qui combine les ressources linguistiques avec un modèle d'apprentissage automatique pour catégoriser les relations candidates. Nous évaluons les deux implémentations sur l'ensemble de données DARES, avant de les évaluer sur le corpus EMONTAL, afin d'étudier l'impact des erreurs d'OCR et des différents styles d'écriture sur notre approche.

Dans le chapitre 5, nous présentons deux cadres généraux pour la conception d'interfaces de recherche augmentées basées sur des cartes et des lignes de temps générées automatiquement. Pour chaque cadre, nous fournissons une description des travaux connexes, puis nous présentons le cadre en détail, avant de discuter de ses applications et extensions potentielles. Enfin, nous présentons les limitations ainsi que la conclusion de ce travail, avant de présenter des suggestions pour des travaux futurs.

Chapitre 1

Contextualisation et état de l'art

Table des matières

Ces dernières années, les bibliothèques et les archives ont entrepris de nombreuses campagnes de numérisation de leurs collections. Bien que ces campagnes aient ouvert l'accès aux documents d'archives à un public plus large, l'exploration et l'exploitation de leur contenu restent des tâches difficiles, en raison du manque fréquent de structure de leur contenu textuel.

Lorsqu'ils accèdent aux documents par l'intermédiaire de moteurs de recherche, les utilisateurs sont limités par des requêtes basées sur des mots-clés et peuvent ainsi être submergé par l'abondance de documents disponibles, mais pas toujours pertinents pour leur recherche. Il est donc nécessaire de structurer le contenu textuel des documents d'archives afin d'améliorer les interfaces de recherche existantes et de faciliter l'exploration et l'exploitation de ces "Big Data du passé" (*Big Data of the Past*, Kaplan (2020)).

Dans ce chapitre, nous abordons le sujet de l'application des méthodes de Traitement Automatique des Langues (TAL) aux documents historiques. Ces méthodes permettent d'enrichir les documents par l'ajout d'annotations sémantiques, qui peuvent être exploitées par les moteurs de recherche pour indexer les documents et permettre des requêtes plus puissantes. De plus, ces annotations peuvent être exploitées pour construire des interfaces de recherche augmentées telles que des cartes ou des frises chronologiaues, qui permettent une lecture proche et une lecture distante du corpus.

Dans la section 1.1, nous expliquons comment les annotations sémantiques produites par les

méthodes de TAL peuvent aider à structurer le contenu textuel des documents historiques. Nous fournissons plusieurs exemples de ces annotations sémantiques, ainsi que plusieurs exemples de la manière dont elles peuvent être exploitées par les interfaces de recherche. Dans la section 1.2, nous présentons plusieurs verrous méthodologiques et technologiques qui doivent être résolus pour un traitement efficace des documents historiques. Enfin, nous décrivons la tâche d'extraction d'informations dans le chapitre 1.3, et nous nous concentrons en particulier sur les tâches de reconnaissance d'entités nommées et d'extraction de relations.

1.1 Enrichissement sémantique des documents historiques

Parmi les travaux appliquant les méthodes de TAL aux documents historiques, de nombreux travaux se sont concentrés sur l'application de méthodes de reconnaissance d'entités nommées (*Named Entity Recognition*, NER). La tâche de NER vise à détecter les mentions d'entités réelles telles qu'une personne, un lieu, une date ou une organisation dans les documents (M. Ehrmann, 2008; M. Ehrmann et al., 2021a). La détection des mentions d'entités est importante pour comprendre le contenu et les thèmes des documents. De plus, ces mentions peuvent être exploitées par les moteurs de recherche pour indexer les documents, ce qui permet d'interroger et de filtrer les documents en fonction de leurs entités. Pour ces raisons, la tâche NER appliquée aux documents historiques a fait l'objet d'une attention particulière ces dernières années, comme le montrent des travaux tels que Boros, Hamdi, et al. (2020), M. Ehrmann, Romanello, Flückiger, and Clematide (2020), and M. Ehrmann et al. (2021b).

La tâche de reconnaissance d'entités nommées peut être utilisée pour d'autres tâches en aval telles que l'établissement de liens entre entités nommées, l'extraction de relations ou l'extraction d'événements. Par exemple, la tâche de résolution des entités nommées (*Named Entity Linking*, NEL) vise à déterminer à quelle entité réelle se réfèrent plusieurs mentions d'entités. En général, cette tâche associe ces mentions d'entités à une seule entrée d'une base de données ou de connaissances telle que Wikidata ou FreeBase. Ainsi, la tâche NEL permet de comprendre et d'interroger le corpus en se basant sur les entités réelles et non uniquement sur les mentions d'entités. En outre, l'utilisation d'une base de connaissances externe telle que Wikidata permet d'ajouter des données contextuelles au corpus, par exemple des informations sur les entités ou le document lui-même (Linhares Pontes et al., 2019b).

D'autre part, la tâche d'extraction de relations (*Relation Extraction*, RE) vise à détecter les relations entre les entités, telles que les relations *lieu de naissance* ou *étudier à*. La tâche d'extraction de relations permet donc d'étudier la manière dont les entités sont liées les unes aux autres. En outre, les relations détectées peuvent être exploitées pour construire des graphes de connaissances à partir des documents, qui peuvent être utilisés pour d'autres tâches en aval telles que la recherche

d'informations (E. Klein et al., 2014; Plum et al., 2022). Enfin, d'autres travaux tels que Boros, Nguyen, et al. (2022), Ide and Woolner (2004), and Smith (2002) se sont concentrés sur la tâche d'extraction d'événements (*Event Extraction*, EE), qui vise à détecter les mentions d'événements dans les documents. Ces méthodes visent également à détecter quand et où l'événement s'est produit, ainsi que les personnes qui y ont participé. Lorsqu'elles sont appliquées à des documents historiques, les méthodes d'extraction d'événements peuvent aider à comprendre les documents ainsi que les entités détectées en construisant des chronologies à partir des événements détectés.

Parmi les autres tâches de traitement du langage naturel, la tâche de modélisation de thématique (*Topic Modelling*, TM) appliquée aux documents historiques a également suscité un grand intérêt ces dernières années, comme le montrent des travaux tels que Marjanen et al. (2020a) and Y. Yan et al. (2009). La tâche de TM vise à détecter de manière non supervisée les sujets présents dans un corpus. La plupart des méthodes sont adaptables à n'importe quel corpus puisqu'elles détectent des sujets latents à partir des documents, au lieu de rechercher des sujets explicites choisis par l'utilisateur. En appliquant ces méthodes, les documents peuvent être indexés en fonction de leurs sujets et permettent de rechercher et de comparer des documents en fonction de leur sujet, ainsi que d'étudier l'évolution d'un sujet dans le temps. D'autres travaux tels que Dominguès et al. (2019), Hamdi et al. (2021), and Sprugnoli et al. (2016) se sont concentrés sur la tâche d'analyse de sentiment, qui vise à mesurer la polarité (c'est-à-dire le sentiment positif, négatif ou neutre) ou les émotions exprimées dans les textes. Appliquées à des documents historiques et combinées à des méthodes de modélisation de thématique, les méthodes d'analyse de sentiment permettent d'étudier la manière dont un sujet est perçu au fil du temps.

Outre leur rôle dans la structuration du contenu textuel des documents, les annotations sémantiques peuvent être exploitées par les moteurs de recherche pour indexer les documents, comme dans les plateformes *impresso*[1], NewsEye[2] ou *Retronews*[3]. Ces plateformes permettent à l'utilisateur de rechercher des documents à l'aide de requêtes basées sur des mots clés et de filtrer les documents en fonction de leurs annotations sémantiques. En outre, les annotations sémantiques produites par les méthodes de TAL peuvent être exploitées pour construire des interfaces de recherche augmentées qui permettent une lecture distante du corpus (Moretti, 2013), tout en permettant de revenir au document original pour une lecture rapprochée (Jatowt, 2021).

Par exemple, des travaux tels que E. Klein et al. (2014) proposent d'étudier le vocabulaire du corpus sous forme de nuages de mots, où la taille des mots dépend de leur fréquence dans le corpus. Ces nuages de mots peuvent être complétés par d'autres annotations telles que les sentiments, comme dans Sprugnoli et al. (2016). D'autres travaux comme M. Ehrmann, Romanello, Clematide, et al. (2020) proposent d'exploiter les plongements sémantiques (*word embeddings*) diachroniques

[1] https://impresso-project.ch/
[2] https://www.newseye.eu/
[3] https://www.retronews.fr/

pour suggérer des termes similaires aux mots-clés utilisés dans une requête et qui sont plus proches de l'état de la langue dans lequel les documents ont été écrits que de l'état moderne. De même, des travaux tels que Cordell and Smith (2017) and Romanello et al. (2020) proposent plusieurs outils de visualisation tels que des graphiques pour observer la réitération d'un contenu textuel similaire dans les documents.

Bertin et al. (2015), Blevins (2014), Dominguès et al. (2019), and Moncla et al. (2019) proposent de générer automatiquement des cartes à partir des mentions de lieux qui ont été détectées par le processus de reconnaissance des entités nommées. La Figure 1.1 présente un exemple de carte générée automatiquement. Pour générer de telles cartes, une étape de géocodage est nécessaire, qui vise à associer des coordonnées à la mention d'un lieu à partir d'un texte (McDonough et al., 2019). Cette étape nécessite des ressources externes connues sous le nom de répertoires géographiques (*gazetteer*), comme Geonames[4] ou Pleiades (Bagnall et al., 2016), qui stocke les coordonnées de lieux existants et anciens connus. Ces cartes peuvent être complétées par des connaissances contextuelles fournies par des sources externes telles que Wikidata, comme dans Gutehrlé et al. (2021).

De même, les événements et les expressions temporelles mentionnés dans les documents peuvent être exploités pour générer automatiquement des frises chronologiques à partir de documents historiques, comme le suggère Gutehrlé et al. (2022). Ces frises peuvent fournir une vue d'ensemble des collections historiques et constituer un nouveau moyen d'accès aux archives d'articles de presse, comme le suggère Duan et al. (2017) and Pasquali et al. (2019a). Nous fournissons plus de détails sur les interfaces de recherche augmentées basées sur des cartes et des chronologies générées automatiquement dans le chapitre 5.

1.2 Verrous méthodologiques et technologiques

Grâce aux campagnes de numérisation menées par les archives et les bibliothèques, de nombreuses collections de documents historiques sont désormais disponibles sous forme d'images, consultables en ligne. Afin de traiter ces documents avec des méthodes de TAL, il est d'abord nécessaire de convertir la structure de ces documents ainsi que leur contenu textuel dans un format lisible par la machine. En raison de la grande quantité de documents disponibles, cette tâche est souvent effectuée automatiquement. La structure physique et l'ordre de lecture du document peuvent être déterminés en appliquant des méthodes d'analyse de la structure physique (*Physical Layout Analysis*, PLA), tandis que l'organisation du document en éléments logiques (en-tête, titre, paragraphes...) peut être déterminée en appliquant des métodes d'analyse de la structure logique (*Logical Layout Analysis*, LLA). Enfin, le contenu textuel des documents peut être transcrit par des méthodes de re-

[4] https://www.geonames.org/

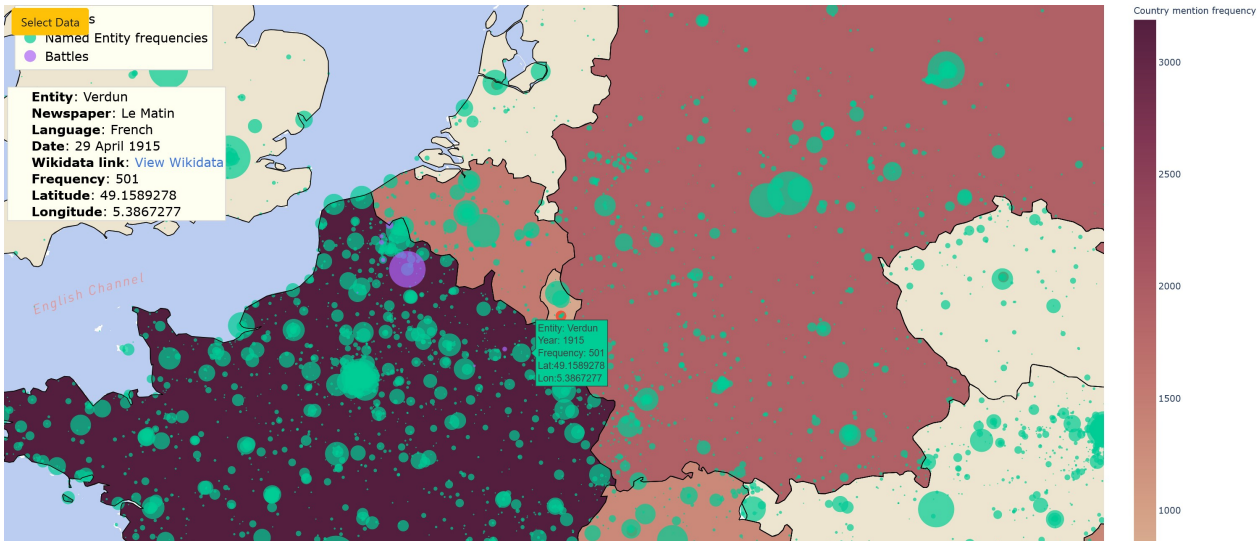


Figure 1.1: Exemple d'interface sémantique : carte générée automatiquement à partir de lieux identifié dans *Le Matin* (1913-1915) (Gutehrlé et al., 2021)

connaissance optique de caractères (*Optical Character Recognition*, OCR), ou par des méthodes de reconnaissance de textes manuscrits (*Handwritten Text Recognition*, HTR) lorsque les documents sont des manuscrits.

Le succès de ces méthodes dépend de multiples facteurs, tels que la qualité des documents originaux et de leur copie numérique, la mise en page du document, la typographie utilisée, ainsi que la différence entre la langue du document et la langue pour laquelle le modèle OCR a été entraîné. Ainsi, les documents en format lisible par machine produits par ces méthodes automatiques contiennent généralement des erreurs, qui ont un impact sur les résultats obtenus par les méthodes de TAL et les moteurs de recherche (Boros, Nguyen, et al., 2022; M. Ehrmann et al., 2021a; Linhares Pontes et al., 2019b). Des méthodes de post-traitement sont généralement nécessaires pour éliminer ces erreurs, tandis que des méthodes TAL robustes aux données bruyantes sont nécessaires pour structurer les documents historiques.

Les collections de documents historiques sont généralement hétérogènes et contiennent des documents de différentes natures (périodiques, livres, ...) et origines (périodiques religieux, littérature, journaux à orientation politique, etc.) De plus, selon la période d'origine des documents, la ou les langues dans lesquelles ils sont écrits peuvent être plus ou moins éloignées de leur forme moderne. La plupart des outils de TAL disponibles aujourd'hui sont conçus pour traiter des documents écrits dans l'état moderne de la langue, et sont généralement dédiés à des styles d'écriture spécifiques. Les performances de ces outils sont donc moindres lorsqu'ils sont appliqués à des documents écrits dans plusieurs langues, dans un style différent ou dans un état plus ancien de la langue (M. Ehrmann, Colavizza, Rochat, & Kaplan, 2016). De plus, les ressources externes telles que Wikidata ou Geonames utilisées pour des tâches telles que le NEL ou les tâches de géocodage

sont généralement construites à partir de données contemporaines. Par conséquent, elles peuvent ne pas être adaptées à la période du corpus étudié et comportent le risque d'introduire des anachronismes (Gutehrlé et al., 2021).

Ainsi, le traitement des documents historiques nécessite soit de construire des modèles dédiés, soit d'adapter des modèles existants, par exemple en appliquant des méthodes d'apprentissage par transfert (*transfer-learning)* comme dans Boroş et al. (2020). De nombreuses ressources dédiées à des tâches telles que la reconnaissance d'entités nommées (M. Ehrmann, Romanello, Flückiger, & Clematide, 2020; M. Ehrmann et al., 2022), l'analyse de la structure logique des documents (Gutehrlé & Atanassova, 2021a) ou la résolution d'entités nommées (Hamdi et al., 2021) ont été publiées ces dernières années, ce qui a permis la création de tels outils. Cependant, ces ressources restent rares pour de nombreuses autres tâches et langues, car leur création est une tâche coûteuse (T. T. H. Nguyen et al., 2021a).

D'autres travaux tels que Piotrowski (2012) ont proposé d'appliquer des méthodes de normalisation, afin que le contenu textuel des documents corresponde à un état adapté à ces outils. Cependant, plus l'état de la langue est éloigné de la langue moderne, plus il est difficile de normaliser les variations orthographiques, d'autant plus pour les manuscrits anciens, où l'écriture est moins standardisée. En outre, la plupart des méthodes se concentrent exclusivement sur la correspondance entre les orthographes anciennes et modernes d'un mot, mais ne s'intéressent pas à leur évolution sémantique. Il y a donc un risque de faire correspondre des termes dont le sens a changé depuis. Ainsi, bien que cette étape facilite les requêtes et l'utilisation des outils de TAL, il est important de s'assurer qu'elle n'introduit pas de nouvelles erreurs.

1.3 Extraction d'informations

La tâche d'extraction d'informations vise à transformer des données non structurées, telles que des textes, en données structurées (Jurafsky & Martin, 2008). Ces données structurées peuvent ensuite être exploitées pour des tâches en aval telles que la recherche d'informations, la constitution de bases de données, etc. L'extraction d'informations implique de nombreuses tâches telles que l'extraction d'événements, la détection d'expressions temporelles ou l'analyse de sentiments. Dans ce chapitre, nous avons choisi de nous concentrer sur les tâches de *reconnaissance d'entités nommées* et d'*extraction de relations*, et plus précisément sur la tâche d'extraction conjointe de relations et d'entités.

La reconnaissance d'entités nommées (*Named Entity Recognition*, NER) consiste à identifier les mentions de noms d'entités telles que des personnes, des lieux, des organisations ou des dates, tandis que l'extraction de relations (RE) vise à déterminer les relations entre les entités. Ces tâches sont généralement appliquées de manière séquentielle, la tâche d'extraction de relations visant à

déterminer les relations entre les entités détectées par la tâche de reconnaissance d'entités nommées. Cependant, des travaux récents se sont concentrés sur la tâche d'extraction conjointe de relations et d'entités (*Joint Extraction of Relations and Entities*, JERE), qui vise à extraire les mentions d'entités et les relations d'une manière conjointe à partir de phrases (Pawar, Palshikar, & Bhattacharyya, 2017). Certaines approches telles que Pawar, Bhattacharya, and Palshikar (2017) and Roth and Yih (2004) de la tâche d'extraction conjointe de relations et d'entités proposent d'entraîner de multiples classificateurs pour détecter séparément les mentions d'entités et de relations dans la phrase, avant de combiner les résultats des classificateurs. Ce résultat est contrôlé par des contraintes de domaine telles que des règles qui peuvent être exprimées par des approches telles que l'optimisation linéaire en nombres entiers ou les champs aléatoire de Markov logiques.

D'autres travaux, tels que Kate and Mooney (2010), proposent de représenter toutes les entités et relations possibles dans une phrase sous la forme d'un graphe arborescent appelé "graphe de pyramide de cartes" (*card-pyramid graph*). Ce graphe est analysé par un algorithme qui combine des règles de production similaires à des grammaires non contextuelles et des classificateurs pour détecter les mentions d'entités et de relations à partir du graphe de pyramide de cartes.

D'autres travaux proposent de considérer la tâche JERE comme un problème de prédiction structurée, où l'objectif est de produire une structure décrivant les types d'entités et leurs relations à partir de la phrase originale. Par exemple, Q. Li and Ji (2014) propose un cadre conjoint incrémental qui produit un graphe où les nœuds représentent les mentions d'entités et les arêtes représentent les relations entre les entités. De même, Miwa and Sasaki (2014) présente la tâche JERE comme un problème de remplissage de table, et propose de générer une matrice symétrique, où les lignes et les colonnes sont des tokens et où la cellule contient les types d'entités et les relations.

Des travaux plus récents proposent des approches de bout en bout, dans lesquelles un réseau neuronal unique apprend à catégoriser les entités et les relations. Ces travaux s'appuient généralement sur l'architecture LSTM (Bekoulis et al., 2018; B. Yu et al., 2020; Zheng et al., 2017), ainsi que sur les réseaux neuronaux en graphe (Fu et al., 2019; Sun et al., 2019) ou les architectures Transformer (Eberts & Ulges, 2019; Shang et al., 2022). Ces modèles apprennent à identifier les mentions d'entités et de relations à partir des plongements de mots (*word embeddings*), bien que certains travaux tels que Miwa and Bansal (2016) intègrent également d'autres caractéristiques linguistiques telles que les partie du discours des mots ou leurs rôles de dépendance syntaxiques.

Les méthodes d'extraction conjointe de relations et d'entités sont généralement évaluées en termes de scors de Précision, de Rappel et F1 sur les mêmes jeux de données que les méthodes d'extraction de relations, tels que les jeux de données ACE 2004 ou ACE 2005 (Doddington et al., 2004). Comme l'indique Pawar et al. (2021), les approches JERE sont capables d'obtenir des scores F1 d'environ 83 % sur la tâche de reconnaissance d'entités nommées. Toutefois, elles obtiennent généralement des scores F1 inférieurs, jusqu'à 57 %, pour la catégorisation conjointe d'entités et

de relations, ce qui témoigne de la difficulté de la tâche.

Chapitre 2

Corpus de documents historiques

Table des matières

Ce chapitre est consacré à la description de notre jeu de données et aux étapes de prétraitement qui ont été nécessaires pour préparer les données à la tâche ultérieure d'extraction conjointe d'entités et de relations. Nous présentons le corpus EMONTAL, qui consiste en des documents périodiques imprimés publiés aux 19ème et 20ème siècles dans les régions de Bourgogne et de Franche-Comté en France. Nous avons choisi de nous concentrer sur les périodiques tels que les journaux, étant donné leur importance en tant que sources primaires reflétant les événements locaux et internationaux (Tibbo, 2007). Ces documents sont donc adaptés à notre étude, qui se concentre sur l'extraction d'informations relatives à des personnes et à des lieux à partir de documents historiques. Nous avons choisi de limiter le champ d'application de ce corpus aux documents imprimés, et d'exclure les documents manuscrits, afin d'éviter les problèmes liés à la transcription des manuscrits avec des méthodes d'OCR. De plus, nous choisissons de limiter le champ de ce corpus aux documents publiés aux 19ème et 20ème siècles, afin de s'assurer que la langue dans laquelle les documents sont écrits est aussi proche que possible de l'état moderne de la langue française.

Le contenu textuel des documents de notre corpus est stocké au format XML ALTO, qui fournit la transcription OCR du document, son ordre de lecture et sa structure physique. En raison du processus OCR, les transcriptions du contenu textuel des documents peuvent contenir des artefacts et des erreurs, qui peuvent avoir un impact sur la qualité de toutes méthodes de TAL appliquées au

contenu du document, telle que la reconnaissance d'entités nommées ou l'extraction de relations. Le nettoyage des artefacts produits par le processus d'OCR est donc une étape nécessaire, afin de minimiser leur impact sur l'application des méthodes de TAL. En outre, le format XML ALTO ne fournit pas la structure logique des documents, ce qui rend les documents de notre corpus inadaptés à l'application des méthodes de TAL. Nous avons donc développé notre propre méthodologie pour déterminer la structure logique des documents de notre corpus, qui est une étape nécessaire pour obtenir des données textuelles structurées.

Le reste de ce chapitre est structuré comme suit : nous décrivons le corpus EMONTAL au chapitre 2.1. Au chapitre 2.2, nous présentons les méthodes que nous appliquons pour nettoyer le contenu textuel des documents de notre corpus. Dans le chapitre 2.3, nous présentons notre approche pour détecter la structure logique des documents à partir de leur représentation au format XML ALTO, ainsi que l'évaluation de cette approche. Enfin, nous proposons dans le chapitre 2.4 un nouveau format XML nommé EMONTAL permettant de décrire la structure logique des documents, qui est adapté à l'application de méthodes de TAL.

2.1 Description du corpus EMONTAL

Les documents du corpus EMONTAL ont été collectés depuis le Fond régional : Franche-Comté et depuis Fond régional : Bourgogne sur Gallica[1], le service numérique de la Bibliothèque Nationale de France (BnF), à l'aide des API mises à disposition. Nous n'avons collecté que les documents de ces collections qui ont été traités avec l'OCR. Les documents du corpus EMONTAL sont stockés au format XML. Les métadonnées des documents sont stockées dans la balise `oai`, tandis que le contenu textuel du document est stocké dans la balise `ocr`, qui suit le format XML ALTO. La liste complète des collections de notre corpus figure à l'annexe 19.4.

La Figure 2.1 montre la distribution des documents par décennies dans les fonds Franche-Comté et Bourgogne du corpus EMONTAL. Les documents que nous avons collectés dans le fonds Franche-Comté ont été publiés entre 1840 et 1940, bien que certains d'entre eux aient également été publiés dans les années 1990. La plupart des documents ont été publiés entre 1900 et 1910, avec un pic de 1 028 publications dans les années 1910. Les documents que nous avons collectés dans le fonds Bourgogne ont été publiés entre 1810 et 1960. La plupart des documents ont été publiés entre 1890 et 1930, avec un pic de 1 236 publications dans les années 1910 et 1920. L'absence de documents publiés avant 1900 s'explique par le manque de transcriptions OCR disponibles sur la plateforme Gallica. La baisse des publications dans les deux fonds dans les années 1940 s'explique par les événements de la Seconde Guerre mondiale, qui ont imposé des restrictions sur le papier. Par ailleurs, les documents publiés après les années 1940 peuvent manquer car ils ne

[1] https://gallica.bnf.fr/html/und/france/bourgogne-franche-comte

sont pas encore tombés dans le domaine public. La Table 2.1 montre la distribution des collections, des documents et des pages, ainsi que des balises `TextBlock`, `TextLine` et `String`, qui sont propres au format XML ALTO, dans le corpus EMONTAL.

	Collection	Numéro	Page	`TextBlock`	`TextLine`	`String`
Fond Franche-Comté	46	2,648	255,670	3,733,845	11,373,606	83,001,454
Fond Bourgogne	113	5,738	637,407	8,117,055	24,183,489	189,266,170
Total	159	8,386	893,077	11,850,900	35,557,095	272,267,624

Table 2.1: Distribution des collections, des documents et des pages, ainsi que des balises `TextBlock`, `TextLine` et `String` dans le corpus EMONTAL

Les documents du corpus EMONTAL appartiennent à huit grandes catégories thématiques, que nous avons identifiées manuellement en analysant les titres et le contenu de chaque collection:

Agriculture : documents traitant de questions agricoles, en particulier liées aux productions locales telles que les vins. Ces documents ont été publiés entre 1860 et 1940, avec un pic de 614 publications dans les années 1920

Combattants et patriotisme : des journaux ou périodiques patriotiques publiés par des combattants, comme les résistants pendant la Seconde Guerre Mondiale. Ces documents ont été publiés entre 1900 et 1940, avec un pic de 113 publications dans les années 1900

Journaux généralistes et partisans : journaux généralistes ainsi que des journaux affiliés à des partis politiques. Ces documents ont été publiés entre 1900 et 1940, avec un pic de 682 publications dans les années 1920

Loisirs : des périodiques traitant de divers loisirs, tels que la photographie ou le sport. Ces documents ont été publiés entre 1900 et 1930, avec un pic de 166 publications dans les années 1920

Pouvoirs locaux et économie : documents traitant de questions économiques, ainsi que des décisions prises par les autorités locales. Ces documents ont été principalement publiés entre 1830 et 1940, avec un pic de 194 publications dans les années 1870, bien que certains documents aient été publiés dans les années 1980 et 1990

Vie religieuse : documents publiés par des organismes religieux tels que les paroisses locales, traitant de la vie religieuse et des événements dans la région. Ces documents ont été publiés entre 1890 et 1930, avec un pic de 1 110 publications dans les années 1910

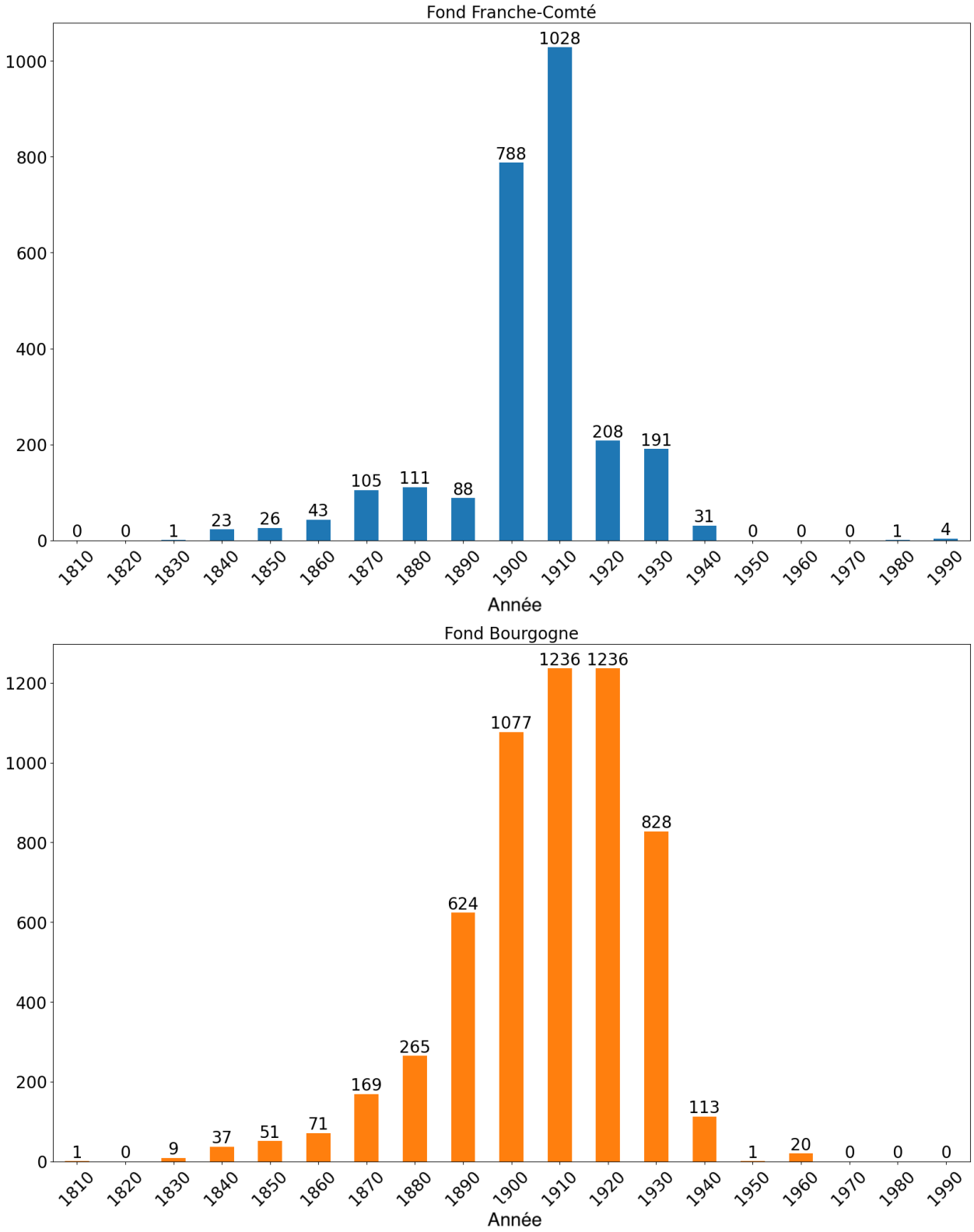


Figure 2.1: Distribution des documents par décennies dans les fonds *Franche-Comté* et *Bourgogne* du corpus EMONTAL

Science et connaissance : documents relatifs aux sciences et à la culture, tels que l'histoire, l'archéologie ou la littérature. Ces documents ont été publiés entre 1830 et 1960, avec un pic de 183 publications dans les années 1930

Femmes dans la société : documents publiés par des femmes ou traitant du statut des femmes dans la société. Ces documents ont été publiés entre 1910 et 1940, avec un pic de 9 publications dans les années 1920.

La Table 2.2 présente des exemples de titres de documents pour chaque catégorie thématique. La liste des collections de documents appartenant à chaque catégorie thématique est présentée dans l'annexe 19.4.

Catégorie	Titre
Agriculture	Bulletin du Comité d'agriculture de Beaune La Bourgogne rurale : revue mensuelle dágriculture, de viticulture et dhorticulture
Combattants et patriotisme	Le Petit écho du 21e Régiment d'infanterie Jeunesse du Maquis
Journaux généralistes et partisans	Le Franc-Comtois de Paris : journal d'information des départements Doubs, Jura, Haute-Saône, Haut-Rhin Le Semeur. Organe régional du Parti communiste
Loisirs	Les Sports de l'Est : hebdomadaire illustré de tous les sports, Lorraine, Franche-Comté, Bourgogne, Champagne Le Collectionneur : bulletin mensuel de publicité, philatélie, cartophilie...
Pouvoirs locaux et économie	Tableaux de l'économie bourguignonne Rapports et délibérations / Conseil général du Doubs
Vie religieuse	Bulletin paroissial de Notre-Dame, Dijon Vers l'avenir : organe mensuel de la Jeunesse catholique de Franche-Comté et du Territoire-de-Belfort
Science et connaissance	Bulletin de la Société archéologique et biographique du canton de Montbard La Vie meilleure : revue sociologique et littéraire
Femmes dans la société	La Voix des femmes de Bourgogne La France féminine : revue mensuelle pendant la guerre : organe de défense des droits de la femme française

Table 2.2: Exemples de documents par catégorie thématique dans le corpus EMONTAL

2.2 Correction des transcriptions OCR des documents

Les documents de notre corpus ayant été traités par OCR, leurs transcriptions peuvent contenir des erreurs et des artefacts qui doivent être corrigés afin de limiter leur impact sur l'application des méthodes de TAL. Dans notre travail, nous avons identifié deux types d'erreurs principales dans les transcriptions OCR : la césure qui apparaît à la fin des lignes, et les caractères mal transcrits qui conduisent à des tokens incorrects.

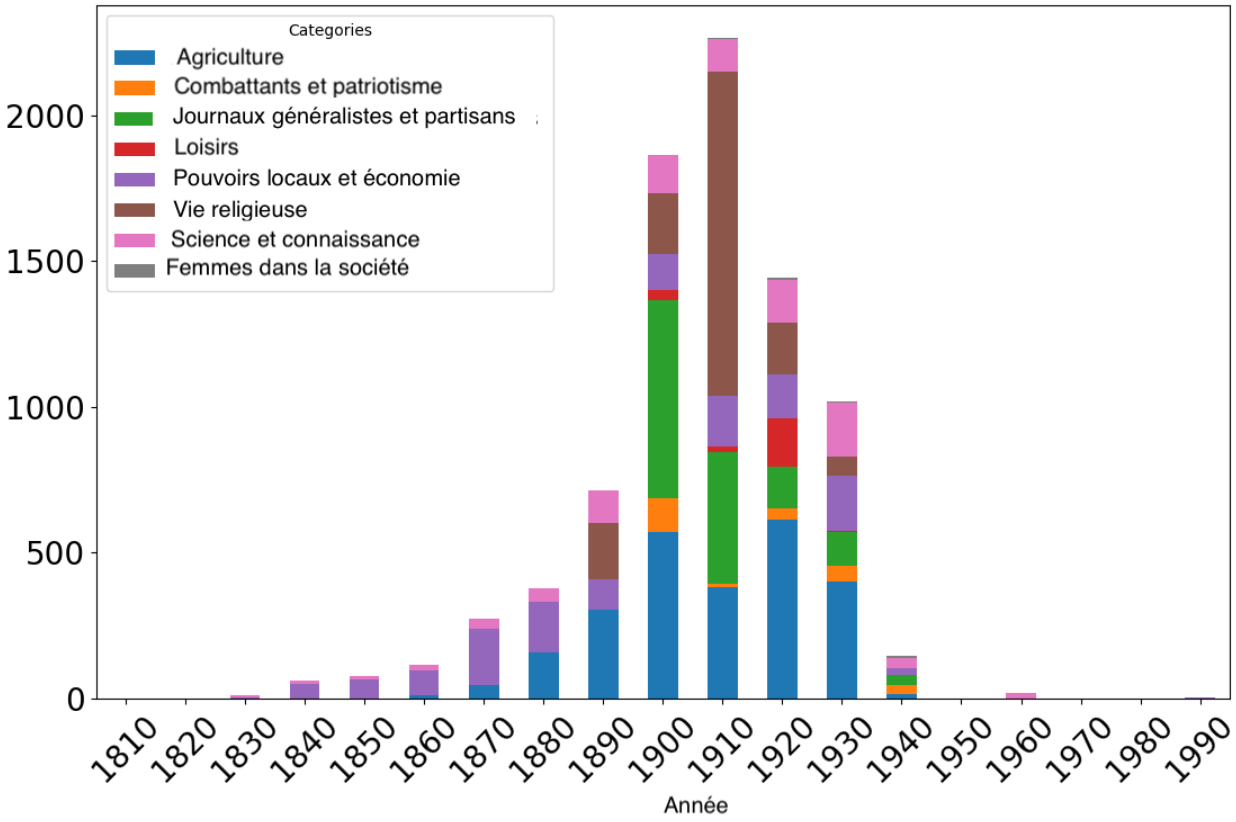


Figure 2.2: Distribution des documents par catégorie thématique et par décennie

En typographie, le procédé de césure consiste à diviser un mot par un trait d'union, de manière à respecter la justification du texte. La césure peut donc introduire des artefacts, surtout si le texte est transcrit à l'aide de méthodes OCR. Il est ainsi nécessaire de procéder à un post-traitement des traits d'union afin d'éviter que ces artefacts n'interfèrent avec le traitement ultérieur. Des règles de traitement simples peuvent être utilisées pour supprimer la césure et corriger les tokens séparés, comme dans Généreux and Spano (2015). Dans les documents XML ALTO, les traits d'union sont représentés par la balise `Hyp`. Les balises `Hyp` sont situées à la fin d'une balise `TextLine` et peuvent apparaître dans deux contextes possibles : soit la balise `TextLine` est suivie d'une autre balise `TextLine`, soit la balise `TextLine` est la dernière balise du `TextBlock` qui la contient.

Nous proposons un méthode simple pour traiter les balises `Hyp` dans les documents XML ALTO. Si la balise `Hyp` apparaît dans le premier contexte, nous supprimons la balise `String` suivante dans la structure XML, afin d'éviter de collecter le contenu textuel d'un mot deux fois par la suite. Nous supprimons ensuite la balise `Hyp` de la structure XML. Si la balise `Hyp` apparaît dans le deuxième contexte, nous la supprimons simplement de l'arbre XML, sans autre traitement. Les principales étapes de cette méthodes sont présentées dans la Figure 2.3.

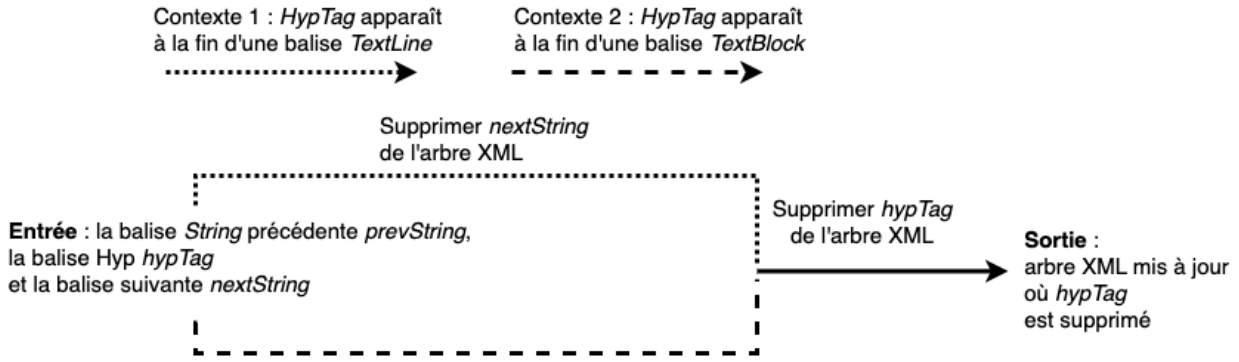


Figure 2.3: Etapes principales pour l'implémentation de la méthode pour supprimer les césures

Le processus d'OCR peut produire deux types d'erreurs : soit des erreurs de non-mot, c'est-à-dire la substitution ou l'omission de caractères qui aboutissent à des mots inexistants, soit des erreurs de mot réel, c'est-à-dire la substitution ou l'omission de caractères qui aboutissent à des mots existants, différents des mots réels contenus dans le document (T. T. H. Nguyen et al., 2021a). Nous proposons une méthode basée sur des règles pour corriger les erreurs de non-mot dans les transcriptions OCR. Notre objectif est de proposer une méthode simple pour nettoyer les transcriptions OCR et limiter l'impact des erreurs de non-mot sur les processus ultérieurs. Nous avons conçu ces règles en observant les erreurs de non-mot dans notre corpus. Cette méthode se compose des quatre étapes, qui sont illustrées dans la Figure 2.4. Ces étapes visent à corriger un document en conservant, supprimant ou corrigeant un mot, sur la base de caractéristiques extraites des mots dans les documents.

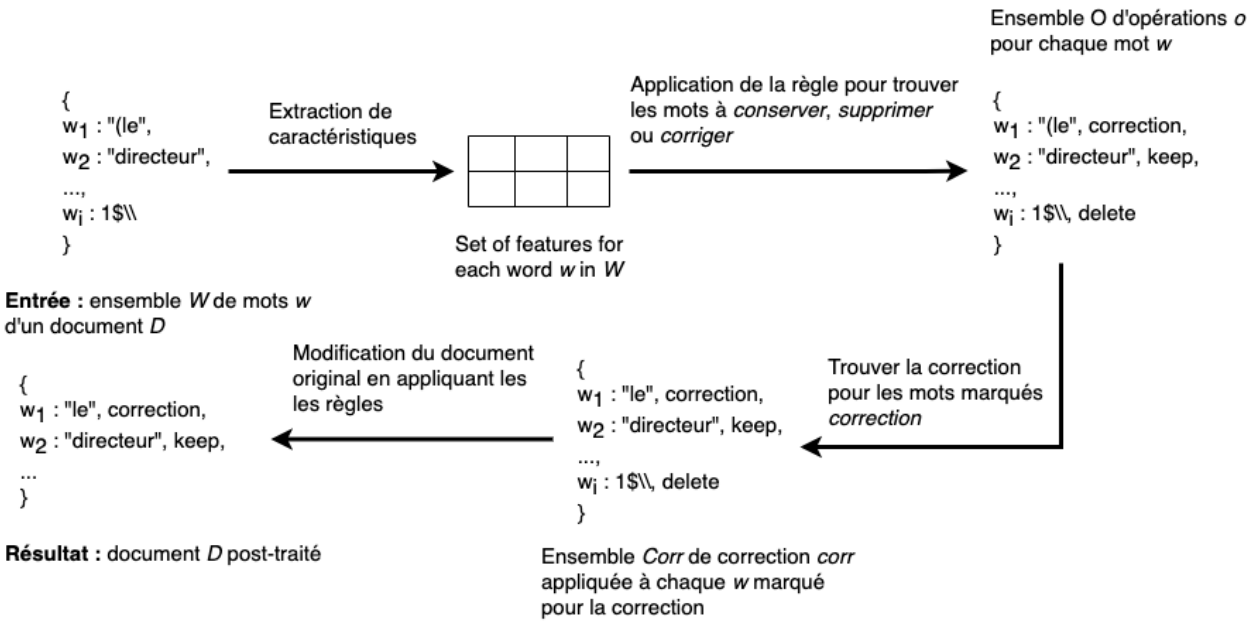


Figure 2.4: Etapes principales pour l'implémentation de la méthode de post-traitement de l'OCR

L'étape d'*extraction de caractéristiques* consiste à extraire les caractéristiques de chaque mot du document. Nous extrayons des caractéristiques morphologiques et sémantiques décrivant la structure des mots. L'ensemble des caractéristiques extraites est décrit dans la Table 2.3.

Caractéristiques	Description
word_length	le nombre de caractères du mot
stw_capital	Vrai si le mot commence par une majuscule, sinon Faux
stw_elision	Vrai si le mot commence par une élision, sinon Faux
non_alpha_prop	la proportion de caractères non alphanumériques dans le mot
ends_punct	Vrai si le mot se termine par un signe de ponctuation, sinon Faux
is_punct	Vrai si le mot est une ponctuation, sinon Faux
is_digit	Vrai si le mot est un chiffre, sinon Faux
is_oneletter_word	Vrai si le mot est dans l'ensemble $\{a,à,y,ô,m\}$, sinon Faux

Table 2.3: Ensemble des caractéristiques extraites de chaque mot dans un document pour le post-traitement de l'OCR

L'étape d'*application des règles* consiste à appliquer l'ensemble des règles pour déterminer quelle opération doit être appliquée à un mot, c'est-à-dire s'il doit être conservé tel quel, supprimé ou corrigé. Par défaut, les mots sont conservés tels quels. Les règles sont décrites dans la Table 2.4, et sont appliquées dans l'ordre indiqué. Les règles 1 et 2 sélectionnent les candidats pour l'opération *Supprimer*, tandis que la règle 3 sélectionne les candidats pour l'opération *Corriger*. La règle 1 stipule que la proportion de caractères non alphanumériques dans un mot doit être supérieure

ou égale à 75. Cette valeur a été déterminée de manière empirique.

Règle	Objectif	Conditions	Opération
1	Supprimer les mots non alphanumériques	W.non_alpha_prop >= 75 et pas W.is_punct	Supprimer
2	Supprimer des mots d'un seul caractère	pas W.is_oneletter_word and W.word_length = 1 et pas W.is_punct et pas W.is_digit	Supprimer
3	Proposer une correction de texte	pas W.stw_capital et pas W.ends_punct et pas W.is_punct et pas W.is_digit et pas W.stw_elision	Corriger

Table 2.4: Règles de post-traitement de l'OCR pour déterminer si un mot doit être conservé, supprimé ou corrigé. L'opération par défaut est *Conserver*

L'étape de *suggestion de correction* consiste à traiter les mots marqués pour correction. Tout d'abord, tout caractère répété trois fois ou plus est remplacé par une seule occurrence. Par exemple, la transcription incorrecte "*mercrediii*" serait corrigée en "*mercredi*". Nous appliquons ensuite un algorithme de correction orthographique pour trouver la correction la plus probable pour le mot. Pour notre implémentation, nous avons choisi d'appliquer l'algorithme SymSpell (Symmetric Delete spelling)[2] as in Huynh et al. (2020). Si aucune correction possible ne peut être trouvée, le mot est conservé tel quel.

Enfin, l'étape de *conversion du document* modifie le document en appliquant les opérations déterminées pour chaque mot.

2.3 Comprendre la structure logique des documents

Le contenu textuel des documents de notre ensemble de données est présenté dans le format XML ALTO, qui décrit l'ordre de lecture et la structure physique du document. Cependant, ce format ne fournit pas la structure logique du document, c'est-à-dire son organisation en éléments logiques tels que les en-têtes, les titres, les paragraphes ou les phrases. Nous proposons une approche pour la tâche d'analyse de la structure logique des documents (*Logical Layout Analysis*, LLA), qui vise à attribuer une étiquette logique aux balises `TextBlock` et `TextLine` dans les documents XML ALTO. Les détails et l'évaluation de cette méthode ont été présentés précédemment dans Gutehrlé and Atanassova (2021b, 2022). Cette méthode consiste en deux étapes principales, qui sont illustrées dans la Figure 2.5.

[2]Disponible sur GitHub: https://github.com/wolfgarbe/SymSpell

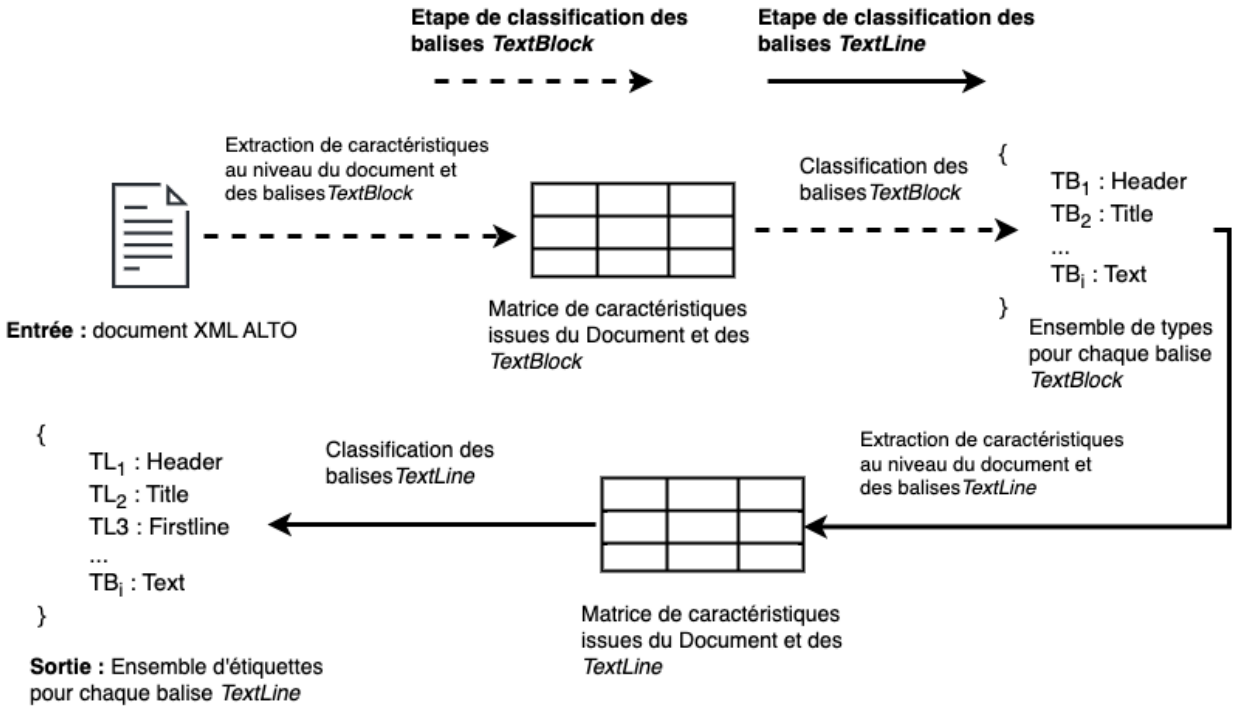


Figure 2.5: Etapes principales pour l'implémentation de notre approche pour la tâche d'analyse de la structure logique des documents (LLA)

L'étape de *classification de TextBlock* attribue des étiquettes logiques aux balises `TextBlock`. Une balise `TextBlock` n'est traitée que si elle ne possède pas déjà un attribut *type*, ce qui est le cas de la plupart des balises `TextBlock` de notre ensemble de données. Cette étape met à jour le document XML ALTO en ajoutant un attribut *type* aux balises `TextBlock` avec les étiquettes logiques prédites. L'étape *TextLine classification* attribue des étiquettes logiques aux balises `TextLine`, sur la base des caractéristiques extraites et des résultats de l'étape d'annotation des balises `TextBlock`.

Nous proposons un jeu de données pour développer et évaluer notre approche de la tâche LLA. Nous avons construit ce jeu de données en sélectionnant des documents du corpus EMONTAL dont l'attribut `nqamoyen`, c'est-à-dire la mesure de la qualité de l'OCR, est supérieur ou égal à 90%. Cet ensemble de données est disponible sur Zenodo (Gutehrlé & Atanassova, 2021a). Pour chaque document XML ALTO sélectionné, nous stockons deux fichiers CSV, l'un pour l'annotation des balises `TextBlock` et l'autre pour l'annotation des balises `TextLine`. L'ensemble de données a été annoté manuellement par un seul annotateur, puis divisé en un jeu d'entraînement et un jeu de test. La Table 2.5 montre la répartition des collections, des numéros, des pages et des balises `TextBlock`, `TextLine` et `String` dans les jeux de données d'entraînement et de test.

Nous avons divisé les documents en trois catégories de mise en page : **1c, 2c** et **3c+**. Notre objectif est d'évaluer la capacité de notre méthode à s'adapter aux différentes mises en page du

	Collection	Numéro	`TextBlock`	`TextLine`	`String`	Page
Train	15	48	4,608	51,815	338,583	368
Test	6	6	1,445	8,836	63,343	52
Total	21	54	6,053	60,651	401,926	420

Table 2.5: Distribution des collections, des numéros, des pages et des balises `TextBlock`, `TextLine` et `String` dans les jeux de données d'entraînement et de test

corpus EMONTAL. La Table 2.6 présente la distribution des documents dans les jeux de données d'entraînement et de test dans les trois catégories de mise en page.

1c : documents dont le texte est affiché sur une seule colonne

2c : documents dont le texte est affiché sur deux colonnes

3c+ : documents dont le texte est affiché sur trois colonnes ou plus

Mise en page	Entraînement	Test
1c	18	2
2c	5	2
3c+	25	2
Total	48	6

Table 2.6: Distribution des documents par catégories de mise en page dans les jeux de données d'entraînement et de test pour la tâche d'analyse de la structure logique des documents

Nous avons défini l'ensemble d'étiquettes présenté dans le tableau 2.7 pour annoter les balises `TextBlock` et `TextLine` dans notre jeu de données. Cet ensemble d'étiquettes permet de déterminer la catégorie logique des documents, en annotant chaque élément `TextBlock` et `TextLine` avec l'une des étiquettes que nous avons définies.

Etiquette	Text	Firstline	Title	Header	Other
`TextBlock`	X		X	X	X
`TextLine`	X	X	X	X	X

Table 2.7: Ensemble d'étiquettes logiques pour les balises `TextBlock` and `TextLine`

L'étiquette *Firstline* doit être comprise comme "première ligne du paragraphe". Ainsi, toute balise `TextLine` étiquetée *Firstline* indique le début d'un paragraphe. Une petite partie des balises `TextBlock` et `TextLine` correspond à des éléments qui ne sont pas pertinents pour

notre étude, tels que des images, des tableaux ou des publicités. Ces éléments sont étiquetés comme *Autre*(*Other*) et sont ignorés pour l'évaluation. La Table 2.8 présente la distribution des étiquettes dans les jeux de données, obtenue après l'annotation manuelle de tous les éléments `TextBlock` et `TextLine`.

		Train		Test	
	Etiquette	Nombre	Proportion	Nombre	Proportion
TextBlock	Header	333	7.380 %	53	3,860 %
	Other	1,686	37.367 %	128	9.322 %
	Text	2,064	45.744 %	1,102	80.262 %
	Title	429	9.507 %	90	6.554 %
TextLine	Firstline	9,785	18.921 %	1,563	17.840 %
	Header	740	1.430 %	115	1.312 %
	Other	3,098	5.990 %	201	2.294 %
	Text	36,272	70.138 %	6,648	75,881 %
	Title	1,820	3.519 %	234	2.670 %

Table 2.8: Distribution des étiquettes logiques pour les balises `TextBlock` and `TextLine` dans les jeux d'entraînement et de test pour la tâche d'analyse de la structure logique des documents

A partir de ce jeu de données, nous extrayons et calculons des ensembles de caractéristiques géométriques, morphologiques et sémantiques des documents XML ALTO aux niveaux des balises `TextLine`, des balises `TextBlock` et du document. Ces caractéristiques, leurs descriptions et leurs niveaux sont présentés dans la Table 2.9.

Nous employons ces ensembles de caractéristiques pour mettre en place et évaluer trois implémentations de notre approches. Une première implémentation repose sur un ensemble de règles conçues manuellement par observation du corpus EMONTAL. Une seconde implémentation repose sur l'algorithme d'apprentissage de règles (*rule-learning*) RIPPER. Enfin, la troisième implémentation repose sur l'algorithme d'apprentissage automatique (*machine-learning*) Gradient Boosting. La Table 2.10 présente les scores moyens de Précision, de Rappel et F1 de chaque implémentation pour les tâches de classification `TextBlock` et `TextLine`.

	Caractéristique	Description	TextLine	TextBlock	Document
1	*page*	numéro de page de la page contenant l'élément	X	X	
2	*blockType*	type de bloc	X		
3	*wordCount*	nombre de mots	X	X	
4	*precedingSpace, followingSpace*	les espaces au-dessus et au-dessous de l'élément	X	X	
5	*height, width*	les valeurs de hauteur et de largeur de la ligne	X		
6	*hpos, vpos*	les coordonnées de la ligne sur la page, c'est-à-dire sa position horizontale et verticale	X		
7	*diffHpos*	différence entre *hpos* et la valeur médiane de *hpos* dans le bloc	X		
8	*firsthpos, firstvpos*	coordonnées de la première ligne du bloc		X	
9	*lasthpos, lastvpos*	coordonnées de la dernière ligne du bloc		X	
10	*linecount*	nombre de lignes dans le bloc		X	
11	*wordRatio*	nombre de mots par ligne		X	
12	*medHeight, medWidth*	la hauteur et la largeur médianes des lignes		X	X
13	*medHpos, medVpos*	valeurs médianes de *hpos* et *vpos* dans le bloc		X	
14	*medWordCount, medLineSpace*	le nombre médian de mots par ligne et l'espace médian entre les lignes dans le bloc		X	X
15	*medBlockHeight, medBlockWidth*	la hauteur médiane des lignes et la hauteur et la largeur des blocs			X
16	*medBlockSpace*	valeur médiane de l'espace entre les blocs			X
17	*thirdQuartileLineSpace*	troisième quartile des valeurs d'interligne dans le document			X
18	*medWordRatio, medLineCount*	nombre médian de mots par ligne et nombre médian de lignes par bloc dans le document			X
19	*capitalProp, digitProp* *nonAlphaProp*	proportion de lettres majuscules et de chiffres proportion de caractères non alphanumériques	X	X	
20	*stwCapital, stwDigit*	Vrai si la ligne commence par une lettre majuscule ou un chiffre, Faux sinon	X		
21	*endsPunct*	Vrai si la ligne se termine par une ponctuation, Faux sinon	X		
22	*headerMark1*	Vrai si l'élément contient le mot "Page" ou un tiret, Faux sinon	X	X	
23	*headerMark2*	Vrai si l'élément contient une date, une devise, une adresse, Faux sinon	X	X	
24	*simTitle*	similarité de la ligne avec le titre du document, calculée par la distance de Levenshtein	X		
25	*simHeaderSet*	la plus grande similarité de la ligne avec les mots contenus dans l'ensemble des mots d'en-tête, calculée par la distance de Levenshtein	X		

Table 2.9: Caractéristiques extraites au niveau des balises `TextLine`, `TextBlock` et du document pour notre approche pour la tâche d'analyse de la structure logique des documents (LLA)

		Text			Title			Firstline			Header		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
TextBlock	RB	**0.959**	0.966	**0.962**	0.600	**0.639**	**0.610**				**0.726**	**0.298**	**0.406**
	RP	0.841	**0.996**	0.912	0.560	0.455	0.480				0.440	0.194	0.268
	GB	0.907	0.938	0.920	**0.665**	0.425	0.492				0.431	0.144	0.215
TextLine	RB	**0.969**	**0.991**	**0.979**	0.595	**0.733**	**0.639**	**0.949**	**0.861**	**0.902**	**0.803**	**0.348**	**0.435**
	RP	0.890	0.912	0.901	0.640	0.255	0.363	0.671	0.788	0.724	0.444	0.039	0.072
	GB	0.846	0.961	0.898	**0.788**	0.254	0.383	0.678	0.482	0.538	0.625	0.108	0.183

Table 2.10: Scores de Précision, Rappel et F1 moyens de l'approche à base de règles (RB), de l'ensemble de modèles RIPPER (RP) et du modèle Gradient Boosting (GB) pour les tâches de classification `TextBlock` et `TextLine`. Les meilleurs scores pour chaque métrique et pour chaque étiquette sont indiqués en gras

Les trois approches obtiennent des résultats relativement bons pour la tâche de classification `TextBlock`. Notre approche basée sur des règles obtient le meilleur score de Précision et le meilleur score F1 de 0,959 et 0,962 respectivement, tandis que RIPPER obtient le meilleur score de Rappel, qui est de 0,996. Les performances des trois approches se situent entre 0,425 et 0,665 pour la classification des blocs Title. Gradient Boosting obtient la meilleure Précision moyenne avec 0,665, tandis que le système à base de règles obtient le meilleur Rappel moyen et le meilleur score F1 de 0,639 et 0,610 respectivement. Enfin, les performances pour la classification des blocs Header varient de 0,144 à 0,726. Notre approche à base de règles obtient les meilleurs scores de Précision, de Rappel et F1, qui sont respectivement de 0,726, 0,298 et 0,406. Le score de Précision est plus élevé que le score de Rappel pour chaque approche, ce qui suggère que la définition de règles exhaustives pour détecter les blocs Header est une tâche difficile.

De même, pour la tâche de classification `TextLine`, les trois approches obtiennent de bons résultats. Les performances pour la classification des lignes de Titres sont variables. Gradient Boosting obtient un score de Précision de 0,788. Cependant, notre approche obtient les meilleurs scores de Rappel et de F1, de 0,733 et 0,639, qui sont plus élevés que ceux des deux autres modèles. Pour la classification Firstline, notre approche obtient les meilleurs scores dans chaque métrique, avec une Précision de 0,949, un Rappel de 0,861 et un F1 de 0,902. Gradient Boosting n'obtient qu'un score F1 moyen de 0,538, ce qui suggère que les paragraphes sont plus faciles à détecter avec des règles simples. Enfin, les performances de la classification des lignes Header varient considérablement. Notre approche obtient les meilleurs scores de Précision, de Rappel et F1, qui sont respectivement de 0,803, 0,348 et 0,435. De même, le score de Précision est plus élevé que le score de Rappel pour chaque modèle.

En conclusion, notre approche basée sur des règles est plus performante que les deux autres approches dans presque toutes les évaluations. Elle obtient en particulier de meilleurs résultats en termes de Rappel, ce qui indique qu'elle couvre plus de types de chaque étiquette logique que les deux autres modèles. En comparant Gradient Boosting et RIPPER, nous constatons qu'en général, Gradient Boosting obtient de meilleurs scores de Précision, tandis que RIPPER obtient de meilleurs scores de Rappel. En outre, nous remarquons que le modèle RIPPER obtient des scores F1 plus faibles pour la tâche de classification `TextLine` que pour la tâche de classification `TextBlock`. L'étape de classification `TextBlock` est conçue comme une étape intermédiaire qui devrait faciliter la classification de balises `TextLine`, qui est le résultat final de l'algorithme. Cependant, contrairement à notre approche, RIPPER n'utilise que peu les résultats de la classification `TextBlock`. Cela pourrait expliquer ses résultats inférieurs pour la tâche de classification des balises `TextLine`.

Les résultats obtenus par notre approche montrent la pertinence de notre méthode à base de règles, que nous avons spécifiquement conçue pour les documents historiques. Cependant, comme le montre l'évaluation comparative, le modèle Gradient Boosting peut atteindre des scores de Précision très élevés. Il serait donc intéressant de combiner notre approche avec l'approche d'apprentissage automatique de manière hybride. Chaque modèle serait entraîné à détecter des étiquettes spécifiques, ce qui permettrait de surmonter les faiblesses des autres systèmes. L'algorithme RIPPER peut apprendre des règles très précises, qui sont parfois plus performantes que nos règles élaborées manuellement. En outre, RIPPER obtient en moyenne de meilleurs scores de Précision que de Rappel. Cela suggère que RIPPER est meilleur pour produire des règles fines plutôt que des règles générales. La production manuelle de règles étant une une tâche fastidueuse, il serait intéressant d'utiliser RIPPER de manière exploratoire afin de produire un ensemble de règles de base avec des scores de Précision élevés. Cet ensemble de règles pourrait ensuite être mis à jour manuellement afin d'améliorer son Rappel. Par ailleurs, la mise en page des documents historiques évolue rapidement, notamment dans les journaux. Ainsi, le modèle RIPPER pourrait s'avérer un outil utile pour construire rapidement un ensemble de règles adapté à des périodes de publication spécifiques.

2.4 Convertir les documents XML

Nous proposons de convertir les documents XML originaux du corpus EMONTAL en un format XML allégé personnalisé appelé EMONTAL. Ce format décrit la structure logique du document et est adapté à l'application de méthodes TAL. Ces documents XML sont structurés selon le schéma présenté dans la Figure 2.6. La définition complète du schéma XML (XSD) figure en annexe 19.4.

Le document XML est contenu dans une balise `document`, qui est divisée en une balise `metadata` et une balise `content`. La balise `metadata` contient les métadonnées des docu-

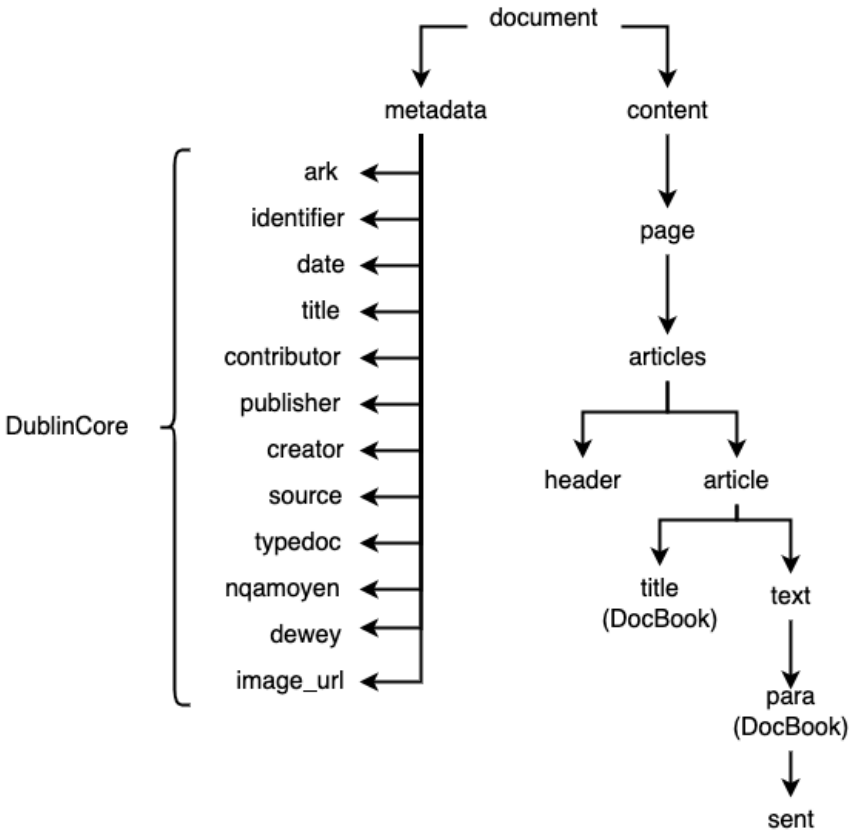


Figure 2.6: Structure du format XML EMONTAL

ments. Les balises contenues dans la balise `metadata` sont un sous-ensemble sélectionné de la balise `oai` du document XML original, et suivent le format Dublin Core. La Table 2.11 fournit une description de chaque balise de la balise `metadata`.

Balise	Description
ark	l'identifiant de l'arche du document, tel qu'il est utilisé dans Gallica
identifier	url du document sur Gallica
date	la date de publication du document. Cette valeur peut être présentée sous deux formats : AAAA-MM ou AAAA-MM-JJ.
title	le titre du document
contributor	personne ayant contribué à la création du document
publisher	l'éditeur original du document
creator	le créateur du document
source	le stockage original du document
typedoc	le type de document
nqamoyen	le score de qualité de l'OCR
dewey	la catégorie du document, selon la classification décimale de Dewey (CDD)
image_url	l'url du document scanné

Table 2.11: Ensemble de balises contenues dans la balise `metadata` du format XML EMONTAL. Ces balises suivent le format Dublin Core

La balise `content` fournit la structure logique du document et son contenu textuel. Elle s'inspire du format XML DocBook, qui permet de décrire la structure logique des documents dans un format léger et flexible. La Table 2.12 fournit une description de chaque balise de la balise *content*. Les balises qui peuvent être trouvées dans le jeu de balises XML DocBook original sont marquées par le symbole *.

Chaque balise, à l'exception de la balise `articles`, possède un attribut *id* qui l'identifie de manière unique dans le document. Les articles présents sur plusieurs pages ont le même attribut *id*. Les balises *para* ont un attribut *block_id* qui les relie à la balise `TextBlock` correspondante dans le document XML ALTO d'origine. La Figure 2.7 et la Figure 2.8 montrent un exemple des balises `metadata` et `content` suivant le schéma XML EMONTAL.

Les balises `ent` ont un attribut *id* et un attribut *type*. L'attribut *id* des balises `ent` est composé de leur numéro d'index dans l'ensemble du document, suivi de leur type, tel que "pers_191". L'attribut *type* indique le type de l'entité, par exemple *Personne* ou *Lieu*. L'attribut *id* est constitué du type de l'entité suivi d'un nombre.

Les balises `rel` sont imbriquées dans les balises `ent` et décrivent une relation dans laquelle une entité est impliquée. La balise `rel` n'a pas de contenu textuel. Chaque balise `rel` possède un attribut *id*, *type*, *score* et *anchor*. L'attribut *id* des balises `rel` est composé du numéro d'index dans la phrase, suivi du type de relation, tel que *2_dateOfBirth*. L'attribut *type* indique le type de relation,

```xml
<?xml version="1.0" encoding="utf-8"?>
<document>
<metadata>
<ark>
bpt6k5839946n
</ark>
<identifier>
https://gallica.bnf.fr/ark:/12148/bpt6k5839946n
</identifier>
<date>
1900
</date>
<title>
Mémoires et documents inédits pour servir à l'histoire de la Franche-Comté / publiés par l'Académie de Besançon
</title>
<contributor/>
<publisher>
Académie de Besançon (Besançon)
</publisher>
<languague/>
<creator>
Académie des sciences, belles-lettres et arts de Besançon et de Franche-Comté. Auteur du texte
</creator>
<source>
Bibliothèque nationale de France, département Collections numérisées, 2008-216649
</source>
<typedoc>
fascicule
</typedoc>
<nqamoyen>
99.98
</nqamoyen>
<dewey>
0
</dewey>
<image_url>
https://gallica.bnf.fr/ark:/12148/bpt6k5839946n/highres
</image_url>
</metadata>
```

Figure 2.7: Extrait de la section des `metadata` d'un document au format XML EMONTAL dans notre corpus (Mémoires et documents inédits pour servir à l'histoire de la Franche-Comté, 1900)

```
<content>
 <page id="1">
  <header block_id="PAG_00000001_TB000005" id="header_1">
   MEMOIRES
   ET
   DOCUMENTS INÉDITS
   POUR SERVIR A L' HISTOIRE
   DE LA FRANCHE-COMTÉ
  </header>
  <articles>
   <article id="article_01">
    <title id="title_01">
     T. IX.
    </title>
    <text id="text_01"/>
   </article>
  </articles>
 </page>
 <page id="3">
  <header block_id="PAG_00000003_TB000004" id="header_3">
   ET
   DOCUMENTS INÉDITS
   POUR SERVIR A L'HISTOIRE
  </header>
  <articles>
   <article id="article_01">
    <title id="title_01">
     T. IX.
    </title>
    <text id="text_01">
     <para block_id="PAG_00000003_TB000001" id="para_1">
      <sent id="sent_1">
       MÉMOIRES
      </sent>
     </para>
```

Figure 2.8: Extrait de la section des `content` d'un document au format XML EMONTAL dans notre corpus (Mémoires et documents inédits pour servir à l'histoire de la Franche-Comté, 1900)

Balise	Description
page	le contenu d'une page du document
header	le contenu de l'en-tête de la page. Cette balise peut souvent être vide
articles	contient tous les articles de la page en cours. Chaque article est contenu dans une balise *article*
article	contient le contenu d'un article
*title	le titre de l'article
text	le contenu textuel de l'article
*para	un paragraphe dans le texte
sent	phrase dans un paragraphe
ent	mention d'une entité nommée
rel	relation dans laquelle une entité est impliquée. Cette balise est imbriquée dans une balise `ent`, et n'a pas de contenu textuel

Table 2.12: Ensemble de balises contenues dans la balise `content` du format XML EMONTAL. Les balises qui peuvent être trouvées dans le jeu de balises XML DocBook original sont marquées par le symbole *

par exemple *educatedAt* ou *spouse*. L'attribut *score* est le score d'association prédit (expliqué dans le chapitre 3) pour cette relation par le méthode ELIJERE. L'attribut *anchor* est le prédicat du patron lexico-syntaxique qui a extrait cette relation, par exemple "naître_VERB". L'attribut *id* des balises `rel` relie les balises `ent` impliquées dans une relation. Par exemple, comme le montre la Figure 2.9, la balise `ent` avec l'attribut *id* "pers_191" décrit une entité de type *Personne*, tandis que la balise `ent` avec l'attribut "*loc_193*" *id* décrit une entité de type *Lieu*. Ces deux balises partagent une balise `rel`, qui possède l'attribut *1_dateOfBirth*. La Figure 2.9 montre un exemple de balises `ent` et `rel` dans un document XML EMONTAL de notre corpus.

```
<sent id="sent_109">
 <ent id="pers_191" type="pers">
  <rel id="1_dateOfBirth" type="dateOfBirth" score="0.6991231441497803" rule="semantic" anchor="naître_VERB"/>
  <rel id="2_placeOfBirth" type="placeOfBirth" score="0.5120583176612854" rule="semantic" anchor="naître_VERB"/>
  Charles Geneviève Louis Auguste Thimolée d'Eon de Beaumont
  </ent>
  naquit à
  <ent id="loc_193" type="loc">
   <rel id="2_placeOfBirth" type="placeOfBirth" score="0.5120583176612854" rule="semantic" anchor="naître_VERB"/>
   Tonnerre
  </ent>
   en
  <ent id="time_194" type="time">
   <rel id="1_dateOfBirth" type="dateOfBirth" score="0.6991231441497803" rule="semantic" anchor="naître_VERB"/>
   1728
  </ent>
  ; il était fils du premier magistrat municipal de cette ville c'était, on le voit un très authentique
  <ent id="loc_195" type="loc">
  Bourguignon
  </ent>
</sent>
```

Figure 2.9: Exemples de balises `ent` et `rel` d'un document XML EMONTAL de notre corpus (Bulletin de la Société d'archéologie et d'histoire de Tonnerre, Juin 1939)

Chapitre 3

Annotation sémantique de documents historiques : présentation de l'approche ELIJERE

Table des matières

La tâche d'annotation sémantique consiste à annoter le contenu textuel des documents avec des catégories sémantiques, telles que les sentiments (ex: positif, négatif), les types d'entités (ex: personne, lieu, temps) ou les relations (ex: lieu de naissance, d'éducation), pour n'en citer que quelques-unes. Les annotations sémantiques permettent de structurer le contenu textuel des documents et peuvent être exploitées par les interfaces de recherche pour faciliter l'exploration d'une collection de documents. Elles peuvent également être exploitées pour construire des interfaces de recherche augmentées, qui permettent une lecture distante de la collection. Nous pouvons diviser les approches pour la tâche d'annotation sémantique en deux catégories principales : les *approches basées sur des règles* et les *approches basées sur l'apprentissage automatique* (*machine-learning*), parmi lesquelles nous considérons les approches basées sur l'apprentissage profond (*deep-learning*).

Les approches basées sur des règles s'appuient sur des ensembles de règles qui permettent généralement d'obtenir des scores de Précision élevés. Cependant, ces approches généralisent mal, et sont particulièrement sensibles aux données bruyantes, telles que les erreurs produites par

l'OCR. Ainsi, elles obtiennent généralement des scores de Rappel plus faibles. Les approches d'apprentissage automatique quant à elle apprennent des règles à partir des données. Elles sont particulièrement efficaces pour généraliser à partir des données, et sont moins sensibles aux entrées bruyantes que les approches par règles.

Les deux types d'approches nécessitent des ressources. Les approches basées sur des règles nécessitent des ensembles de règles qui énoncent les conditions qu'un contenu textuel doit remplir pour être annoté avec une catégorie sémantique donnée. L'ajout et la mise à jour de l'ensemble de règles permettent à ces approches de détecter et de traiter de nouvelles catégories sémantiques. Cependant, plus un ensemble de règles est grand, plus il est difficile de le maintenir et de le mettre à jour. De plus, l'élaboration de règles est une tâche longue qui nécessite une expertise dans le domaine d'application. Les modèles utilisés dans les approches d'apprentissage automatique quant à elles doivent être entraînés sur de grands jeux de données annotées. La constitution de tels jeux de données reste une tâche longue et coûteuse. De plus, le modèle doit être entraîné à nouveau sur une quantité suffisante de données annotées pour apprendre à identifier de nouvelles catégories sémantiques. Ainsi, le maintien de la qualité et de la cohérence interne des annotations dans les ces jeux de données peut également constituer un défi.

Les approches récentes d'apprentissage profond, en particulier celles basées sur l'architecture *Transformer*, permettent de limiter la quantité de données nécessaires en affinant des modèles pré-entraînés sur des ensembles de données plus petits dans le domaine (*transfer-learning*). Toutefois, ces approches nécessitent des ressources informatiques importantes et coûteuses telles que des cartes graphiques (GPU) ou des *Tensor Processing Unit* (TPU) pour l'apprentissage et l'inférence. De plus, les approches basées sur des de réseaux de neurones sont souvent considérées comme des "boîtes noires" (Barredo Arrieta et al., 2020) en raison de la difficulté d'interpréter leurs processus. A l'inverse, les approches à base de règles sont interprétables, puisqu'elles reposent sur des règles explicitement énoncées, tandis que les approches traditionnelles d'apprentissage automatique peuvent être plus faciles à interpréter, en fonction de la complexité de l'architecture sous-jacente. De plus, les approches à base de règles et l'apprentissage automatique nécessitent moins de ressources informatiques, et peuvent fonctionner sur des processeurs (CPU).

Afin d'extraire les mentions de personnes, de lieux et des relations qui les lient dans notre corpus, nous proposons l'approche ELIJERE (*Extensible, Lightweight and Interpretable Joint Extraction of Relations and Entities*). Notre approche repose sur un ensemble de deux ressources linguistiques :

- un *Index Syntaxique*, qui décrit la relation qu'un patron syntaxique exprime, ainsi que le type d'entités impliquées dans la relation

- un *Index Lexical*, qui décrit comment une relation est exprimée lexicalement

Nous exploitons ces deux ressources pour extraire et catégoriser à partir de phrases les mentions d'entités impliquées dans des relations. Les ressources linguistiques sont construites à partir de patrons lexico-syntaxiques collectés depuis un ensemble de phrases, qui sont faiblement annotées à l'aide de la méthode de supervision distante (Mintz et al., 2009). Cette méthode suppose que, si deux entités participent à une relation dans une base de connaissances, alors au moins une phrase qui mentionne les deux entités dans un document exprime cette relation (Riedel et al., 2010). En appliquant cette méthode, nous pouvons rapidement construire un jeu de données annotées de phrases exprimant n'importe quel concept stocké dans une base de connaissances, à partir desquelles nous pouvons construire nos ressources linguistiques. Notre approche est ainsi extensible et interprétable, puisqu'elle s'appuie sur des ressources linguistiques explicites et extensibles. Elle est également légère, puisqu'elle nécessite peu de ressources informatiques et peut fonctionner sur des processeurs. Les principales étapes de notre approche sont présentées dans la Figure 3.1.

Le reste du chapitre est organisé comme suit : dans la section 3.1, nous présentons la méthode DARES (*Distant Annotation of Relations and Entities in Sentences*), pour construire un ensemble de phrases annotées avec les mentions d'entités et de relations. Dans la section 3.2, nous décrivons notre méthode pour construire les ressources linguistiques sur lesquelles l'approche ELIJERE s'appuie. Enfin, dans la section 3.3, nous décrivons notre approche de la tâche d'extraction conjointe de Relations et d'Entités (*Joint Extraction of Relations and Entities*, JERE).

3.1 Annotation distante des relations et des entités dans les phrases (DARES)

La méthode DARES (*Distant Annotated Relations and Entities in Sentences*) permet de construire un ensemble de phrases annotées avec les mentions d'entités et de relations. Cette méthode s'appuie sur la méthode de supervision distante (Riedel et al., 2010) qui permet de construire rapidement un ensemble de phrases annotées exprimant n'importe quel concept stocké dans une base de connaissances. Les ressources linguistiques sur lesquelles s'appuie l'approche ELIJERE sont construites à partir d'un ensemble de phrases, où les mentions de relations et d'entités sont annotées. Pour constituer cet ensemble de phrases annotées, nous suivons les étapes décrites dans la Figure 3.2 :

L'étape de *collection de déclarations* permet d'extraire des déclarations, c'est-à-dire des triples, contenues dans une base de connaissances telle que Wikidata ou Freebase. Un triple décrit dans un format lisible par une machine une propriété que possède une entité . Il se présente sous la forme `(sujet ; prédicat ; objet)`, où `sujet` est l'entité sur laquelle porte la déclaration, `prédicat` est la relation ou la propriété décrivant le sujet, et `objet` est la valeur de la

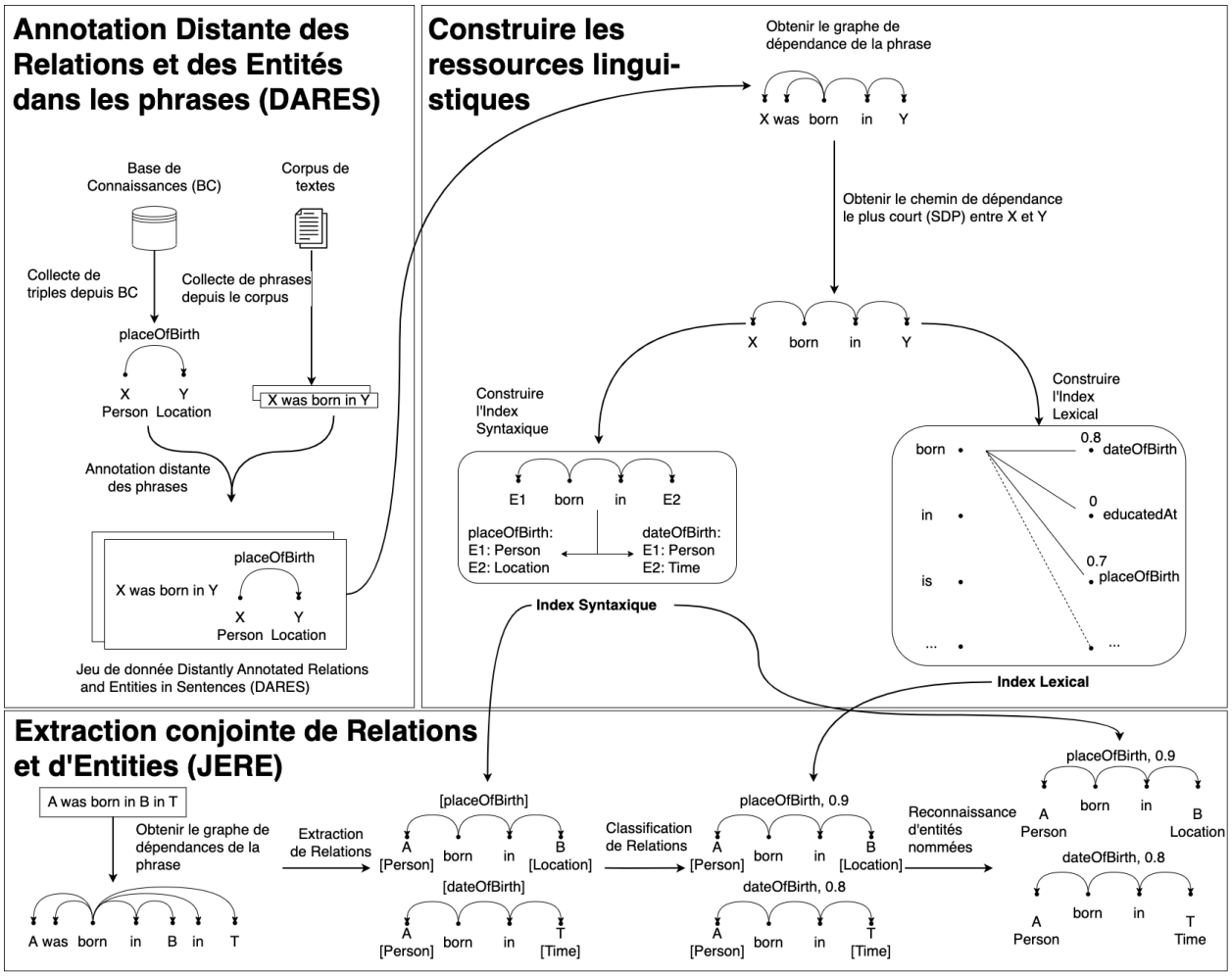


Figure 3.1: Etapes principales de l'approche ELIJERE

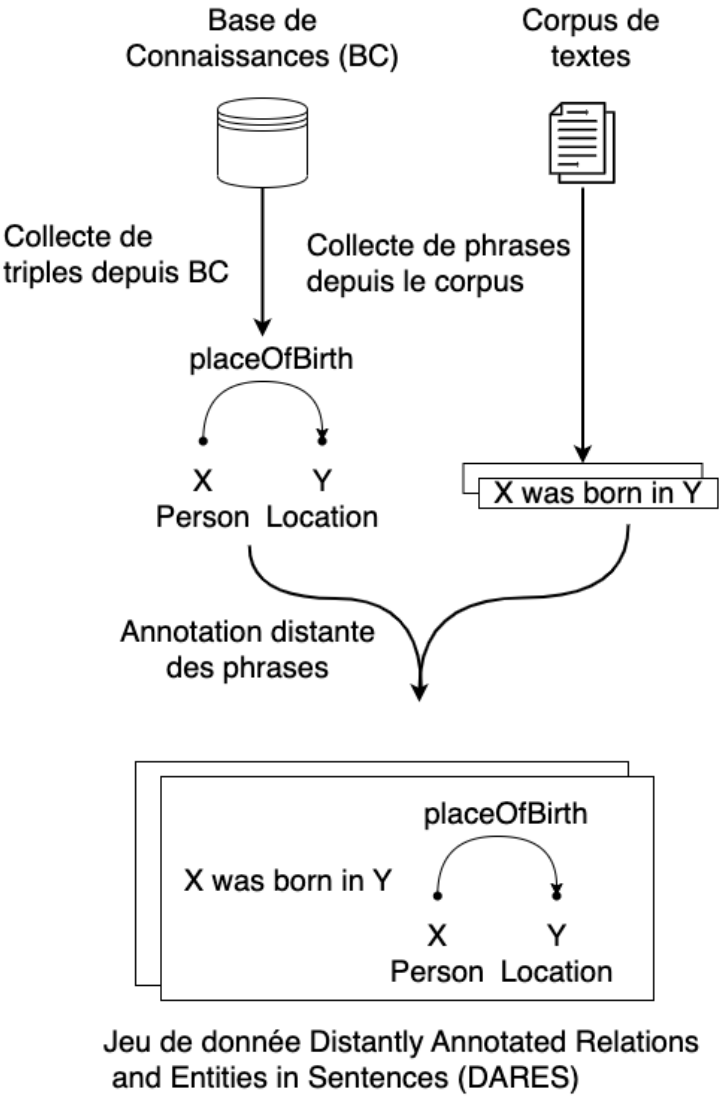


Figure 3.2: Etapes principales de la méthode DARES

relation ou de la propriété. Par exemple, étant donné le sujet "Douglas Adams", l'objet "1952" et le prédicat "date de naissance", le triple `(Douglas Adams ; date de naissance ; 1952)` indique que "Douglas Adams" est né en 1952. Ci-dessous d'autres exemples de triples pouvant être extraits de la base de connaissances Wikidata :

1. `(Marie Curie; métier; physicienne)`
2. `(George Washington; lieu de mort; Mont Vernon)`
3. `(France; partage sa frontière avec; Belgique)`

L'étape de *collection de phrases* collecte toutes les phrases d'un corpus de documents. Dans notre travail, nous collectons toutes les phrases d'un ensemble d'articles issus de Wikipedia. Cependant, dans une approche plus générale, les phrases peuvent potentiellement être extraites de documents d'origines diverses tels que le World Wide Web. Étant donné un corpus d'articles provenant de Wikipedia, nous extrayons, par exemple, les phrases suivantes :

1. Marie Curie était une physicienne.
2. George Washington est mort au Mount Vernon.
3. La France partage sa frontière avec la Belgique au nord.

Enfin, l'étape d'*annotation distante* applique la méthode de supervision distante pour associer les phrases aux déclarations. Une phrase est associée à une déclaration si elle contient des mentions des deux entités impliquées dans la déclaration, c'est-à-dire des mentions du `sujet` et de l' `objet` du triple. Par exemple, en appliquant la méthode de supervision distante à l'ensemble des exemples mentionnés ci-dessus, nous associerions les exemples précédents de phrases et d'énoncés comme suit :

1. Marie Curie était une physicienne. : `(Marie Curie; métier; physicienne)`
2. George Washington est mort à Mount Vernon. : `(George Washington; lieu de mort; Mont Vernon)`
3. La France partage sa frontière avec la Belgique au nord. : `(France; partage sa frontière avec; Belgique)`

3.2 Construire les ressources linguistiques

Pour extraire et catégoriser les relations et les entités mentionnées dans une phrase, notre approche s'appuie sur deux ressources linguistiques principales, appelées Index Syntaxique et Index Lexical, L'Index Syntaxique décrit les relations qu'un modèle exprime, ainsi que le type d'entités impliquées dans la relation, tandis que L'Index Lexical décrit comment une relation est exprimée lexicalement. Les deux indexes sont construits à partir de patrons lexico-syntaxiques exprimant une relation entre des entités collectées depuis le jeu de données DARES. Pour collecter les patrons lexico-syntaxiques, nous nous appuyons sur l'hypothèse de R. Bunescu and Mooney (2005) selon laquelle la relation entre deux entités se trouve dans le chemin de dépendance le plus court entre elles (*Shortest Dependency Path*). Les principales étapes de la construction de ces ressources à partir de le jeu de données DARES sont présentées dans la Figure 3.3.

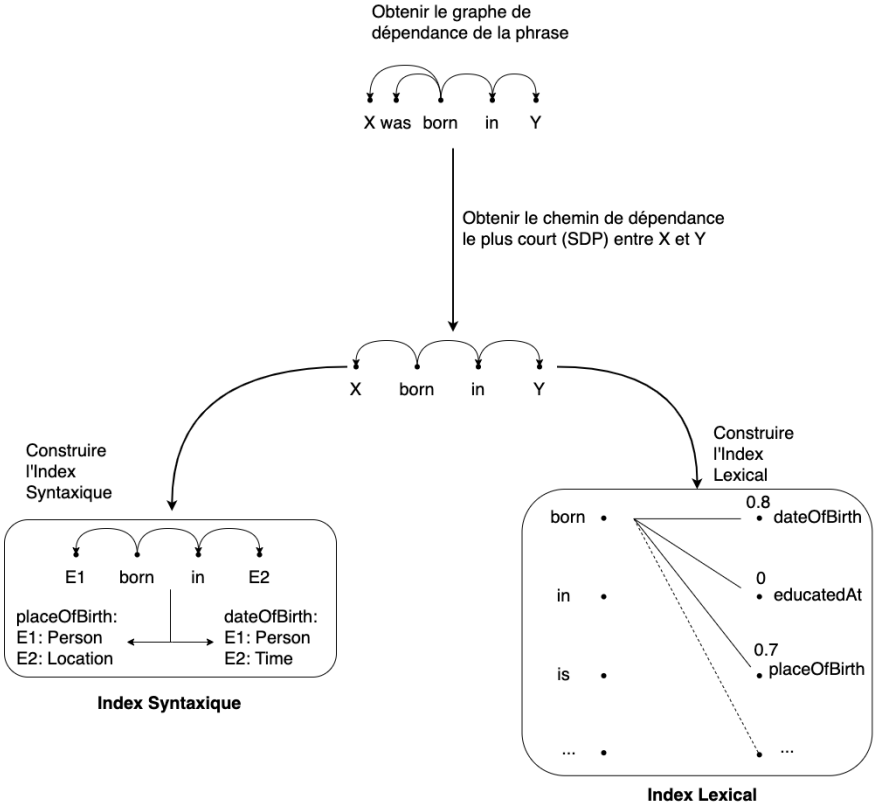


Figure 3.3: Principales étapes pour la constitution des ressources linguistiques sur lesquelles l'approche ELIJERE repose

Nous obtenons les structures syntaxiques des phrases sous la forme de graphes de dépendance, comme illustré dans la Figure 3.4. Nous choisissons de travailler au niveau syntaxique puisqu'il est plus facile de collecter des modèles exprimant une relation entre des entités à ce niveau qu'au niveau de la surface de la phrase. En effet, les patrons exprimant une relation dans une phrase peuvent être

contigus ou non contigus. Les patrons contigus impliquent une séquence de mots qui se suivent, tandis que les patrons non contigus impliquent des séquences de mots qui peuvent être séparés par d'autres éléments de la phrase. Par exemple, la phrase "Douglas Adams est né à Cambridge en 1952" exprime les relations *lieu de naissance* et *date de naissance*. Le patron exprimant la relation *lieu de naissance* implique la séquence contiguë de mots "Douglas Adams est né à Cambridge", tandis que le patron exprimant la relation *date de naissance* implique la séquence non contiguë de mots "Douglas Adams est né [...] en 1952". L'extraction de ces patrons non contigus nécessite de travailler au niveau de la structure syntaxique de la phrase, et ne peut se faire en ne considérant que la représentation de surface de la phrase.

Pour collecter les patrons lexico-syntaxiques qui expriment une relation entre deux entités, nous extrayons les graphes du plus court chemin de dépendance (*Shortest Dependency Path*, SDP) entre les entités. La Figure 3.5 présente deux exemples de graphes SDP extraits de cette phrase.

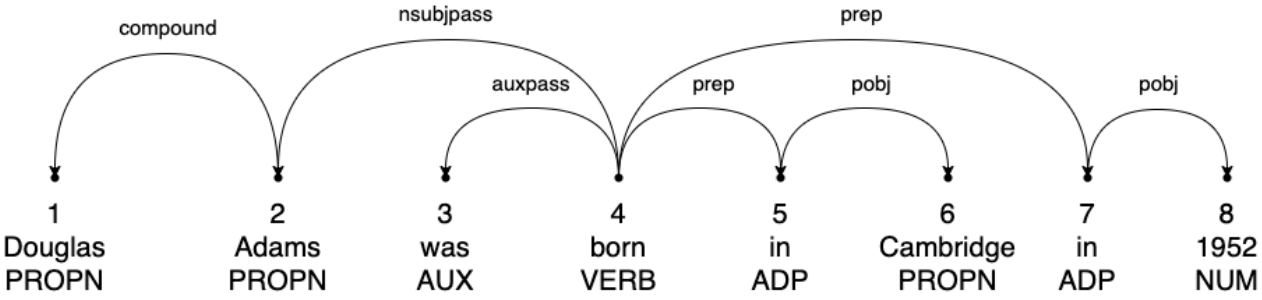


Figure 3.4: Graphe de dépendance de la phrase "*Douglas Adams was born in Cambridge in 1952*" ("Douglas Adams est né à Cambridge en 1952")

L'Index Syntaxique consiste en un ensemble de classes de patrons lexico-syntaxiques exprimant une ou plusieurs relations entre deux entités. Chaque patron indique également les types d'entités impliquées dans chaque relation qu'il peut exprimer. Les patrons lexico-syntaxiques collectés peuvent différer les uns des autres par leur vocabulaire et le type d'entités impliquées dans la relation exprimée. Cependant, ils peuvent partager la même structure syntaxique. Ainsi, pour construire l'Index Syntaxique, nous identifions des classes de patrons basées sur les conditions suivantes : les patrons lexico-syntaxiques appartiennent à la même classe s'ils partagent la même structure syntaxique, les mêmes partie de discours (POS) et le même prédicat. Le prédicat est la tête du patron, c'est-à-dire le noeud qui régit tous les autres noeuds. Le prédicat est généralement un verbe, mais il peut appartenir à d'autres catégories grammaticales, comme le nom. Par exemple, les deux patrons lexico-syntaxiques de la Figure 3.5 partagent la même structure syntaxique et le verbe "né" comme prédicat. Cependant, le second patron diffère du premier par ses parties du discours. Les deux modèles appartiennent donc à des classes différentes.

L'Index Lexical consiste en une liste de vocabulaire où à chaque mot sont associées les relations possibles dans lesquelles ce mot peut apparaître. Le vocabulaire est extrait des mêmes patrons

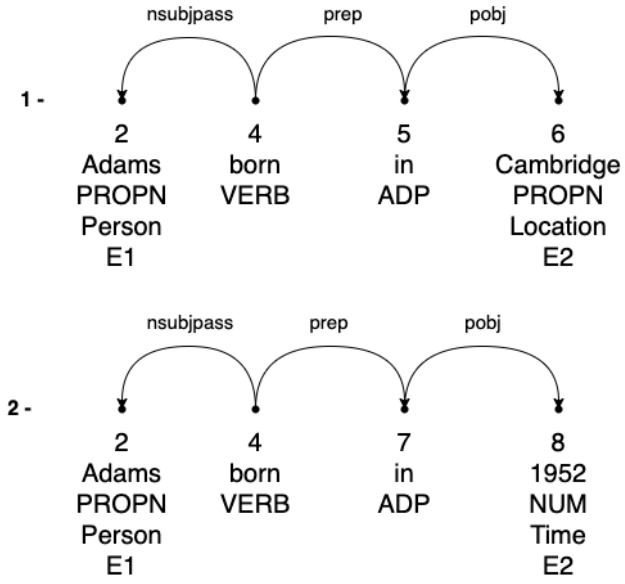


Figure 3.5: Graphes du plus court chemin de dépendance (SDP) entre les entités ("Adams", "Cambridge") et ("Adams", "1952"), extraits du graphe de dépendance illustré dans la Figure 3.4

lexico-syntaxiques sur lesquels l'Index Syntaxique est construit. L'Index Lexical est construit en mesurant un score d'association entre les vocabulaires des patrons lexico-syntaxiques et les relations. Ces scores d'association sont d'abord basés sur les fréquences des mots apparaissant dans les schémas lexico-syntaxiques, avant d'être normalisés en poids TF-IDF. Par exemple, le mot "né" a un score d'association élevé avec les relations *lieu de naissance* et *date de naissance*, puisqu'il apparaît essentiellement dans les patrons qui expriment ces relations. En revanche, il a des scores d'association plus faibles ou nuls avec d'autres relations dans lesquelles il est rarement ou jamais utilisé, telles que *éduqué à* ou *lieu de mort*. Nous pouvons fixer un seuil afin d'éliminer de l'index les mots dont le poids d'association est trop faible. L'Index Lexical s'inspire de l'index inversé pondéré (*weighted inverted index*) de la méthode d'analyse sémantique explicite (*Explicit Semantic Analysis*, ESA) proposée par Gabrilovich and Markovitch (2007). L'index inversé pondéré représente un concept de Wikipédia, c'est-à-dire un article Wikipédia, par un vecteur de mots apparaissant dans le contenu de l'article. Chaque mot se voit attribuer un poids TF-IDF, de manière à indiquer la force de l'association entre le mot et les concepts. Contrairement à l'index inversé pondéré de la méthode ESA, notre approche utilise les relations entre les entités comme concepts, au lieu des articles de Wikipédia.

3.3 Extraction conjointe des Relations et des Entités

Notre approche de la tâche d'extraction conjointe des Relations et des Entités exploite l'Index Syntaxique et l'Index Lexical pour extraire et catégoriser les entités et les relations mentionnées

dans une phrase. Notre approche consiste en trois étapes principales illustrées dans la Figure 3.6 :

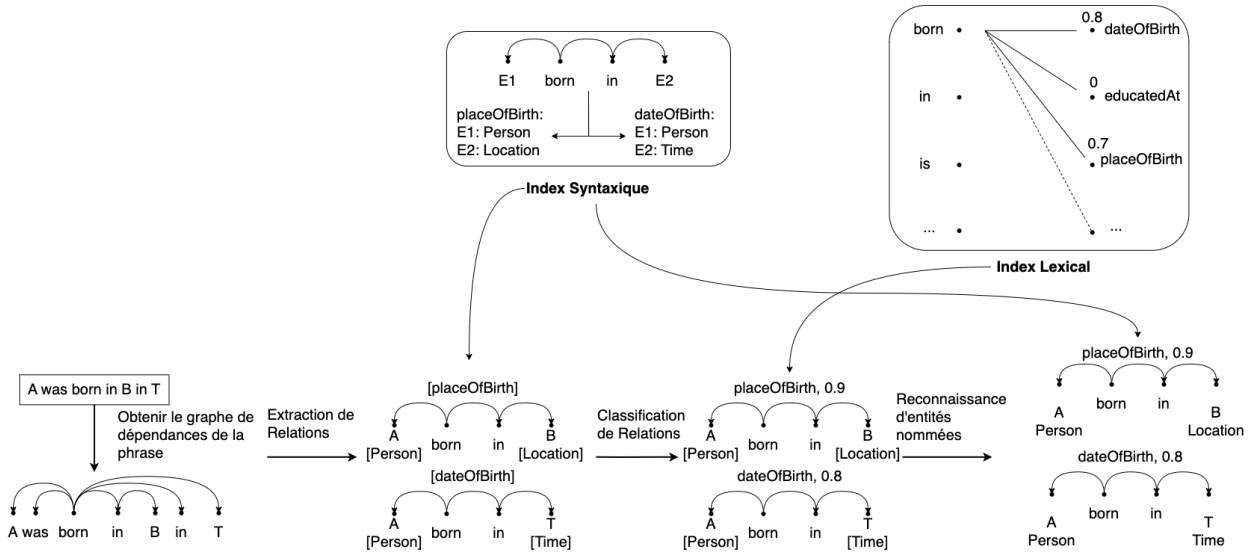


Figure 3.6: Principales étapes de notre approche de la tâche d'extraction conjointe des Relations et des Entités

L'étape d'*extraction de relations* extrait les mentions de relations dans la phrase. Nous appliquons les patrons contenus dans l'Index Syntaxique au graphe de dépendance de la phrase pour y trouver des sous-graphes candidats qui correspondent au patron. Chaque sous-graphe candidat est associé à un ensemble de relations possibles, qui sont les relations que le patron peut exprimer selon l'Index Syntaxique. Par exemple, à partir du graphe de dépendance de la phrase *The Republic of Guinea borders the Atlantic ocean to the west and Senegal to the north.* ("La République de Guinée est bordée à l'ouest par l'océan Atlantique et au nord par le Sénégal") illustré dans la Figure 3.7, nous trouvons les sous-graphes candidats illustrés dans la Figure 3.8 en appliquant les patrons contenus dans l'Index Syntaxique.

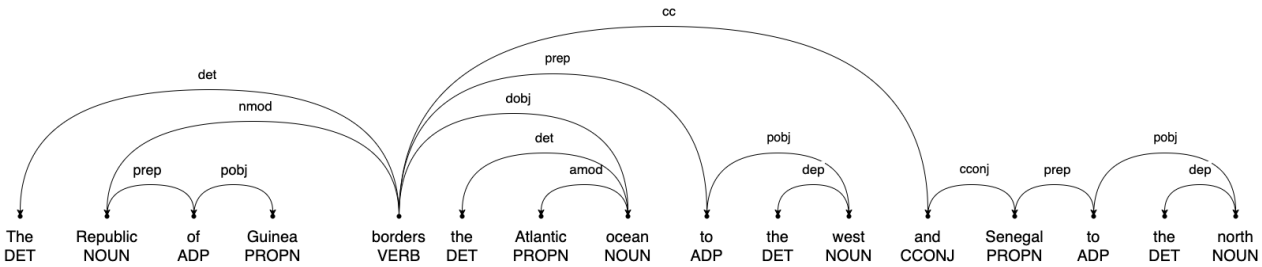


Figure 3.7: Graphe de dépendance de la phrase "*The Republic of Guinea borders the Atlantic ocean to the west and Senegal to the north.*" ("La République de Guinée est bordée à l'ouest par l'océan Atlantique et au nord par le Sénégal")

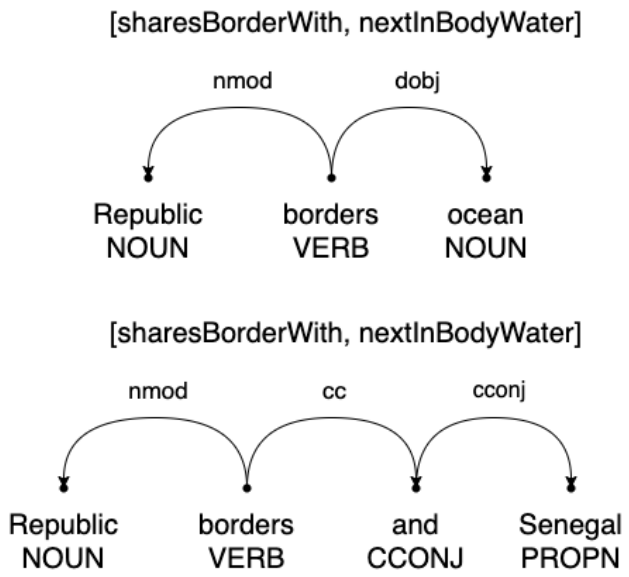


Figure 3.8: Sous-graphes candidats extraits lors de l'étape d'*extraction de relations*. Chaque sous-graphe est associé à un ensemble de relations possibles

L'étape de *catégorisation des relations* permet de classer les sous-graphes candidats extraits lors de l'étape d'*extraction des relations*. Un sous-graphe candidat est catégorisé en fonction de son vocabulaire et des scores d'associations contenus dans l'Index Lexical. Parmi les relations possibles exprimées par un sous-graphe candidat, nous sélectionnons la relation avec laquelle ce candidat a le score d'association le plus élevé. Ce score est calculé comme la moyenne harmonique des scores d'association du vocabulaire du sous-graphe. Nous pouvons fixer un seuil, appelé seuil sémantique, pour nous assurer que le score d'association moyen du sous-graphe candidat est suffisamment élevé. Par exemple, comme le montre la Figure 3.9, compte tenu de leur vocabulaire et de leurs étiquettes possibles, les sous-graphes candidats illustrés dans la Figure 3.8 sont respectivement classés comme *proche d'un cours d'eau* (*nextInBodyWater*) et *partage ses frontières avec* (*sharesBorderWith*).

Enfin, l'étape de *reconnaissance des entités nommées* détermine les types d'entités impliquées dans la relation. Les types d'entités sont déterminés par le patron lexico-syntaxique ayant extrait le sous-graphe candidat dans l'étape d'*extraction de relation*, et la relation prédite dans l'étape de *catégorisation des relations*. Par exemple, comme le montre la Figure 3.10, chaque entité impliquée dans les relations exprimées par les deux graphes candidats est de type *Location*, en raison du patron qui a extrait ces candidats, et des étiquettes qui ont été prédites.

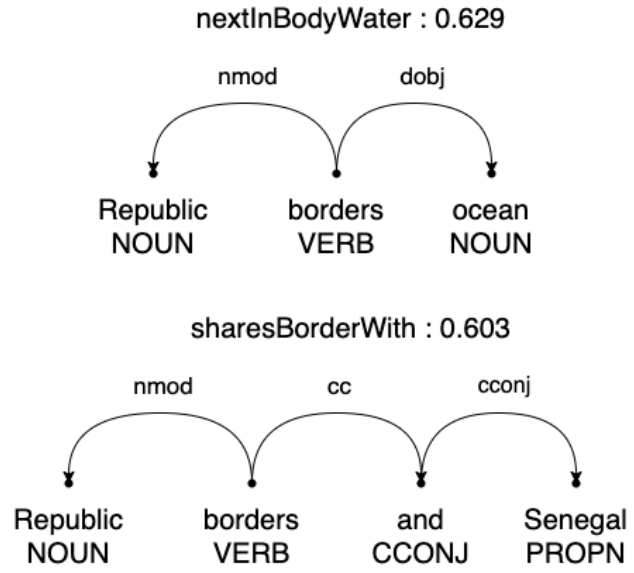


Figure 3.9: Sous-graphes candidats classés lors de l'étape de *classification des relations*

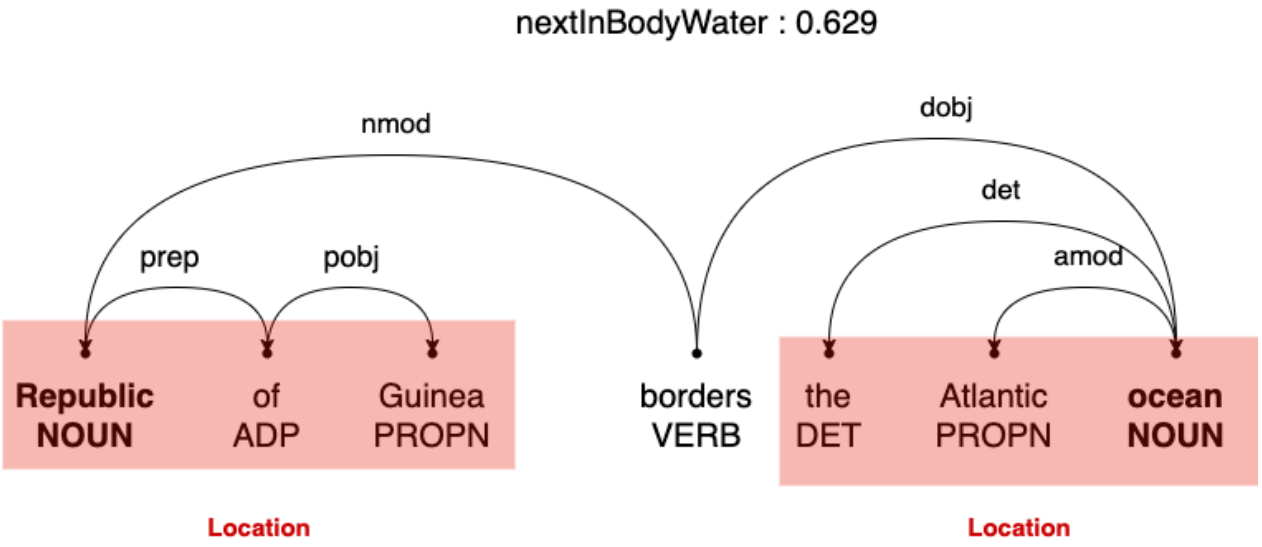


Figure 3.10: Entités impliquées dans les relations catégorisées lors de l'étape *reconnaissance des entités nommées*

Chapitre 4

Évaluation et discussion de l'approche ELIJERE

Table des matières

Dans ce chapitre, nous proposons une évaluation ainsi qu'une discussion de l'approche ELIJERE. Nous proposons une implémentation de notre approche nommée *modèle ELIJERE de base*, qui s'appuie sur l'Index Syntaxique pour extraire à partir d'une phrase les sous-graphes candidats exprimant une relation, et sur l'Index Lexical pour catégoriser ces candidats en fonction de leur vocabulaire. Afin de comparer notre implémentations à d'autres approches, nous proposons une seconde implémentation, nommée *modèle ELIJERE hybride*. Celle-ci remplace l'Index Lexical par un classificateur reposant sur des méthodes d'apprentissages automatiques pour catégoriser les candidats sur la base de leur vocabulaire. Nous évaluons d'abord ces deux implémentations sur le jeu de données DARES, avant de les évaluer sur le corpus EMONTAL, afin d'étudier l'impact des erreurs produites par le processus d'OCR, ainsi que par les différences de styles d'écriture sur notre approche.

Le reste de ce chapitre est structuré comme suit : nous présentons d'abord les jeux de données ainsi que le protocole d'évaluation des deux implémentations dans la section 4.1, avant de présenter une évaluation comparative des deux implémentations dans la section 4.2

4.1 Jeux de données et protocole d'évaluation

La distribution des relations et des entités dans la partie test du jeu de données DARES est présentée respectivement dans la Table 4.1 et la Table 4.2. Environ 18 % des relations de ce jeu de test sont

étiquetées comme *Autre* (*Other*). Nous ne tenons pas compte de cette étiquette pour l'évaluation de l'approche ELIJERE.

Etiquette	Nombre	Proportion
capitalOf	140	11.91 %
country	183	15.57 %
dateOfBirth	71	6.042 %
dateOfDeath	39	3.319 %
educatedAt	13	1.106 %
headOfGoverment	15	1.276 %
inception	12	1.021 %
memberOf	53	4.510 %
nextInBodyWater	47	4.000 %
occupation	125	10.638 %
Other	213	18.127 %
placeOfBirth	67	5.702 %
sharesBordersWith	174	14.808 %
spouse	23	1.957 %
Total	1,175	100 %

Table 4.1: Distribution des étiquettes de relation dans la partie test du jeu de données DARES

Etiquette	Nombre	Proportion
Person	391	19.628 %
Location	1,337	67.118 %
Time	128	6.425 %
Misc	136	6.827 %
Total	1,992	100 %

Table 4.2: Distribution des étiquettes d'entités dans la partie test du jeu de données DARES

De plus, afin d'évaluer les performances de l'approche ELIJERE sur des documents historiques, nous avons construit à partir du corpus EMONTAL un ensemble de phrases annotées contenant les mentions de relations et d'entités. Ce jeu de données suit la même structure que le jeu de données DARES. Nous avons produit cet ensemble de données de manière semi-automatique : tout d'abord, nous avons appliqué la méthode ELIJERE pour collecter des phrases du corpus EMONTAL et annoter automatiquement les mentions d'entités et de relations qu'elles contiennent. Nous avons ensuite revu et corrigé manuellement les annotations de ces phrases, et ajouté les relations et entités que le modèle avait pu manquer. De plus, nous avons ajouté manuellement des éléments négatifs dans le jeu de données en ajoutant des graphes qui n'expriment aucune relation. Nous avons classé ces échantillons négatifs avec l'étiquette *Autre* (*Other*). Enfin, nous avons sélectionné au hasard 299

phrases dans cet ensemble de phrases corrigées. La Table 4.3 montre la distribution des étiquettes de relations, tandis que la Table 4.4 montre la distribution des étiquettes d'entités dans le jeu de données de test EMONTAL.

Etiquette	Nombre	Proportion
capitalOf	3	0.711%
country	3	0.711%
dateOfBirth	40	9.503%
dateOfDeath	36	8.550%
educatedAt	15	3.563%
inception	18	4.276%
memberOf	17	4.035%
occupation	63	14.964%
Other	112	26.600%
placeOfBirth	55	13.067%
spouse	33	7.837%
sharesBordersWith	26	6.174 %
Total	421	100 %

Table 4.3: Distribution des étiquettes de relations dans le jeu de données de test EMONTAL

Etiquette	Nombre	Proportion
Person	551	60.152 %
Location	195	21.288 %
Time	102	11.135 %
Misc	68	7.423 %
Total	916	100 %

Table 4.4: Distribution des étiquettes d'entités dans le jeu de données de test EMONTAL

Nous évaluons les deux implémentations de l'approche ELIJERE selon le protocole d'évaluation suivant : Nous évaluons chaque modèles sur les tâches d'extraction de relations (RE) et de reconnaissance d'entités nommées (NER) de manière séparée. Nous évaluons la capacité des modèles à catégoriser la relation exprimée par un graphe candidat en termes de scores de Précision, de Rappel et de F1. De même, nous évaluons également la capacité du modèle à déterminer le type et les frontières des entités impliquées dans la relation en termes de de scores de Précision, de Rappel et de F1. Pour l'évaluation de la tâche NER, nous employons les cadres d'évaluation type, partiel, strict et exact, comme décrit dans le schéma d'évaluation SemEval (Segura-Bedmar, Martínez, & Herrero-Zazo, 2013). Pour chaque modèle, nous effectuons dix évaluations, en fixant à chaque fois un seuil sémantique entre 0 et 0,9. Notre objectif est d'étudier l'impact du seuil sémantique sur

les performances du modèle. Lors de la catégorisation de la relation du graphe candidat, le modèle produit l'étiquette *Autre* si l'une des conditions suivantes est remplie :

- Le modèle ne trouve pas d'entrée dans l'Index Syntaxique correspondant au prédicat du graphe candidat
- Le modèle ne trouve pas de patron dans l'Index Syntaxique correspondant au graphe candidat
- Le score sémantique prédit par le modèle est inférieur au seuil sémantique

4.2 Discussion

La Table 4.5 montre le score moyen obtenu par les modèles ELIJERE de *base* et *hybride* sur la tâche d'extraction de relations. Sur le jeu de données DARES, le modèle *hybride ELIJERE* obtient les meilleurs résultats, avec une Précision moyenne de 0,675, un Rappel moyen de 0,137 et un score F1 moyen de 0,218. De même, Le modèle *hybride ELIJERE* obtient également les meilleurs résultats sur le corpus EMONTAL, avec une Précision moyenne de 0,469, un Rappel moyen de 0,166 et un score F1 moyen de 0,218.

		P	R	F1
DARES	*modèle ELIJERE de base*	0.518	0.120	0.189
	modèle ELIJERE hybride	**0.675**	**0.137**	**0.218**
EMONTAL	*modèle ELIJERE de base*	0.352	0.154	0.199
	modèle ELIJERE hybride	**0.469**	**0.166**	**0.218**

Table 4.5: Scores moyens par seuil sémantique obtenus par les modèles ELIJERE de *base* et *hybride* sur le jeu de données DARES et le corpus EMONTAL

La Table 4.6 compare la distribution moyenne des types d'erreurs produites par les modèles *base* et *hybride ELIJERE* sur le jeu de données DARES et le corpus EMONTAL. Sur le jeu de données DARES, le modèle hybride produit 883 erreurs alors que le modèle de *base* en produit 900. Le modèle hybride produit moins d'erreurs de type *score sémantique trop faible*, ce qui suggère que le modèle produit des scores plus sûrs que le modèle de *base*. Sur le corpus EMONTAL, le modèle *hybride* produit 235 erreurs tandis que le modèle *base* en produit 243. Le modèle *hybride* produit moins d'erreurs de type *sémantique trop faible*, ce qui suggère que le modèle produit des scores plus sûrs que le modèle de *base*, et est peut être moins affecté par les erreurs d'OCR. D'après ces résultats, le modèle *hybride ELIJERE* semble être plus adapté à la tâche d'extraction de relations que le modèle de *base*.

	Seuil	Absence de Prédicat Compte	Absence de Prédicat Proportion	Absence de Patron Compte	Absence de Patron Proportion	Mauvaise étiquette Compte	Mauvaise étiquette Proportion	Score sémantique trop faible Compte	Score sémantique trop faible Proportion	Total
DARES	*modèle ELIJERE de base*	**509**	**56.572** %	**326**	**36.232** %	**17**	**2.000** %	47	5.193 %	900
	modèle ELIJERE hybride	**509**	57.630 %	**326**	36.910 %	26	2.965 %	**22**	**2.491** %	**883**
EMONTAL	*modèle ELIJERE de base*	128	**52.909** %	84	**34.722** %	6	**2.746** %	24	9.622 %	243
	modèle ELIJERE hybride	128	54.572 %	84	35.813 %	8	3.458 %	**15**	**6.154** %	**235**

Table 4.6: Types d'erreurs moyennes et distribution des échantillons mal étiquetés par le modèle ELIJERE de base et *hybride* sur le jeu de données DARES et le corpus EMONTAL

La Table 4.7 compare les scores moyens par seuils sémantiques pour chaque cadre d'évaluation obtenus par les deux modèles sur la tâche de reconnaissance d'entités nommées. Sur le jeu de données DARES, le modèle *hybride ELIJERE* obtient les scores les plus élevés, bien que les deux modèles obtiennent des scores similaires dans tous les contextes d'évaluation. Le modèle *hybride* obtient des scores de Précision de 0,980 dans la configuration de type, 0,846 dans la configuration partielle, 0,681 dans la configuration stricte et 0,693 dans la configuration exacte. Le modèle ELIJERE de *base* obtient un score de Précision moyen de 0,982 dans le cadre type, ce qui est légèrement supérieur au score obtenu par le modèle *hybride*. Le modèle *hybrid* obtient des scores de Rappel inférieurs, avec un score de 0,129 dans le cadre type, un score de 0,111 dans le cadre partiel, un score de 0,090 dans le cadre strict et un score de 0,091 dans le cadre exact. Par conséquent, il obtient des scores F1 plus faibles, avec 0,226 dans le cadre type, 0,195 dans le cadre partiel, 0,157 dans le cadre strict et 0,160 dans le cadre exact.

Sur le corpus EMONTAL, le modèle *hybride* obtient également les scores les plus élevés, bien que les deux modèles obtiennent des scores similaires dans tous les contextes d'évaluation. Le modèle obtient des scores de Précision de 0,827 dans le cadre type, 0,818 dans le cadre partiel, 0,506 dans le cadre stricte et 0,637 dans le cadre exact. Le modèle de *base* obtient un score de Précision moyen de 0,856 dans le cadre type, ce qui est légèrement supérieur au score obtenu par le modèle *hybride*. Cependant, le modèle *hybrid* obtient des scores de Rappel inférieurs, avec un score de 0,160 dans les cadre type et partiel, un score de 0,097 dans le cadre strict et 0,124 dans le cadre exact. Par conséquent, il obtient des scores F1 inférieurs, avec 0,265 dans le cadre type, 0,264 dans le cadre partiel, 0,161 dans le cadre strict et 0,205 dans le cadre exact. D'après ces résultats, le modèle *hybrid* semble être légèrement plus adapté à la tâche de reconnaissance d'entités nommées sur le corpus EMONTAL que le modèle de base.

Les évaluation de l'approche ELIJERE confirment que notre approche de la tâche d'extraction conjointe de Relations et d'Entités est extensible, légère et interprétable. En annotant le jeu de données DARES de manière distante, notre méthode est capable de collecter rapidement des patrons

	Seuil	Cadre	P	R	F1
DARES	*modèle ELIJERE de base*	type	**0.982**	0.111	0.199
		partiel	**0.829**	0.095	0.169
		strict	**0.649**	0.074	0.133
		exact	**0.659**	0.076	0.135
	modèle ELIJERE hybride	type	**0.980**	0.129	0.226
		partiel	**0.846**	0.111	0.195
		strict	**0.681**	0.090	0.157
		exact	**0.693**	0.091	0.160
EMONTAL	*modèle ELIJERE de base*	type	**0.856**	0.133	0.225
		partiel	**0.800**	0.132	0.220
		strict	**0.488**	0.076	0.129
		exact	**0.600**	0.099	0.167
	modèle ELIJERE hybride	type	**0.827**	0.160	0.265
		partiel	**0.818**	0.160	0.264
		strict	**0.506**	0.097	0.161
		exact	**0.637**	0.124	0.205

Table 4.7: Scores de Précision, Rappel et F1 moyens obtenus par les modèles ELIJERE de *base* et *hybride* pour la tâche de reconnaissance d'entités nommées sur le jeu de données DARES et le corpus EMONTAL

pour l'extraction des entités impliquées dans une relation. Cette méthode permet à notre approche de collecter des patrons d'extraction pour toute relation et toute entité dans n'importe quelle langue disponible sur Wikidata et Wikipedia. De plus, le résultat et le processus de notre approche sont interprétables, puisqu'ils s'appuient sur des patrons d'extraction syntaxiques explicites.

Les faibles scores de Rappel et F1 sont la principale faiblesse de l'approche ELIJERE. Ils sont dus à la faible variation des patrons d'extraction dans le jeu de données sur lequel les Indexes Syntaxiques et Lexicaux sont construits. Nous envisageons plusieurs moyens pour accroître la diversité des patrons contenus dans les Indexes : nous pourrions construire les Indexes sur un jeu de données composé de documents d'origines diverses, et non pas d'une origine unique. Nous pourrions également demander à un Grand Modèle de Langage tel que ChatGPT de reformuler les phrases connues ou même d'en suggérer de nouvelles, afin de collecter de nouveaux patrons.

Les scores élevés de Précision obtenus par les deux modèles sur la tâche de reconnaissance d'entités nommées dans le cadre *type* montrent la capacité du système à détecter correctement les types d'entités. Cependant, les faibles scores obtenus dans les cadres strict, exact et partiel indiquent que les modèles peinent à identifier correctement les frontières des entités. Actuellement, les frontières d'une entité sont détectées à l'aide de règles conçues manuellement, et qui s'appuient sur les parties du discours et les rôles de dépendance des mots. Afin d'améliorer la détection des frontières, nous pourrions, par exemple, apprendre les frontières des entités impliquées dans une relation à partir des graphes SDP pendant le processus de construction de l'Index Syntaxique.

La catégorisation incorrecte d'un graphe candidat peut provenir d'annotations incorrectes dans

le jeu de données à partir duquel les ressources linguistiques sont construites. Ces annotations incorrectes sont causées par la méthode de supervision distante. Nous devons donc ajouter une étape de post-traitement pour corriger ces annotations afin d'améliorer les performances de notre approche. Cette étape de post-traitement peut être effectuée manuellement ou automatiquement, par exemple en s'appuyant sur des méthodes de regroupement (*clustering*).

Les scores obtenus par le modèle ELIJERE de *base* sont en moyenne inférieurs à ceux obtenus par le modèle *hybride*. Cependant, la sortie du modèle *base* pourrait être plus facile à interpréter, puisque chaque mot est associé à des relations explicites. Ainsi, nous pouvons déterminer comment chaque mot a contribué à la catégorisation d'un graphe candidat. D'autre part, la classification des relations à l'aide de modèles d'apprentissage automatique tels que XGBoost peut être plus difficile à interpréter, en fonction de la complexité du modèle.

L'évaluation de l'approche ELIJERE sur le corpus EMONTAL montre que notre approche est affectée par les artefacts et les transcriptions incorrectes produits par le processus d'OCR. Une meilleure correction de ces erreurs est donc nécessaire pour améliorer les performances de notre approche sur les documents historiques. Notre approche pourrait également s'appuyer sur des méthodes de recherche approximatives (*fuzzy matching*) afin de devenir flexible et de s'adapter à l'entrée bruyante des documents historiques. En rendant notre approche plus flexible, nous augmenterions ses scores de Rappel, au prix d'une baisse de ses scores de Précision. Notre approche est également affectée par les styles d'écriture des documents du corpus EMONTAL, qui sont différents des articles de Wikipedia à partir desquels les ressources linguistiques de l'approche ELIJERE sont construites. Nous envisageons d'employer une méthode de *bootstraping* inspirée de l'approche SnowBall (Agichtein & Gravano, 2000), afin de collecter des patrons lexico-syntaxiques propres au corpus cible.

Chapitre 5

Interfaces de recherche augmentées

Table des matières

Outre leur rôle dans la structuration du contenu textuel des documents, les annotations sémantiques peuvent être exploitées par les moteurs de recherche pour indexer les documents. Ainsi, les moteurs de recherche peuvent permettre aux utilisateurs de rechercher des documents à l'aide de requêtes basées sur des mots-clés et de filtrer les résultats en fonction des annotations sémantiques des documents. En outre, ces annotations sémantiques peuvent être exploitées pour construire des interfaces de recherche augmentées, telles que des cartes, des frises chronologiques ou des réseaux d'entités, pour n'en citer que quelques-unes, qui viendraient compléter les moteurs de recherche traditionnels (Jatowt, 2021). Ces interfaces permettraient une lecture distante (Moretti, 2013) des documents, et permettraient de révéler des modèles cachés dans les données et de conduire à de nouvelles questions de recherche, tout en gardant l'accès aux documents originaux afin de permettre une lecture rapprochée (Gutehrlé & Atanassova, 2023).

Dans cette partie, nous nous concentrons sur les interfaces de recherche augmentées qui sont basées sur des cartes et des frises chronologiques. Tout d'abord, nous proposons un cadre pour la construction d'une interface permettant d'étudier comment les "imaginaires spatiaux" (Watkins, 2015) sont représentés dans un jeu de données de journaux historiques publiés pendant la Première Guerre mondiale. Cette interface fournit plusieurs modules, tels qu'une carte ou un tableau de concordance, qui permettent des lectures rapprochées et distantes du jeu de données sous-jacent.

Deuxièmement, nous proposons un cadre conceptuel pour générer automatiquement des frises chronologiques informatives, lisibles et interprétables à partir d'une collection de documents his-

toriques. Nous abordons les défis liés à la génération de frises chronologiques à partir de tels documents, avant de suggérer plusieurs méthodes pour implémenter ce cadre, et de discuter de ses extensions et applications potentielles. Ces cadres n'ont pas été appliqués au corpus EMONTAL, puisqu'ils sont le résultat de projets externes sur le traitement de documents historiques qui ont été réalisés dans le cadre de cette thèse.

Le reste de ce chapitre est structuré comme suit : nous décrivons notre cadre pour construire une interface pour l'étude des imaginaires spatiaux dans le chapitre 5.1, avant de présenter notre cadre conceptuel pour générer automatiquement des frises chronologiques dans le chapitre 5.2.

5.1 Cartes

Les imaginaires spatiaux sont "des histoires et des façons de parler de lieux et d'espaces qui transcendent le langage en tant que performances incarnées par des personnes dans le monde matériel" (Watkins, 2015), qui sont partagées par de grands groupes de personnes, ou une société dans son ensemble (Davoudi et al., 2018). L'interface web ainsi que notre code sont accessibles au public[1][2]. Ce travail a été présenté dans Gutehrlé et al. (2021), et a été initié à l'origine par l'équipe SpaceWars lors du Helsinki Digital Humanities Hackathon 2021[3]. Les résultats obtenus par l'équipe SpaceWars lors de ce hackathon sont présentés dans un article de blog, disponible sur le site web de NewsEye[4].

Notre interface est composée d'un module cartographique et d'un module de concordance, qui permettent de passer facilement d'une lecture distante à une lecture rapprochée du jeu des données. Ces modules permettent d'observer comment les imaginaires spatiaux sont exprimés dans le jeu de données ou dans un sous-ensemble sélectionné. Le jeu de données peut être filtré en fonction de la langue, du titre ou de la date de publication des journaux.

Le module carte montre les mentions de lieux présentes dans le jeu de données sélectionné. Plus un lieu est fréquent, plus il apparaît en grand sur la carte. La fond de carte montre les frontières et les capitales des pays entre 1913 et 1920. Ce fond de carte s'appuie sur les données fournies par Schvitz et al. (2021), qui combine plusieurs sources pour représenter les frontières des pays de 1886 à 2019. Chaque pays de la carte est associé à une couleur : plus la couleur est foncée, plus le pays est mentionné fréquemment dans le jeu de données sélectionné. En outre, la carte présente des informations contextuelles telles que les capitales et les batailles qui se sont déroulées au cours de la période sélectionnée. Les données relatives aux capitales et aux batailles proviennent des données que nous avons extraites de Wikidata. Comme pour les mentions de lieux, plus une

[1] http://spacewars.newseye.eu/
[2] https://github.com/dhh21/SpaceWars
[3] https://www.helsinki.fi/en/digital-humanities/dhh-hackathon/helsinki-digital-humanities-hackathon-2021-dhh21
[4] https://www.newseye.eu/blog/news/where-did-it-happen-spatial-imaginaries-of-world-war-i/

bataille a duré longtemps, plus elle apparaît grande sur la carte. En survolant ou en cliquant sur une entité, l'utilisateur peut afficher les métadonnées qui s'y rapportent, telles que sa fréquence, le journal dans lequel elle apparaît ou le lien vers la ressource Wikidata associée à cette mention d'entité, à condition que ce lien ait été trouvé[5].

La Figure 5.1 montre les résultats pour *Le Matin* entre 1913 et 1915 sur le module cartographique. Chaque lieu mentionné dans ce sous-ensemble apparaît sous la forme d'un point vert, tandis que les batailles qui se sont déroulées pendant cette période apparaissent sous la forme d'un point violet. De même, la capitale de chaque pays est représentée par un point orange. La légende de la carte apparaît dans le coin supérieur gauche de l'interface. Les données relatives à un lieu sélectionné, en l'occurrence "Verdun", sont affichées dans le cadre située dans le coin gauche de l'interface.

Le module de concordance fournit une table de concordance où chaque mention d'un lieu dans le document original est entourée de son contexte gauche et droit. Par défaut, ces contextes sont limités à 5 mots avant et après la mention. Chaque mention dans la table est associée à un lien qui redirige l'utilisateur vers la plateforme NewsEye, qui stocke les scans originaux des documents. Cette fonctionnalité permet à l'utilisateur de prolonger sa recherche en retournant au document original. De plus, la table de concordance est liée à un graphique qui montre la distribution des mentions d'entités dans le temps, par langue ou par journal. Le tableau de concordance et le graphique sont présentés dans la Figure 5.2, qui montre les occurrences de "Verdun" dans le temps et en contexte dans *Le Matin* entre 1913 et 1915.

Pour tester notre interface, nous avons observé comment les imaginaires spatiaux sont exprimés dans les numéros de *Le Matin* publiés entre 1913 et 1915. Cette étude de cas a révélé que la plupart des lieux mentionnés dans ce sous-ensemble sont situés en France, en Allemagne, en Suisse et en Belgique, comme le montre le module cartographique de la Figure 5.1. Ces pays apparaissent également plus fréquemment que les autres pays environnants, comme le montre leur couleur plus foncée. Cela suggère que *Le Matin* avait une vision eurocentrique de la guerre et se concentrait sur le front occidental, même si les combats s'étendaient également à l'Afrique et à l'Asie. Il s'agit là d'un exemple de biais pertinents pour la recherche historique et qui pourraient être facilement découverts grâce à notre interface.

Par ailleurs, nous avons choisi de nous concentrer sur les imaginaires spatiaux liés à la ville de Verdun. En raison de sa proximité avec la frontière allemande, Verdun a été une position clé dans le conflit. Cela est confirmé par les mentions de ce lieu dans *Le Matin* entre 1913 et 1915, qui apparaissent presque exclusivement dans les rapports de guerre ou les articles liés au conflit, comme le montre la Figure 5.2. Il est intéressant de noter que d'autres mentions de lieux liés à Verdun renforcent la relation de la ville avec le thème de la guerre. Par exemple, Verdun est parfois mentionnée comme "*place de Verdun*". "*place de*" indique généralement une place de la ville, mais

[5]Plus de détails sur l'interface peuvent être vus dans notre vidéo tutorielle:https://youtu.be/iIpEvM9IFaM

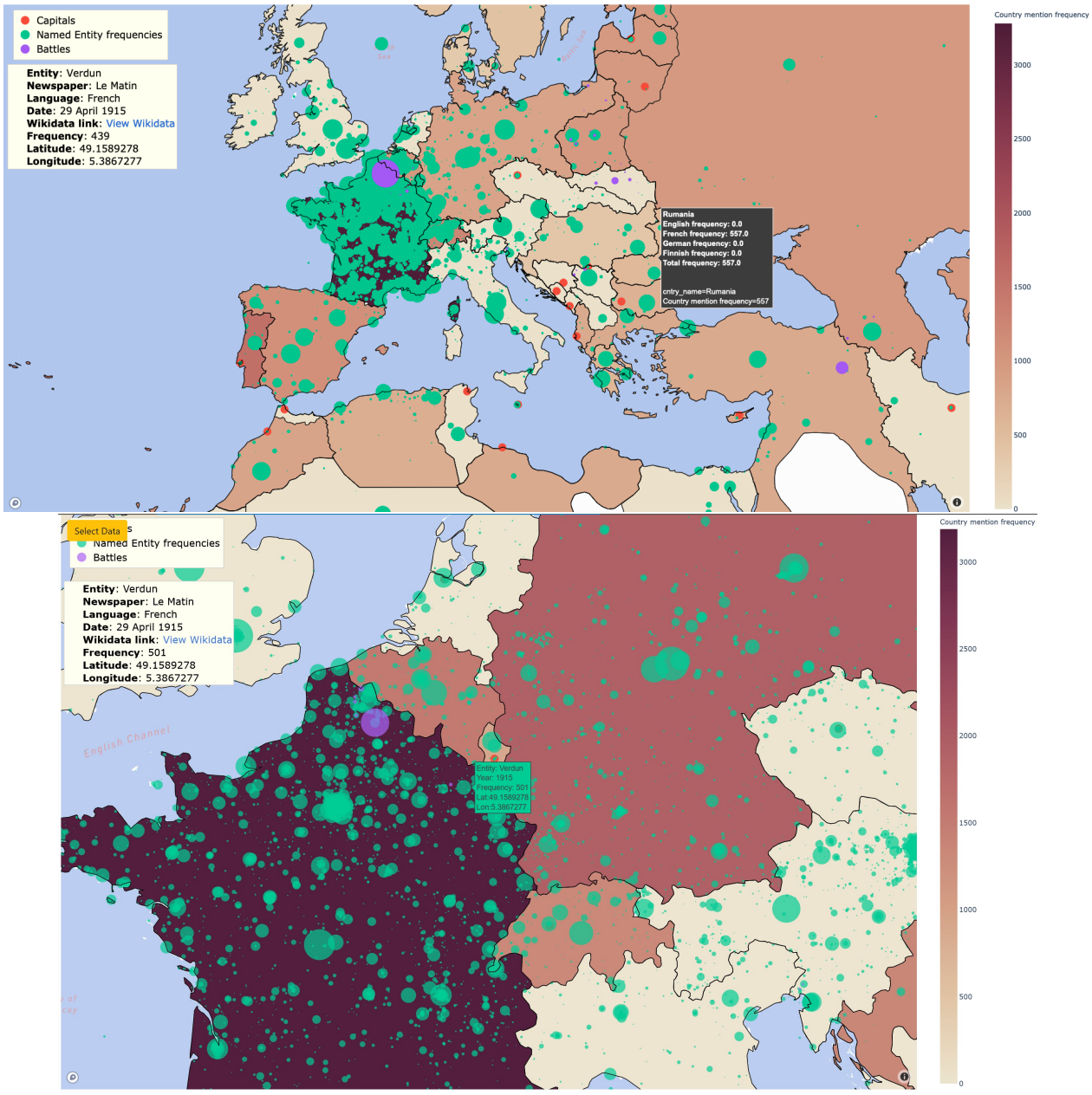


Figure 5.1: Exemple de module cartographique. Tous les lieux mentionnés dans *Le Matin* (1913-1915) sont affichés. La légende de la carte apparaît dans le coin supérieur gauche de l'interface. Les données relatives au lieu sélectionné "Verdun" sont également affichées dans le cadre situé dans le coin gauche de l'interface

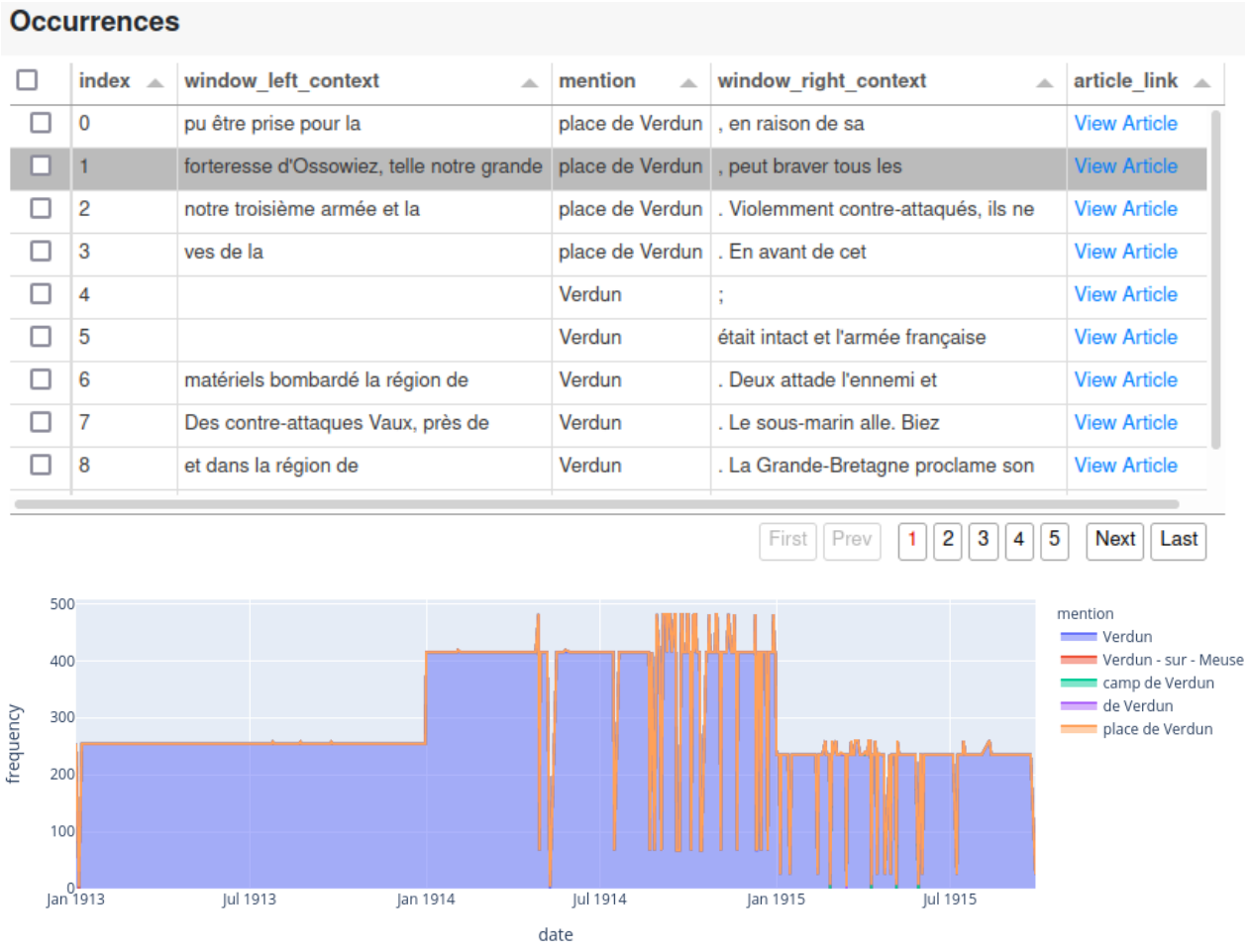


Figure 5.2: Occurrences de "Verdun" dans le temps et en contexte dans *Le Matin* entre 1913 et 1915 dans le module de concordance

dans ce cas, il s'agit toujours de la ville de *Verdun* elle-même. Il existe également des mentions du camp militaire de Verdun "*camp de Verdun*", mais uniquement dans des articles publiés en 1913 et faisant référence à la guerre franco-prussienne. Ces deux mentions insistent sur l'importance de la ville en tant que position défensive, même avant le début de la Première Guerre mondiale. La seule exception est la mention de "*Verdun-sur-Meuse*", qui était le nom officiel de la ville entre 1801 et 1970 et qui n'est mentionné que pour indiquer le lieu de naissance d'une personne. Ce sont des exemples d'éléments qui ont pu être identifiés grâce à une lecture distante du jeu de données, et qui ont été confirmés par une lecture rapprochée du document.

Ces études de cas montrent la validité d'une telle interface. Le module de carte que nous proposons permet une lecture distante du corpus, ce qui peut aider à révéler des informations cachés dans les données, tandis que le module de concordance permet de lire le jeu de données de manière rapprochée, et d'examiner les éléments qui ont pu être identifiés grâce au module cartographique. Cependant, la conception de telles interfaces dédiées à l'analyse des jeux de données historiques reste entravée par le manque de de données historiques dans un format machine, même pour des périodes étudiées de manière exhaustive telles que la Première Guerre mondiale. Ce manque de ressources dédiées oblige à s'appuyer sur d'autres ressources telles que Wikidata ou Geonames qui sont adaptées au traitement de données contemporaines, et pas entièrement adaptées au traitement de documents historiques.

5.2 Frises chronologiques

Nous proposons un cadre conceptuel pour la tâche d'*Archive TimeLine Summarisation* (ATLS), qui vise à générer automatiquement des frises chronologiques informatives, lisibles et interprétables à partir de collections d'archives de documents historiques. Ces frises chronologiques pourraient servir d'outil de lecture distante et de première étape dans l'exploration d'un jeu de données en fournissant une vue d'ensemble de ses événements clés. Le cadre ATLS est une extension de la tâche *TimeLine Summarisation*, et aborde les défis auxquels les méthodes TLS standard sont confrontées lorsqu'elles sont appliquées à des collections d'archives, telles que la rareté des données, les problèmes d'OCR, les changements de contexte et les changements linguistiques au fil du temps, afin de générer des frises chronologiques basées sur ces jeux de données. Ce travail a déjà été présenté dans Gutehrlé et al. (2022), et est le résultat d'un séjour de recherche de trois mois entre avril et juillet 2022 au Digital Science Center (DiSC) de l'Université d'Innsbruck (Autriche). Ce séjour de recherche a été encadré par le professeur Adam Jatowt (Department of Computer Science & DiSC, Deputy Head of Digital Science Center, Université d'Innsbruck, Autriche) et le professeur Antoine Doucet (laboratoire L3i, Université de La Rochelle).

Le cadre *Archive TimeLine Summarisation* prend en entrée un jeu de données longitudinales

composé de documents datés, tels que des articles de presse provenant d'une collection de journaux historiques. Ce jeu de données peut être conçu de manière autonomone ou à partir de documents renvoyés par un moteur de recherche pour une requête donnée $\mathcal{Q}$. Le jeu de données peut être au format brut ou avoir été prétraité. Nous suggérons au moins les deux étapes de prétraitement suivantes : tout d'abord, nous recommandons de nettoyer le jeu de données s'il a été traité par OCR, manuellement ou semi-automatiquement, car la qualité de l'OCR aura un impact sur les étapes suivantes (T. T. H. Nguyen et al., 2021b). Deuxièmement, nous recommandons de détecter les expressions temporelles, puisqu'elles constituent un bon indicateur des mentions d'événements. Les expressions temporelles sont soit explicites (par exemple "17 février 1995"), soit implicites (par exemple "hier", "le mois prochain"). On peut utiliser des outils tels que HeidelTime (Strötgen & Gertz, 2010) ou SUTime (Chang & Manning, 2012) pour détecter les expressions temporelles dans le texte et les résoudre dans un format de date absolue, ce qui simplifie leur utilisation dans le processus ATLS. Cependant, la détection des expressions temporelles, en particulier les expressions implicites, reste une tâche difficile. En outre, les outils disponibles tels que ceux-ci ont été principalement conçus pour des données contemporaines, et peuvent donc ne pas fonctionner aussi bien sur des données historiques. De plus, le jeu de données d'entrée pourrait être prétraité en appliquant des méthodes TAL telles que la reconnaissance d'entités nominales (NER), la modélisation thématique (*Topic Modelling*, TM), l'extraction d'événements (EE), l'extraction de relations (RE), l'extraction de mots-clés (KE) ou la génération de mots-clés (KG).

Le cadre ATLS se compose d'une étape de *génération de frise chronologique* et d'une étape de *présentation de frise chronologique*, qui sont illustrées dans la Figure 5.3. La première étape extrait du jeu de données les éléments textuels décrivant un événement et leur attribue un score d'importance. La première étape ne doit être exécutée qu'une seule fois sur l'ensemble des données traitées, puisqu'elle vise à détecter les éléments composant la frise chronologique à générer. La deuxième étape génère la frise chronologique en filtrant les événements et en sélectionnant leur description. La deuxième étape peut être exécutée plusieurs fois pour mettre à jour la frise chronologique.

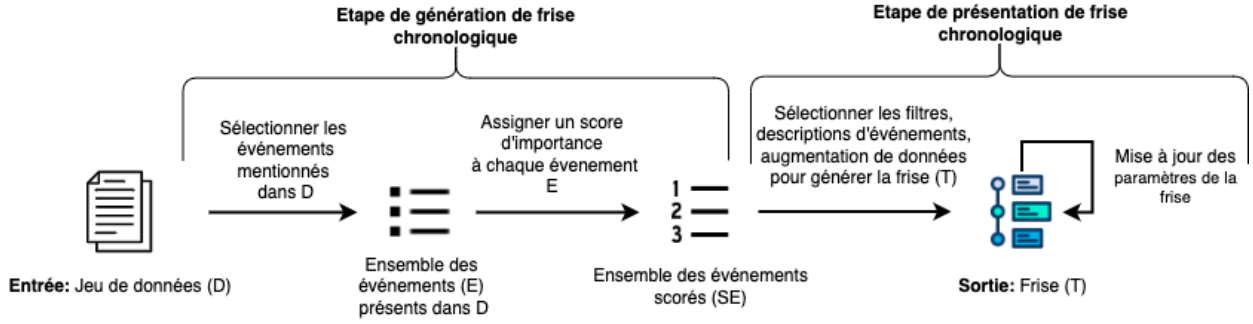


Figure 5.3: Etapes principales pour construire le système ATLS

L'étape *génération de frise chronologique* extrait les mentions d'événements de l'ensemble de données et leur attribue un score d'importance. Bien que les événements puissent être définis de différentes manières, une définition communément acceptée est "quelque chose qui *se produit* ou qui est vrai dans une circonstance donnée", comme indiqué dans le guide TimeML (Saurí et al., 2006).

Les événements peuvent être détectés de multiples façons : on peut les détecter par l'analyse statistique du corpus (Chieu & Lee, 2004; Pasquali et al., 2019b). Cependant, ces méthodes statistiques sont particulièrement adaptées aux ensembles de données homogènes, mais peuvent ne pas fonctionner aussi bien sur des ensembles de données hétérogènes ou fragmentaires. On pourrait également entraîner un modèle d'apprentissage par classement sur des résumés créés par des experts afin de détecter les phrases importantes, comme dans G. B. Tran et al. (2013). Cela nécessiterait cependant des données d'entraînement qui restent rares, même lorsqu'il s'agit de données contemporaines. Il est également possible d'utiliser un modèle de détection d'événements pour détecter et annoter les événements dans l'ensemble de données, comme dans Chasin (2010) and N. K. Nguyen et al. (2020). Cependant, l'entraînement de ces modèles nécessite des ressources annotées qui font souvent défaut, en particulier pour les données historiques. De plus, les résultats de ces modèles sont influencés par la qualité de l'OCR des documents.

Enfin, nous pourrions sélectionner comme événement toute phrase contenant au moins une expression temporelle, explicite ou implicite, comme dans Duan et al. (2019) and K.-H. Nguyen et al. (2014). Cette sélection pourrait être affinée en prenant les phrases qui contiennent également une entité nommée, comme dans Abujabal and Berberich (2015) and Bedi et al. (2017). On pourrait ensuite appliquer des algorithmes tels que Affinity Propagation (Frey & Dueck, 2007) ou Chinese Whispers (Biemann, 2006) pour rassembler les phrases décrivant le même événement, comme dans Rusu et al. (2014), Steen and Markert (2019), and Y. Yu et al. (2021).

Quelle que soit la méthode utilisée pour les détecter, les événements doivent tous être associés à une valeur temporelle. Il peut s'agir des expressions temporelles associées aux mentions d'événements, ou de la date de création du document (DCD) si aucune expression temporelle n'est présente dans le contenu du document.

Nous suggérons donc de mesurer le score d'importance de manière non supervisée en extrayant des caractéristiques de l'ensemble de données, comme dans Campos et al. (2018), Chieu and Lee (2004), and K.-H. Nguyen et al. (2014). Certaines des caractéristiques qui, selon nous, pourraient aider à mesurer ce score d'importance sont énumérées ci-dessous, avec des suggestions sur la manière de les calculer :

Redondance : plus un événement est mentionné fréquemment, plus il devrait être important. On peut alors simplement compter les occurrences des événements ou, comme alternative, leur attribuer des poids d'importance en calculant leurs scores TF-IDF sur toutes les unités de

temps. Toutefois, comme les données peuvent être fragmentaires dans les jeux de données d'archives, il est préférable de ne pas utiliser cette caractéristique seule

Références contemporaines : un événement peut être important à un moment donné si d'autres événements se produisant à la même période y font référence. Ainsi, pour évaluer cette caractéristique, nous pourrions compter le nombre de fois qu'un événement est mentionné dans les descriptions d'autres événements dans une courte période donnée autour de cet événement

Références rétrospectives : De même, un événement est susceptible d'être important si les documents continuent à le mentionner quelque temps après qu'il s'est produit. Pour évaluer ce type de référence à l'événement dans le temps, on pourrait compter la fréquence (et peut-être la durée) à laquelle un événement est mentionné par d'autres événements survenus après une période donnée.

Causalité : un événement est susceptible d'être important s'il est la cause d'autres événements qui se sont produits après lui. Pour évaluer la causalité d'un événement, on peut utiliser *graphes de référence de date* comme dans G. Tran, Herder, and Markert (2015), qui mesurent la fréquence des références, l'influence thématique et l'influence temporelle entre deux événements pour déterminer un lien de causalité. Il est également possible d'utiliser des méthodes d'extraction de relations causales (CRE), telles que celles présentées dans Gao et al. (2019) par exemple. Toutefois, la tâche d'extraction de relations causales est loin d'être résolue et peut nécessiter un prétraitement supplémentaire des jeux de données

Sens commun : certains événements sont clairement plus importants que d'autres, par exemple la naissance d'un enfant ou le mariage d'un partenaire sont généralement des événements plus importants dans l'histoire d'une famille que le fait de repeindre une maison. Pour représenter ce type de connaissance de sens commun et calculer cette caractéristique, il peut être nécessaire de créer un ensemble de données d'événements jugés importants pour former un classificateur à une classe (1CC) comme dans Duan et al. (2019) ou un modèle d'apprentissage par classement comme dans Ge et al. (2015). Il convient de noter que si les événements importants peuvent être collectés à partir de manuels historiques, la collecte d'événements non importants peut être moins aisée et plus problématique ; la solution pourrait donc consister à s'appuyer sur une tâche 1CC.

En utilisant ces caractéristiques, une formule simple pour calculer l'importance d'un événement pourrait être la suivante :

$$\alpha \cdot F1 + \beta \cdot F2 + \gamma \cdot F3 + \delta \cdot F4 + \epsilon \cdot F5$$

où $F1, F2, F3, F4, F5$ sont les valeurs mises à l'échelle des caractéristiques décrites ci-dessus et $\alpha, \beta, \gamma, \delta, \epsilon$ sont des hyperparamètres dont la valeur est définie par l'utilisateur ou les gardiens des archives documentaires. Comme pour la détection d'événements, l'utilisateur peut être invité à sélectionner l'une de ces caractéristiques pour calculer ce score.

Certaines périodes peuvent contenir beaucoup plus de documents que d'autres. Par exemple, moins de documents peuvent être disponibles en temps de guerre en raison de la censure ou des restrictions de papier. Ce manque de documents peut conduire à des événements beaucoup plus ou beaucoup moins mentionnés que d'autres et fausser les caractéristiques basées sur la fréquence telles que *redondance*, *contemporain* et *références rétrospectives*. Ces caractéristiques doivent donc être normalisées avant d'être incorporées. En outre, nous suggérons ces caractéristiques parce qu'elles sont faciles à calculer, mais nous reconnaissons également qu'elles peuvent ne pas être suffisantes pour mesurer l'importance d'un événement du point de vue d'un expert tel qu'un historien. La formule de calcul du score d'importance étant modulaire, il est possible d'incorporer d'autres caractéristiques en collaboration avec des experts.

L'étape de *présentation de frise chronologique* génère la frise chronologique à partir des événements qui ont été détectés et notés à l'étape précédente. Un jeu de données peut mentionner des centaines ou des milliers d'événements . Il est donc nécessaire de sélectionner ceux qui seront ajoutés à la frise. Pour ce faire, nous pouvons utiliser des filtres tels que ceux décrits ci-dessous. Le poids de ces filtres pourrait être modifié dans l'interface utilisateur, ce qui permettrait aux utilisateurs de mettre à jour instantanément la frise chronologique.

Top N : les N événements les plus importants sont retenus

Seuil d'Importance (IT) : seuls les événements dont le score d'importance est supérieur à un seuil préétabli IT sont retenus. Des seuils individuels pour les caractéristiques présentées précédemment qui composent le score d'importance peuvent également être définis

Seuil de Diversité Thématique ($TopDT$) : supprime les mentions d'événements redondants et garantit la diversité thématique de la frise chronologique. La diversité thématique peut être mesurée simplement à l'aide de la pertinence marginale maximale (MMR) (Goldstein-Stewart & Carbonell, 1998) ou de la métrique de blocage *n*-gramme comme dans Liu (2019)

Seuil de Diversité Temporelle ($TempDT$) : garantit que chaque unité de temps sur la frise chronologique générée est représentée de manière égale en fixant un nombre minimum et maximum d'événements qui peuvent apparaître à chaque unité de temps

Il y a plusieurs façons de représenter un événement sur une frise chronologique. On peut sélectionner une phrase qui décrit l'événement. Si cette phrase est trop longue, on peut utiliser les

méthodes de compression de phrases (Filippova & Strube, 2008) pour ne conserver que la partie la plus importante. Comme indiqué précédemment, un événement peut être représenté par un ensemble de phrases. L'utilisateur peut donc sélectionner une phrase parmi ce groupe ou générer un nuage de termes de toutes les phrases qu'il contient, comme dans Duan et al. (2019). On peut également utiliser les titres si les documents cibles sont des articles, comme dans Pasquali et al. (2019b) and G. Tran, Alrifai, and Herder (2015).

Enfin, nous pourrions également utiliser un système de génération de langage naturel (NLG) comme dans Steen and Markert (2019), puisque les textes générés sont souvent plus faciles à comprendre que les textes extraits des documents. Cependant, les méthodes abstractives telles que celles-ci peuvent souffrir d'inexactitudes ou d'hallucinations, c'est-à-dire générer des informations qui ne sont pas présentes dans les documents originaux. Ainsi, les méthodes abstractives peuvent générer des descriptions d'événements inappropriées et perdre le lien avec les documents originaux. D'autre part, un inconvénient courant des méthodes purement extractives est que les phrases sélectionnées peuvent nécessiter un certain contexte ou au moins un post-traitement pour que les utilisateurs puissent les comprendre correctement (par exemple, les pronoms peuvent avoir besoin d'être résolus ou nous devons ajouter des définitions ou des descriptions de certaines entités ou événements).

Pour les comprendre correctement, certains événements peuvent nécessiter des connaissances contextuelles qui sont absentes de l'ensemble des données traitées. Cela peut notamment se produire si l'utilisateur n'est pas un expert du domaine. Ces connaissances contextuelles peuvent être trouvées dans des bases de connaissances telles que Wikidata ou les pages de l'année de Wikipedia (voir par exemple N. K. Tran et al., 2015). Ainsi, les frises générées par un système ATLS pourraient être complétées par des données contextuelles fournies par des bases de connaissances externes, comme dans Ceroni et al., 2014.

L'évaluation d'un système TLS est une tâche difficile en raison du manque de jeux de données d'évaluation et de la subjectivité inhérente à la tâche. Afin d'évaluer les résultats, nous suggérons d'évaluer manuellement les chronologies produites, soit en suivant certains critères d'évaluation comme dans Duan et al. (2017), soit en les comparant avec des ressources créées par des experts telles que des chronologies dérivées de livres d'histoire comme dans Bedi et al. (2017). On pourrait également utiliser ce cadre pour créer un jeu de données d'évaluation spécifique au corpus donné, en vue d'une évaluation automatique.

Les chronologies sont généralement représentées de manière linéaire, où chaque unité de temps a la même taille (généralement un jour ou une année). Cependant, la granularité optimale des unités temporelles peut varier lors de la génération d'une ligne temporelle sur une longue période. De plus, les événements mentionnés dans les documents historiques ne sont pas toujours enregistrés avec la même précision temporelle (par exemple, certains événements peuvent avoir des dates manquantes,

les dates peuvent être imprécises ou difficiles à déduire). Une solution possible serait de générer des lignes temporelles logarithmiques, où la granularité de l'unité de temps change avec le temps, comme suggéré dans Jatowt and Au Yeung (2011).

Si les documents des jeux de données sont annotés avec des entités nommées, il est possible de générer des chronologies basées sur les entités. Cela pourrait aider à comprendre l'histoire d'une entité spécifique telle qu'une personne ou un lieu, comme dans Duan et al. (2019).

Discussion et Limitations

Dans ce chapitre, nous discutons de certaines des limites de la recherche présentée dans ce manuscrit.

Tout d'abord, les résultats et évaluations présentés dans cette thèse ne portent que sur les contributions principales de notre travail, à savoir l'approche proposée pour la tâche d'analyse de la structure logique des documents (LLA), le jeu de données DARES et l'approche ELIJERE. En parallèle, nous avons fait le choix de présenter une pipeline complète de traitement des documents historiques, depuis les images traitées par OCR pour obtenir leur structure textuelle au format XML ALTO, jusqu'à la construction des interfaces de recherche augmentée. Pour ce faire, et afin de pouvoir achever ce travail dans le cadre d'un doctorat, certaines implémentations et méthodes n'ont pas été évaluées dans ce manuscrit. En particulier, nous n'avons pas évalué les règles proposées pour le post-traitement des transcriptions OCR du corpus EMONTAL. Comme nous l'avons expliqué dans le chapitre 2.2, notre objectif n'était pas de proposer une correction complète des transcriptions OCR, ce qui aurait nécessité une pipeline de post-traitement beaucoup plus complexe. Au contraire, notre objectif était de proposer une méthode simple pour limiter l'impact des erreurs de non-mot sur les processus ultérieurs appliqués au corpus, tels que notre approche de la tâche LLA ou l'approche ELIJERE. En même temps, l'évaluation de l'approche ELIJERE sur le corpus EMONTAL a clairement montré que les erreurs restantes dans les transcriptions OCR ont un fort impact sur sa performance. Ainsi, une évaluation de nos règles de post-traitement OCR pourrait être nécessaire afin d'évaluer leur efficacité et de déterminer si d'autres méthodes de post-traitement s'avéreraient mieux adaptées pour nettoyer les transcriptions du corpus EMONTAL.

En outre, comme expliqué au chapitre 2.3, nous avons construit un jeu de données afin d'évaluer notre approche de la tâche LLA. Ce jeu de données a été constitué en sélectionnant manuellement des documents du corpus EMONTAL avec différentes mises en page et une qualité de transcription OCR jugée satisfaisante. L'étiquette logique correspondant à chaque bloc et ligne de texte dans le jeu de données a ensuite été ajoutée manuellement par un seul annotateur. Cependant, ce processus d'annotation manuelle n'est pas l'objet de notre travail, et nous ne fournissons donc pas ici les guides d'annotation qui ont été utilisés ou le protocole détaillé. Ceux-ci pourraient faire l'objet de publications futures dans un souci de reproductibilité de la construction de ces jeux de données. De plus, cette construction manuelle du jeu de données peut avoir introduit un biais que nous n'avons

pas pris en compte pour l'évaluation de notre approche de la tâche LLA. Ainsi, afin de limiter l'impact de ce biais, il serait nécessaire d'évaluer l'approche que nous proposons pour la tâche LLA sur d'autres jeux de données composés de documents historiques sélectionnés de manière aléatoire.

Comme indiqué dans le chapitre 4, l'approche ELIJERE que nous proposons dans ce travail souffre de plusieurs limitations. Tout d'abord, les modèles lexico-syntaxiques appris manquent de diversité et peuvent être incorrectement annotés par le processus de supervision à distance. Ils dépendent directement de la diversité des expressions et de la richesse du jeu de données DARES, sur lequel les modèles lexico-syntaxiques sont collectés. Deuxièmement, l'approche ELIJERE a du mal à s'adapter aux styles d'écriture du corpus cible et aux entrées bruitées telles que les transcriptions OCR incorrectes. La disparité des styles entre ceux du jeu de données DARES et ceux du corpus cible est un facteur important pour les scores inférieurs obtenus lors de l'évaluation sur le corpus EMONTAL. De plus, le jeu de données DARES et les patrons lexico-syntaxiques sur lesquels s'appuie l'approche ELIJERE sont construits à partir d'une base de connaissances et d'un corpus connexe qui sont limités. Bien que l'approche ELIJERE puisse théoriquement fonctionner avec n'importe quelle base de connaissances et n'importe quel corpus, elle repose actuellement sur la structure de la base de connaissances Wikidata et sur le corpus Wikipédia. Nous devons donc étudier l'adaptabilité de l'approche ELIJERE à d'autres bases de connaissances et à d'autres corpus, et évaluer l'impact de ce changement sur ses performances.

L'évaluation de notre approche pourrait également être améliorée en adaptant notre méthode pour prendre en compte d'autres jeux de données existants. Pour l'instant, nous avons évalué l'approche ELIJERE sur des jeux de données personnalisés construits à partir de Wikipedia et du corpus EMONTAL. Des jeux de données standard tels que DocRed pour l'évaluation des approches sur la tâche d'extraction de relations pourraient également être considérés. Cette question sera abordée dans de futurs travaux afin d'évaluer correctement l'efficacité de notre approche et de la comparer à d'autres approches pour les tâches d'extraction de relations et d'extraction conjointe de relations et d'entités.

Enfin, dans le chapitre 5, nous avons proposé deux cadres conceptuels pour construire des interfaces sémantiques riches basées sur des cartes et des frises chronologiques dédiées à l'exploration de collections d'archives. Nous avons décrit leurs motivations, leurs implémentations et donné des exemples basés sur d'autres jeux de données et projets. Cependant, ces cadres sont le résultat de projets externes sur le traitement de documents historiques que nous avons réalisés dans le cadre de cette thèse de doctorat, et ils n'ont pas été directement appliqués au corpus EMONTAL. Nous prévoyons d'implémenter et d'appliquer ces cadres au corpus EMONTAL, ce qui nous permettrait de montrer la pertinence et l'efficacité de ces interfaces pour l'exploration des collections d'archives. Le développement de ces interfaces web contribuerait également à ouvrir l'accès au

corpus EMONTAL à un public plus large, tel que les chercheurs en sciences humaines et au grand public, et à sensibiliser à l'importance de l'exploration des collections d'archives et de leur contenu. Plus précisément, l'utilité d'une telle interface pourrait être évaluée en permettant aux utilisateurs d'effectuer des recherches historiques liées à des sujets spécifiques, ou de mener des études historiques, telles que des travaux généalogiques ou prosopographiques. Ces travaux permettraient, par exemple, d'identifier les événements, les lieux et les moments liés à un individu, afin de construire une carte de ses déplacements, une chronologie des événements les plus importants survenus dans sa vie ou d'établir son réseau social.

Conclusion

Dans cette thèse, nous avons proposé de nouvelles approches en TAL au problème de l'exploration et de l'exploitation des collections d'archives de documents historiques. Grâce aux campagnes de numérisation entreprises par les archives et les bibliothèques ces dernières années, les collections d'archives de documents historiques sont devenues accessibles à un public plus large. Cependant, la recherche d'informations dans ces collections reste un défi en raison du manque de structuration de leur contenu textuel. Dans ce contexte, nous avons proposé plusieurs méthodes pour structurer ces documents en annotant leur contenu textuel avec des annotations sémantiques.

Comme nous l'avons souligné dans cette thèse, le prétraitement des documents pour les préparer à l'application des méthodes de TAL est une étape importante. Dans ce travail, nous nous sommes concentrés sur trois étapes de prétraitement : premièrement, nous avons proposé une approche pour nettoyer les transcriptions des documents obtenues par les méthodes d'OCR. Cette approche est basée sur des règles heuristiques pour éliminer les artefacts générés par le processus de césure et pour nettoyer les transcriptions erronées. Notre priorité avec cette approche était de limiter l'impact des transcriptions incorrectes sur le traitement ultérieur.

Deuxièmement, nous avons proposé une approche de l'analyse de la structure logique des documents (LLA). Notre approche est basée sur des caractéristiques géométriques, morphologiques et sémantiques extraites de documents au format XML ALTO. Nous avons comparé les performances de trois implémentations de cette approche : un modèle basé sur des règles élaborées manuellement, une implémentation d'apprentissage de règles basée sur le modèle RIPPER, et une implémentation d'apprentissage automatique basée sur l'algorithme Gradient Boosting. L'évaluation de ces implémentations montre que les règles que nous avons conçues spécifiquement pour notre corpus sont plus performantes que les deux autres méthodes. Cependant, cette évaluation suggère également que la combinaison de méthodes de manière hybride peut permettre d'obtenir des résultats supérieurs. Elle suggère également que les méthodes d'apprentissage de règles peuvent aider à la construction d'ensembles de règles initiales, ou à la construction de règles dédiées à des mises en page spécifiques.

Troisièmement, nous avons appliqué notre approche à la tâche LLA pour convertir les documents au format XML ALTO de notre corpus au format EMONTAL. Ce nouveau format s'inspire

des formats XML Dublin Core et XML DocBook et décrit la structure logique des documents. Il se prête donc à l'application de méthodes de TAL. Nous avons proposé un ensemble de balises pour intégrer les annotations sémantiques liées à la tâche d'extraction conjointe de Relations et d'Entités ((Joint Extraction of Relations and Entities, JERE). Cependant, ce format pourrait être étendu pour intégrer des annotations sémantiques liées à d'autres méthodes de TAL telles que l'extraction d'événements, l'analyse de sentiments, etc.

Considérant le traitement sémantique des documents historiques, une contribution majeure de cette thèse est l'approche ELIJERE (*Extensible, Lightweight and Interpretable Joint Extraction of Relations and Entities*), une nouvelle approche de la tâche JERE. Contrairement à la plupart des approches de la tâche JERE, qui reposent sur des méthodes d'apprentissage profond, notre approche s'appuie sur deux ressources linguistiques, appelées Index Syntaxique et Index Lexical, pour extraire et catégoriser les relations et les entités mentionnées dans les phrases. Ces ressources sont construites à partir de patrons lexico-syntaxiques qui expriment une relation entre des entités. L'Index Syntaxique décrit les relations qu'un patron lexico-syntaxique exprime, ainsi que le type d'entités impliquées dans les relations. L'Index Lexical décrit comment une relation est exprimée lexicalement. Pour collecter ces patrons, nous avons introduit le jeu de données DARES, qui consiste en des phrases faiblement annotées collectées à partir d'articles de Wikipedia. Ces phrases sont faiblement annotées en appliquant la méthode de supervision distante (Mintz et al., 2009), tandis que les patrons lexico-syntaxiques sont collectés en extrayant le chemin de dépendance le plus court (R. Bunescu & Mooney, 2005) entre les entités impliquées dans une relation. En appliquant cette méthode, nous pouvons rapidement construire un ensemble de données annotées de phrases exprimant n'importe quel concept stocké dans une base de connaissances, sur lequel nous pouvons construire nos ressources linguistiques. Ainsi, notre approche est extensible et interprétable puisqu'elle est basée sur des ressources linguistiques explicites et extensibles. Elle est également légère car elle nécessite peu de ressources informatiques et peut fonctionner sur des processeurs (CPUs). Nous avons évalué une implémentation de base ainsi qu'une implémentation hybride de l'approche ELIJERE sur le jeu de données DARES et le corpus EMONTAL. Les évaluations sur les deux jeux de données montrent que l'implémentation hybride est largement plus performante que l'implémentation de base. Dans des travaux futurs, nous avons l'intention d'évaluer l'approche ELIJERE sur des jeux de données standard tels que les jeux de données du New York Times ou TACRED, afin de la comparer à d'autres approches de la tâche JERE, en particulier les approches par apprentissage profond.

L'évaluation sur le jeu de données DARES montre que notre approche doit être améliorée dans plusieurs directions. Étant donné que l'annotation obtenue par la méthode de supervision distante peut être incorrecte, une étape de post-traitement est nécessaire pour nettoyer ces erreurs et améliorer les performances du système. Nous avons fait plusieurs suggestions pour le post-

traitement de ces annotations, comme l'application de méthodes de regroupement (*clustering*). Bien que notre approche permette de collecter des patrons précis, ces patrons manquent de diversité. Nous avons fait plusieurs suggestions pour collecter des patrons plus divers, comme s'appuyer sur des patrons de surface au lieu de patrons lexico-syntaxiques, collecter les patrons à partir d'un corpus qui n'est pas uniquement composé d'articles de Wikipedia, ou demander à un Grand Modèle de Langage tel que ChatGPT de générer de nouvelles phrases à partir desquelles nous pouvons collecter des patrons. L'évaluation montre également que notre approche est capable de catégoriser les entités impliquées dans une relation, mais peine à déterminer correctement leurs frontières. De même, nous avons proposé plusieurs solutions à ce problème, comme l'apprentissage des frontières des entités avec les patrons lexico-syntaxiques depuis le jeu de données DARES, au lieu de s'appuyer sur des règles élaborées manuellement.

L'évaluation sur le corpus EMONTAL montre également que notre approche est affectée par les erreurs produites par le processus d'OCR, et qu'elle peine à s'adapter aux différents styles d'écriture du corpus. Nous avons fait plusieurs suggestions pour adapter les ressources linguistiques au corpus cible, par exemple en appliquant une méthode de bootstrapping. De même, nous avons fait plusieurs suggestions pour rendre le modèle plus robuste au bruit introduit par des transcriptions OCR incorrectes, par exemple en s'appuyant sur des méthodes de recherches approximatives au lieu de méthodes de recherches exactes. De plus, notre approche pour nettoyer les erreurs produites par le processus d'OCR reste simple et ne peut pas, par exemple, nettoyer les erreurs qui nécessitent de prendre en compte le contexte. Bien que nous n'ayons pas évalué notre approche pour nettoyer les transcriptions OCR, l'impact de ces erreurs sur l'approche ELIJERE suggère que la méthode de nettoyage n'est pas suffisante et devrait être améliorée.

Enfin, nous avons proposé deux cadres conceptuels qui exploitent les annotations sémantiques des documents pour construire des interfaces de recherche augmentée avec des fonctionnalités de lecture proche et distante. Premièrement, nous avons proposé un cadre pour la construction d'une interface qui permet d'étudier l'expression des imaginaires spatiaux dans les documents. Cette interface se compose d'un module cartographique et d'un module de concordance qui fournissent des perspectives différentes sur le corpus sous-jacent. Nous avons décrit les étapes nécessaires à la construction de cette interface, avant de la tester sur un jeu de données de journaux européens publiés pendant la période de la Première Guerre mondiale. Deuxièmement, nous avons proposé le cadre *Archive TimeLine Summarisation* (ATLS) pour la génération automatique de frises chronologiques sur des collections de documents longitudinales, hétérogènes et fragmentées, telles que les collections d'archives. Nous avons fait plusieurs propositions pour mettre en œuvre ce cadre et générer des frises chronologiques informatives, lisibles et interprétables, avant de discuter de ses applications et extensions potentielles.

De telles interfaces sémantiquement riches, combinées à des moteurs de recherche, peuvent

s'avérer des outils utiles pour fournir une vue d'ensemble des collections historiques, ainsi que pour servir de nouveaux moyens d'accès à l'information dans les archives. Bien que nous n'ayons pas appliqué ces cadres à notre corpus, les méthodes d'annotation sémantique que nous avons proposées résolvent certains des principaux problèmes associés à la mise en œuvre de telles interfaces. Notre travail futur consistera à appliquer ces cadres à notre corpus et à construire des interfaces sémantiquement riches dédiées aux collections de Bourgogne et Franche-Comté. Nous avons également l'intention de demander à des chercheurs en sciences humaines (historiens, archivistes, etc.) d'évaluer la qualité des cartes et des frises générées, et d'évaluer l'efficacité de ces cadres pour l'étude des collections d'archives.

En expérimentant un corpus de périodiques historiques imprimés d'origines diverses publiés aux 19ème et 20ème siècles dans les régions Bourgogne et Franche-Comté en France, nous avons développé un processus pour le traitement de tels documents. Les approches développées dans ce travail ont de nombreuses applications dans le domaine des Humanités Numériques. Tout d'abord, les moteurs de recherche peuvent être améliorés en exploitant les annotations sémantiques ajoutées aux données textuelles par nos méthodes. Deuxièmement, des interfaces plus riches, basées sur des modules visuels tels que des frises chronologiques et des cartes, peuvent être créées pour permettre l'accès aux collections par lecture distante. Cependant, la difficulté de la recherche d'information dans les documents concerne également les corpus d'autres origines et périodes, tels que les pages web, les documents financiers, les publications scientifiques... Dans ce contexte, les méthodes proposées dans cette thèse ont été conçues de manière à pouvoir être appliquées à des corpus d'autres origines que les fonds d'archives.

Comme nous l'avons souligné dans ce travail, l'annotation sémantique de documents nécessite des ressources telles que des ensembles de règles ou des ensembles de données annotées, qui restent rares aujourd'hui. Pour cette raison, nous nous sommes concentrés sur des méthodes telles que l'apprentissage de règles ou la supervision distante pour accélérer la construction de telles ressources. Dans nos travaux futurs, nous avons l'intention de nous concentrer sur des méthodes telles que celles-ci qui soutiennent la création de ressources requises pour les approches basées sur les règles et l'apprentissage automatique. En outre, l'évaluation de notre approche pour la tâche LLA, ainsi que l'évaluation de l'approche ELIJERE, suggèrent que les approches hybrides qui combinent des méthodes basées sur des règles et des méthodes d'apprentissage automatique peuvent obtenir de meilleurs résultats que les approches qui s'appuient sur une seule méthode. C'est pourquoi nous avons l'intention d'explorer plus avant ces approches hybrides dans nos travaux futurs.

PARTIE II

Information extraction from unstructured documents for the valorisation of historical periodicals.
Application to the heritage of the Bourgogne Franche-Comté Region in France

Introduction

The large-scale exploitation of historical documents stored in archives, such as newspapers, registers, maps or certificates to name a few, is a major problem in the field of Digital Humanities, and remains challenging in many aspects. In recent years, digitisation campaigns led by archives and libraries have made it possible to preserve the content of historical documents in a digitised format and make them accessible to the general public. Moreover, thanks to automatic methods such as Optical Character Recognition (OCR), these digitised documents can be converted into machine-readable formats that can be exploited by search engines, allowing a wider audience to access and search these collections. However, the textual content of historical documents extracted by automatic methods such as OCR is generally poorly structured. While this allows for keyword searches in the documents, it does not allow for the development of more advanced interfaces for visualisation and exploitation of the data. When dealing with large collections, users performing keyword searches can quickly become overwhelmed by a multitude of results that are not always relevant to their search.

As a result, a significant problem remains: historical documents contain vast amounts of information that can be potentially useful to both the general public and researchers in the Humanities; however, extracting specific information from large collections, while possible, requires considerable effort. This makes it difficult to envisage all possible applications in this area and to explore the full potential of the exploitation of historical documents. In this context, Kaplan and di Lenardo (2017) define the "Big Data of the Past", referring to the phenomenon of intensive and rapid production of historical data as a result of the mass digitisation of historical collections, and the need to develop new methods and tools to deal with these data.

A cornerstone in the creation of interfaces for the efficient exploration of historical datasets and the improvement of search interfaces is the structuring of the textual content of archival documents. The textual structure itself, i.e. the division of text elements into titles, headers, paragraphs, etc., conveys part of the meaning, allows differentiated indexing of the different types of elements, and thus the creation of more efficient search engines and interfaces. The field of Natural Language Processing (NLP) offers various methods and tools in this regard, which need to be improved and adapted to the types of documents in historical document collections. Furthermore, NLP tools can be used to annotate the textual content of documents with semantic categories such as sentiments (e.g. positive, negative), entity types (e.g. person, place, time) or relations (e.g. place of birth, educated at), to name a few. Such semantic annotations can be integrated into the indices of search interfaces to help explore the collections of documents beyond the possibilities of keyword search (Atanassova, 2012), and also to build augmented search interfaces, such as maps or timelines, which would allow a distant reading (Moretti, 2013) of the collections.

The semantic annotation of historical documents remains a challenge for several reasons. The documents in machine-readable formats obtained by automatic methods such as OCR usually con-

tain errors that have been shown to affect the results obtained by search engines and NLP methods (Boros, Nguyen, et al., 2022; M. Ehrmann et al., 2021a; Linhares Pontes et al., 2019a). Therefore, post-processing methods are usually required to remove these errors. In addition, most NLP tools available today are designed to process documents written in the modern state of the language and are generally dedicated to specific writing styles. As a result, these tools are usually not adapted to process historical documents, which may be written in different styles and in an earlier state of the language. Although resources for processing historical documents have appeared in recent years, such as annotated datasets (M. Ehrmann, Romanello, Flückiger, & Clematide, 2020; M. Ehrmann et al., 2022; Gutehrlé & Atanassova, 2021a; Hamdi et al., 2021), such resources remain rare.

In this PhD thesis, we address the question of the exploration and exploitation of collections of historical documents through methods and tools in NLP. Our major contribution in this field consists of three parts:

1. proposing a new method to determine the logical structure of historical newspapers using a rule-based approach

2. proposing a method to extract entities and relations about people and locations that are mentioned in texts. Our approach is called Extensible, Lightweight and Interpretable Joint Extraction of Relations and Entities (ELIJERE). It relies on linguistic resources that have been obtained by distant supervision

3. proposing two general frameworks for the design of augmented search interfaces, based respectively on automatically generated maps and timelines, which allow easy switching between close and distant reading of the documents

For this work, we have built a corpus of printed periodicals of various origins published in the 19th and 20th centuries in the *Bourgogne* and *Franche-Comté* regions in France. We choose to focus on periodicals, given their importance as primary sources of information to the general public reflecting local and international events at the time. The documents in our corpus address various topics such as war, religion or science, and are written in various styles. We have collected the corpus from Gallica, the digital archive of the *Bibliothèque Nationale de France (BNF)*[6]. Gallica provides the content of the document in the XML ALTO format, which provides the OCR transcription of the document, its reading order and its physical organisation. We focus on the processing of the textual content of these documents rather than on other types of content such as images, advertisements or tables, as most information in our corpus are conveyed through textual means.

[6] https://gallica.bnf.fr/

This work is part of the EMONTAL project (*Extraction et Modélisation ONTologique des Acteurs et Lieux pour la valorisation du patrimoine de Bourgogne Franche-Comté*)[7] and is funded by the *Bourgogne Franche-Comté* region in France for the years 2020-2023.

The rest of this work is structured as follows: in Part I, we address the question of applying NLP methods to structure the textual content of historical documents. We first provide an overview of the various NLP methods that can be applied, as well as an overview of other projects similar to ours. We then describe the Document Layout Analysis task, which is necessary to convert the documents into a machine-readable format. Finally we describe the Semantic Annotation task, and focus in particular on the Named Entity Recognition and Relation Extraction tasks. In Part II, we present the documents in the EMONTAL corpus and the topic they address. We then describe our approach to clean the errors produced by the OCR process, as well as our approach to determine the logical structure of documents.

We present the ELIJERE approach in Part III. We first describe the DARES dataset, which consists in sentences weakly annotated with relations and entities collected from Wikipedia and Wikidata. Then we describe our method to build the linguistic resources from the DARES dataset, before presenting how we exploit these resources for the Joint Extraction of Relations and Entities task. In Part IV, we evaluate the ELIJERE approach. We evaluate a base implementation of this approach, which rely on the two linguistic resources to extract and categorise candidate relations. Moreover, we evaluate a hybrid implementation which combines the linguistic resources with a machine-learning classifier to categorise the candidate relations. We evaluate both implementations on the DARES dataset, before evaluating them on the EMONTAL corpus, in order to study the impact of OCR errors and various writing styles on our approach.

In Part V, we present two general frameworks for the conception of augmented search interfaces based on automatically generated maps and timelines. For each framework, we provide a description of related works, then present the framework in detail, before discussing its potential applications and extensions. Finally we discuss the limitations of this work, before presenting the conclusion of this work and suggestions for future works.

[7] http://tesniere.univ-fcomte.fr/projet-emontal/

Part I

Contextualisation and state of the art

In recent years, libraries and archives have undertaken numerous digitisation campaigns of their collections. While these campaigns have open the access of archival documents to a wider audience, the exploration and exploitation of their contents remain challenging tasks, as they are usually poorly structured. When accessed through search engines, the user is limited by queries based on keywords and can be overwhelmed by the abundance of documents available but not always relevant to their search. Thus, it is necessary to structure the textual content of archival documents in order to improve existing search interfaces and facilitate the exploration and exploitation of these "Big Data of the Past" (Kaplan & di Lenardo, 2017).

In this part, we address the topic of applying Natural Language Processing (NLP) methods to historical documents. These methods allow to enrich documents through the addition of semantic annotations, which can be exploited by search engines to index the documents and allow for more powerful queries. Moreover, these annotations can be exploited to build augmented search interfaces such as maps or timelines, which allow a close reading and a distant reading of the corpus.

The rest of this part is structured as follows: in Chapter 6, we give a general presentation of Natural Language Processing methods applied to collections of historical documents. In Chapter 7, we describe the Document Layout Analysis task, which aims at converting documents into a machine-readable format. Finally, we describe the Information Extraction task in Chapter 8, and focus in particular on the Named Entity Recognition and Relation Extraction tasks.

Chapter 6

Natural Language Processing applied to historical documents

Table of contents

The digitisation campaigns carried out by libraries and archives in recent years have facilitated access to documents in their collections. However, exploring and exploiting these documents remain difficult tasks due to the sheer quantity of documents available for consultation.

In this chapter, we address the question of the application of Natural Language Processing methods on collections of historical documents. Rather than proposing an exhaustive state of the art, in the following sections, we only outline the major research problems and advances in this field.

In Section 6.1, we explain how the semantic annotations produced by NLP methods can help in structuring the textual content of historical documents. We provide multiple examples of such semantic annotations, as well as several examples of how they can be exploited by search interfaces. In Section 6.2, we present multiple methodological and technological locks that must be addressed

for the efficient processing of historical documents. Finally in Section 6.3, we describe in details several Digital Humanities projects which have applied Natural Language Processing methods to structure collections of historical documents and have built augmented search interfaces to assist in the exploration and exploitation of these collections.

6.1 Semantic Enrichment of historical documents

Among the works applying NLP methods to historical documents, an important body of work has focused on the application of Named Entity Recognition (NER) methods. The Named Entity Recognition (NER) task aims at detecting the mentions of real entities such as a person, a place, a date or an organisation in documents (M. Ehrmann, 2008; M. Ehrmann et al., 2021a). Detecting the mentions of entities is important to understand the content and topics of documents. Moreover, these mentions can be exploited by search engines to index the documents, thus allowing to query and filter documents according to their entities. For these reasons, the NER task applied to historical documents has been the subject of an important focus in the recent years, as shown by works such as Boros, Hamdi, et al. (2020), M. Ehrmann, Romanello, Flückiger, and Clematide (2020), and M. Ehrmann et al. (2021b).

The Named Entity Recognition task can be used for other downstream tasks such as Named Entity Linking, Relation Extraction, or Event Extraction. For instance, the Named Entity Linking (NEL) task aims at determining to which real entity multiple entity mentions are referring to. Usually, this task associates these entity mentions to a single entry from a database or knowledge base such as Wikidata or FreeBase. Thus, the NEL task allows to understand and query the corpus based on real entities and not solely on entity mentions. Moreover, the use of an external knowledge base such as Wikidata allows to add contextual data to the corpus, for instance information about the entities or the document itself (Linhares Pontes et al., 2019b).

On the other hand, the Relation Extraction (RE) task aims at detecting the relations between entities, such as the *place of birth* or *educated at* relations. Thus, the RE task allows to study how entities are related to each other. Moreover, the detected relations can be exploited to build knowledge graphs from the documents, which can be used for other downstream tasks such as information retrieval (E. Klein et al., 2014; Plum et al., 2022). Finally, other works such as Boros, Nguyen, et al. (2022), Ide and Woolner (2004), and Smith (2002) have focused on the Event Extraction (EE) task, which aims to detect the mentions of events in documents. These methods also aim to detect when and where the event occurred, as well as who participated in it. When applied to historical documents, Event Extraction methods can assist in understanding the documents as well as the detected entities by building timelines from the detected events.

Among other Natural Language Processing tasks, the Topic Modelling (TM) task applied to

historical documents has also received an important interest in the recent years, as shown by works such as Marjanen et al. (2020a) and Y. Yan et al. (2009). The TM task aims at detecting in an unsupervised manner the topics discussed in a corpus. Most methods are adaptable to any corpus since they detect latent topics from the documents, instead of searching for explicit topics chosen by the user. By applying such methods, the documents can be indexed according to their topics and allow to search and compare documents according to their topic, as well as to study the evolution of a topic over time. Other works such as Dominguès et al. (2019), Hamdi et al. (2021), and Sprugnoli et al. (2016) have focused on the Sentiment Analysis task, which aims to measure the polarity (i.e. positive, negative or neutral sentiment) or the emotions expressed in texts. When applied to historical documents and combined with Topic Modelling methods, Sentiment Analysis methods allow to study how a topic was perceived over time.

Apart from their role to structure the textual content of documents, the semantic annotations can be exploited by search engines to index the documents, as in the *impresso*[1], *NewsEye*[2] or *Retronews*[3] platforms. Such platforms allow the user to search documents using keyword-based queries, as well as filtering documents according to their semantic annotations. Moreover, the semantic annotations produced by NLP methods can be exploited to build augmented search interfaces which allow a distant reading of the corpus (Moretti, 2013), while allowing to return to the original document for close reading (Jatowt, 2021).

For instance, works such as E. Klein et al. (2014) propose to study the vocabulary of the corpus as word clouds, where the size of the words depend on their frequency in the corpus. These word clouds can be augmented with other annotations such as sentiments as in Sprugnoli et al. (2016). Other works such as M. Ehrmann, Romanello, Clematide, et al. (2020) propose to exploit diachronic word embeddings to suggest terms that are similar to keywords used in a query and which are closer to the state of language in which the documents are written than the modern state. Similarly, works such as Cordell and Smith (2017) and Romanello et al. (2020) propose several visualisation tools such as graphs to observe the reiteration of similar textual content across documents.

Bertin et al. (2015), Blevins (2014), Dominguès et al. (2019), and Moncla et al. (2019) propose to automatically generate maps from the mentions of locations that have been detected by the Named Entity Recognition process. An example of such map is shown in Figure 6.1. To generate such maps, a geocoding step is necessary, which aims at associating coordinates to a mention of a location from a text (McDonough et al., 2019). This step requires external resources known as gazetteers such as Geonames[4] or Pleiades (Bagnall et al., 2016), which stores the coordinates

[1] https://impresso-project.ch/
[2] https://www.newseye.eu/
[3] https://www.retronews.fr/
[4] https://www.geonames.org/

of known existing and ancient places. These maps can be augmented with contextual knowledge provided by external sources such as Wikidata, as in Gutehrlé et al. (2021).

Similarly, the events as well as temporal expressions mentioned in documents can be exploited to automatically generate timelines from historical documents, as suggested in Gutehrlé et al. (2022). Such timelines can provide an overview of historical collections as well serve as a novel information access means to news article archives, as in Duan et al. (2017) and Pasquali et al. (2019a). We provide more details on augmented search interfaces based on automatically generated maps and timelines in Part V.

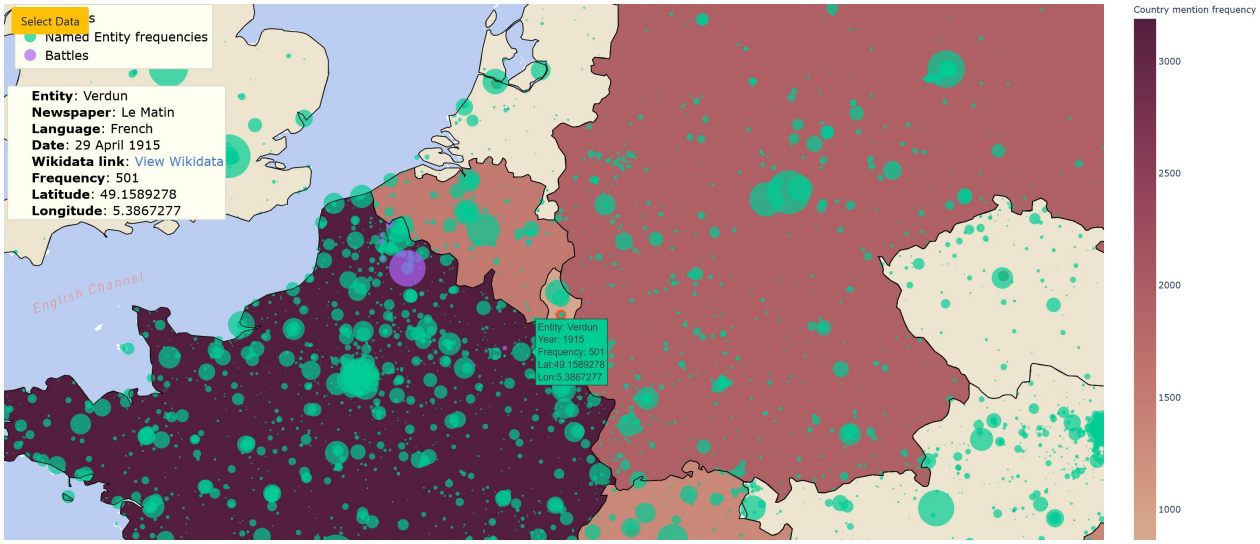


Figure 6.1: Example of a semantic interface: automatically generated map from identified location in *Le Matin* (1913-1915) (Gutehrlé et al., 2021)

6.2 Methodological and technological locks

Thanks to digitisation campaigns led by archives and libraries, many collections of historical documents are now available as images, that can be consulted online. In order to process these documents with NLP methods, it is first necessary to convert the structure of these documents as well their textual content into a machine-readable format. Due to the large amount of available documents, this task is often performed automatically. The physical structure and reading order of the document can be determined by applying Physical Layout Analysis (PLA) methods, whereas the organisation of the document into logical elements (header, title, paragraphs...) can be determined by applying Logical Layout Analysis (LLA). Finally, the textual content of documents can be transcribed by Optical Character Recognition (OCR) methods, or by Handwritten Text Recognition (HTR) methods when the documents are manuscripts.

The success of these methods depends on multiple factors, such as the quality of the original documents and its numeric copy, the layout of the document, the typography used, as well as the difference between the language of the document and the language for which the OCR model has been trained. Thus, the documents in machine-readable format produced by Document Understanding methods usually contain errors, that have been shown to impact the results obtained by NLP methods and search engines (Boros, Nguyen, et al., 2022; M. Ehrmann et al., 2021a; Linhares Pontes et al., 2019b). Post-processing methods are usually required to remove these errors. NLP methods that are robust to the noisy input are necessary to structure historical documents. We provide a more detailed presentation of the methods and challenges related to the Document Understanding task in Chapter 7.

Collections of historical documents are usually heterogeneous and contain documents of different natures (periodicals, books, ...) and origins (religious periodicals, literature, politically-oriented newspapers, etc.) written in various languages and styles. Moreover, depending on the period of origin of the documents, the language or languages in which the documents are written may be more or less distant from their modern form. Most NLP tools available today are designed to process documents written in the modern state of the language, and are usually dedicated to specific writing styles. The performance of these tools is therefore lower when they are applied to documents written in several languages, in a different style or in an older state of the language (M. Ehrmann, Colavizza, Rochat, & Kaplan, 2016). Moreover, the external resources such as Wikidata or Geonames used for tasks such as the Named Entity Linking or geocoding tasks are generally built from contemporary data. Therefore, they may not be adapted to the period of the corpus under study, and carry the risk of introducing anachronisms and biases into the interpretation of results (Gutehrlé et al., 2021).

Thus, the processing of historical documents requires to either build dedicated models or to adapt existing models, for instance by applying transfer learning methods as in Boroş et al. (2020). Multiple resources dedicated to tasks such as Named Entity Recognition (M. Ehrmann, Romanello, Flückiger, & Clematide, 2020; M. Ehrmann et al., 2022), Logical Layout Analysis (Gutehrlé & Atanassova, 2021a) or Named Entity Linking (Hamdi et al., 2021) have been published in the recent years, which allow for the creation of such tools. However, such resources remain rare for many other tasks and languages, as their creation is a costly task (T. T. H. Nguyen et al., 2021a).

Other works such as Piotrowski (2012) have suggested to apply normalisation methods, so that the textual content of documents corresponds to a state adapted for these tools. However, the further the state of the language is from the modern language, the more difficult it is to normalise orthographic variations. Moreover, most methods focus exclusively on the correspondence between the ancient and modern spellings of a word, but do not deal with their semantic evolution. There is therefore a risk of matching terms whose meaning has since changed. Thus, although this step

facilitates queries and the use of NLP tools, it is important to ensure that it does not introduce errors. We provide a more detailed presentation of the Information Extraction task applied to historical documents and its challenges in Chapter 8.

6.3 Related projects

In the following sub-sections, we present several Digital Humanities projects similar to ours that have applied Natural Language Processing methods to structure the textual content of historical documents. We first provide a detailed presentation of the *Venice Time Machine*, *impresso: Media Monitoring of the Past* and *NewsEye: A digital investigator for historical newspapers* projects, before giving an overview of three other similar projects. We focus on these projects in particular, as they have proposed new methods and tools for the efficient exploration and exploitation of historical documents.

6.3.1 Venice Time Machine

The *Venice Time Machine*[5] project is a project initiated in 2012 by the Ecole Polytechnique Fédérale de Lausanne (EPFL) and the University Ca'Foscari of Venice which aims to build a multidimensional model of Venice and its evolution over a period of more than 1,000 years (Kaplan, 2015). The project is supported by the *READ European eInfrastructure* project, the SNSF project *Linked Books*, the ANR-SNSF Project *GAWS* during its first phase, the *Parcels of Venice* SNSF Project for its second phase, as well as by the philanthropic support of the Lombard Odier Fondation.

The first phase of the project ran between 2012 and 2019 and aimed to create the largest geo-historical database of Venice by digitising the millions of register pages and photographs in the State Archive in Venice and at the Fondazione Giorgio Cini (Kaplan, 2020). To do so, the project proposed to apply deep-learning approaches to digitise, transcribe and structure these "Big Data of the Past" (Kaplan & di Lenardo, 2017).

Due to the ancient and fragile nature of the documents, novel digitisation methods have been developed to digitise the documents without opening them (Albertin et al., 2015a, 2015b). Moreover, several approaches such as Colavizza et al. (2019) have been proposed to assist in the conception of digitisation strategies, and determine what collections to digitise in priority.

Deep-learning approaches have been applied to determine the structure of documents (Oliveira et al., 2018), as well as transcribing their textual content (Oliveira & Kaplan, 2018). These approaches have also been applied to convert non-textual documents such as cadastral maps into machine-readable format (Ares Oliveira et al., 2019).

[5] https://www.epfl.ch/research/domains/venice-time-machine/inbrief/

The information extracted from the documents is stored in a semantic graph of linked data and accessible via a historical geographical information system (Kaplan, 2015), which allows textual queries, as well as visual and geo-historical queries (di Lenardo et al., 2016).

Several projects have been initiated from the Venice Time Machine project. Among them, the *Garzoni* project[6] proposes to build an information system from the *Accordi dei Garzoni* and apply computational methods to assist in historical research on the questions of apprenticeship in the early modern Venetian society (M. Ehrmann, Colavizza, Topalov, et al., 2016). Similarly, the *Linked Books* project[7] proposes to study the history of Venice by applying machine and deep-learning approaches to build a network of citations from books and journals on the history of Venice (Colavizza et al., 2018). Finally, the *Replica* project[8] proposes to digitise historical photographic archives and to apply deep-learning methods to build a search engine dedicated to the exploration of artistic collections (di Lenardo, 2022).

The Venice Time Machine project has entered its second phase in 2020, which aims to develop the Venice Mirror World, a 4D model of the city. Moreover, the project is now part of the Time Machine Europe program[9], which includes similar Time Machine projects such as *Amsterdam Time Machine*[10] or *Antwerp Time Machine*[11].

6.3.2 impresso: Media Monitoring of the Past

The *impresso: Media Monitoring of the Past* project is an interdisciplinary research project which aims to leverage machine and deep-learning methods to assist in the exploration and exploitation of historical media across time and languages. The project is led by a consortium of designers, digital humanists, computational linguists and historians from the Ecole Polytechnique Fédérale de Lausanne, the University of Zurich and the University of Luxembourg for its phase, and also from the University of Lausanne for its second phase. It is funded by the Swiss National Science Foundation as part of the Sinergia funding programme for its first phase, and funded by the Swiss National Science Foundation and the Luxembourg National Research Fund as part of the Sinergia / INTER funding programmes for its second phase.

The first phase of the project ran from 2017 to 2020 and aimed at developing critical text-mining tools and search interfaces to explore and study large-scale newspaper archives. These methods were applied to a multilingual corpus of Swiss and Luxembourgish digitised newspapers written in French, German and English (Romanello et al., 2020). The textual content and the

[6] https://www.epfl.ch/labs/dhlab/projects/garzoni/
[7] https://www.epfl.ch/labs/dhlab/projects/linkedbooks/
[8] https://www.epfl.ch/labs/dhlab/projects/replica/
[9] https://www.timemachine.eu/
[10] https://www.amsterdamtimemachine.nl/
[11] https://www.uantwerpen.be/en/projects/antwerp-time-machine/

structure of the documents have been extracted by applying custom Optical Character Recognition and Optical Layout Recognition methods (Barman et al., 2020). Errors produced by OCR methods have been corrected by applying neural translation models (Amrhein & Clematide, 2018). Several NLP methods have been applied to enrich the documents at the semantic level. The documents have been processed with linguistics analysis such as part-of-speech tagging and lemmatised (Makarov & Clematide, 2018a, 2018b). The mentions of persons, places and organisation have been detected by Named Entity Recognition methods. Moreover, Named Entity Linking methods have been applied to solve entity mentions and to add contextual knowledge (M. Ehrmann et al., 2021b; Ehrmann et al., 2020). The content and the topics of documents have been determined by the application of keyphrase extraction, text classification, and multilingual Topic Modelling methods. Finally, several methods have been proposed to find meaningful reiteration of similar texts across documents, time and languages (Düring et al., 2023; Romanello, 2018).

The documents are accessible through the *impresso* web app[12]. From this interface, the user can perform keyword-based queries as well as filter documents according to their semantic annotations. Moreover, the interface allows to perform queries based on images, and to create subcorpus from the entire collection of documents. This interface also exploits diachronic and multilingual word embeddings (Schelb et al., 2022) to allow the user to query documents across languages, as well as to suggest keywords. The interface also proposes several tools to visualise the corpus metadata and the semantic annotations, such as the detected topics or reused texts. Examples of the *impresso* web app are shown in Figures 6.2, 6.3, 6.4 and 6.5.

The second phase of the *impresso* project has started in 2023. This phase proposes to extend the newspapers collection to Western European radio sources, as well as extending the capabilities of the web app (Düring et al., 2023). Moreover, this phase will propose the *impresso* data lab[13] which will provide researchers with data access and annotation services.

6.3.3 NewsEye: A digital investigator for historical newspapers

The *NewsEye: A digital investigator for historical newspapers* project is a project funded by the European Union's Horizon 2020 research and innovation programme and led by a consortium of the national libraries of Austria, France and Finland, and research groups in humanities, social science and computer science from the universities of La Rochelle, Helsinki, Innsbruck, Rostock, Paul-Valéry Montpellier and Vienna between 2018 to 2021. The project proposes to leverage deep-learning and big data approaches to develop methods and tools for the efficient exploration and exploitation of historical newspapers. These methods are applied to a corpus of historical newspapers published in Finland, France, Germany published in the 19th and 20th centuries (Doucet et al.,

[12] https://impresso-project.ch/app/
[13] https://impresso-project.ch/the-app/

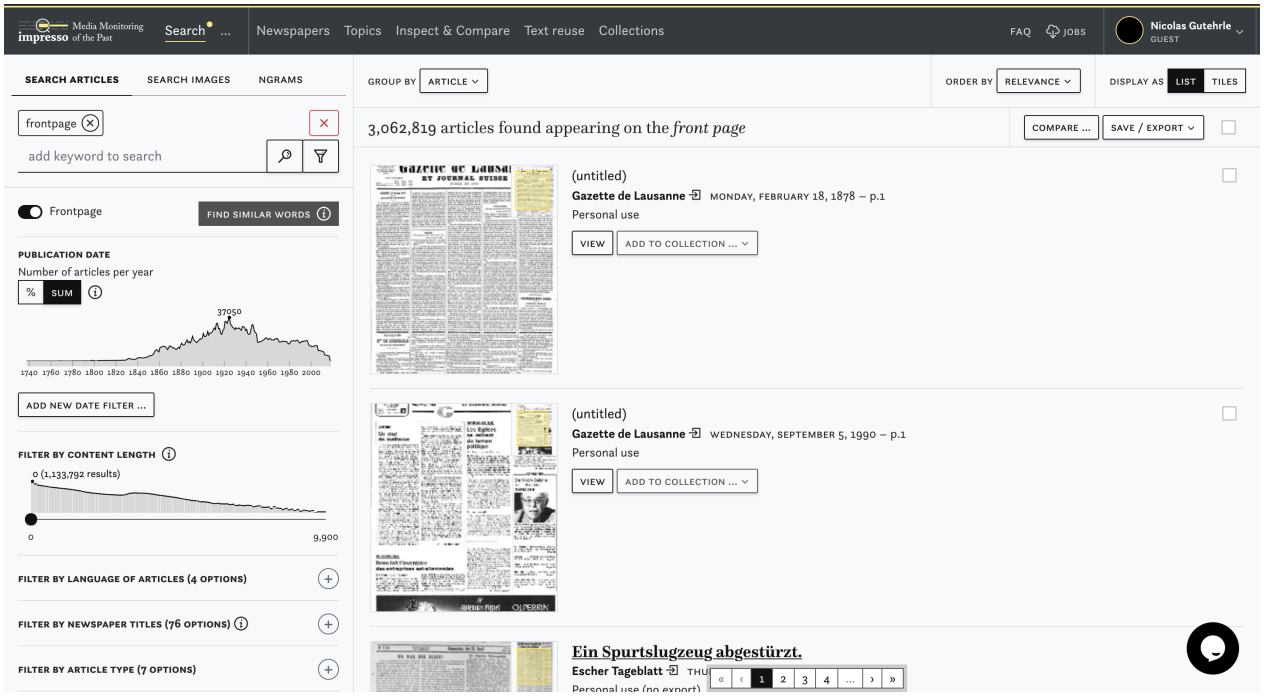


Figure 6.2: Search tab in the *impresso* web app

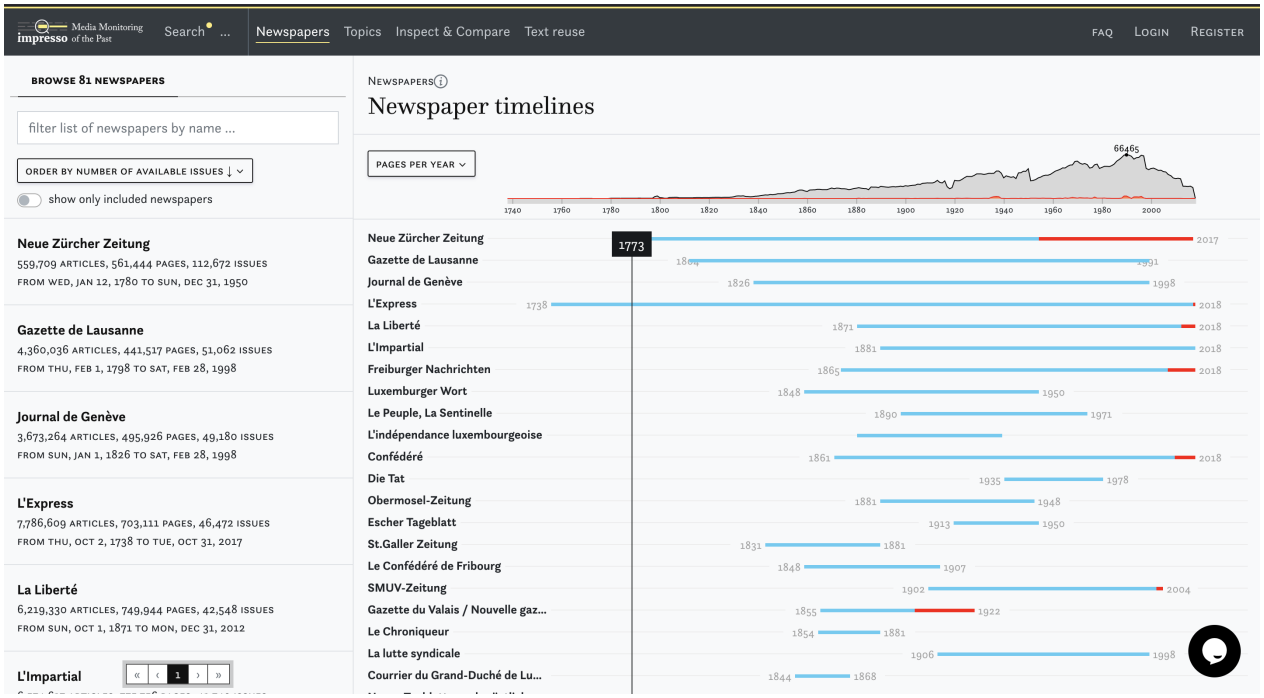


Figure 6.3: Newspapers tab in the *impresso* web app

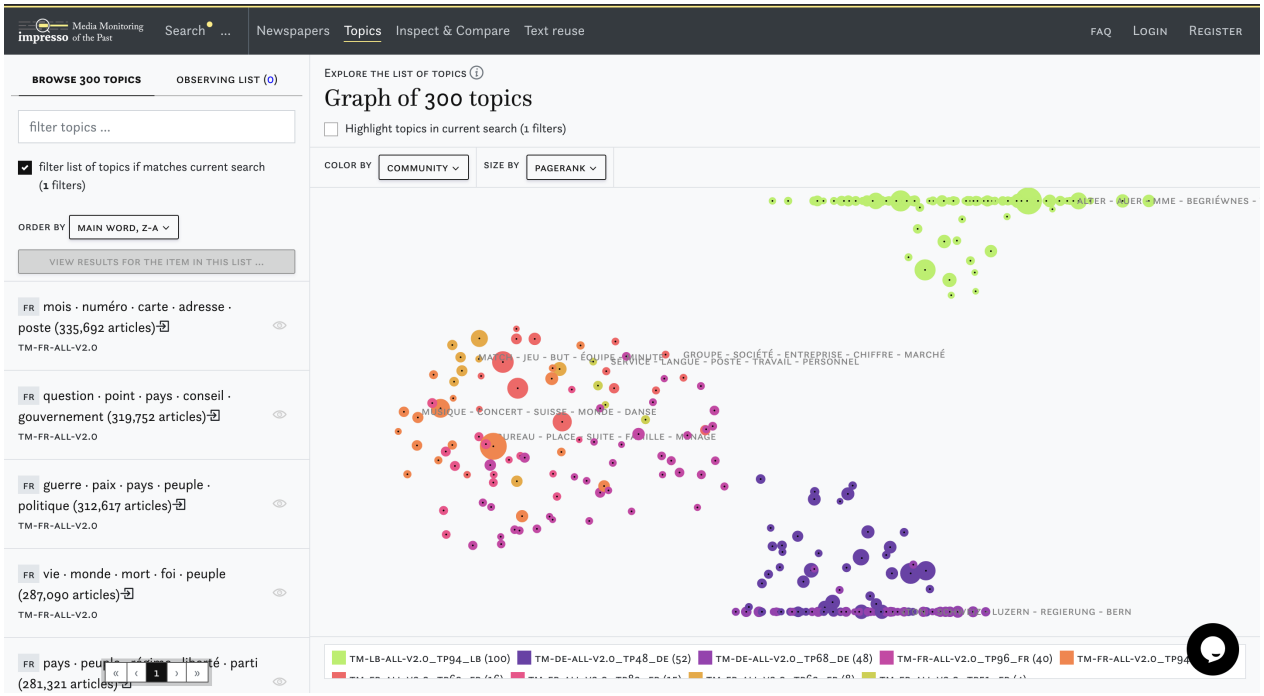


Figure 6.4: Topics tab in the *impresso* web app

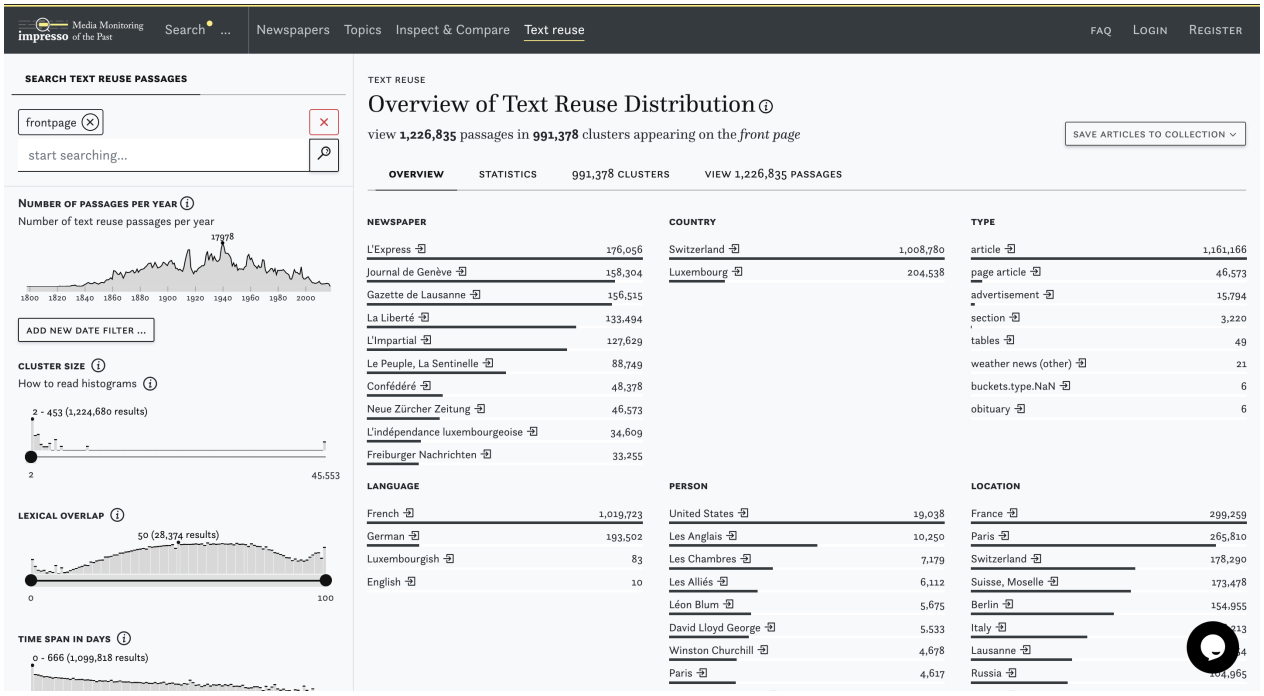


Figure 6.5: Text reuse tab in the *impresso* web app

2020).

Deep-learning approaches to the OCR and OLR tasks have been applied to obtain the textual content and the structure of the documents in the corpus (Michael et al., 2019). Similar approaches have been applied to post-process and correct errors produced by the OCR methods (Duong et al., 2020). Moreover, the textual content of documents has been enriched with semantic annotations, as in the *impresso* project. Deep-learning approaches to the Named Entity Recognition and Named Entity Linking tasks have been applied to detect and solve the mentions of persons, locations, organisations and human productions in the documents (Boros, Hamdi, et al., 2020; Boros, Pontes, et al., 2020; Linhares Pontes et al., 2022; Pontes et al., 2020). Similarly, deep-learning approaches to the Event Detection task that are language-independent have been applied to detect the mentions of events in the documents, as well as their time of occurrence and the persons involved in them (Bernard et al., 2021; Boros, Moreno, & Doucet, 2022; Boros et al., 2021).

A combination of lexicon-based and machine-learning based approaches have been applied to detect the stance of documents, i.e. to determine the bias of the author of a text (Hamdi et al., 2021). The documents have also been processed with multilingual and dynamic topic models to allow contextualised and contrastive content analysis. These methods allow to study the evolution of a topic or a discourse across time and languages (Marjanen et al., 2019, 2020b; Zosa & Granroth-Wilding, 2019). Moreover, several methods have been proposed to generate explicit labels and descriptions of the topics produced by the topic models (Zosa et al., 2022).

The documents in the corpus are available through a search interface which allows to perform keyword queries, as well as to filter documents according to their semantic annotations. An example of this interface is shown in Figure 6.6. Moreover, this interface integrates a personal research assistant which allows to perform automatic and explicit search of the corpus. This assistant is composed of an *Investigator* component which performs the search in an automatic way, as well as a *Reporter* and an *Explainer* component which provide an explanation of the results as well as an explanation of the process leading to these results. The explanations are expressed in natural language by applying Natural Language Generation methods (Leppänen & Toivonen, 2021; Pivovarova et al., 2020).

6.3.4 Other projects

Apart from the *Venice Time Machine, impresso: Media Monitoring of the Past* and *NewsEye* projects, there has been numerous other Digital Humanities projects which rely on NLP methods to structure historical documents. We present some of these projects below.

The *Trading Consequences* project (E. Klein et al., 2014) is a collaborative project between environmental historians in Canada and computational linguists and computer scientists in the UK

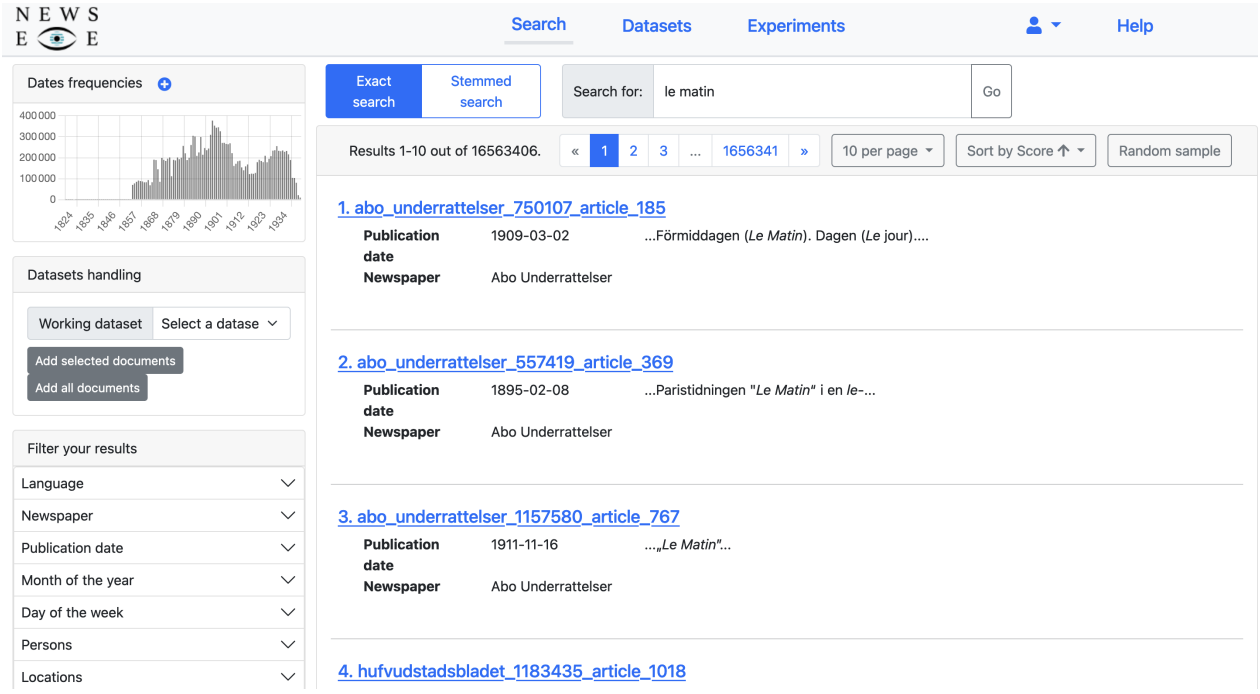


Figure 6.6: Search tab of the *NewsEye* web app

which proposes to apply text-mining methods to structure and assist in the study of a two million pages corpus related to the commodity trade in the 19th century. Rule-based and lexicon-based approaches have been applied to detect and annotate the mentions of persons, locations, organisation, monetary values, dates and commodities. Similarly, a rule-based approach has been applied to extract the relations between the mentions of locations and commodities occurring in a sentence. These semantic annotations are exploited by several visualisation tools which allow a distant reading of the corpus. For instance, a map allows to observe the locations mentioned in the corpus, word clouds to visualise important commodities as well as important locations and networks to observe the relations between commodities.

The *Viral Texts* project (Cordell & Smith, 2017) aims to develop theoretical models to understand the textual and thematic qualities that made news stories, short fiction, and poetry in nineteenth-century newspapers and magazines go viral. This project proposes approaches to the Text Reuse Detection task as well as visualisation tools such as graphs to observe the reiteration of textual content in the corpus. This project has integrated the *Oceanic Exchange* project (Exchanges, 2017), which aims at determining how information flew across national and linguistic boundaries through computational analysis of 19th century newspapers.

The *Living with Machines* project (Ahnert et al., 2023) is led by a consortium of historians, data scientists, geographers, computational linguists, and curators. This project proposes to apply computational analysis to a corpus of 19th century to study how the first industrial revolution

impacted the British society. Several approaches to the Semantic Change Detection and Sense Disambiguation tasks relying on word embeddings and language models have been proposed to observe the evolution of the English language in the 19th century (Beelen et al., 2021; Pedrazzini & McGillivray, 2022). Zero-shot approaches to the Named Entity Recognition task as well as approaches to the Toponym Resolution task applied to historical documents have also been proposed (Ardanuy et al., 2020; De Toni et al., 2022). Moreover, this project proposes visualisation tools such as the MapReader tool (Hosseini et al., 2022), which allows to interact with a collection of over digitised 15,000 19th ordnance survey maps, or the PressPicker[14], which allows to visualise the lineage of newspapers.

Our project is similar to others in that we propose to to structure the textual content of documents with semantic annotations in order to improve current search interfaces. However, in this work we focus on extracting information related to people and places mentioned in the documents. We believe that extracting such information is important in order to fully understand the history of an entity such as a person. This goal can be assimilated to the Joint Extraction of Entities and Relation (JERE) task, which to our knowledge has not been extensively studied in the context of historical documents and for which dedicated resources are lacking. Thus, the methods and resources we propose in this thesis would make it possible to advance the JERE task applied to historical documents, and could encourage future work in this area. Furthermore, we propose two conceptual frameworks for the design of augmented search interfaces, based respectively on automatically generated maps and timelines, that would exploit the semantic annotations of documents and allow easy switching between close and distant reading approaches.

[14] https://tinyurl.com/4vdmfah8

Chapter 7

Converting documents to a machine-readable format

Table of contents

As explained in the previous chapter, many collections of historical documents are available online as scans of the original documents. In order to apply NLP methods to these documents, it is first necessary to convert these images into a machine-readable format. In this chapter, we present three tasks which allow this conversion: *Physical Layout Analysis, Logical Layout Analysis* and *Optical Character Recognition*, which we describe in Section 7.1, 7.2 and 7.3 respectively.

The *Physical Layout Analysis* task (PLA) aims at identifying the physical elements composing a document, such as blocks of texts, figures or tables, as well as the reading order of the document. The *Logical Layout Analysis* task (LLA) aims at determining the logical structure of a document, i.e. determining its organisation in logical elements such as title, header, paragraph, table, etc. Such logical elements can integrate one or more regions in the document that have been identified by PLA. Finally, the *Optical Character Recognition* (OCR) task aims at transcribing the textual content of the document.

7.1 Physical Layout Analysis

Physical Layout Analysis approaches can be divided into *bottom-up* approaches, *top-down* approaches and *hybrid* approaches.

Bottom-up approaches such as the Docstrum algorithm (O'Gorman, 1993) determine the physical structure of an image by grouping visual elements together. These approaches start at finest level of the image, usually at the pixel level, in order to group similar elements together. On the other hand, top-down approaches determine the physical structure of an image by dividing it into smaller elements. These approaches start with the complete image and divide it based on specific conditions, such as the presence of a white space (Antonacopoulos & Ritchings, 1995) or by distinguishing textual regions from non-textual regions (Bukhari et al., 2011). Finally, other works such as Barlas et al. (2014) propose hybrid approaches which combine bottom-up and top-down methods to determine the structure of a document.

Bottom-up approaches are the most commonly adopted approaches, as they are able to adapt to multiple layouts. However, these approaches are slower than top-down approaches. Top-down approaches are more adapted to regular layout and clean data, such as contemporary documents. Moreover, hybrid method are less frequently used than bottom-up or top-down approaches (Binmakhashen & Mahmoud, 2019). Machine and deep-learning approaches such as Shih and Chen (1996), Wei et al. (2013), and Wick and Puppe (2018) to this task have also been proposed, which require less post-processing steps than other methods. However, they require sufficient data to be trained, and usually suffer from the imbalanced nature of the datasets.

Several pre-processing steps are required to prepare the input image for the application of the methods described above. For instance, the *binarization* pre-processing step (Pratikakis et al., 2017) converts the image into binary colour pixels, i.e. black and white pixels, in order to affirm the boundaries of the visual elements of the image, whereas the *de-skewing* step (Shafii & Sid-Ahmed, 2015; Vasilopoulos & Kavallieratou, 2016) removes the potential angle of image.

The Physical Layout Analysis task has attracted a lot of interest in the recent years, as shown by the existence of multiple competitions such as ICDAR (Simistira et al., 2017) or ICFHR (Antonacopoulos et al., 2011). Multiple datasets of printed documents in multiple languages such as the PRIMA (Antonacopoulos et al., 2009) or the BCE-Arabic (Saad et al., 2016) datasets, as well as handwritten documents such as GW20 (Lavrenko et al., 2004) are available. The processing of historical documents in various languages has also earned attention, as shown by the existence of datasets such as the Saint Gall and Parzival datasets (Fischer et al., 2011), as well as the existence of dedicated competitions (Simistira et al., 2017).

Physical Layout Analysis approaches are usually evaluated according to element-counting based metrics. A common framework is the the Pixel-Level evaluation framework, which evaluates the

model based on the amount of matching pixel between the segmentation results and the ground-truth. Other evaluation frameworks such as the Region-Level evaluation framework (Vasilopoulos & Kavallieratou, 2016), as well as more customizable frameworks have been proposed (Antonacopoulos & Bridson, 2007).

7.2 Logical Layout Analysis

Existing Logical Layout Analysis systems make use of various methods that go from heuristic systems to more recent architectures using neural networks.

Some heuristic systems use grammars such as stochastic or attributed grammars, where the document is represented as a string of symbols, as in Namboodiri and Jain (2007). In their work, the grammar describes multiple production rules, each associated with a logical label. The string of symbols is then parsed by the grammar in order to extract logical labels. Other systems, such as LA-PDFText (Ramakrishnan et al., 2012) or DeLoS (Niyogi & Srihari, 1995), use rules that state the condition a physical block must meet to be given a logical label. For instance, the DeLoS system uses first-order predicates in order to infer the logical category of a physical block.

Recent approaches to the Logical Layout Analysis task rely on clustering methods (Riedl et al., 2019), machine-learning model such as Conditional Random Field (CRF) (Councill et al., 2008; Hébert et al., 2014; Luong et al., 2012), as well as neural-network architecture such as Long Short-Term Memory (LSTM) architectures (Akl et al., 2019; Prasad et al., 2018). These approaches cast the Logical Layout Analysis task as a sequence labelling or as a sequence-to-sequence task. These models are usually trained on either visual features, textual features or a combination of both (Rangoni et al., 2012; Zulfiqar et al., 2019). When relying on textual features, these models benefit from the use of static and contextual word-embeddings to encode the textual content of documents, as in Prasad et al. (2018) and Zulfiqar et al. (2019). Hybrid methods which combine rule-based and machine-learning approaches such as have been proposed. These hybrid methods usually combine heuristic rules with unsupervised-learning models such as k-means or Hierarchical Agglomerative Clustering (HAC), as in Gatos et al. (1999) and Klampfl and Kern (2013).

Logical Layout Analysis methods are usually evaluated in terms of Precision, Recall and F1 measure, although other metrics such as the accuracy metric or based on a cost function have been proposed (Mao et al., 2003). A few datasets of contemporary documents are available for physical and Logical Layout Analysis purposes, such as the Publaynet data set (Zhong et al., 2019) or the Medical Articles Record Groundtruth (MARG).

Some works have focused on the application of the Logical Layout Analysis task to historical documents. These works include rule-based approaches and machine-learning approaches (Gutehrlé & Atanassova, 2022; Hébert et al., 2014), as well as deep-learning approaches (Barman

et al., 2020). Among them, Barman et al. (2020) propose a system that goes beyond usual logical labels by labelling physical block as either Serial, Weather Forecast, Death Notice and Stock Exchange Table. To do so, their system combines visual and textual features using the word-embedding representation of each word and its coordinates on the page. Their results show that combining textual and visual features provide better results in most cases than using just one of them. Textual features are also more efficient to deal with the diachronic aspect of documents because they are more stable over time than visual features.

Several small data sets exist such as the DIVA-HISDB data set (Simistira et al., 2016) which contains 150 annotated pages of three different medieval manuscripts or the European Newspapers Project data set (Clausner et al., 2015) which contains 528 documents. Other data sets in non-European languages exist, such as the PHIBD data set (Nafchi et al., 2013), which contains images of 15 Persian historical and old manuscripts, and the HJ dataset (Shen et al., 2020), which contains 2,271 Japanese newspapers published in 1953, which was generated in a semi-automatic way.

7.3 Optical Character Recognition

The OCR process consists in a *text detection* step which aims at detecting blocks of texts in the document, and a *text transcription* step which transcribes the found textual contents (Subramani et al., 2021). These steps are usually performed sequentially, although recent works proposed to performed them in a joint manner (Feng et al., 2019; H. Li et al., 2017). Most recent approaches to the OCR task rely on deep-learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM). Recent works such as Transkribus (Kahle et al., 2017) or eScriptorium (Kiessling et al., 2019) have also focused on the Handwritten Text Recognition task, which aims at automatically transcribing handwritten documents. Similarly, these works rely on deep-learning architectures. Several formats such as ALTO[1], PAGE [2] or hOCR[3] have been conceived to store the output of OCR methods as well as the output of Physical Layout Analysis methods.

The quality of the transcriptions obtained by the OCR process depend on various factors, such as the quality of the original documents and its numeric copy, the layout of the document, the typography used, as well as the difference between the language of the document and the language for which the OCR model has been trained. The incorrect transcriptions and errors produced by the OCR process have been shown to have an impact on results of search queries, as well as on the quality of linguistic analysis such as part-of-speech tagging, dependency analysis or NLP processes

[1] https://www.loc.gov/standards/alto/
[2] https://github.com/PRImA-Research-Lab/PAGE-XML
[3] http://kba.github.io/hocr-spec/1.2/

such as Named Entity Recognition, Topic Modelling or Event Detection (Boros, Nguyen, et al., 2022; Linhares Pontes et al., 2019a; Mutuvi et al., 2018). Thus, it is usually necessary to post-process the output of the OCR process in order to clean the text from incorrect transcriptions and artefacts (T. T. H. Nguyen et al., 2021a).

The OCR post-processing task consists in an *error detection* step which aims to identify incorrect transcription of a word, and a *error correction* step, which aims at proposing a correction of an incorrectly transcribed word detected in the previous step. Moreover, errors produced by OCR processes can be divided into two types: either *non-word errors*, i.e. substitution or omission of characters that result in non existing words, or real-word errors, i.e. substitution or omission of characters that result in existing words, which are different from the real ones contained in the document. T. T. H. Nguyen et al. (2021a) propose *manual* and *(semi)-automatic* OCR post-processing methods.

Because of the large amount of textual data, the manual corrections of incorrect transcriptions are usually realised in crowd-sourcing manner, through platforms such as *Trove* (Holley, 2008), *Kokos* (Clematide et al., 2016) or *DigitalKoot* (Chrons & Sundell, 2011). Other crowd-sourcing methods have been proposed, such as repurposing the Word-Wide web security measure CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart), to have users write the transcription of a word (Von Ahn et al., 2008). The crowdsourcing approaches are efficient and cost-effective. However, they require to find a sufficient amount of volunteers. Moreover, experts must validate the proposed corrections to ensure the quality of the data.

Automatic methods consist in automatically correcting the errors and artefacts produced by the OCR process. These methods can be divided into *isolated-word approaches* and *context-depend approaches*. In *isolated-word approaches*, the correction of a word is only based on the word itself, and does not take its context into consideration. Such approaches are easier to implement and can correct non-word errors. However, they are unable to correction real-word errors. In *context-dependent approaches*, the correction of a word is based on the word itself, as well as its context. Such approaches are more complex than *isolated-word approaches*, but are able to correct both *non-word* and *real-word* errors.

A first method consists in obtaining multiple transcription of the same text, either by using the same OCR engine with different parameters, or by using multiple OCR engines. The best transcription of a word is then selected by a voting policy as in Reul et al. (2018) or by classifier such as Conditional Random Field (CRF) or a Long Short-Term Memory (LSTM) as in Al Azawi et al. (2015) and Lund et al. (2013).

Another method consists in using dictionaries to find a candidate correction that is most similar to an incorrectly transcribed word. Works such as Estrella and Paliza (2014) rely on the Damerau-Levenshtein (Damerau, 1964; Taghva et al., 1998) edit distance to measure the similarity between a

word and a candidate correction, whereas other works such as Hammarström et al. (2017) and Jean-Caurant et al. (2017) rely on their word embedding representation. Some works such as Fonseca Cacho and Taghva (2020) propose to train classifiers such as SVM on features extracted from the word as well as features such as bigram or trigram frequency extracted from its context. The use of local and contextual features allow to improve the selection of candidate corrections, which can be selected from a lexicon (Khirbat, 2017) or generated by multiple-character edits (T.-T.-H. Nguyen et al., 2018). Although such lexical approaches are easy to implement, they are only able to correct non-word errors. Moreover, they are heavily dependent on the coverage of the dictionaries upon which they rely.

Other works such as Borovikov et al. (2004), Llobet et al. (2010), Perez-Cortes et al. (2000), and Richter et al. (2018) rely on probabilistic models such as character language models, Weighted Finite State Machine (WFSM) or Hidden Markov Model (HMM) to suggest a correction of a word. Given an incorrect word, these model generate the most probable sequence of characters that is similar to the incorrect word. Similarly, Evershed and Fitch (2014), Hládek et al. (2017), and Ringlstetter et al. (2007) use statistical language models such as word n-gram language model whereas Amrhein and Clematide (2018), Bengio et al. (2000), and D'hondt et al. (2016) use neural-network language models trained on implicit features such as word embeddings or character embeddings to suggest the correction of a word based on its context.

Recent approaches cast the OCR post-processing task as a sequence-to-sequence task, which aims at converting a sequence of elements into another sequence of elements. More specifically, many works consider the OCR post-processing task as a translation task, which aims at translating an incorrect transcription into a correct transcription. Works such as Afli et al. (2016) and Schulz and Kuhn (2017) rely on statistical machine translation models, whereas works such as Rigaud et al. (2019) and Schaefer and Neudecker (2020) rely on neural-network architectures such as LSTM or BERT. Such methods have especially been applied in the *impresso* and *NewsEye* projects (Amrhein & Clematide, 2018; T. T. H. Nguyen et al., 2020). Although efficient, methods based on statistical and neural machine translation models require a large amount of data to be trained.

Several frameworks for statistical and neural machine translations have been developed, such as GIZA++ (Och & Ney, 2003), MosesSMT (Koehn et al., 2007), OpenNMT (G. Klein et al., 2018) or Nematus (Sennrich et al., 2017). Moreover, several methods and tools such as OCRSpell (Bassil & Alwani, 2012) or PoCoTo (Vobl et al., 2014) have been proposed to post-process the OCR transcriptions in a semi-automatic way, where models are used to assist the user in detecting and suggesting corrections to the user.

OCR post-processing approaches are usually evaluated on the error detection task in terms of Precision, Recall and F1 scores. On the other hand, the evaluation of these approaches on the error correction task is usually based on metrics such the error rate, accuracy, or BLEU (Papineni et al.,

2002). The OCR task as well as the OCR post-processing tasks have received an important interest in recent years, as shown by the multiple available datasets such as Text+Berg (Göhring & Volk, 2011), GT4HistOCR (Springmann et al., 2018) or RETAS (Yalniz & Manmatha, 2011), and the dedicated competitions such as the ICDAR2017 and ICDAR2019 (Chiron et al., 2017; Rigaud et al., 2019). These datasets and competitions focus on contemporary documents, as well as historical documents.

Chapter 8

Information Extraction

Table of contents

The Information Extraction task aims at turning unstructured data such as texts into structured data (Jurafsky & Martin, 2008). These structured data can then be exploited for downstream tasks such as information retrieval, database population, etc. Information Extraction imply many tasks such as Event Extraction, Temporal Expression Detection or Sentiment Analysis. In this chapter, we choose to focus on two specific Information Extraction tasks: the *Named Entity Recognition* and the *Relation Extraction* tasks, which we present in Section 8.1 and 8.2 respectively.

Named Entity Recognition (NER) consists in identifying mentions of entity names such as persons, locations, organisations or dates, whereas Relation Extraction (RE) aims at determining the relations between entities. These tasks are usually applied sequentially, as the Relation Extraction task aims at determining the relation between entities detected by the Named Entity Recognition task. However, recent works have focused on the Joint Extraction of Relations and Entities (JERE) task, which aims at extracting entity mentions and relations in a join manner from sentences (Pawar, Palshikar, & Bhattacharyya, 2017). We provide a more detailed description of the JERE task in Section 8.3.

8.1 Named Entity Recognition

8.1.1 Approaches

Approaches to the NER task can be divided into three main categories: *rule-based*, *feature-engineered* and *feature-inferred*. Rule-based approaches such as Appelt et al. (1993), Farmakiotou et al. (2000), J.-H. Kim and Woodland (2000), and Sekine and Nobata (2004) rely on lexicon of known entities as well as on handwritten rules to identify mentions of entities in texts. Thus, these approaches can achieve very high Precision scores, especially when they are configured for specific domains. However, they usually achieve lower Recall scores. Moreover, building the lexicon and the rule sets to detect entity mentions require linguistic and domain expertise. Thus, rule-based systems are not easily adapted to other domains.

Feature-engineered systems include traditional machine-learning based systems, which consist in learning a statistical model from sets of features. The models are trained on features extracted from the documents at different levels such as word-level (capitalization, part-of-speech, dependency role...), document-level (word frequency, coocurrences, syntactic context...) or static word embeddings such word2vec, GloVe, fastText. The task is usually cast as either a a multi-class classification task or a sequence-labelling task. The most common algorithms used for the multi-class classification task are Decision Trees (Sekine, 1998) and Support Vector Machine (SVM) (Asahara & Matsumoto, 2003), whereas the most common algorithms used for the sequence-labelling task are Maximum Entropy Models (ME) (Borthwick et al., 1998), Hidden Markov Models (HMM) (Bikel et al., 1998) and Conditional Random Field (CRF) (McCallum & Li, 2003; Ritter et al., 2011; Rocktäschel et al., 2012). As for other task, these approaches required large annotated dataset of entities to train the classifiers.

Bootstrapping and distant supervision methods have been proposed to tackle the issue of required large annotated dataset. Bootstrapping approaches such as He and Sun (2017), Liao and Veeramachaneni (2009), Mishra and Diesner (2016), and Thenmalar et al. (2015) use entity seeds or high Precision extraction patterns to find a first set of entities from an unlabelled dataset. The system then extracts the context of these entities in order to build new extraction patterns. This process is repeated until no new entities can be extracted or another condition is met. Distant supervision methods such as Hedderich et al. (2021), Liang et al. (2020), and X. Wang et al. (2020, 2021) rely on an external lexicon or knowledge base to add weak annotations to a dataset. These annotations are then exploited to train a classifier or collect extraction patterns and build a rule set. Both methods are particularly useful to rapidly build a NER system dedicated to a specific domain or to a low-resource language. However, bootstrapping approaches tend to extract over specific patterns and thus have low Recall scores, whereas the distant supervision may produce incorrect annotations.

Works such as Ji et al. (2019), Rocktäschel et al. (2012), Shaalan and Oudah (2014), and Srivastava et al. (2011) propose hybrid systems which combine rule-based and machine-learning based systems. Such systems usually use a lexicon or a set of rules to extract precise entities whereas statistical models such as CRF extract more complex entities. Rules can also be applied to control and eventually correct the output of the statistical model. Hybrid models have been shown to outperform the performance of independent rule-based or statistical models. Furthermore, they are easier to adapt than rule-based models.

Finally, works such as Agerri and Rigau (2016), Collins and Singer (1999), Etzioni et al. (2005), Nadeau et al. (2006), and S. Zhang and Elhadad (2013) propose unsupervised approaches to the NER task. Entity mentions are extracted from an unlabelled dataset by applying extraction patterns. These mentions are then clustered according to their similarity, which is determined through lexical, syntactic and semantic features. Unsupervised approaches are extensible and can adapt to new domains. Moreover, they can help bootstrap an annotated dataset upon which supervised method can be trained. However, the performance of unsupervised approaches is usually much lower than that of supervised methods. Moreover, a post-processing step is necessary to label the entity clusters.

Recent approaches to the NER task rely on deep learning architectures and infer features from raw data (J. Li et al., 2020; Yadav & Bethard, 2019). Instead of being represented by engineered features, documents are represented by either static or contextual word embeddings (Sheikhshab et al., 2018; Zhai et al., 2019) at word-level (Collobert et al., 2011; Huang et al., 2015), character-level (Y. Kim et al., 2016; Kuru et al., 2016) or a combination of both (Lample et al., 2016; Ma & Hovy, 2016). The task is generally cast as sequence labelling task. Most approaches rely on the Recurrent Neural Network architecture such as Long-Short Term (LSTM) or Bidirectional LSTM (BLSTM) (Chiu & Nichols, 2016; Limsopatham & Collier, 2016; D. Zeng et al., 2017) or on the Transformer architecture (Hakala & Pyysalo, 2019; Souza et al., 2019). Deep-learning approaches have been shown to largely outperform rule-based and feature-engineered models while requiring less time to prepare the data. However, they require very large datasets in order to be trained. Pre-trained Transformer models such as BERT can be fine-tuned on the NER classification task, thus limiting the amount of necessary training data. Fine-tuning pre-trained language models on the NER task is particularly useful when training data is scarce, and has been shown to outperform other deep learning approaches (H. Yan et al., 2019). Similarly, several recent works such as Bogdanov et al. (2024), Jung et al. (2024), S. Wang et al. (2023), and Ye et al. (2024) have proposed to leverage the reasoning capabilities of pretrained Large Language Models (LLM) to perform the Named Entity Recognition. Most of these works rely on prompt engineering and few-shot learning approaches to circumvent the issue of lacking training resources, which have been shown to reach satisfactory performance.

The performances of NER systems built on contemporary data is highly degraded when applied to historical documents, because of the domain shift, the noisy input, as well as the language dynamic and semantic shifts (M. Ehrmann, Colavizza, Rochat, & Kaplan, 2016). In order to deal with noisy input and the spelling variation issues, rule-based systems such as Borin et al. (2007), Broux and Depauw (2015), Crane and Jones (2006), Grover et al. (2008), Milanova et al. (2019), and Platas et al. (2020) must be flexible, for instance by applying fuzzy matching methods instead of exact matching methods. Moreover, normalising the language of documents so that it corresponds to the current state of the language has been shown to improve the performance (Piotrowski, 2012).

Machine-learning based approaches relying on Ensemble methods (Packer et al., 2010; Won et al., 2018), Maximum Entropy (Nissim et al., 2004), or CRF (Aguilar et al., 2016; Neudecker et al., 2014; Ritze et al., 2014) have also been proposed. Some of these systems have been trained on contemporary datasets such as the COnLL dataset, whereas others have been trained on custom historical data. More recently, deep-learning approaches based on the BiLSTM architecture (Hubková, 2019; Riedl & Padó, 2018; Rodrigues Alves et al., 2018) or on the Transformer architecture (Boros, Pontes, et al., 2020; Labusch et al., 2019; Rouhou et al., 2022) have been proposed. These models can reach very high results, even when pre-trained on contemporary data and fine-tuned on small historical datasets. The textual contents of documents are encoded with either static embeddings (Rodrigues Alves et al., 2018; Sprugnoli et al., 2018) or contextual embeddings (Boros, Pontes, et al., 2020; Kew et al., 2019; Kristanti & Romary, 2020; Labusch et al., 2019). Using pre-trained in-domain embeddings has been shown to improve results. Moreover, using character embeddings and sub-word embeddings allows to deal with the noisy input produced by the OCR process and with the spelling variations issues (M. Ehrmann et al., 2021b).

8.1.2 Datasets and evaluation schemes

There exist multiple datasets for the NER task covering multiple domains such as newswire articles, Wikipedia articles, web pages (J. Li et al., 2020), social networks (Baldwin et al., 2015) in many languages such as English, French (Sang & De Meulder, 2003), Arabic (Shaalan, 2014), German (Benikova et al., 2014), Slavic languages (Piskorski et al., 2017) or Asian languages (Singh, 2008). Most available datasets provide examples of entities of the types *Person*, *Location* and *Organisation*, but are not limited to these types only. There are however domain specific datasets such as biomedical datasets (Collier & Kim, 2004; Segura-Bedmar, Martínez Fernández, & Herrero Zazo, 2013) or biology datasets (Bossy et al., 2013; Deléger et al., 2016) with entity types specific to their domain (e.g. protein, DNA, drug types...), as well as datasets such as Krallinger et al. (2015) and Ohta et al. (2002) which annotate the mentions of nested and discontinuous entities.

Historical NER datasets have been published in the recent years. These datasets contains doc-

uments published between the 18th and the 20th century in multiple languages such as French, English, German, Czech, Finnish (M. Ehrmann, Romanello, Flückiger, & Clematide, 2020; Hamdi et al., 2021; Hubková, 2019; Neudecker, 2016; Ritze et al., 2014; Ruokolainen & Kettunen, 2018). These datasets are usually monolingual and contain newspapers and literary documents. In general, they contain annotations of *Person*, *Location* and *Organisation*, although some datasets cover other genres such as medical documents, and cover older periods and languages such as Coptic (Schroeder & Zeldes, 2016) or Latin (Erdmann et al., 2016).

NER systems are usually evaluated in terms of Precision, Recall and F1-score. The evaluation of NER approaches focus on the ability of the systems to extract the correct spans of texts and to classify these spans with the right types. However in certain contexts, extracting the right boundaries might be less important than classifying the correct types, and vice versa. Thus, multiple evaluation schemes have been proposed in order to evaluate systems under different evaluation settings, such as the MUC, COnLL, SemEval and ACES evaluation schemes.

The MUC evaluation scheme (Grishman & Sundheim, 1996) propose to evaluate the system under the *type* and *text* axes. The *type* evaluation setting, where only the ability of the system to find the correct type is evaluated, and the *text* evaluation setting, where only the ability of the system to extract the correct boundaries is evaluated. Both axes are evaluated in terms of Precision, Recall and F1 scores. The COnLL scheme (Sang & De Meulder, 2003; Tjong Kim Sang, 2002) proposes the exact-match evaluation setting, where both the entity span and type must be correct. The system is evaluated in terms of Precision, Recall and micro F measure.

The SemEval scheme (Segura-Bedmar, Martínez, & Herrero-Zazo, 2013) evaluates NER systems under the *strict, exact, partial* and *type* evaluation settings: the *strict* setting evaluates the correctness of entity spans and types. The *exact* setting only evaluates the predicted boundaries, regardless of the predicted type. The *partial* settings evaluates if the predicted boundaries have some overlap with the true span, regardless of the predicted type. Finally, the *type* setting only evaluates the predicted type regardless of the predicted boundaries. Each setting evaluates the system in terms of Precision, Recall and micro F measure. Finally, the ACE scheme (Doddington et al., 2004) proposes a more complex evaluation scheme where the user must define weights for each entity type, so that each type contributes differently to the final score. The final score is 100% minus the percentage of each error (type error, span error...).

8.2 Relation Extraction

8.2.1 Approaches

Relation Extraction approaches are usually divided into rule-based approaches and supervised approaches. Rule-based approaches such as Arnold and Rahm (2014), Hearst (1992), and Nebhi (2013) detect and classify the relations between entities by applying manually or semi-automatically crafted rules. These systems usually have high Precision scores but low Recall scores. Moreover, the conception of these rules is a time-consuming task which requires expertise.

Bootstrapping methods (Brin, 1998) are a potential solution to conceive such rules more quickly. Given a set of initial seeds of entity pairs involved in a given relation, a first set of patterns is extracted from an unlabelled dataset. These new patterns are then used to find new pairs of entities. This process is repeated until no new patterns of entity pairs are found. Works such as Agichtein and Gravano (2000) and Phi et al. (2018) evaluate the extracted patterns at each iteration and add constraints, such as entity types, in order to select the best possible patterns. The bootstrapping method allows to rapidly extract patterns expressing a relation if the initial seeds are of good quality. However, the precision of the extracted patterns tends to be low. Moreover, the collected patterns tend to suffer from semantic drift issue, i.e. the relation they express may actually differ from the relation expressed by the initial seed relation.

Supervised approaches cast the Relation Extraction task as a multiclass classification task. These approaches are generally divided into kernel-based and feature-based methods. Kernel-based methods (R. Bunescu & Mooney, 2005; R. C. Bunescu & Mooney, 2005; Collins & Duffy, 2001; Huiyu Sun, 2022) use a kernel function to compute the similarity between samples and classify them with a Support Vector Machine (SVM). Feature-based methods categorise a relation based on features such as lexical, syntactic or semantic features as well as static or contextual word embeddings. Works such as Chan and Roth (2011), Kambhatla (2004), D. P. Nguyen et al. (2007), and Rink and Harabagiu (2010) rely on classifiers such as SVM or Maximum Entropy, whereas more recent works rely on deep-learning architectures such as LSTM, Graph Neural Networks or Transformer-based models (Han et al., 2018a; H. Wang et al., 2022; Y. Zhang et al., 2018). As for the Named Entity Recognition task, recent works such as X. Li et al. (2024), B. Zhang and Soh (2024), and H. Zhou et al. (2024) have also proposed to leverage the reasoning capabilities of pretrained Large Language Model to perform the Relation Extraction task, usually by applying a few-shot or a prompt engineering approaches.

Supervised approaches require large annotated datasets of relations, which are costly and time-consuming to produce. A potential solution to this issue is the distant supervision method proposed by Mintz et al. (2009). This method assumes that if two entities participate in a relation in a knowledge base such as Wikidata or FreeBase, then at least one sentence which mentions both

entities in a document expresses that relation. Thus, this method allows to quickly build a dataset of weakly annotated samples, which can be used to train classifiers. However, the same entity pair from a knowledge base may be involved in various relations. Thus, there is a high risk to incorrectly annotate samples with incorrect relations.

Most approaches to the Relation Extraction task focus on extracting intra-sentences relations, i.e. extracting relations between entities in the same sentence. However, a relation between entities may be expressed at the document level, through multiple sentences. Recent works such as Christopoulou et al. (2018), Quirk and Poon (2017), Swampillai and Stevenson (2011), W. Zeng et al. (2017), and W. Zhou et al. (2020) focus on document-level Relation Extraction, and aim at extracting intra-sentences and inter-sentences relations, i.e. relations between entities in different sentences. Document-level Relation Extraction is however a more difficult task than sentence-level Relation Extraction since the same entity pair may participate in multiple semantic relations. Moreover, the contextual information needed to classify the relation is more difficult to find at the document level than at the sentence level. Document-level Relation Extraction also requires to pre-process the documents with more steps such as co-reference resolution or discourse analysis (Han et al., 2020).

The task of Open Information Extraction (OIE) is a similar task to Relation Extraction which aims at extracting all possible facts from documents in an unsupervised manner (Yates et al., 2007). OIE systems are domain-independent, and must be able to extract facts regardless of the type of document. The extracted facts are represented as `(arg1;relation;arg2)` triples. OIE systems are usually rule-based, self-supervised or unsupervised. Rule-based methods as in Akbik and Löser (2012), Fader et al. (2011), and White et al. (2016) extract facts by applying hand-crafted extraction patterns based on linguistic features such as part-of-speech or dependency trees. Self-supervised methods train a classifier on a set of patterns that have been automatically extracted from an unlabelled dataset, as in Bhutani et al. (2016), Del Corro and Gemulla (2013), Mausam et al. (2012), Wu and Weld (2010), Yahya et al. (2014), and Yates et al. (2007). Most recent works rely on neural network architectures (S. Zhou et al., 2022) and propose to cast the OIE task as a sequence-labelling task (Stanovsky et al., 2018) or a sequence-generation task (Cui et al., 2018).

OIE methods allow to extract vast amount of information from documents. They are usually scalable and can process large amount of data such as the World Wide Web. However, these approaches tend to extract facts that are either too general or too specific, thus limiting their usefulness (Pawar, Palshikar, & Bhattacharyya, 2017). Moreover, most methods are language-dependent: most available systems such as (Fader et al., 2011) are tuned to English. Only a few works such as (Zhila & Gelbukh, 2014) are applied to other languages, whereas works such as (White et al., 2016) are language-agnostic or multilingual (Ro et al., 2020).

Unsupervised relation extraction, sometimes called relation discovery (Hasegawa et al., 2004),

is another form of OIE, where clustering methods identify sets of patterns expressing the same concept from an unlabelled dataset. First, entity pairs and their contexts are extracted and transformed into features, which are usually morphological, lexical and semantic, before being clustered by similarity (Alfonseca et al., 2012; Hasegawa et al., 2004; D. Lin & Pantel, 2001; Y. Yan et al., 2009; L. Yao et al., 2011, 2012). Some works add constraints such as entity types constraint (Hasegawa et al., 2004; L. Yao et al., 2011) or external knowledge (Lopez de Lacalle & Lapata, 2013) in order to orient the clustering step. Similarly, recent works (Ali et al., 2021; Elsahar et al., 2017) rely on word embeddings or language models such as BERT (Devlin et al., 2019) to encode sentences into dense-vectors, then cluster them by similarity. Since they do not rely on labelled datasets, unsupervised relation classifiers are extensible and are able to identify new relations from the datasets. However, they are often less precise than supervised methods. Moreover, a further step to label the clusters is necessary.

To our knowledge, only a few works have focused on the task of Relation Extraction applied to historical documents. Due to the lack of available annotated resources to train models, some of these works rely on rule-based approaches, pre-existing NLP tools and pre-existing resources. For instance, E. Klein et al. (2014) detect the relation between entities of type *commodities* and *locations* occurring in the same sentence based on prepositions such as *from, to, etc.* De Paiva et al. (2014) apply NLP tools such as FreeLing as well as resources such as lexicon and WordNet to add semantic annotation to an extant dictionary of historical biographies written in Portuguese and originally published in 1984. These annotations aim at assisting historians and researchers exploring this document. Similarly, (Quaresma & Finatto, 2020) rely on the FreeLing NLP tool to perform tasks such as POS tagging, dependency parsing, Semantic Role Labelling, SVO triple extraction on a book written in Spanish and published in 1735. These annotations aim at extracting information related to entities and events from this document, in order to populate an ontology which follows the Simple Event Model. Other works rely on machine and deep-learning approaches, such as Efremova et al. (2015) where several methods such as SVM and Hidden Markov Model are compared to extract family relationship extraction from historical documents written in Dutch. Similarly, Toledo et al. (2019) propose to integrate a CNN with a BLSTM-based language model to annotate handwritten historical documents with semantic annotations. Unlike other approaches which require first to obtain the transcription of the textual content of documents, this approach propose to annotate the document directly from the scan.

8.2.2 Dataset and evaluation schemes

Many datasets have been created in the recent years to help conceive and evaluate these approach. Most available datasets such as CoNLL2004 (Roth & Yih, 2004), ACE (Doddington et al., 2004),

Sem-Eval (Girju et al., 2007; Hendrickx et al., 2010), the New York Times (NYT) dataset (Riedel et al., 2010) or TACRED (Y. Zhang et al., 2017) have been created on news articles and web pages. Others are domain-specific, such as the ScienceIE (Augenstein et al., 2017), SciERC (Luan et al., 2018) or SemEval2018 (Gábor et al., 2018) datasets which were made on scientific documents. Others have been created on Wikipedia such as the FewRel (Han et al., 2018b), mLAMA (Kassner et al., 2021) or SMiLER (Seganti et al., 2021). Most of these datasets are monolingual and contain examples in English. There are however some multilingual datasets such as mLAMA (Kassner et al., 2021), or SMiLER (Seganti et al., 2021). Moreover, most datasets focus on Relation Extraction at the sentence level. There are however a few datasets such as DocRED (Y. Yao et al., 2019) and DWIE (Zaporojets et al., 2021) which also focus on RE at the document level.

Relation Extraction systems as well as OIE systems are generally evaluated in terms of Precision, Recall and F1 scores. OIE and unsupervised methods are evaluated on the same datasets as Relation Extraction by comparing the discovered relations and facts with those from the datasets. Since OIE systems must be domain-independent, building a dataset to evaluate them is a difficult task (Niklaus et al., 2018). Thus, most work evaluate their methods by examining a small sample of their results, although recent works by Schneider et al. (2017) and Stanovsky and Dagan (2016) have proposed benchmarks to evaluate OIE system.

8.3 Joint Extraction of Relations and Entities

Some approaches such as Pawar, Bhattacharya, and Palshikar (2017) and Roth and Yih (2004) to the Joint Extraction of Relations and Entities task propose to train local classifiers to separately detect mentions of entities and relations in the sentence, before combining the output of the classifiers. This output is controlled through domain constraints such as rules that can be expressed through approaches such as Integer Linear Programming or Markov Logic Networks.

Other works such as Kate and Mooney (2010) propose to represent all possible entities and relations in a sentence as a tree-like graph called *card-pyramid graph*. This graph is parsed by an algorithm which combines production rules similar to Context Free Grammar (CFG) and classifiers to detect the mentions of entities and relations from the card-pyramid graph.

Other works propose to cast the JERE task as a structured prediction problem, where the aim is to produce a structure describing the entity types and their relations from the original sentence. For instance, Q. Li and Ji (2014) propose an incremental joint framework which outputs a graph where nodes represent entity mentions and edges represent the relations between entities. Similarly, Miwa and Sasaki (2014) cast the JERE task as a table-filling problem, and propose to generate symmetric matrix, where rows and columns are tokens and where the cell contains the entity types and relations.

More recent works propose end-to-end approaches, where a single neural network learns to categorise entities and relations with a single model. These works usually rely on the LSTM architecture (Bekoulis et al., 2018; B. Yu et al., 2020; Zheng et al., 2017), as well as the Graph Neural Networks (Fu et al., 2019; Sun et al., 2019) or the Transformer architectures (Eberts & Ulges, 2019; Shang et al., 2022). These model learns to identify the mentions of entities and relations from the word embeddings representation of words, although some works such as Miwa and Bansal (2016) also incorporate other linguistic features such as part-of-speech tags or dependency roles.

Joint Extraction of Relations and Entities methods are usually evaluated in terms of Precision, Recall an F1 scores on the same datasets as the Relation Extraction methods such as the ACE 2004 or ACE 2005 datasets. As reported by Pawar et al. (2021), JERE approaches are able to achieve F1 scores around 83 % on the Named Entity Recognition task. However, they usually achieve lower F1 scores up to 57 % on the Joint categorisation of entities and relations, suggesting the difficulty of the task.

Part II

Pre-processing historical documents

This part is dedicated to the description of our dataset and the pre-processing steps that have been necessary to prepare the data for the subsequent Joint Extraction of Entities and Relations task. We introduce the EMONTAL corpus, which consists of printed periodical documents published in the 19th and 20th centuries in the Bourgogne and Franche-Comté regions of France. We choose to focus on periodicals such as newspapers, given their importance as primary sources reflecting local and international events. The textual content of documents in our corpus is stored in the XML ALTO format, which provides the OCR transcription of the document, its reading order and its physical structure.

Because of the OCR process, the transcriptions of the textual content of documents may contain artefacts and errors, which may have an impact on the quality of any NLP method applied to the content of the document, such as Named Entity Recognition or Relation Extraction. Thus, cleaning artefacts produced by the OCR process is a necessary step, in order to minimise their impact on the application of NLP methods. Moreover, the XML ALTO format does not provide the logical structure of documents, making the documents in our corpus unsuited for the application of NLP methods. Thus, we developed our own methodology for determining the logical structure of the documents in our corpus, which is a necessary step to obtain the structured textual data.

This part is organised as follows: we describe the EMONTAL corpus in Chapter 9. In Chapter 10, we present the methods we apply to clean the textual content of the documents in our corpus. Finally in Chapter 11 and Chapter 12, we present and evaluate our approach to determine the logical structure of the documents in the EMONTAL corpus.

Chapter 9

Corpus of historical documents

Table of contents

In this chapter, we present the EMONTAL corpus, a dataset of printed periodical documents published in the 19th and 20th century in the *Bourgogne* and *Franche-Comté* regions in France. We choose to work on periodical documents such as newspapers, since they are primary source of information, and reflect events that occurred at local and international levels (Tibbo, 2007). Thus, these documents are adapted to our study, which focuses on the extraction of information related to people and places from historical documents.

We choose to limit the scope of this corpus to printed documents, and exclude manuscript documents, to avoid the issues related to the transcriptions of manuscripts with OCR methods. Moreover, we choose to limit the scope of this corpus to documents published in the 19th and 20th centuries, to ensure that the language in which the documents are written is as close as possible to the modern state of the French language.

The rest of this chapter is structured as follows: In Section 9.1, we provide a statistical analysis of the corpus, as well as a description of the format in which the documents are stored. Finally, we propose a thematic analysis of the documents in the EMONTAL corpus in Section 9.2.

9.1 Description of the EMONTAL corpus

We have collected the documents in the EMONTAL corpus from the *Fond régional: Franche-Comté* (Regional fund: Franche-Comté) and the *Fond régional: Bourgogne* (Regional fund: Bour-

gogne) collections[1] on Gallica[2], the digital archive of the *Bibliothèque Nationale de France* (National Library of France, BnF) with the provided APIs. We have only collected documents from these collections that have been processed with OCR. Figures 9.1, 9.2, 9.3, 9.4, 9.5 show examples of documents collected from both funds, showing the diversity of layouts and OCR quality in the EMONTAL corpus. The complete list of collections in our corpus is shown in Annexe 19.4.

Figure 9.6 shows the distribution of documents by decades in the *Franche-Comté* and *Bourgogne* funds in the EMONTAL corpus. The documents we have collected in the *Franche-Comté* fund have been published between 1840 and 1940, although some of them have also been published in the 1990s. Most documents have been published in between 1900 and 1910, with a peak of 1,028 publication in the 1910s. The documents we have collected in the *Bourgogne* fund have been published between 1810 and 1960. Most documents have been published between 1890 and 1930, with a peak of 1,236 publication in the 1910s and 1920s. The gap of available documents published before 1900 can be explained by a lack of available OCR transcriptions on the Gallica platform. The drop of publications in both funds in the 1940s can be explained by the events of WWII, which imposed restrictions on papers. Moreover, documents published after the 1940s may be lacking because they may not yet be in the public domain.

We store the collected documents in the XML format. The main tags of each document are the following:

- `oai`: metadata of the document. These metadata follow the Dublin Core format
- `image_url`: url to the scanned document
- `toc`: table of contents of the document
- `pagination`: pagination of the document
- `num_pages`: number of pages in the document
- `ocr`: OCR output for the document. This output follows the XML ALTO format

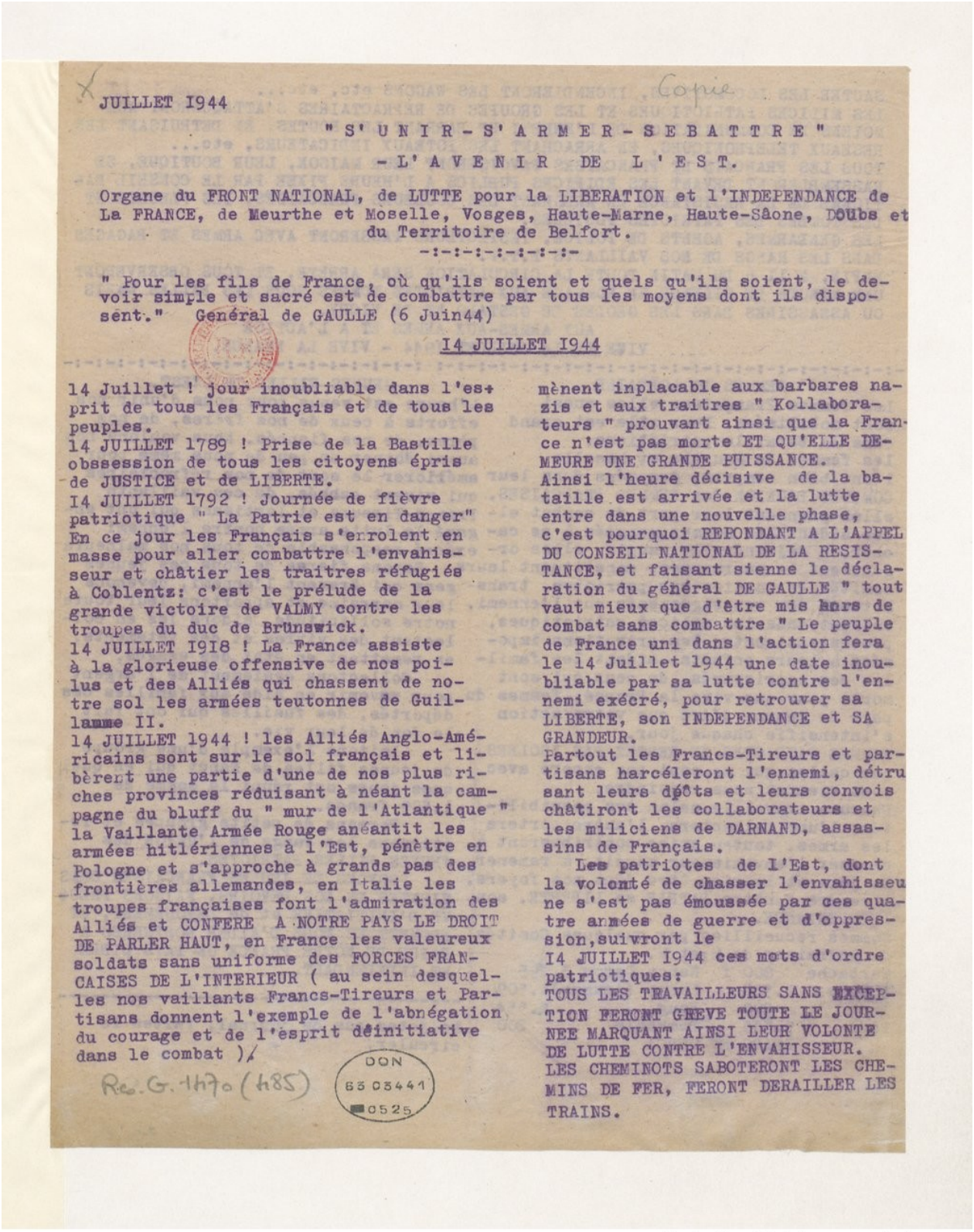
The `oai` tag contains metadata about the document itself, as well as metadata related to the OCR process and the storage on Gallica. Figure 9.7 shows an example of the content of the `oai` tag.

The main elements of the `oai` tag are the following:

- `dc:identifier`: the url to the document on the Gallica platform

[1] https://gallica.bnf.fr/html/und/france/bourgogne-franche-comte
[2] https://gallica.bnf.fr

Figure 9.1: Excerpt of the first page of the second issue of the communist newspaper *Le Semeur* (The Sower) published on the 23rd of April 1932



JUILLET 1944

Copie

" S' U N I R - S' A R M E R - S E B A T T R E "

- L' A V E N I R DE L ' E S T.

Organe du FRONT NATIONAL, de LUTTE pour la LIBERATION et l'INDEPENDANCE de La FRANCE, de Meurthe et Moselle, Vosges, Haute-Marne, Haute-Saône, Doubs et du Territoire de Belfort.

-:-:-:-:-:-:-:-

" Pour les fils de France, où qu'ils soient et quels qu'ils soient, le devoir simple et sacré est de combattre par tous les moyens dont ils disposent." Général de GAULLE (6 Juin44)

I4 JUILLET I944

14 Juillet ! jour inoubliable dans l'esprit de tous les Français et de tous les peuples.
14 JUILLET 1789 ! Prise de la Bastille obsession de tous les citoyens épris de JUSTICE et de LIBERTE.
I4 JUILLET 1792 ! Journée de fièvre patriotique " La Patrie est en danger" En ce jour les Français s'enrolent en masse pour aller combattre l'envahisseur et châtier les traitres réfugiés à Coblentz: c'est le prélude de la grande victoire de VALMY contre les troupes du duc de Brunswick.
14 JUILLET 1918 ! La France assiste à la glorieuse offensive de nos poilus et des Alliés qui chassent de notre sol les armées teutonnes de Guillaume II.
14 JUILLET 1944 ! les Alliés Anglo-Américains sont sur le sol français et libèrent une partie d'une de nos plus riches provinces réduisant à néant la campagne du bluff du " mur de l'Atlantique " la Vaillante Armée Rouge anéantit les armées hitlériennes à l'Est, pénètre en Pologne et s'approche à grands pas des frontières allemandes, en Italie les troupes françaises font l'admiration des Alliés et CONFERE A NOTRE PAYS LE DROIT DE PARLER HAUT, en France les valeureux soldats sans uniforme des FORCES FRANCAISES DE L'INTERIEUR (au sein desquelles nos vaillants Francs-Tireurs et Partisans donnent l'exemple de l'abnégation du courage et de l'esprit d'initiative dans le combat),

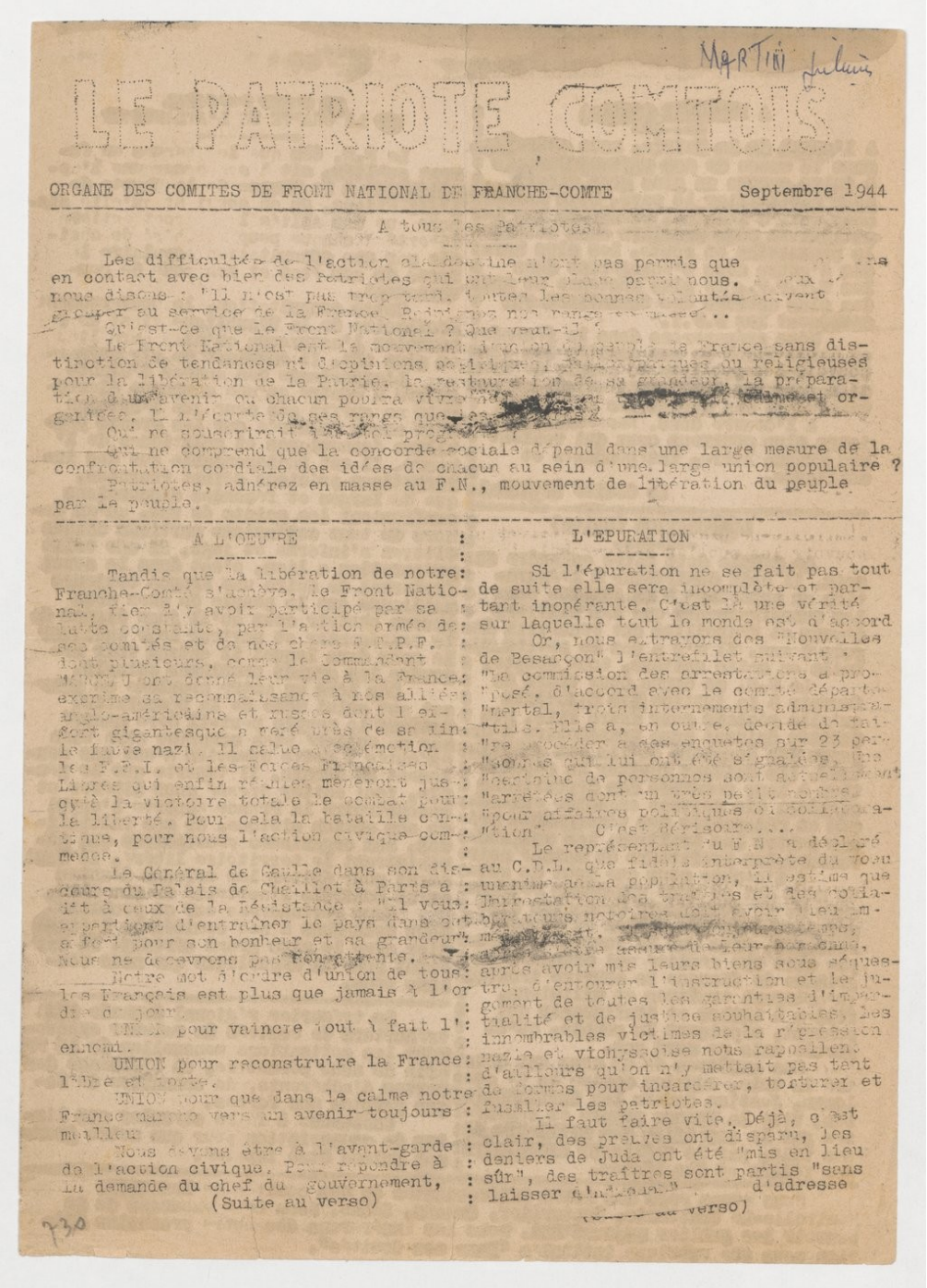
Re.G.1470 (485)

DON 63 03441 0525

mènent inplacable aux barbares nazis et aux traitres " Kollaborateurs " prouvant ainsi que la France n'est pas morte ET QU'ELLE DEMEURE UNE GRANDE PUISSANCE.
Ainsi l'heure décisive de la bataille est arrivée et la lutte entre dans une nouvelle phase, c'est pourquoi REPONDANT A L'APPEL DU CONSEIL NATIONAL DE LA RESISTANCE, et faisant sienne le déclaration du général DE GAULLE " tout vaut mieux que d'être mis hors de combat sans combattre " Le peuple de France uni dans l'action fera le 14 Juillet 1944 une date inoubliable par sa lutte contre l'ennemi exécré, pour retrouver sa LIBERTE, son INDEPENDANCE et SA GRANDEUR.
Partout les Francs-Tireurs et partisans harcéleront l'ennemi, détruisant leurs dépôts et leurs convois châtiront les collaborateurs et les miliciens de DARNAND, assassins de Français.

Les patriotes de l'Est, dont la volonté de chasser l'envahisseur ne s'est pas émoussée par ces quatre années de guerre et d'oppression, suivront le I4 JUILLET I944 ces mots d'ordre patriotiques:
TOUS LES TRAVAILLEURS SANS EXCEPTION FERONT GREVE TOUTE LE JOURNEE MARQUANT AINSI LEUR VOLONTE DE LUTTE CONTRE L'ENVAHISSEUR.
LES CHEMINOTS SABOTERONT LES CHEMINS DE FER, FERONT DERAILLER LES TRAINS.

Figure 9.2: Excerpt of the first page of the *L'avenir de l'est* (The future of the east) published in July 1944 by Resistance fighters



LE PATRIOTE COMTOIS

ORGANE DES COMITES DE FRONT NATIONAL DE FRANCHE-COMTE Septembre 1944

A tous les Patriotes

Les difficultés de l'action clandestine n'ont pas permis que [illegible] en contact avec bien des patriotes qui [illegible] parmi nous. [illegible] nous disons : "Il n'est pas trop tard. Toutes les bonnes volontés doivent s'unir au service de la France. Rejoignez nos rangs serrés..."

Qu'est-ce que le Front National ? Que veut-il ?

Le Front National est le mouvement d'union qui groupe la France sans distinction de tendances ni d'opinions politiques, philosophiques ou religieuses pour la libération de la Patrie, la restauration de sa grandeur, la préparation d'un avenir où chacun pourra vivre [illegible] organisés. Il n'écarte de ses rangs que [illegible]

Qui ne souscrirait à [illegible]

Qui ne comprend que la concorde sociale dépend dans une large mesure de la confrontation cordiale des idées de chacun au sein d'une large union populaire ?

Patriotes, adhérez en masse au F.N., mouvement de libération du peuple par le peuple.

A L'OEUVRE

Tandis que la libération de notre Franche-Comté s'achève, le Front National, fier d'y avoir participé par sa lutte constante, par l'action armée de ses comités et de nos chers F.T.P.F. dont plusieurs, comme le Commandant MARIN, ont donné leur vie à la France, exprime sa reconnaissance à nos alliés anglo-américains et russes dont l'effort gigantesque a percé très loin le fauve nazi. Il salue avec émotion les F.F.I. et les Forces Françaises Libres qui enfin réunies mèneront jusqu'à la victoire totale le combat pour la liberté. Pour cela la bataille continue, pour nous l'action civique commence.

Le Général de Gaulle dans son discours du Palais de Chaillot à Paris a dit à ceux de la Résistance : "Il vous appartient d'entraîner le pays dans cet effort pour son bonheur et sa grandeur. Nous ne décevrons pas notre patrie."

Notre mot d'ordre d'union de tous les Français est plus que jamais à l'ordre du jour.

UNION pour vaincre tout à fait l'ennemi.

UNION pour reconstruire la France libre et forte.

UNION pour que dans le calme notre France marche vers un avenir toujours meilleur.

Nous devons être à l'avant-garde de l'action civique. Pour répondre à la demande du chef du gouvernement,

(Suite au verso)

L'EPURATION

Si l'épuration ne se fait pas tout de suite elle sera incomplète et partant inopérante. C'est là une vérité sur laquelle tout le monde est d'accord

Or, nous extrayons des "Nouvelles de Besançon" l'entrefilet suivant :

"La commission des arrestations a proposé, d'accord avec le comité départemental, trois internements administratifs. Elle a, en outre, décidé de faire procéder à des enquêtes sur 23 personnes qui lui ont été signalées. Une centaine de personnes sont actuellement arrêtées dont un très petit nombre pour affaires politiques ou collaboration" C'est dérisoire....

Le représentant du F.N. a déclaré au C.D.L. que fidèle interprète du vœu unanime de la population, il estime que l'arrestation des traîtres et des collaborateurs notoires doit avoir lieu immédiatement. [illegible] après avoir mis leurs biens sous séquestre, d'entourer l'instruction et le jugement de toutes les garanties d'impartialité et de justice souhaitables. Les innombrables victimes de la répression nazie et vichyssoise nous rappellent d'ailleurs qu'on n'y mettait pas tant de formes pour incarcérer, torturer et fusiller les patriotes.

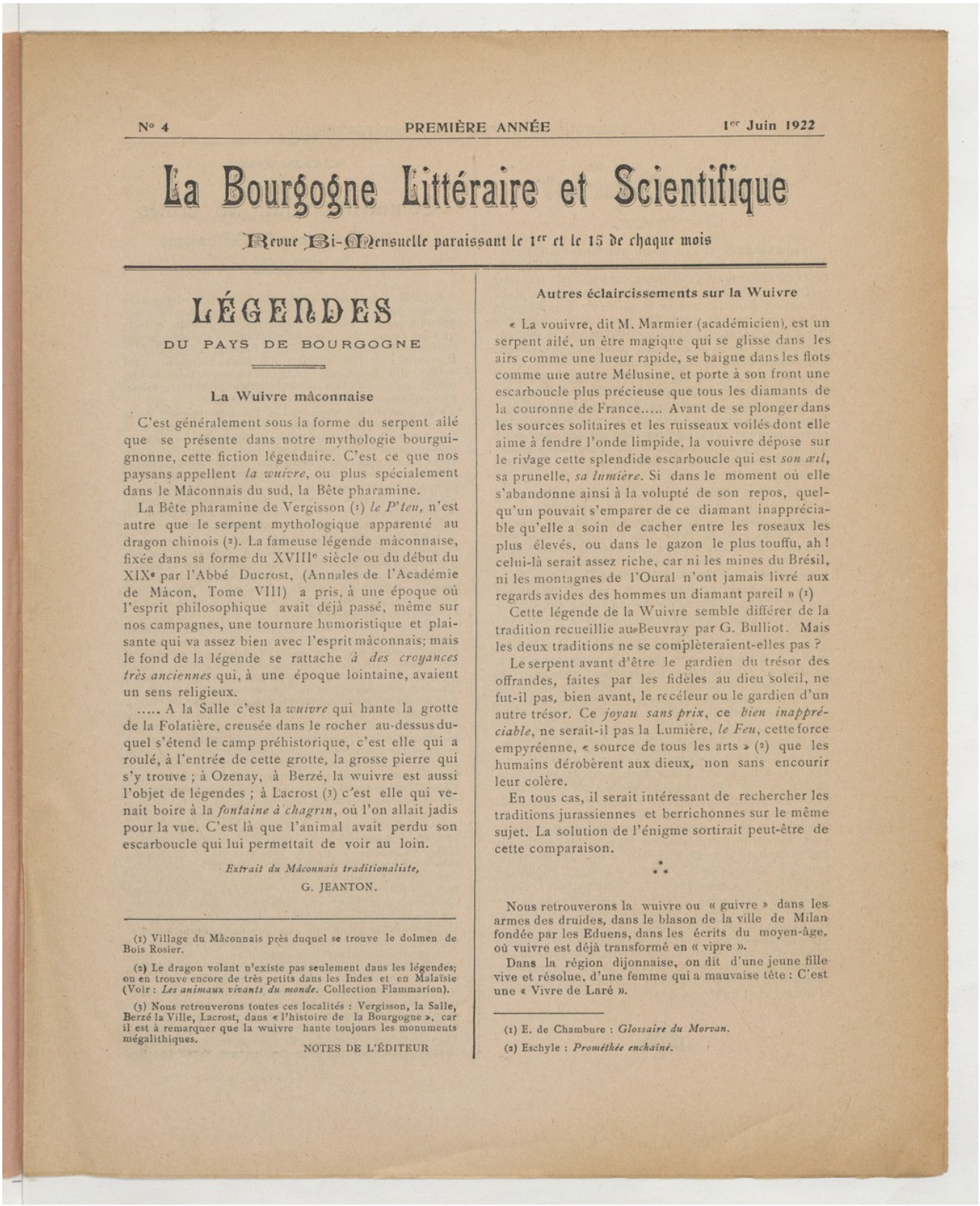
Il faut faire vite. Déjà, c'est clair, des preuves ont disparu, les deniers de Juda ont été "mis en lieu sûr", des traîtres sont partis "sans laisser d'adresse

(Suite au verso)

930

Source gallica.bnf.fr / Musée de la Résistance nationale / Champigny-sur-Marne

Figure 9.3: Excerpt of the first page of the *Le patriote Comtois* (The Comtois patriot) published in September 1944 by Resistance fighters



N° 4 PREMIÈRE ANNÉE 1er Juin 1922

La Bourgogne Littéraire et Scientifique

Revue Bi-Mensuelle paraissant le 1er et le 15 de chaque mois

LÉGENDES

DU PAYS DE BOURGOGNE

La Wuivre mâconnaise

C'est généralement sous la forme du serpent ailé que se présente dans notre mythologie bourguignonne, cette fiction légendaire. C'est ce que nos paysans appellent *la wuivre*, ou plus spécialement dans le Mâconnais du sud, la Bête pharamine.

La Bête pharamine de Vergisson (1) *le P'teu*, n'est autre que le serpent mythologique apparenté au dragon chinois (2). La fameuse légende mâconnaise, fixée dans sa forme du XVIIIe siècle ou du début du XIXe par l'Abbé Ducrost, (Annales de l'Académie de Mâcon, Tome VIII) a pris, à une époque où l'esprit philosophique avait déjà passé, même sur nos campagnes, une tournure humoristique et plaisante qui va assez bien avec l'esprit mâconnais; mais le fond de la légende se rattache *à des croyances très anciennes* qui, à une époque lointaine, avaient un sens religieux.

..... A la Salle c'est la *wuivre* qui hante la grotte de la Folatière, creusée dans le rocher au-dessus duquel s'étend le camp préhistorique, c'est elle qui a roulé, à l'entrée de cette grotte, la grosse pierre qui s'y trouve ; à Ozenay, à Berzé, la wuivre est aussi l'objet de légendes ; à Lacrost (3) c'est elle qui venait boire à la *fontaine à chagrin*, où l'on allait jadis pour la vue. C'est là que l'animal avait perdu son escarboucle qui lui permettait de voir au loin.

Extrait du Mâconnais traditionaliste,

G. JEANTON.

(1) Village du Mâconnais près duquel se trouve le dolmen de Bois Rosier.

(2) Le dragon volant n'existe pas seulement dans les légendes; on en trouve encore de très petits dans les Indes et en Malaïsie (Voir : *Les animaux vivants du monde.* Collection Flammarion).

(3) Nous retrouverons toutes ces localités : Vergisson, la Salle, Berzé la Ville, Lacrost, dans « l'histoire de la Bourgogne », car il est à remarquer que la wuivre hante toujours les monuments mégalithiques.

NOTES DE L'ÉDITEUR

Autres éclaircissements sur la Wuivre

« La vouivre, dit M. Marmier (académicien), est un serpent ailé, un être magique qui se glisse dans les airs comme une lueur rapide, se baigne dans les flots comme une autre Mélusine, et porte à son front une escarboucle plus précieuse que tous les diamants de la couronne de France..... Avant de se plonger dans les sources solitaires et les ruisseaux voilés dont elle aime à fendre l'onde limpide, la vouivre dépose sur le rivage cette splendide escarboucle qui est *son œil*, sa prunelle, *sa lumière*. Si dans le moment où elle s'abandonne ainsi à la volupté de son repos, quelqu'un pouvait s'emparer de ce diamant inappréciable qu'elle a soin de cacher entre les roseaux les plus élevés, ou dans le gazon le plus touffu, ah ! celui-là serait assez riche, car ni les mines du Brésil, ni les montagnes de l'Oural n'ont jamais livré aux regards avides des hommes un diamant pareil » (1)

Cette légende de la Wuivre semble différer de la tradition recueillie au Beuvray par G. Bulliot. Mais les deux traditions ne se compléteraient-elles pas ?

Le serpent avant d'être le gardien du trésor des offrandes, faites par les fidèles au dieu soleil, ne fut-il pas, bien avant, le receleur ou le gardien d'un autre trésor. Ce *joyau sans prix*, ce *bien inappréciable*, ne serait-il pas la Lumière, *le Feu*, cette force empyréenne, « source de tous les arts » (2) que les humains dérobèrent aux dieux, non sans encourir leur colère.

En tous cas, il serait intéressant de rechercher les traditions jurassiennes et berrichonnes sur le même sujet. La solution de l'énigme sortirait peut-être de cette comparaison.

∴

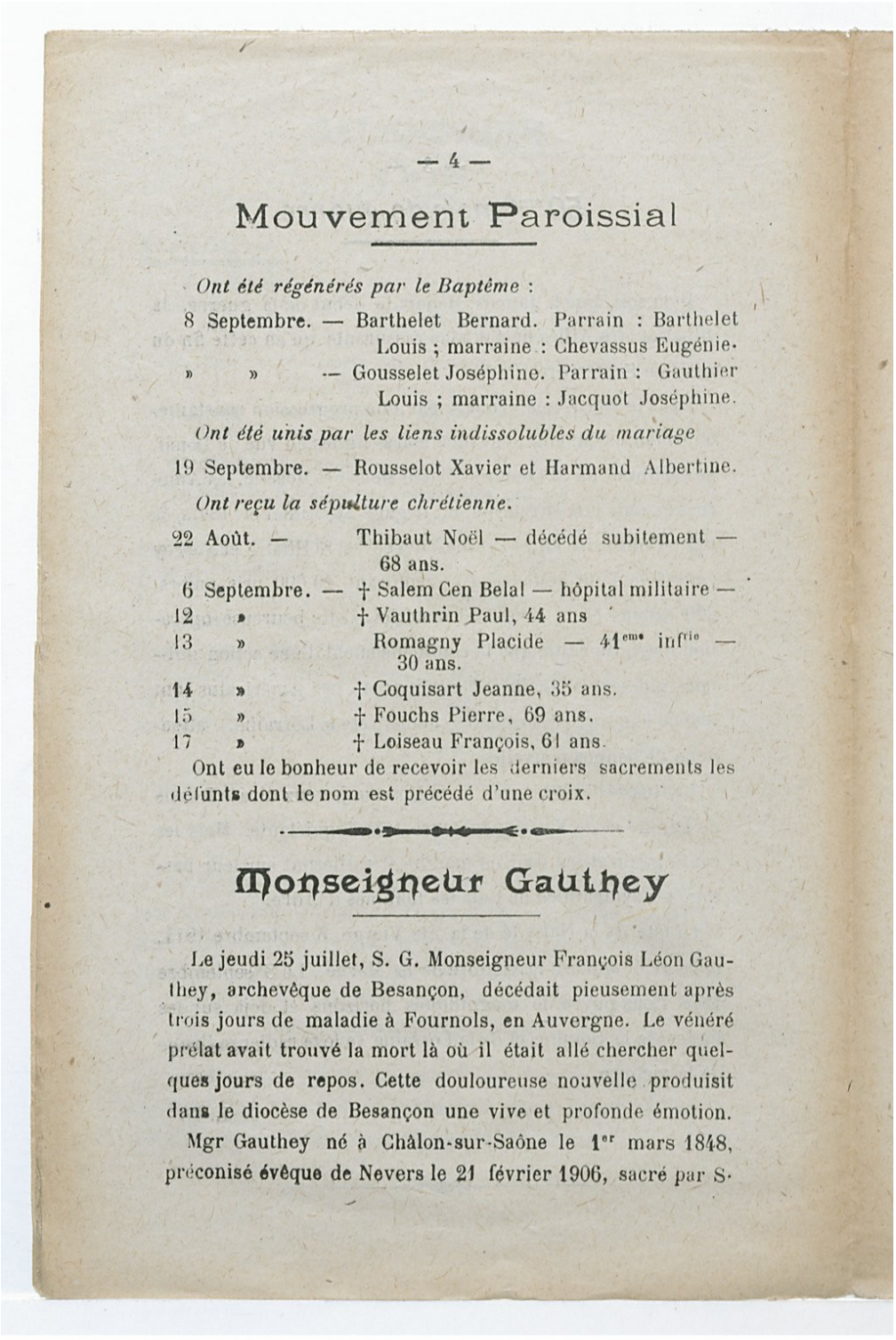
Nous retrouverons la wuivre ou « guivre » dans les armes des druides, dans le blason de la ville de Milan fondée par les Eduens, dans les écrits du moyen-âge, où vuivre est déjà transformé en « vipre ».

Dans la région dijonnaise, on dit d'une jeune fille vive et résolue, d'une femme qui a mauvaise tête : C'est une « Vivre de Laré ».

(1) E. de Chambure : *Glossaire du Morvan.*

(2) Eschyle : *Prométhée enchaîné.*

Figure 9.4: Cover of the fourth issue of the scientific and literary periodical *La Bourgogne* (Burgundy) published on the 1st of June 1922



— 4 —

Mouvement Paroissial

Ont été régénérés par le Baptême :

8 Septembre. — Barthelet Bernard. Parrain : Barthelet Louis ; marraine : Chevassus Eugénie.
» » — Gousselet Joséphine. Parrain : Gauthier Louis ; marraine : Jacquot Joséphine.

Ont été unis par les liens indissolubles du mariage

19 Septembre. — Rousselot Xavier et Harmand Albertine.

Ont reçu la sépulture chrétienne.

22 Août. — Thibaut Noël — décédé subitement — 68 ans.
6 Septembre. — † Salem Cen Belal — hôpital militaire —
12 » † Vauthrin Paul, 44 ans
13 » Romagny Placide — 41ème inf^rie — 30 ans.
14 » † Coquisart Jeanne, 35 ans.
15 » † Fouchs Pierre, 69 ans.
17 » † Loiseau François, 61 ans.

Ont eu le bonheur de recevoir les derniers sacrements les défunts dont le nom est précédé d'une croix.

Monseigneur Gauthey

Le jeudi 25 juillet, S. G. Monseigneur François Léon Gauthey, archevêque de Besançon, décédait pieusement après trois jours de maladie à Fournols, en Auvergne. Le vénéré prélat avait trouvé la mort là où il était allé chercher quelques jours de repos. Cette douloureuse nouvelle produisit dans le diocèse de Besançon une vive et profonde émotion.

Mgr Gauthey né à Châlon-sur-Saône le 1er mars 1848, préconisé évêque de Nevers le 21 février 1906, sacré par S·

Figure 9.5: First page of the 130th issue of the parochial bulletin of Arc-Les-Gray (small town in the Franche-Comté region), published on October 1918

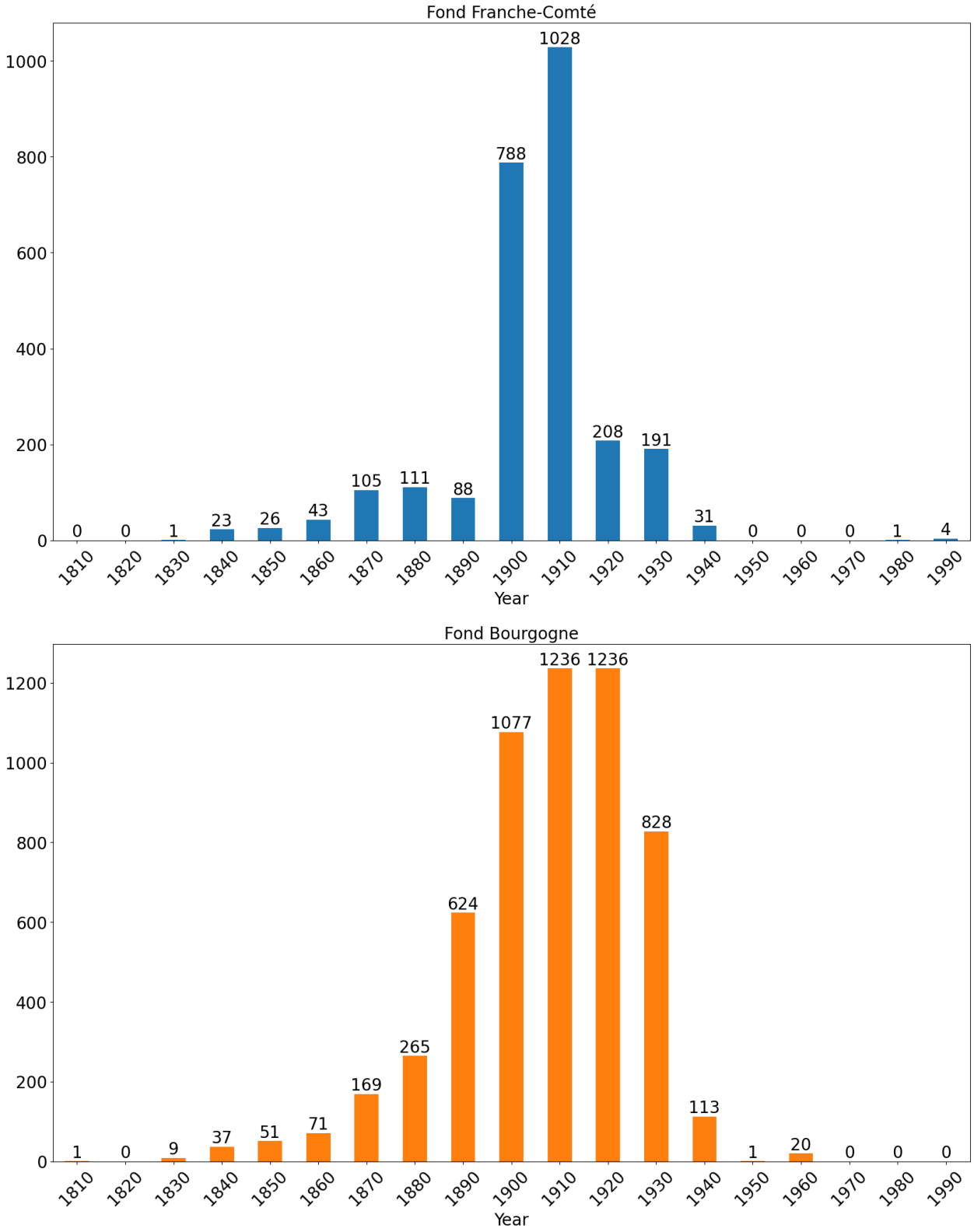


Figure 9.6: Document distribution by decades in the *Franche-Comté* and *Bourgogne* funds of the EMONTAL corpus

- `dc:date`: the original publication date of the document. There are two possible formats: either YYYY-MM or YYYY-MM-DD
- `dc:title`: the title of the document
- `dc:creator`: the original creator of the document
- `dc:publisher`: the original publisher of the document
- `dc:language`: the language of the text
- `dc:relation`: the url to the collection the document belongs to
- `provenance`: the domain name, which is always *bnf.fr*
- `source`: the source of the document, usually set to *Bibliothèque nationale de France* followed by the department where the document can be found
- `typedoc`: the document type, i.e. book, fascicule, etc. Since we have collected periodicals, the *typedoc* of our documents is always fascicule
- `mode_indexation`: the indexation method, usually text
- `nqamoyen`: the average OCR quality
- `title`: the title of the document (as in the `oai` tag)
- `date`: the publication date of this document
- `first_indexation_date`: date of the first indexation of the document
- `request_url`: the url to get the OAI information

The `ocr` tag provides the textual content and physical layout of documents, by following the XML ALTO format. Lines of text are contained in `TextLine` tags, which in their turn contain `String` tags for words and `SP` tags for spaces. `TextLine` tags are grouped into blocks in `TextBlock` tags. Sometimes, `TextBlock` tags are grouped into `ComposedBlock` tags. Other tags such as `Illustration`, `polygon` or `Shape` describe graphical elements such as images or line separation. An example of the content of this tag is shown in Figure 9.8.

`TextBlock` and `TextLine` tags have the following attributes:

id: the tag's identifier

height, width: the text's height and width in the original document

```
<metadata>
<oai_dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/"
   xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
   http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
 <dc:identifier>
  https://gallica.bnf.fr/ark:/12148/bpt6k881591v
 </dc:identifier>
 <dc:date>
  1944-08
 </dc:date>
 <dc:description>
  août 1944
 </dc:description>
 <dc:description>
  1944/08.
 </dc:description>
 <dc:description/>
 <dc:title>
  La Haute-Saône libre : organe départemental du Front national
 </dc:title>
 <dc:creator>
  Front national de lutte pour la libération et l'indépendance de la France. Auteur du texte
 </dc:creator>
 <dc:publisher>
  [s.n.]
 </dc:publisher>
 <dc:type xml:lang="fre">
  texte
 </dc:type>
 <dc:type xml:lang="eng">
  text
 </dc:type>
 <dc:type xml:lang="fre">
  publication en série imprimée
 </dc:type>
 <dc:type xml:lang="eng">
  printed serial
 </dc:type>
 <dc:language>
  fre
 </dc:language>
 <dc:language>
  français
 </dc:language>
 <dc:relation>
  Notice du catalogue : http://catalogue.bnf.fr/ark:/12148/cb32786048z
 </dc:relation>
 <dc:source>
  Bibliothèque nationale de France, département Réserve des livres rares, RES-G-1470 (626)
 </dc:source>
```

Figure 9.7: Excerpt of the metadata collected from Gallica for the document *La Haute-Saône Libre* published in August 1944 in our corpus

vpos: the vertical position of the text on the page. The higher the value, the lower the word is on the page

hpos: the horizontal position of the text on the page. The higher the value, the further on the right the text is on the page

language: the language of the text (only present in `TextBlock` tags)

content: the textual content of the String tag

Table 9.1 shows the distribution of collections, documents and pages, as well as `TextBlock`, `TextLine` and `String` tags in the EMONTAL corpus.

	Collections	Issues	Pages	TextBlocks	TextLines	Strings
Fond Franche-Comté	46	2,648	255,670	3,733,845	11,373,606	83,001,454
Fond Bourgogne	113	5,738	637,407	8,117,055	24,183,489	189,266,170
Total	159	8,386	893,077	11,850,900	35,557,095	272,267,624

Table 9.1: Distribution of collections, documents, pages and `TextBlock`, `TextLine`, and `String` tags in the EMONTAL corpus

Some `TextBlock` tags also have a `type` attribute. This attribute appears most often for `TextBlock` tags which describe tables or advertisements blocks. However, this attribute is rare in our corpus: as shown in Table 9.2, nearly 85 % of the `TextBlock` tags in the EMONTAL corpus do not have a `type` attribute.

9.2 Thematic analysis of the EMONTAL corpus

The documents in the EMONTAL corpus belong to eight broad thematic categories, which we have manually identified by analysing the titles and content of each collection.We have identified eight different themes, as follows:

Agriculture : documents addressing agricultural matters, especially related to local productions such as wines. These documents have been published between 1860 and 1940, with a peak of 614 publications in the 1920s

Fighters and patriotism : patriotic newspapers or periodicals published by fighters, such as Resistance fighters. These documents have been published between 1900 and 1940, with a peak of 113 publications in the 1900s

```
<page accuracy="85.9" height="3639" id="PAG_00000001" physical_img_nr="1" quality="OK" width="2875">
 <topmargin height="359" hpos="0" id="PAG_00000001_TopMargin" vpos="0" width="2875">
 </topmargin>
 <leftmargin height="2978" hpos="0" id="PAG_00000001_LeftMargin" vpos="359" width="232">
 </leftmargin>
 <rightmargin height="2978" hpos="2630" id="PAG_00000001_RightMargin" vpos="359" width="245">
 </rightmargin>
 <bottommargin height="302" hpos="0" id="PAG_00000001_BottomMargin" vpos="3337" width="2875">
 </bottommargin>
 <printspace height="2978" hpos="232" id="PAG_00000001_PrintSpace" vpos="359" width="2398">
  <illustration height="387" hpos="243" id="PAG_00000001_IL000001" vpos="359" width="2361">
   <shape>
    <polygon points="253,366 2613,366 2613,747 253,747 253,366">
    </polygon>
   </shape>
  </illustration>
  <composedblock height="2496" hpos="232" id="PAG_00000001_CB000001" type="container" vpos="841" width="2398">
   <shape>
    <polygon points="237,848 2631,848 2631,3338 237,3338 237,848">
    </polygon>
   </shape>
   <composedblock height="2492" hpos="232" id="PAG_00000001_CB000002" type="container" vpos="845" width="778">
    <shape>
     <polygon points="237,848 1011,848 1011,3338 237,3338 237,848">
     </polygon>
    </shape>
    <textblock height="111" hpos="240" id="PAG_00000001_TB000001" language="fr" vpos="845" width="763">
     <shape>
      <polygon points="248,848 1011,848 1011,957 248,957 248,848">
      </polygon>
     </shape>
     <textline height="36" hpos="262" id="PAG_00000001_TL000001" vpos="848" width="650">
      <string content="M" height="18" hpos="262" id="PAG_00000001_ST000001" vpos="863" wc="0" width="17">
      </string>
      <sp hpos="280" id="PAG_00000001_SP000001" vpos="851" width="6">
      </sp>
      <string content="lj.lt" height="32" hpos="287" id="PAG_00000001_ST000002" vpos="851" wc="0.4280000031" width="41">
      </string>
      <sp hpos="329" id="PAG_00000001_SP000002" vpos="863" width="6">
      </sp>
      <string content="fui" height="32" hpos="336" id="PAG_00000001_ST000003" vpos="852" wc="0.4133333266" width="43">
      </string>
      <sp hpos="380" id="PAG_00000001_SP000003" vpos="852" width="7">
      </sp>
      <string content="l.'jü" height="26" hpos="388" id="PAG_00000001_ST000004" vpos="858" wc="0.4180000126" width="56">
      </string>
      <sp hpos="445" id="PAG_00000001_SP000004" vpos="862" width="15">
```

Figure 9.8: Excerpt of the OCR content collected from Gallica for the document *La Haute-Saône Libre* published in August 1944 in our corpus. The tags have been converted to lower case to simplify their processing

	Franche-Comté fund		*Bourgogne* fund	
Type attribute	Count	Percentage	Count	Percentage
No type	3,145 740	84.249 %	6,913 778	85.176 %
table	372,071	9.965 %	681,773	8.399 %
illegible	72,914	1.953 %	229,811	2.831 %
table illegible	57,174	1.531 %	105,538	1.300 %
unworkable illegible	39,008	1.045 %	32,638	0.402 %
advertisement	21,466	0.575 %	128,095	1.578 %
unworkable	11,158	0.299 %	9,157	0.113 %
table unworkable	8,673	0.232 %	3,230	0.040 %
table unworkable illegible	2,731	0.073 %	6,031	0.074 %
titre1	1,496	0.040 %	3,082	0.038 %
textStamped	1,095	0.029 %	1,040	0.013 %
illegible unworkable	282	0.008 %	394	0.005 %
table illegible unworkable	33	0.001 %	94	0.001 %
scriptFonts	4	0.000 %	55	0.001 %
Total	3,733,845	100 %	8,114,716	100 %

Table 9.2: Distribution of the *type* attribute in `TextBlock` tags in the EMONTAL corpus

Generalist and partisan newspapers : general newspapers as well as newspapers published by groups such as political parties. These documents have been published between 1900 and 1940, with a peak of 682 publications in the 1920s

Leisure : periodicals addressing various leisure activities, such as photography or sports. These documents have been published between 1900 and 1930, with a peak of 166 publications in the 1920s

Local powers and economy : documents addressing economical matters, as well as decisions made by local authorities. These documents have been mainly published between 1830 and 1940, with a peak of 194 publications in the 1870s, although some documents have been published in the 1980s and 1990s

Religious : documents published by religious organisms such as local parishes, which address the religious life and events in the area. These documents have been published between 1890 and 1930, with a peak of 1,110 publications in the 1910s

Science and culture : documents related to the topics of sciences and arts, such as history, archaeology or literature. These documents have been published between 1830 and 1960, with a peak of 183 publications in the 1930s

Women in society : documents either published by women or addressing the status of women in

society. These documents have been published between 1910 and 1940, with a peak of 9 publications in the 1920s

Table 9.3 shows examples of document titles for each thematic categories. The list of document collections that belong to each thematic category is presented in Annexe 19.4.

Category	Title
Agriculture	*Bulletin du Comité d'agriculture de Beaune* (Bulletin of the Beaune Agricultural Committee) "*La Bourgogne rurale : revue mensuelle dágriculture, de viticulture et dhorticulture*" (The Rural Burgundy: monthly periodical of agriculture, viticulture and horticulture)
Fighters and patriotism	*Le Petit écho du 21e Régiment d'infanterie* (The Little Echo of the 21st Infantry Regiment) "*Jeunesse du Maquis*" (Youth of the Maquis)
Generalist and partisan newspapers	*Le Franc-Comtois de Paris : journal d'information des départements Doubs, Jura, Haute-Saône, Haut-Rhin* (The Franc-Comtois in Paris: newspaper for the Doubs, Jura, Haute-Saône and Haut-Rhin departments) *Le Semeur. Organe régional du Parti communiste* (The Sower. Organe régional du Parti communiste)
Leisure	*Les Sports de l'Est : hebdomadaire illustré de tous les sports, Lorraine, Franche-Comté, Bourgogne, Champagne* (The Sports of the East: weekly illustrated sports magazine, Lorraine, Franche-Comté, Burgundy, Champagne) *Le Collectionneur : bulletin mensuel de publicité, philatélie, cartophilie...* (Le Collectionneur: monthly newsletter on advertising, philately, cartophily...)
Local powers and economy	*Tableaux de l'économie bourguignonne* (Tables of the economy of Burgundy) *Rapports et délibérations / Conseil général du Doubs* (Reports and deliberations / Doubs County Council)
Religious	*Bulletin paroissial de Notre-Dame, Dijon* (Parish bulletin of Notre-Dame, Dijon) *Vers l'avenir : organe mensuel de la Jeunesse catholique de Franche-Comté et du Territoire-de-Belfort* (Towards the future: monthly organ of the Catholic Youth of Franche-Comté and the Territoire-de-Belfort)
Science and culture	*Bulletin de la Société archéologique et biographique du canton de Montbard* (Bulletin of the Archaeological and Biographical Society of the Canton of Montbard) *La Vie meilleure : revue sociologique et littéraire* (The Better Life: sociological and literary review)
Women in society	*La Voix des femmes de Bourgogne* (The Voice of Burgundy Women) "*La France féminine : revue mensuelle pendant la guerre : organe de défense des droits de la femme française*" (The Feminine France: monthly magazine during the war: an organisation defending the rights of French women)

Table 9.3: Examples of documents by thematic categories in the EMONTAL corpus

The distributions of documents by thematic categories and by decades are shown in Table 9.4 and Figure 9.9. Overall, the vast majority of the documents in our corpus (more than 87 %) was published between 1890 and 1940.

Category	1810	1820	1830	1840	1850	1860	1870	1880	1890
Agriculture	0	0	0	0	0	10	45	158	305
Fighters and patriotism	0	0	0	0	0	0	0	0	0
Generalist and partisan newspapers	0	0	0	0	0	0	0	0	0
Leisure	0	0	0	0	0	0	0	0	0
Local powers and economy	0	0	3	50	63	86	194	172	105
Religious life	1	0	0	0	0	0	0	0	192
Science and knowledge	0	0	7	10	14	18	35	46	110
Women in society	0	0	0	0	0	0	0	0	0
Total	1	0	10	60	77	114	274	376	712

Category	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990
Agriculture	572	381	614	402	13	0	0	0	0	0
Fighters and patriotism	113	10	36	51	31	0	0	0	0	0
Generalist and partisan newspapers	682	454	143	117	38	0	0	0	0	0
Leisure	33	18	166	6	0	0	0	0	0	0
Local powers and economy	122	176	151	189	21	0	0	0	1	4
Religious life	210	1,110	178	65	0	0	0	0	0	0
Science and knowledge	133	110	147	185	36	1	20	0	0	0
Women in society	0	5	9	4	5	0	0	0	0	0
Total	1,865	2,264	1,444	1,019	144	1	20	0	1	4

Table 9.4: Distribution of documents by thematic category, 1810-1890 and 1900-1990

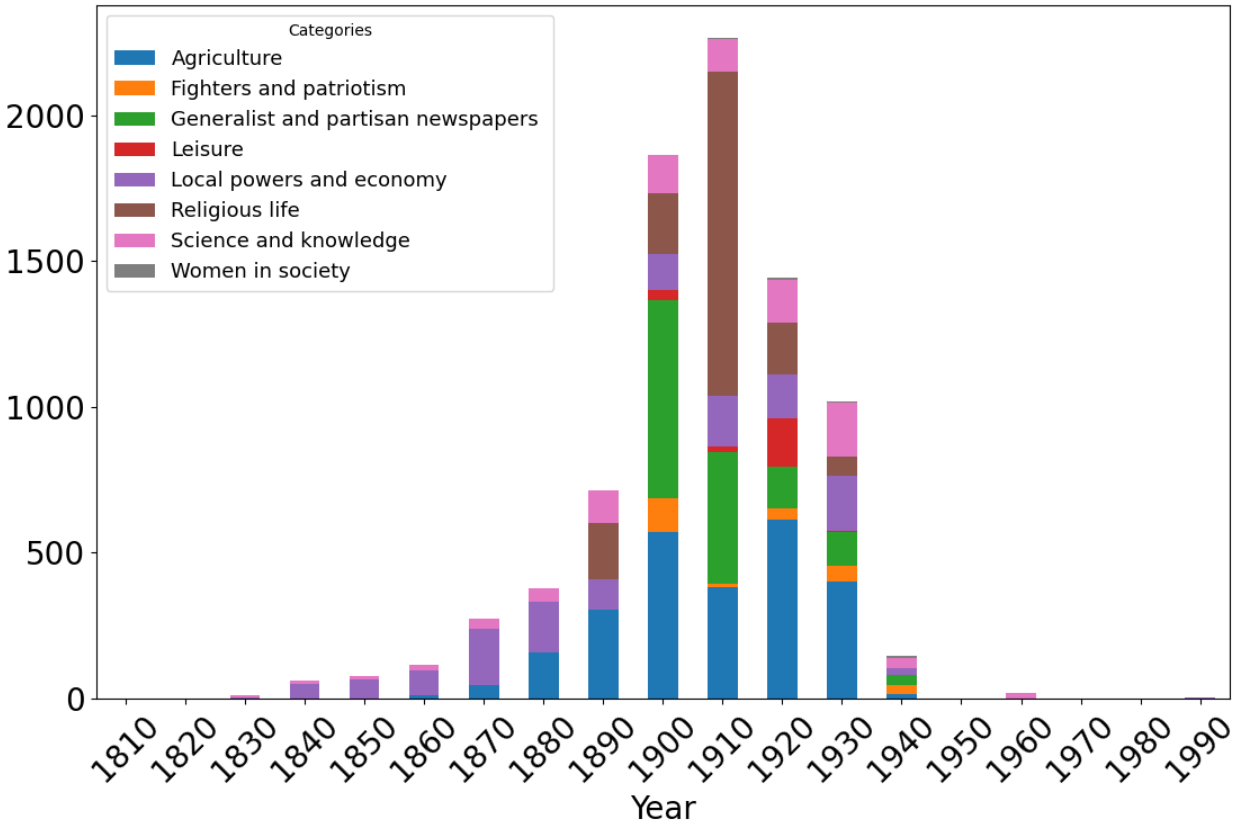


Figure 9.9: Distribution of documents by thematic category and decade

Chapter 10

Cleaning the OCR transcription of documents

Table of contents

As explained in Chapter 9, the documents of the EMONTAL corpus are stored in the XML format. The metadata of documents are stored in the `oai` tag, whereas the textual content of the document is stored in the `ocr` tag, which follows the XML ALTO format. The XML ALTO format provides the textual content of documents, which has been obtained by OCR methods. These transcriptions may contain errors as well as artefacts produced by processes such as the hyphenation process. Correcting these errors and artefacts is necessary, in order to limit their impact on the application of NLP methods.

In our work we have identified two main error types in the OCR transcriptions: hyphenation that appears at the end of lines, and wrongly transcribed characters that lead to incorrect tokens. We have designed methods, specifically tailored for the EMONTAL dataset, to correct these errors. In Section 10.1, we describe our method to remove hyphens from the textual content of documents, and in Section 10.2, we describe our method to correct the OCR transcriptions of documents.

10.1 Hyphen Removal

In typography, hyphenation consists in dividing a word by a hyphen, so that it respects the justification of the text. Hyphenation can therefore introduce artefacts, especially if the text is transcribed using OCR methods. Thus, it is necessary to post-process hyphens so as to prevent such artefacts to interfere with subsequent processing. Simple processing rules can be used to remove hyphenation and correct the separated tokens, as in Généreux and Spano (2015). In XML ALTO documents, hyphens are represented by the `Hyp` tag, which has four attributes:

content : the textual content of the hyphen, i.e. the hyphen symbol ("-")

hpos : the horizontal position of the hyphen in the document

vpos : the vertical position of the hyphen in the document

width : the width of the hyphen in the document

`Hyp` tags are located at the end of a `TextLine` tag and may appear in two possible contexts: either the `TextLine` tag is followed by another `TextLine` tag, or the `TextLine` tag is the last tag of the `TextBlock` that contains it.

In the first context, which is shown in Figure 10.1, the `String` tags $prevString$ and $nextString$, which respectively precede and follow the `Hyp` tag, have the following attributes:

content : either the first part of the word including the hyphen for $prevString$, or its second part for $nextString$

subs_content : the full word, without segmentation

subs_type : either HypPart1 for $prevString$ or HypPart2 for $nextString$

To collect the textual content of documents, we must collect the value of the attribute that contains the complete word, i.e. we must collect the value of either the *content* attribute or of the *subs_content* attribute when it is available. Thus, when the `Hyp` tag appears in the first context, we must remove $nextString$ from the XML tree, so as to avoid collecting the value of the *subs_content* attribute twice.

In the second context, which is shown in Figure 10.2, neither $prevString$ nor $nextString$ have a *subs_content* attribute. The *content* attribute of $prevString$ contains the first half of the word, with a trailing hyphen, whereas the *content* attribute of $nextString$ contains the second half of the word, with no hyphen. This second context is rare, and usually indicates an error in the `TextBlock` segmentation step during the OCR process. In such a context, we must only process the `Hyp` tags and the trailing hyphen in the *content* attribute of $prevString$.

```
<string content="com-" height="23" hpos="1464" id="PAG_00000489_ST000086" stylerefs="TXT_3"
       subs_content="compété" subs_type="HypPart1" vpos="1021" wc="0.99" wd="false" width="89">
</string>
<hyp content="-" hpos="1540" vpos="1029" width="13">
</hyp>
</textline>
<textline height="42" hpos="484" id="PAG_00000489_TL000011" stylerefs="TXT_3" vpos="1064" width="1071">
<string content="pété" height="41" hpos="484" id="PAG_00000489_ST000087" stylerefs="TXT_3"
       subs_content="compété" subs_type="HypPart2" vpos="1064" wc="1" wd="false" width="77">
</string>
```

Figure 10.1: Example of the first possible context of occurrence of the `Hyp` tag in an XML ALTO document from our corpus

```
<string content="l'his-" height="32" hpos="2791" id="PAG_00000003_ST002247" stylerefs="TXT_1" vpos="6707" wc="0.97" width="95"></string>
<hyp content="-" hpos="2886" vpos="6739" width="30"></hyp>
</textline>
</textblock>
<textblock height="419" hpos="2964" id="PAG_00000003_TB000069" language="fr" stylerefs="TXT_1" vpos="5121" width="1202">
<textline height="46" hpos="2964" id="PAG_00000003_TL000284" stylerefs="TXT_1" vpos="5121" width="1195">
<string content="toire" height="30" hpos="2964" id="PAG_00000003_ST002248" stylerefs="TXT_1" vpos="5122" wc="1" width="88"></string>
```

Figure 10.2: Example of the second possible context of occurrence of the `Hyp` tag in an XML ALTO document from our corpus

We propose a simple pipeline to process `Hyp` tags in XML ALTO documents If the `Hyp` tag appears in the first context, we remove the following `String` tag from the XML tree, so as to avoid collecting the textual content of a word twice later. We then remove the `Hyp` tag from the XML tree. Otherwise, if the `Hyp` tag appears in the second context, we simply remove it from the XML tree, without any other processing. We do not remove the trailing hyphen symbol in the *content* attribute of $prevString$, since we might remove the hyphen in compound words and introduce errors. For instance in Figure 10.2, the hyphen in the `String` tag "*Lieutenant-*" should not be removed, since "*Lieutenant-Colonel*" is an existing compound word. Errors that may be produced by not removing trailing hyphens are corrected later by our OCR post-processing method, which we describe in Section 10.2. The main stages of this implementation are shown in Figure 10.3. We have implemented this pipeline in the Python programming language, and used the library *BeautifulSoup* to manipulate XML files. The implementation is available on GitHub[1].

Figure 10.4 shows an example of an XML ALTO document before and after applying the hyphen removal pipeline in the first occurring context. Similarly, Figure 10.5 shows an example of an XML ALTO document before and after applying this pipeline in the second occurring context.

[1] https://github.com/nicolasgutehrle/emontalproject

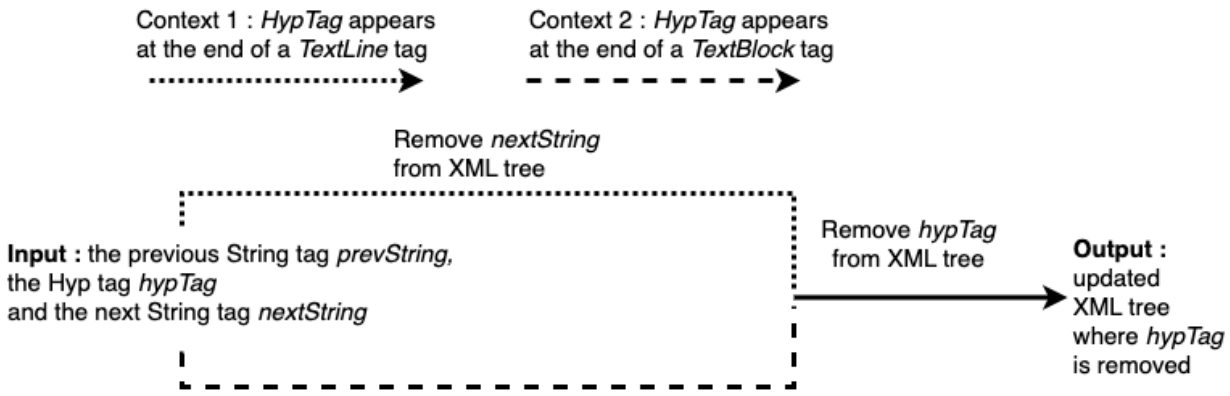


Figure 10.3: Main stages of our implementation of the Hyphen Removal pipeline

Before applying the hyphen removal pipeline

```
<string content="com-" height="24" hpos="1797" id="PAG_00000002_ST000048" stylerefs="TXT_1" subs_content="commerce" subs_type="HypPart1" vpos="1648" wc="1" width="109"></string>
<hyp content="-" hpos="1906" vpos="1672" width="30"></hyp>
</textline>
<textline height="66" hpos="489" id="PAG_00000002_TL000008" stylerefs="TXT_1" vpos="1702" width="1418">
<string content="merce" height="26" hpos="489" id="PAG_00000002_ST000049" stylerefs="TXT_1" subs_content="commerce" subs_type="HypPart2" vpos="1731" wc="1" width="141"></string>
```

After applying the hyphen removal pipeline

```
<string content="com-" height="24" hpos="1797" id="PAG_00000002_ST000048" stylerefs="TXT_1" subs_content="commerce" subs_type="HypPart1" vpos="1648" wc="1" width="109"></string>
</textline>
<textline height="66" hpos="489" id="PAG_00000002_TL000008" stylerefs="TXT_1" vpos="1702" width="1418">
<sp hpos="630" id="PAG_00000002_SP000042" vpos="1726" width="41"></sp>
```

Figure 10.4: Occurrence of the hyphen in the first context, before and after applying the Hyphen Removal pipeline

Before applying the hyphen removal pipeline

```
<string content="Lieutenant-" height="43" hpos="1622" id="PAG_00000006_ST000063" stylerefs="TXT_1" vpos="820" wc="1" width="301">
</string>
<hyp content="-" hpos="1923" vpos="863" width="30">
</hyp>
</textline>
</textblock>
<textblock height="411" hpos="420" id="PAG_00000006_TB000003" language="fr" stylerefs="TXT_1" vpos="877" width="1503">
<textline height="60" hpos="432" id="PAG_00000006_TL000008" stylerefs="TXT_1" vpos="877" width="1486">
<string content="Colonel" height="43" hpos="432" id="PAG_00000006_ST000064" stylerefs="TXT_1" vpos="879" wc="1" width="198">
```

After applying the hyphen removal pipeline

```
<string content="Lieutenant-" height="43" hpos="1622" id="PAG_00000006_ST000063" stylerefs="TXT_1" vpos="820" wc="1" width="301">
</string>
</textline>
</textblock>
<textblock height="411" hpos="420" id="PAG_00000006_TB000003" language="fr" stylerefs="TXT_1" vpos="877" width="1503">
<textline height="60" hpos="432" id="PAG_00000006_TL000008" stylerefs="TXT_1" vpos="877" width="1486">
<string content="Colonel" height="43" hpos="432" id="PAG_00000006_ST000064" stylerefs="TXT_1" vpos="879" wc="1" width="198">
</string>
```

Figure 10.5: Occurrence of the hyphen in the second context, before and after applying the Hyphen Removal pipeline

10.2 OCR post-processing

Because of the OCR process, the transcription of the textual content of our corpus may contain two types of errors: either non-word errors, i.e. substitution or omission of characters that result in non existing words, or real-word errors, i.e. substitution or omission of characters that result in existing words, which are different from the real ones contained in the document (T. T. H. Nguyen et al., 2021a). Since the quality of the transcription has an impact on further processes, such as Named Entity Recognition or Relation Extraction, it is necessary to post-process the OCR transcriptions to remove these types of errors as much as possible.

10.2.1 Method

We propose a rule-based method to make corrections to non-word errors in OCR transcriptions. Our aim is to propose a simple method to clean OCR transcriptions and limit the impact of non-word errors on further processes. We designed these rules by observing the non-word errors in our corpus. This pipeline consists in the four following steps:

1. *feature extraction*
2. *rule application*
3. *correction suggestion*
4. *document conversion*

These steps aim at correcting a document by either keeping, deleting or correcting a word, based on the extracted features.

The *feature extraction* step consists in extracting features from each word in the document. We extract morphological and semantic features describing the structure of the words. The set of extracted features is described in Table 10.1.

The *rule application* step consists in applying the rule set to determine which operation must be applied to a word, i.e. if it is either kept as it is, deleted or corrected. By default, words are kept as they are. The rules are described in Table 10.2, and are applied in the given order. Rule 1 and 2 select candidates for the *Delete* operation, whereas rule 3 selects candidates for the *Correct* operation. Rule 1 states that the proportion of non-alphanumeric characters in a word must be superior or equal to 75%. We have empirically selected this value after multiple tests.

The *correction suggestion* step consists in processing the words marked for correction. First, any character that is repeated three or more times is replaced by one occurrence only. For instance,

Feature name	Description
word_length	the number of characters in the word
stw_capital	True if the word starts with a capital letter, else False
stw_elision	True if the word starts with an elision, else False
non_alpha_prop	the proportion of non-alphanumeric characters in the word
ends_punct	True if the word ends with a punctuation sign, else False
is_punct	True if the word is a punctuation, else False
is_digit	True if the word is a digit, else False
is_oneletter_word	True if the word is in the set $\{a,à,y,ô,m\}$, else False

Table 10.1: Set of features extracted from each word in a document for OCR post-processing

Rule	Objective	Conditions	Operation
1	Remove non-alphanumeric words	W.non_alpha_prop >= 75 and not W.is_punct	Delete
2	Remove one character words	not W.is_oneletter_word and W.word_length = 1 and not W.is_punct and not W.is_digit	Delete
3	Suggest a text correction	not W.stw_capital and not W.ends_punct and not W.is_punct and not W.is_digit and not W.stw_elision	Correct

Table 10.2: OCR post-processing rules to determine if a word must undergo either the Keep, Delete or Correct operation. The default operation is Keep

the incorrect transcription "*mercrediii*" (Wednesdayyy) would be corrected to "*mercredi*" (Wednesday). Then, we apply a spelling correction algorithm to find the most likely correction to the word. If no possible correction can be found, the word is kept as is.

Finally, the *document conversion* step modifies the document by applying the operations determined for each word.

10.2.2 Implementation

We propose an implementation of our OCR post-processing method in the Python programming language. The main stages of this implementation are shown in Figure 10.6.

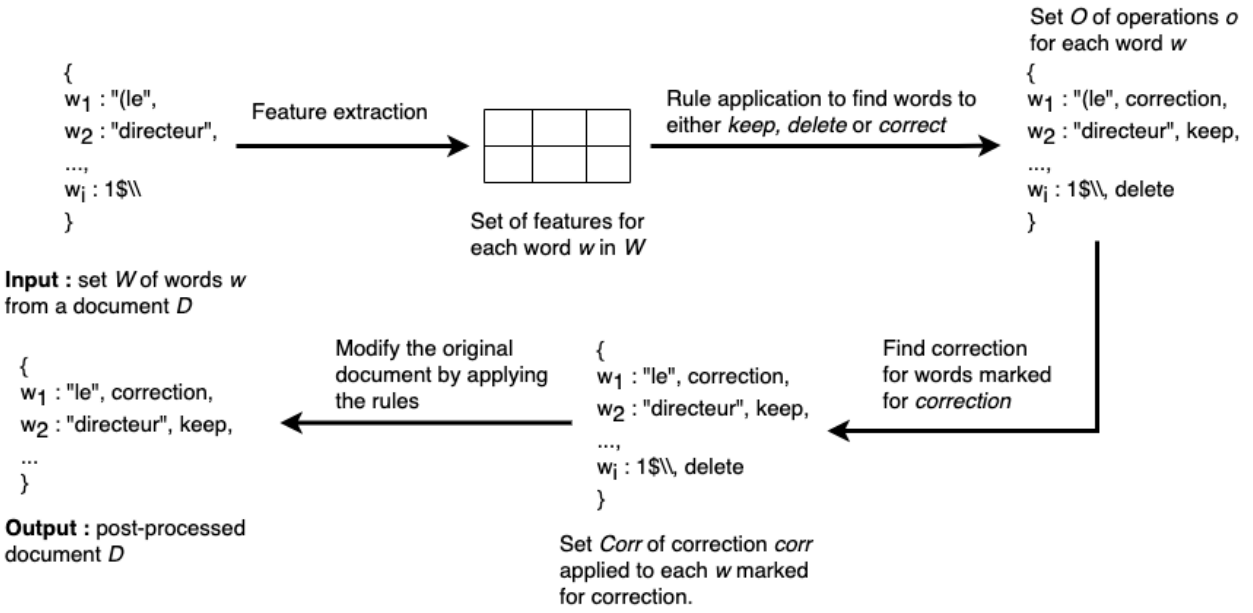


Figure 10.6: Main stages of the implementation of the OCR post-processing pipeline

We apply the OCR post-processing pipeline and the rule set described in the previous Section to the value of *content* attributes of `String` tags in the XML ALTO documents from the EMONTAL corpus. We write a dedicated function for each feature to extract. We use the *pandas* library to store the extracted features in a DataFrame, i.e. a matrix. An "operation" column in the DataFrame stores the operation applied to a `String` tag, i.e. either *Keep*, *Delete* or *Correct*. We store the possible correction of a word marked for correction in a "correction" column. Each rule is implemented as a set of logical operations to select rows from the feature DataFrame. For each matching row, we assign the operation for that rule in the "operation" column. We apply the `([a-z])\1{2,}` regular expression implemented using the *re* library to find characters that are repeated three or more times.

We apply the Symmetric Delete spelling (SymSpell) corrector algorithm[2] as in Huynh et al. (2020) to find candidate corrections to words marked for the *Correct* operation. SymSpell applies the Damerau-Levenshtein edit distance (Damerau, 1964) and requires a term-frequency dictionary to suggest candidate corrections. When processing words marked for correction, we select the most probable candidate according to SymSpell as the word's correction. We ignore any candidate for which the edit distance is superior to 1. The word is not corrected if SymSpell cannot find candidates suiting these conditions.

We rely on three resources to find candidate corrections. First, we rely on the French dictionary provided by SymSpell, which has been created by combining the French part of the Google Books N-gram viewer dataset[3] with the French Hunspell dictionary[4]. Secondly, we rely on a dictionary of French words from 19th and 20th century documents, produced from the Ground Truth dataset of the ICDAR2017 competition for Post-OCR Text Correction (Chiron et al., 2017). We produced this dictionary with the dictionary creator tool provided in the Python implementation of the SymSpell algorithm. Finally, we rely on a list of French stopwords available on GithHub[5]. The code of our implementation is available on GitHub[6].

[2]The algorithm is available on GitHub: https://github.com/wolfgarbe/SymSpell
[3]http://storage.googleapis.com/books/ngrams/books/datasetsv2.html
[4]https://hunspell.github.io/
[5]https://github.com/stopwords-iso/stopwords-fr
[6]https://github.com/nicolasgutehrle/emontalproject

Chapter 11

Understanding the logical structure of documents

Table of contents

The textual content of the documents in our dataset is presented in the XML ALTO format, which describes the reading order and the physical structure of the document. However, this format does not provide the logical structure of the document, i.e. its organisation into logical elements such as headers, titles, paragraphs, or sentences.

We propose an approach to the Logical Layout Analysis (LLA) task, which aims at assigning a logical label to `TextBlock` and `TextLine` tags in XML ALTO documents. The details and the evaluation of this method were previously presented in Gutehrlé and Atanassova (2021b, 2022). We describe the dataset which we built from the EMONTAL corpus in order to develop and evaluate our method in Section 11.1, before presenting our approach and its implementation in Section 11.2 and in Section 11.3 respectively. Furthermore, we present in Section 11.4 the EMONTAL XML format, a light XML format which describes the logical structure of documents. We convert the XML documents in the EMONTAL corpus in this new format, in order to prepare them for the application of NLP processes. The implementation of the methods we propose is available on GitHub[1].

[1] https://github.com/nicolasgutehrle/emontalproject

11.1 Dataset

We propose a dataset to develop and evaluate our approach to the LLA task. We have built this dataset by selecting documents from the EMONTAL corpus where the `nqamoyen` attribute, i.e the OCR quality measure, is greater or equal than 90%. This dataset is publicly available on Zenodo (Gutehrlé & Atanassova, 2021a). For each selected XML ALTO document, we store two CSV files, one for the annotation of `TextBlock` tags and another for the annotation of `TextLine` tags. The dataset was manually annotated by a single annotator, then split into a train and a test set. Table 11.1 shows the distribution of collections, documents, pages and `TextBlock`, `TextLine` and `String` tags in the train and test sets.

	Collections	Issues	`TextBlocks`	`TextLines`	`Strings`	Pages
Train	15	48	4,608	51,815	338,583	368
Test	6	6	1,445	8,836	63,343	52
Total	21	54	6,053	60,651	401,926	420

Table 11.1: Distribution of collections, documents, pages and `TextBlock`, `TextLine` and `String` tags in the train and test sets

We divided the documents into three layout categories: **1c, 2c** and **3c+**. Our aim is to evaluate how our processing pipeline is able to adapt to the various layouts in the EMONTAL corpus. Table 11.2 shows the distribution of documents in the train and test datasets across the three layout categories.

1c : documents where the text is displayed in one column, as in Figure 9.5

2c : documents where the text is displayed in two columns, as in Figure 9.2, 9.3 and 9.4

3c+ : documents where there are at least 3 columns of text, as in Figure 9.1

Layout	Train	Test
1c	18	2
2c	5	2
3c+	25	2
Total	48	6

Table 11.2: Document distribution per layout categories in the Logical Layout Analysis train and test sets

We have defined the annotation tagset presented in Table 11.3 to annotate the `TextBlock` and `TextLine` tags of the dataset. This tagset allows to determine the logical layout of the documents,

by annotating each `TextBlock` and `TextLine` elements with one of the labels that we have defined.

Label	Text	Firstline	Title	Header	Other
`TextBlock`	X		X	X	X
`TextLine`	X	X	X	X	X

Table 11.3: `TextBlock` and `TextLine` tags annotation tagsets

The label *Firstline* must be understood as "first line of the paragraph". Thus, any `TextLine` tag labelled *Firstline* indicates the beginning of a paragraph. A small portion of the `TextBlock` and `TextLine` tags correspond to elements that are not relevant for our study, such as images, tables or advertisement. Those elements are labelled as *Other* and are ignored for the evaluation. Table 11.4 shows the label distribution in the datasets, that was obtained after the manual annotation of all `TextBlock` and `TextLine` elements.

		Train		Test	
	Label	Count	Proportion	Count	Proportion
`TextBlock`	Header	333	7.380 %	53	3,860 %
	Other	1,686	37.367 %	128	9.322 %
	Text	2,064	45.744 %	1,102	80.262 %
	Title	429	9.507 %	90	6.554 %
`TextLine`	Firstline	9,785	18.921 %	1,563	17.840 %
	Header	740	1.430 %	115	1.312 %
	Other	3,098	5.990 %	201	2.294 %
	Text	36,272	70.138 %	6,648	75,881 %
	Title	1,820	3.519 %	234	2.670 %

Table 11.4: `TextBlock` and `TextLine` tags label distribution in the Logical Layout Analysis train and test sets

11.2 Method

Our approach to the LLA task consists in two main stages:

1. *TextBlock classification*

2. *TextLine classification*

The *TextBlock classification* stage attributes logical labels to `TextBlock` tags. A `TextBlock` tags is only processed if it doesn't already have a *type* attribute, which is the case of most `TextBlock` tags in our dataset, as explained in Section 9.1. This stage updates the XML ALTO document by adding a *type* attribute to `TextBlock` tags with the predicted logical labels. The *TextLine classification* stage attributes logical labels to `TextLine` tags, based on the extracted features and the output of the `TextBlock` annotation step.

We extract and calculate sets of geometric, morphological and semantic features from the XML ALTO document at the `TextLine`, `TextBlock` and Document levels. These features, their descriptions and levels are presented in Table 11.5. This feature set is then used for in the rule-based implementation of our LLA approach which we describe in the following section. We also exploit this feature set in the rule-learning and machine-learning implementation of of our approach, which we describe in Chapter 12.

11.3 Implementation

We propose a rule-based implementation of our approach to the LLA task. Its main stages are shown in Figure 11.1.

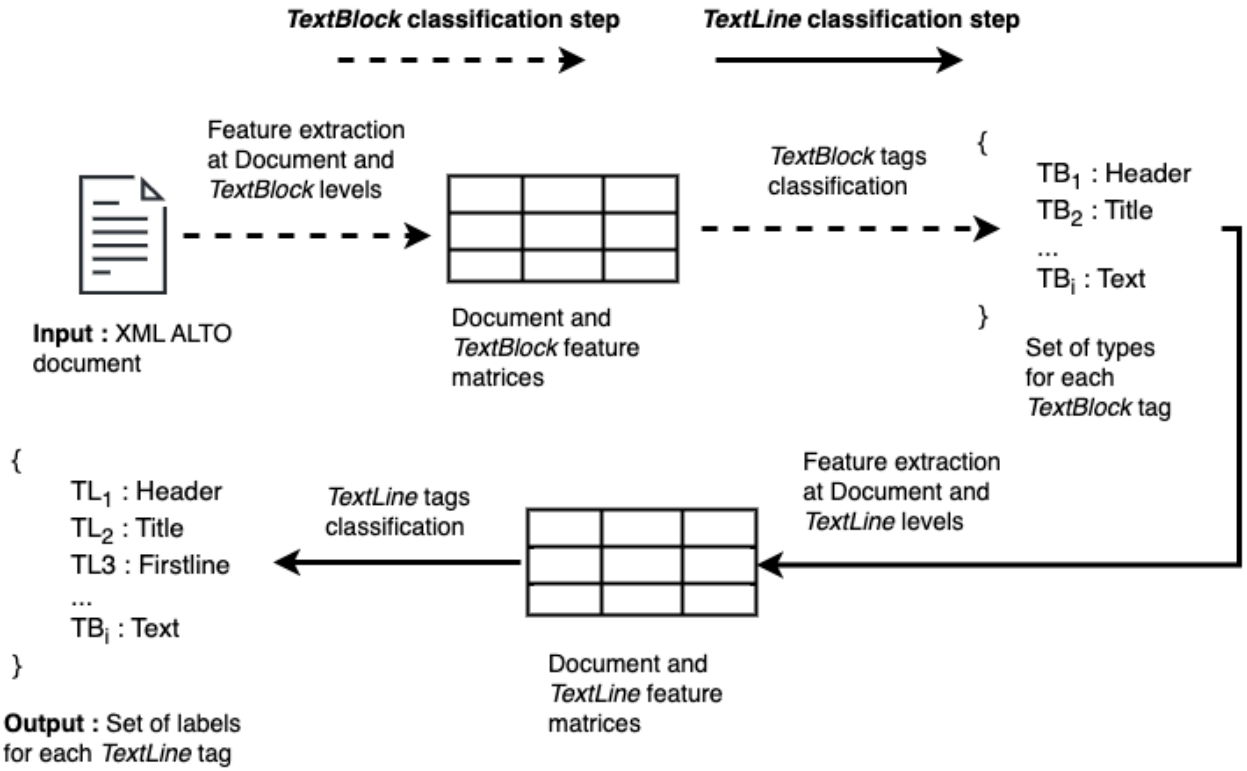


Figure 11.1: Main stages of the implementation of our Logical Layout Analysis approach

	Feature	Description	TextLine	TextBlock	Document
1	*page*	page number of the page containing the element	X	X	
2	*blockType*	type of the block	X		
3	*wordCount*	number of words	X	X	
4	*precedingSpace, followingSpace*	spaces above and below the element	X	X	
5	*height, width*	height and width values of the line	X		
6	*hpos, vpos*	coordinates of the line on the page, i.e. its horizontal and vertical position	X		
7	*diffHpos*	difference between *hpos* and the median *hpos* value in the block	X		
8	*firsthpos, firstvpos*	coordinates of the first line of the block		X	
9	*lasthpos, lastvpos*	coordinates of the last line of the block		X	
10	*linecount*	number of lines in the block		X	
11	*wordRatio*	number of words by line		X	
12	*medHeight, medWidth*	median line height and line width		X	X
13	*medHpos, medVpos*	median *hpos* and *vpos* values in the block		X	
14	*medWordCount, medLineSpace*	median number of words by line and the median space between lines in the block		X	X
15	*medBlockHeight, medBlockWidth*	median line height and block height and width			X
16	*medBlockSpace*	median space value between blocks			X
17	*thirdQuartileLineSpace*	third quartile of line space values in the document			X
18	*medWordRatio, medLineCount*	median number of words by line and median number of line by block in the document			X
19	*capitalProp, digitProp* *nonAlphaProp*	proportion of capital letters and digits proportion of non-alphanumeric characters	X	X	
20	*stwCapital, stwDigit*	True if the line starts either by a capital letter or a number, False otherwise	X		
21	*endsPunct*	True if the line ends with a punctuation, False otherwise	X		
22	*headerMark1*	True if the element contains the word "Page" or a dash sign. False otherwise.	X	X	
23	*headerMark2*	True if the element contains a date, a currency, an address, False otherwise	X	X	
24	*simTitle*	similarity of the line with the title of the document, calculated by the Levenshtein distance	X		
25	*simHeaderSet*	highest similarity of the line with the words contained in the header words set, calculated by the Levenshtein distance	X		

Table 11.5: Features extracted at the `TextLine`, `TextBlock` and Document level for our approach to the Logical Layout Analysis task

Our implementation is written in the Python programming language. We use the *Beautiful-Soup* library to manipulate documents in the XML ALTO format. The features we extract for the `TextBlock` and `TextLine` classification tasks are stored in *pandas* DataFrame. The textual content of `TextLine` tags is extracted from their *content* attribute, or from their *subs_content* attribute when available. The textual content of `TextBlock` tags is defined as the concatenation of the textual content of all the `TextLine` tags they contain.

We extract the *page, height, width, hpos* and *vpos* features from the attributes of `TextBlock` or `TextLine` tags. We calculate the *precedingSpace* and *followingSpace* features as follows: given a `TextLine` $line$ and the following `TextLine` $nextline$, the interline space is measured as the difference between the *vpos* values of $line$ and $nextline$.

We use regular expressions to extract the *headerMark1, headerMark2*, *stwCapital*, *stwDigit* and *endsPunct* features. The *headerMark1* feature is True if a `TextLine` or `TextBlock` tag contains the mention of a page. Similarly, the *headerMark2* feature is True if a `TextLine` or `TextBlock` contains the mention of either a date, a currency or an address. We use the following regular expressions to detect these mentions:

Page : `([---][^\d]*\d[^---]*[---]|Page\s*\d)`

Date : `(janvier|fevrier|mars|avril|mai|juin|juillet|aout|septembre|octobre|novembre|decembre)\s`

Currency : `(centime|franc[s\s:\.,]|frs?[\.\s]|fr$)`

Address : `(rue|avenue|boulevard|impasse|faubourg|quai)`

stwCapital : `([^a-zA-Z\d\s:]*[a-z]?)\s*[A-ZÈÉÊËÏÎÔÂÀ]`

stwDigit : `^\d`

endsPunct : `[\.!?\)]$`

We calculate the *simTitle* and *simHeaderSet* features by applying the fuzzy matching method. Fuzzy matching is an approximate string matching method based on fuzzy logic, which measures the similarity between two textual patterns such as strings. We use the implementation of the fuzzy matching method proposed by the *fuzzywuzzy* library for this task. To measure the *simTitle* feature, we calculate the similarity of each `TextLine` tag with the title of the document. Similarly, to measure the *simHeaderSet* feature, we calculate the similarity of the words found in `TextLine` tags with any of the following expressions. These expressions were found by manually observing the headers in the LLA dataset:

- *Rubrique Locale* (Local rubric)

- *Gérant* (Manager)
- *Publicité* (Advertising)
- *Abonnement* (Subscription)
- *Envoyez les fonds* (Send the money)
- *Conservez chaque numéro* (Keep each issue)
- *Rédacteur* (Editor)
- *Directeur* (Director)
- *Numéro* (Issue)
- *Chèque postal* (Postal cheque)
- *Dépôt* (Depot)
- *Achat-Vente-Echange* (Buy-Sell-Exchange)
- *Annonce* (Advertisement)
- *Imprimerie* (Printing)
- *En vente partout* (On sale everywhere)
- *Paraissant* (Published)

We normalize numerical features by rounding the value to the lower five. For instance, if the *precedingSpace* feature equals 62, it is normalized to 60. Similarly, if the *precedingSpace* feature equals 67, it is normalized to 65. We normalize every numerical features, excepted the *wordCount, capitalProp, digitProp, nonAlphaProp, simTitle, simHeaderSet* features.

The rule-based implementation of our approach rely on two rule-sets, one for the categorisation of `TextBlock` tags and another for the categorisation of `TextLine` tags, which we have manually conceived by observing the annotated instances of each label in the LLA dataset. These rule-sets exploit the set of features described in the previous section to determine the logical label of `TextBlock` and `TextLine` tags.

Table 11.6 presents the rules for annotating `TextBlock` tags and solve conflicting annotations, where B is a `TextBlock` in a document D. Similarly, Table 11.7 presents the rules for annotating `TextLine` tags and solving conflicting annotations, where B is a `TextBlock` in a document D, and L is a `TextLine` in B.

Each rule is applied independently from each other when processing a `TextBlock` or a `TextLine` tag. Thus, the same tag may be annotated by more than one rule, resulting in conflicting annotations. For instance, a `TextBlock` tag may be categorised as *Title* by one rule, and categorised as *Header* by another. Each conflicting annotation is processed by a set of conflict resolution rules, in order to keep only one possible logical label for a tag.

We implement the rules for `TextBlock` and `TextLine` classification presented in Table 11.6 and Table 11.7 as sets of logical operations on the *pandas* DataFrames. Conflicting annotation rules are applied at the end.

Rule	Condition	Label
1	*(B.linecount > D.medLineCount)* or *(B.wordCount > D.medWordCount / 3)*	Text
2	Previous and next `TextBlocks` are Text and *(B.linecount < D.medLineCount)* and *(B.medHeight < D.medBlockHeight)*	Text
3	Previous and next `TextBlocks` are Text and B is not Text and *(B.linecount < 4)* and *(B.precedingSpace > D.medBlockSpace) or (B.followingSpace > D.medBlockSpace)*	Title
4	*B.page = 1* and for any of the first 30 lines of B: *simHeaderSet > 0.9* or *simTitle > 0.9* or *headerMark1* or *headerMark2* or *ctnTotal*	Header
5	*B.page > 1* and for any of the first 4 lines of B: *simHeaderSet > 0.9* or *simTitle > 0.9* or *headerMark1*	Header
6	**Conflicting annotation** between Header and Text or Title:	
	(B.linecount < 15) and *(B.wordCount < 50)*	Header
	Otherwise	Text / Title
7	**Conflicting annotation** between Text and Title:	
	B.medHeight > D.medBlockHeight / 2	Title
	Otherwise	Text

Table 11.6: `TextBlock` tag annotation and conflict resolution rules

We have taken into consideration the following observations while designing the rules:

- Text blocks contain relatively more lines and more words than Title or Header blocks
- A Text block should have a smaller height than a Title block. As such, if there is a confusion between Text and Title block, we should use the height of the block to distinguish between the two
- Title blocks contain few lines, usually not more than three. The role of a title is to introduce the topic of a text section, thus a Title block should be surrounded by Text blocks. The space around that block should also be important, in order to stand out with the surrounding blocks
- Title lines are surrounded by relatively more space in order to stand out from other text sections. The smaller the title is, the less important the space around it is

Rule	Condition	Label
1	L.*blockType* = Header	Header
2	L.*blockType* = Title	Title
3	L.*precedingSpace* = 0 and L.*followingSpace* > D.*medLineSpace* and L.*simTitle* < 60 and L.*simHeaderSet* < 60 and L.*stwCapital*	Title
4	L.*wordCount* < B.*medWordCount* and L.*precedingSpace* > D.*thirdQuartileLineSpace* and L.*followingSpace* > D.*thirdQuartileLineSpace*	Title
5	L.*capitalProp* > 10 and L.*wordCount* < B.*medWordCount* and L.*height* < B.*medHeight* and (L.*precedingSpace* > D.*thirdQuartileLineSpace* or L.*followingSpace* > D.*thirdQuartileLineSpace*)	Title
6	L.*diffHpos* > 104 and L.*capitalProp* > 0 and L.*precedingSpace* > D.*medLineSpace* and L.*followingSpace* > D.*medLineSpace*	Title
7	L.*hpos* > B.*medHpos* and L.*diffHpos* < 105 and (L.*stwCapital* or L.*stwDigit*)	Firstline
8	L.*width* < B.*medWidth* and L.*wordCount* < B.*medWordCount* and L.*hpos* < B.*medHpos*	Lastline
9	Previous `TextLine` is LastLine and L.*stwCapital* and L.*followingSpace* < B.*medLineSpace*	Firstline
10	Previous `TextLine` is not Lastline and L.*stwCapital* and L.*precedingSpace* > B.*medLineSpace* and L.*followingSpace* < B.*medLineSpace*	Firstline
11	Previous `TextLine` is not Lastline and L.*stwCapital* and L.*hpos* > B.*medHpos*	Firstline
12	None of the rules 1-9 above is True	Text
13	**Conflicting annotation** between Header and other label: Previous `TextLine` is Header and next `TextLine` is Header	 Header
14	**Conflicting annotation** between Title and FirstLine: L.*followingSpace* < B.*medLineSpace* and L.*capitalProp* < 15 Otherwise	 Title Firstline

Table 11.7: `TextLine` tag annotation and conflict resolution rules

- Small titles (e.g. sections in an article) usually contain more capital letters then the rest of the text and are center-aligned
- Header blocks are only located at the top of a page, generally in the first four lines of a page. Small blocks at the top of a page are most likely to be Headers
- On the first page of a document, the header can be much longer because it contains more information. We consider that it can be up to 30 lines. This value has been set empirically
- Headers contain very specific information about the document, such as its title, its price, the date or the publisher's name. This information is displayed with keywords and sentences that are recurrent across multiple pages and documents.
- The first line of a page or immediately after a Title should be labelled Firstline, if it starts with a capital letter
- The first line of a paragraph always starts with a capital letter, and is often indented, i.e. the value of their horizontal position is greater than other `TextLines` in the block
- Firstlines that are not indented can be identified if the line that precedes them is shorter, indicating the end of the previous paragraph
- `TextLine` tags that appear between two Header lines should also be annotated as Header
- `TextLine` tags that are contained in a Title or Header block should inherit this annotation

11.4 Converting XML documents

We propose to convert the original XML documents in the EMONTAL corpus into a custom lighter XML format called EMONTAL. This format describes the logical structure of document, and is adapted to the application of NLP methods. These lighter XML documents are structured according to the schema shown in Figure 11.2. The complete XML Schema Definition (XSD) is shown in Annexe 19.4.

The XML document is contained in a `document` tag, which is divided into a `metadata` tag and a `content` tag. The `metadata` tag contains the metadata of the documents. The tags contained in the `metadata` tag are a selected subset of the `oai` tag from the original XML document, and follow the Dublin Core format. Table 11.8 provides a description of each tag in the `metadata` tag.

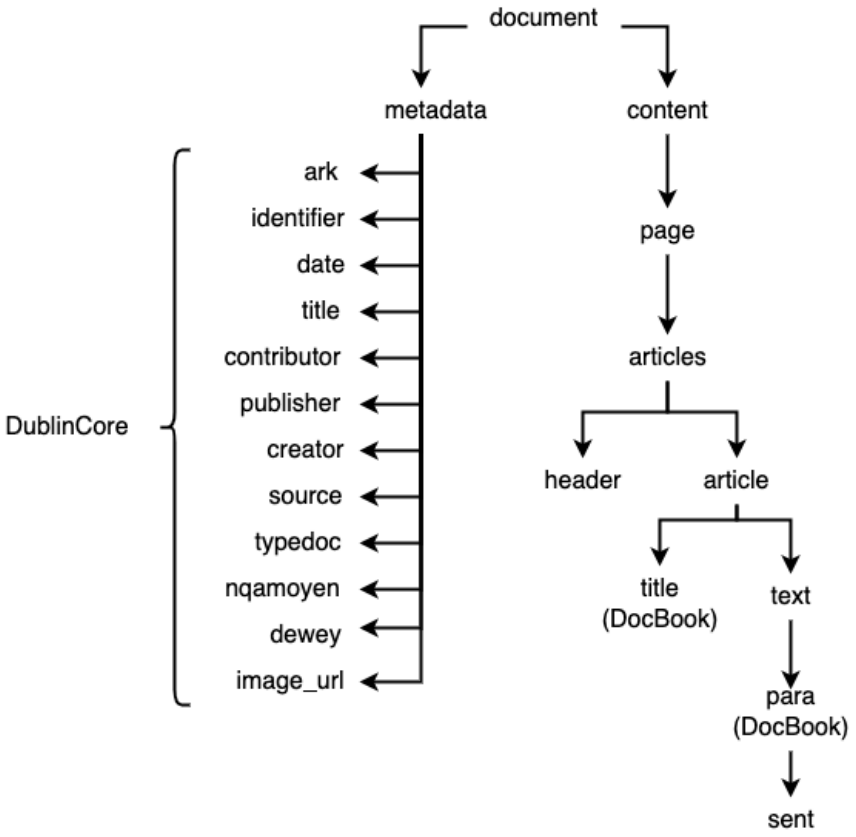


Figure 11.2: Structure of the EMONTAL XML format

Tag name	Description
ark	the ark identifier of the document, as used in Gallica
identifier	url of the document on Gallica
date	the publication date of the document. This value can be in two formats, either YYYY-MM or YYYY-MM-DD
title	the title of the document
contributor	person who contributed to the creation of the document
publisher	the original publisher of the document
creator	the creator of the document
source	the original storage of the document
typedoc	the type of the document
nqamoyen	the OCR quality score
dewey	the category of the document, according to the Dewey Decimal Classification (DDC)
image_url	the url to the scanned document

Table 11.8: Set of tags contained in the `metadata` tag from the EMONTAL XML format. These tags follow the Dublin Core format

The `content` tag provides the logical structure of the document and its textual content. It is inspired by the XML DocBook format[2], which allows to describe the logical structure of documents in a light and flexible format. Table 11.9 provides a description of each tag in the content tag. Tags that can be found in the original XML DocBook tagset are marked by the * symbol.

Every tag, except the `articles` tag, has an *id* attribute that identifies it in a unique way in the document. Articles that are present in multiple pages have the same *id* attribute. *para* tags have a *block_id* attribute, which links them to their corresponding `TextBlock` tag in the original XML ALTO document.

Tag name	Description
page	the content of one page of the document.
header	the content of the page's header. This tag may often be empty
articles	contains all articles in the current page. Each article is contained in an *article* tag
article	contains the content of an article
*title	the title of the article
text	the textual content of the article
*para	a paragraph in the text
sent	sentence in a paragraph

Table 11.9: Set of tags contained in the `content` tag from the EMONTAL XML format. Tags that can be found in the original XML DocBook tagset are marked by the * symbol

We propose a conversion pipeline written in the Python programming language to convert the original XML documents of the EMONTAL corpus into the XML format. This pipeline consists in four steps, which are shown in Figure 11.3.

1. *logical layout analysis* step
2. *article identification* step
3. *xml construction* step
4. *sentence segmentation* step

The *logical layout analysis* step applies our approach for LLA described in Section 11.2 on the `ocr` tag of XML documents, i.e. on XML ALTO documents. To collect the textual content of the XML ALTO document, we extract the values of the *content* and *subs_content* attributes of `String` tags.

The *article identification* step aims at segmenting the document into articles. The periodicals in our corpus are made of multiple articles. However, the XML ALTO documents do not indicate

[2] https://docbook.org/

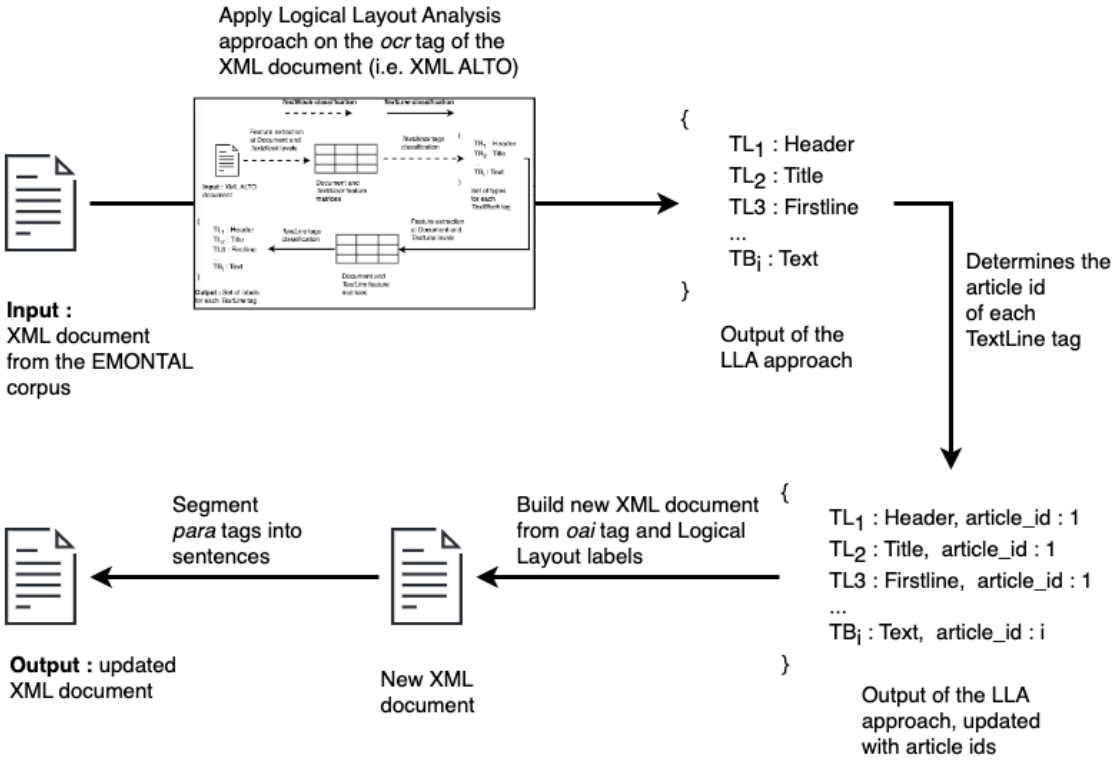


Figure 11.3: Main steps of the XML conversion pipeline

the beginning nor the ending of these articles. We assume that an article consists in a title, followed by textual content. Thus, we first assign a unique article identifier to any `TextLine` tag or any group of consequent `TextLine` tags categorised as Title during the *logical layout analysis* step. Secondly, we propagate the article identifier to any subsequent `TextLine` tags, until we reach another article identifier. By propagating the values, we may accidentally add an article identifier to a Header line, which are not part of any articles. Thus, we remove the article identifier from every Header line.

The *xml construction* step builds the new XML document. We start by creating the `document`, `metadata` and `content` tags. We fill the `metadata` tag by adding it the tags described in Table 11.8, which we extract from the `oai` tag of the original XML document. Secondly, we add the tags described in Table 11.9 to the `content` tag.

The conversion step may produce contiguous sequences of identical tags, such as contiguous sequences of `Text` or `Title` tags. We merge these continuous sequences of tags into a single tag. For instance, a sequence of `Title` tags will be merged into a single `Title` tag.

The *sentence segmentation* step produces the `sent` tags by segmenting the textual content of `para` tags into sentences. NLP frameworks such as *spaCy* perform the sentence segmentation task on the dependency analysis of the text. However, since our corpus contains errors from the OCR process, the dependency analysis may yield incorrect results and impact the sentence segmentation

step. Thus, our approach to the sentence segmentation task is rule-based.

By default, we segment the content of the paragraph into sentences where any of the *.!?* punctuations occur. A set of four exception rules are given to prevent the algorithm from splitting the text at an incorrect position. These exception rules were crafted by observing the EMONTAL corpus. Table 11.10 shows an example of each exception rule, as well as its implementation.

Rule	Example	Implementation
Nota Bene	Nota., N.-B., N.B.	`(N\.\-?B\.|Nota.)`
Post-scriptum	P.S., P.-S., p.s., p.-s.	`\[Pp\]\.\-?\[Ss\]\.`
Mister, miss	M., Mr., Mme., Mlle., MM.	`( )\[Mm\](me|lle|M)? ?\.`
Stop not followed by a capital letter	. après, . monsieur	`\w\. ?(\[a-z\]|,)`

Table 11.10: Exception rules for sentence segmentation

Figure 11.4 and Figure 11.5 show an example of the `metadata` and `content` tags following the EMONTAL XML schema.

```xml
<?xml version="1.0" encoding="utf-8"?>
<document>
 <metadata>
  <ark>
   bpt6k5839946n
  </ark>
  <identifier>
   https://gallica.bnf.fr/ark:/12148/bpt6k5839946n
  </identifier>
  <date>
   1900
  </date>
  <title>
   Mémoires et documents inédits pour servir à l'histoire de la Franche-Comté / publiés par l'Académie de Besançon
  </title>
  <contributor/>
  <publisher>
   Académie de Besançon (Besançon)
  </publisher>
  <languague/>
  <creator>
   Académie des sciences, belles-lettres et arts de Besançon et de Franche-Comté. Auteur du texte
  </creator>
  <source>
   Bibliothèque nationale de France, département Collections numérisées, 2008-216649
  </source>
  <typedoc>
   fascicule
  </typedoc>
  <nqamoyen>
   99.98
  </nqamoyen>
  <dewey>
   0
  </dewey>
  <image_url>
   https://gallica.bnf.fr/ark:/12148/bpt6k5839946n/highres
  </image_url>
 </metadata>
```

Figure 11.4: Excerpt of the metadata section of a document in the EMONTAL XML format in our corpus (*Mémoires et documents inédits pour servir à l'histoire de la Franche-Comté*, 1900)

```
<content>
 <page id="1">
  <header block_id="PAG_00000001_TB000005" id="header_1">
   MEMOIRES
   ET
   DOCUMENTS INÉDITS
   POUR SERVIR A L' HISTOIRE
   DE LA FRANCHE-COMTÉ
  </header>
  <articles>
   <article id="article_01">
    <title id="title_01">
     T. IX.
    </title>
    <text id="text_01"/>
   </article>
  </articles>
 </page>
 <page id="3">
  <header block_id="PAG_00000003_TB000004" id="header_3">
   ET
   DOCUMENTS INÉDITS
   POUR SERVIR A L'HISTOIRE
  </header>
  <articles>
   <article id="article_01">
    <title id="title_01">
     T. IX.
    </title>
    <text id="text_01">
     <para block_id="PAG_00000003_TB000001" id="para_1">
      <sent id="sent_1">
       MÉMOIRES
      </sent>
     </para>
```

Figure 11.5: Excerpt of the content section of a document in the EMONTAL XML format in our corpus (*Mémoires et documents inédits pour servir à l'histoire de la Franche-Comté*, 1900)

Chapter 12

Evaluation of the Logical Layout Analysis approach

Table of contents

In this chapter, we evaluate our approach to the Logical Layout Analysis task described in Section 11.3. In the previous chapter, we presented a set of features and a set of manually created rules for the LLA task. We evaluate the performance of our approach in the following manner:

- we use as evaluation dataset the test set that is presented in Section 11.1
- we evaluate our rule-based approach in terms of Precision, Recall and F1 score (see Section 12.1)

As the set of manually created rules that we propose might not be optimal, we evaluate two other implementations of our approach to the LLA task: a rule-learning approach (see Section 12.2.1)

and a machine-learning (ML) approach exploiting the same feature set (see 12.2.4). Our aim is to show how the results of our rule-based method compare to those of other possible approaches based on machine-learning. Figure 12.1 gives an overview of the structure of this chapter.

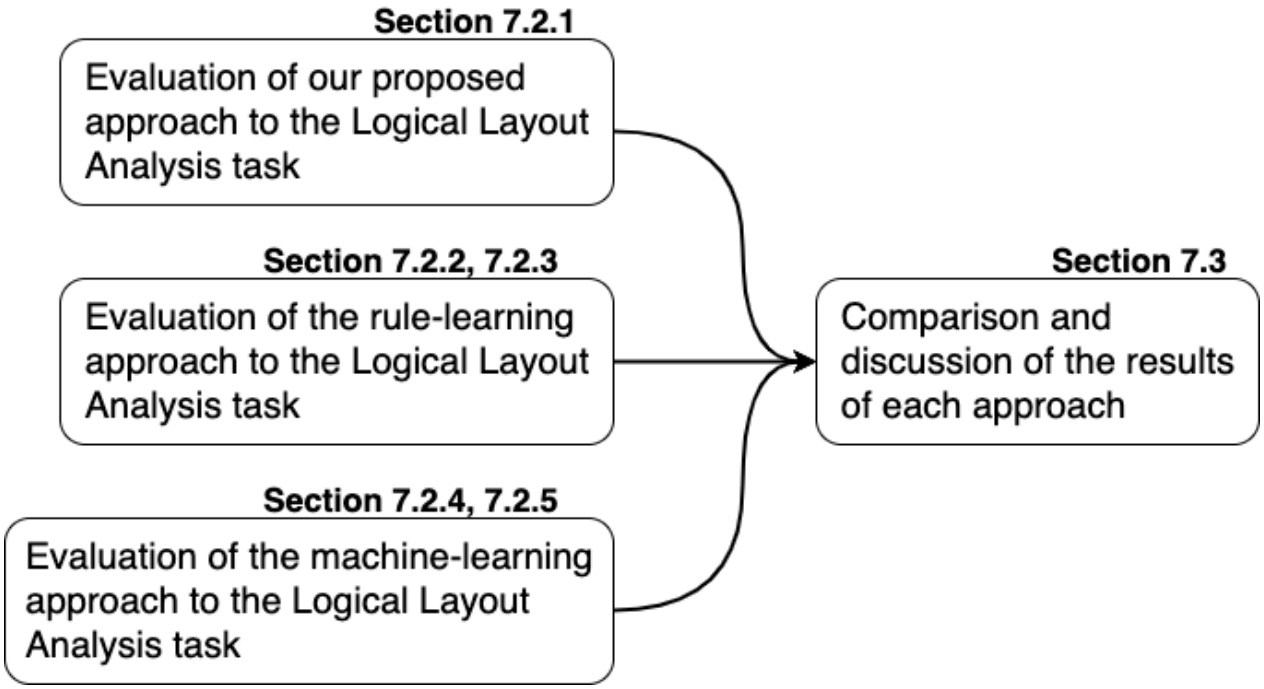


Figure 12.1: Structure of Chapter 12

12.1 Evaluation of our rule-based Logical Layout Analysis approach

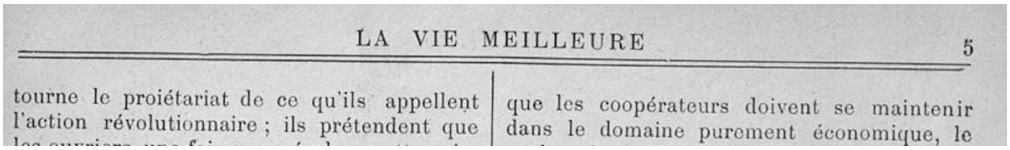
We evaluate our approach to the LLA task described in Section 11.3 on the test set described in Section 11.1. Table 12.1 shows the Precision, Recall and F1 scores of the system on `TextLine` and `TextBlock` annotation tasks.

		Text			Title			Firstline			Header		
	Cat	P	R	F1	P	R	F1	P	R	F1	P	R	F1
TextBlock	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
TextLine	1c	**0.979**	0.986	**0.983**	0.354	0.720	0.473	0.943	0.854	0.895	0.909	**0.598**	**0.721**
	2c	0.961	**0.995**	0.978	**0.746**	**0.765**	**0.747**	**0.955**	0.859	0.902	**1.000**	0.118	0.197
	3c+	0.975	0.992	**0.983**	0.703	0.702	0.702	0.952	**0.877**	**0.913**	0.500	0.400	0.444
	Mean	0.969	0.991	0.979	0.595	0.733	0.639	0.949	0.861	0.902	0.803	0.348	0.435

Table 12.1: Precision, Recall and F1 scores of the rule-based system on the `TextBlock` and `TextLine` classification tasks

The `TextBlock` classification rules perform the best on documents from the 2c layout category. The classification of Title blocks achieves an F1 score of 0.947 on documents from the 2c category and a mean F1 score of 0.610. The classification of Header blocks achieves a good Precision score of 0.726, but achieves a lower Recall score of 0.298. The `TextLine` classification rules perform also the best on documents from the 2c category. The classification of Title lines performs the worse on 1c documents, and achieves a mean F1 score of 0.639 for all layouts. The classification of Firstline achieves very good F1 score that are above 0.902. Finally, the classification of Headers achieves a good Precision score of 0.803 but a lower Recall score of 0.348.

The low Recall scores for the classification of Title and Header labels suggest that the current rules are insufficient and must be extended to capture more types of Titles and Headers. A first common error in the `TextLine` classification step comes from errors in the `TextBlock` classification step, as any line in a Title or a Header block is automatically annotated with the same label. For instance, Figure 12.2 shows two elements of a Header block that were incorrectly annotated: the Header "*La Vie Meilleure* (The Better Life), which is the name of the document, is annotated as Text, whereas the page number "5" is incorrectly annotated as a Title. We can correct such errors by refining or adding more rules to identify Header blocks. More specifically, we can improve the overall performance of the system by first improving the `TextBlock` classification rules.


LA VIE MEILLEURE 5

tourne le proiétariat de ce qu'ils appellent l'action révolutionnaire ; ils prétendent que

que les coopérateurs doivent se maintenir dans le domaine purement économique, le

Figure 12.2: Example of two incorrect annotations: the header "*La Vie Meilleure*" (The Better Life), which is the name of the document, is annotated as Text, whereas the page number "5" is incorrectly annotated as a Title

A second common type of error is the confusion between Titles lines and Firstlines. Most Titles mislabelled as Firstline are in fact short subsection titles. As such, they are similar to other text lines in terms of typography, and are hard to detect with the features we currently use. This confusion happens mainly in documents from the **2c** and **3c+** categories. Figure 12.3 shows an example of such an error, where the Title line "*Mort du Cardinal Ferrata*" (Death of Cardinal Ferrata) was incorrectly annotated as Firstline. The font of this title is different from that of the surrounding text, and it is written in bold. However, we currently do not extract such features. Thus, the rule-based system is not able to properly detect such small titles.

Figure 12.3: Example of incorrect annotation: the title line "*Mort du Cardinal Ferrata*" (Death of Cardinal Ferrata) is incorrectly annotated as Firstline

Other lines mislabelled as Titles are one-line paragraphs such as greetings or signatures, or the beginning of a text section. Such lines have properties similar to Titles, being surrounded by important spaces and being either center or right-aligned. Figure 12.4 shows an example of such an error, where the Text line "*Le Secrétaire*" (The Secretary) is incorrectly annotated as a Title line. This error is caused by the important space surrounding the text, as well as its right-alignment. Extracting features about the font style of the line (bold, italics) and its alignment (left, center, right-aligned) could help solve these errors.

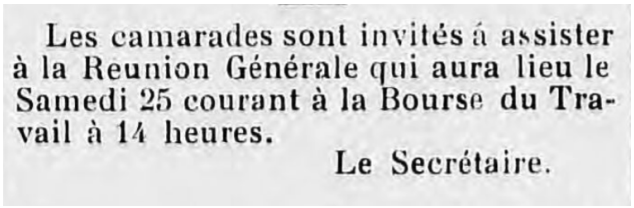


Figure 12.4: Example of incorrect annotation: the Text line "*Le Secrétaire*" (The Secretary) is incorrectly annotated as a Title line

In conclusion, the results obtained by our rule-based approach to the LLA task show its ability to identify the logical structure of documents from their representation in the XML ALTO format. Determining the logical label of `TextBlock` tags seems to be helping in determining the logical label of `TextLine` tags. However, an incorrect categorisation of a `TextBlock` tag implies an incorrect categorisation of the `TextLine` tags within. Thus, improving the `TextBlock` categorisation rules could greatly improve the general performances of our rule-based approach. Moreover,

our approach seems to struggle to identify titles and headers in the document. These issues could be solved by extracting new features and adding more rules to detect these logical labels.

12.2 Comparison of our approach with existing approaches

We compare our rule-based approach to the LLA task to two existing approaches: a rule-learning approach and a machine-learning approach. Our aim is to show how our method compares to pre-existing machine-learning approaches.

Rule-learning algorithms are machine-learning algorithms which aim at learning a rule-set from data. For our experiment, we have selected the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm (Cohen, 1995). We select this algorithm for its ability to produce human-readable rule-sets, and for being considered as the state of the art of rule-learning algorithms (Sammut & Webb, 2017). For a detailed explanation of the RIPPER algorithm, see Cohen (1995) and Gutehrlé and Atanassova (2022). We present the implementation of the rule-learning approach to the Logical Layout task in Section 12.2.1, and its evaluation in Section 12.2.3.

In Section 12.2.4, we present the implementation and the evaluation of the machine-learning approach to the Logical Layout task. We first compare the performances of the Support Vector Machine (SVM), XGBoost, Random Forest, AdaBoost and Gradient Boosting models on the LLA test set. Our aim is to determine which model among these five is the most adapted to the LLA task. Then, we perform a grid search on the best model to find its best set of hyper-parameters, before evaluating it on the LLA test set. We present the implementation of the machine-learning approach to the Logical Layout task in Section 12.2.4, and its evaluation in Section 12.2.5.

12.2.1 Rule-learning approach

We use the implementation of the RIPPER algorithm from the *wittgenstein* Python library. We use the set of features described in Section 11.5 to train the RIPPER model on the train set of the LLA dataset described in Section 11.1. We exclude feature 12 to 18 from the training set, since they are calculated from other features, and thus do not provide any new data.

The RIPPER implementation in *wittgenstein* only deals with binary classification. Thus, we have trained three RIPPER models for the `TextBlock` annotation task and four RIPPER models for the `TextLine` annotation task. Each model is trained to recognise a logical label for the `TextBlock` and `TextLine` annotation tasks, or to predict the label Other instead. We combine these RIPPER models in an ensemble classifier to predict the probability of a `TextBlock` or `TextLine` tag belonging to each label. The most probable label is then assigned to the tag.

For each trained RIPPER model, we performed a grid search with the hyper-parameters shown

in Table 12.2. Table 12.3 shows the best hyper-parameter combinations for each RIPPER model we have trained for the `TextBlock` and `TextLine` annotation tasks.

Hyper-parameter	Description	Default	Grid search values
prune_size	the size of the prune set	.33	.25, .33, .50
k	the number of optimisations to run	2	1, 2
dl_allowance	the allowed size for description length	64	32, 64, 128
n_discretize_bins	specific to the *wittgenstein* implementation. Automatically detects and discretises continuous features in the training set. Controls the size of each bin.	10	5, 10, 20, 30

Table 12.2: Hyper-parameters and grid search values for training the RIPPER models

	Hyper-parameter	Header	Title	Firstline	Text
TextBlock	prune_size	.5	.33		.25
	k	2	1		2
	dl_allowance	64	64		64
	n_discretize_bins	10	10		10
TextLine	prune_size	.33	.25	.33	.5
	k	1	1	1	2
	dl_allowance	64	64	64	64
	n_discretize_bins	10	10	10	10

Table 12.3: Best hyper-parameter combinations for the RIPPER models trained on each logical label for the `TextBlock` and `TextLine` annotation tasks

12.2.2 Analysis of the learned rules and comparison with our rule-based approach

Table 12.4 shows the annotation rules learned by the RIPPER algorithm for the `TextBlock` annotation task, where B is a `TextBlock` in a document D. RIPPER learned 26 rules, whereas we produced 7 in our rule-based system, including the conflict resolution rules.

Rule	Condition	Label
1	*B.capitalProp => 55* and *80 <= B.followingSpace <= 135* and *135 <= B.width <= 368.25*	Title
2	*B.height => 95* and *801.25 <= B.width <= 985* and *5 <= B.capitalProp <= 12.5*	Title
3	*B.capitalProp => 55* and *940 <= B.firstvpos <= 1545* and *B.height <= 30*	Title
4	*B.height => 95* and *485 <= B.firstvpos <= 940*	Title
5	*B.linecount <= 2* and *B.height => 95* and *940 <= B.firstvpos <= 1545* and *79 <= B.wordRatio <= 96*	Title
6	*B.capitalProp => 55* and *B.wordRatio => 149* and *B.digitProp <= 5* and *B.lastvpos <= 900*	Title
7	*B.capitalProp => 55* and *80 <= B.followingSpace <= 135* and *B.linecount <= 2* and *4 <= B.wordCount <= 9*	Title
8	*B.linecount <= 2* and *B.digitProp <= 5* and *B.height => 95* and *175 <= B.followingSpace <= 225*	Title
9	*B.linecount <= 2* and *B.digitProp <= 5* and *135 <= B.followingSpace <= 175* and *65 <= B.height <= 95*	Title
10	*B.linecount <= 2* and *B.digitProp <= 5* and *65 <= B.height <= 95*	Title
11	*B.linecount <= 2* and *B.capitalProp => 55* and *595 <= B.lasthpos <= 1130*	Title
12	*B.linecount <= 2* and *B.digitProp <= 5* and *B.capitalProp => 55* and *80 <= B.followingSpace <= 135* and *5849 <= B.firstvpos <= 7470*	Title
13	*B.linecount <= 2* and *B.digitProp <= 5* and *B.followingSpace => 225* and *B.height => 95* and *4 <= B.wordCount <= 9* and *B.precedingSpace <= 5*	Title
14	*B.height => 95* and *B.followingSpace => 225* and *940 <= B.firstvpos <= 1545* and *4 <= 4 B.wordCount <= 9*	Title
15	*B.linecount <= 2* and *B.digitProp <= 5* and *B.followingSpace => 225* and *B.height => 95* and *61 <= B.wordRatio <= 70*	Title
16	*B.linecount <= 2* and *B.digitProp <= 5* and *135 <= B.followingSpace <= 175*	
17	*B.linecount <= 2* and *B.digitProp <= 5* and *B.followingSpace => 225* and *B.height => 95* and *5 <= B.capitalProp <= 12.5*	Title
18	*B.lastvpos <= 900* and *5 <= B.digitProp <= 80*	Header
19	*B.lastvpos <= 900* and *3015 <= B.firstHpos <= 3640* and *B.capitalProp >= 55*	Header
20	*B.lastvpos <= 900* and *135 <= B.width <= 368.25* and *B.firstvpos <= 485*	Header
21	*B.lastvpos <= 900* and *B.digitProp => 80*	Header
22	*B.lastvpos <= 900* and *12.5 <= B.capitalProp <= 55* and *595 <= B.lasthpos <= 1130*	
23	*B.digitProp <= 5* and *B.capitalProp <= 5* and *B.wordCount => 272*	Text
24	*50 <= B.wordCount <= 272*	Text
25	*23 <= B.wordCount <= 50* and *B.capitalProp <= 5*	Text
26	*B.digitProp <= 5* and *B.capitalProp <= 5* and *45 <= B.precedingSpace <= 80*	Text

Table 12.4: `TextBlock` annotation rules for each `TextBlock` label learned by the three RIPPER models

RIPPER learned 17 different rules to identify Title blocks whereas our rule-based system uses only one. The $linecount$ and $height$ features are the two most frequently used features, as they occur in 11 rules. A recurring condition is that $linecount$ must be inferior to three. This condition is similar to the condition in our rule-based system, where a block must have no more than four lines. This confirms our intuition that a small number of lines is an important factor to detect Title blocks. The $capitalProp$ and $digitProp$ features are also important, as they respectively appear in eight and nine rules. The $capitalProp$ feature must always be above 50 % whereas the $digitProp$ feature must always be inferior or equal to 5 %. Finally, as in our rule-based approach, the $followingSpace$ feature is important, as it appears in nine rules and must always be above a certain threshold. However, RIPPER never uses the $precedingSpace$ feature. This suggests that the RIPPER model only needs to measure the space after a block to identify Title blocks.

RIPPER learned five rules to identify Header blocks, whereas our rule-based system contains two rules. The condition $lastvpos <= 900$ is present in every rule, indicating that the block must be on the top part of the document. The $digitProp$ feature is always greater than 5 %, which suggests that Header blocks always contain numerical values. To detect Header blocks, the RIPPER model exclusively relies on geometric and morphological features, unlike our rule-based system which mostly relies on semantic features such as the $simHeaderset$, $simTitle$, $headerMark1$ or $headerMark2$ features.

RIPPER learned four rules to identify Text blocks, whereas our rule-based system contains two rules. The most important feature is $wordCount$, which appears in three rules. As in our system, this feature must be greater than a small threshold. Here it must be greater than 23 words, whereas it must be greater than one third of the $medWordCount$ feature in our rule-based system. Unlike our system, RIPPER never uses the $linecount$ feature to detect Text blocks. Instead, it uses the $capitalProp$ and $digitProp$ features, which respectively appear in three and two rules. In every rule they appear, these features must be lower or equal than 5 %. Finally, RIPPER uses the $precedingSpace$ feature in only one rule, where the feature must be lower or equal than 80 pixels. This suggests the space before a Text block must be small.

Table 12.5 shows the annotation rules learned by the RIPPER algorithm for the `TextLine` annotation task, where L is a `TextLine` in a document D and B is the `TextBlock` that contains L. RIPPER learned 19 rules whereas we produced 12 in our rule-based sytem, including the conflict resolution rules.

Rule	Condition	Label
1	*L.followingSpace => 75* and *L.stwCapital is True* and *L.blockType=title* and *745 <= L.width <= 970* and *L.vpos <= 985*	Title
2	*L.followingSpace => 75* and *L.stwCapital is True* and *L.blockType=title* and *L.capitalProp <= 5*	Title
3	*L.followingSpace => 75* and *L.stwCapital is True* and *L.blockType=title*	Title
4	*L.precedingSpace => 75* and *L.followingSpace => 75* and *335 <= L.width <= 645* and *20 <= L.simTitle <= 29*	Title
5	*L.followingSpace => 75* and *L.stwCapital is True* and *L.endsPunct is False* and *L.capitalProp <= 5* and *745 <= L.width <= 970* and *L.nonAlphaProp <= 5*	Title
6	*L.precedingSpace => 75* and *L.followingSpace => 75* and *L.stwCapital is True* and *335 <= L.width <= 645*	Title
7	*L.capitalProp => 10* and *335 <= L.width <= 645* and *L.height <= 35* and *505 <= L.hpos <= 1070* and *L.stwCapital is True*	Title
8	*L.precedingSpace => 75* and *L.followingSpace => 75* and *745 <= L.width <= 970*	Title
9	*L.precedingSpace => 75* and *L.followingSpace => 75* and and *L.stwCapital is True* and *645 <= L.width <= 745*	Title
10	*L.capitalProp => 10* and *L.vpos <= 985* and *L.blockType=header*	Header
11	*L.capitalProp => 10* and *L.simHeaderSet => 85.5* and *L.followingSpace => 75* and *335 <= L.width <= 645*	Header
12	*L.capitalProp => 10* and *L.vpos <= 985* and *54 <= L.simHeaderSet <= 60* and *L.precedingSpace <= 35*	Header
13	*L.capitalProp => 10* and *L.simTitle => 42.86* and *L.nonAlphaProp <= 5* and *L.precedingSpace <= 35* and *2550 <= L.hpos <= 3419*	Header
14	*L.capitalProp => 10* and *L.vpos <= 985* and *L.simTitle => 42.86* and *L.nonAlphaProp <= 5* and *335 <= L.width <= 645*	Header
15	*505 <= L.hpos <= 1070* and *L.simHeaderSet => 85.5* and *335 <= L.width <= 645*	Header
16	*L.stwCapital is True* and *970 <= L.width <= 1010*	Firstline
17	*L.stwCapital is True* and *L.capitalProp <= 5*	Firstline
18	*L.stwCapital is False* and *L.stwDigit is False*	Text
19	*1030 <= L.width <= 1050*	Text

Table 12.5: `TextLine` annotation rules for each `TextLine` label learned by the fours RIPPER models

RIPPER learned nine rules to identify Title lines, whereas our rule-based system contains four rules. The most important feature is the $followingSpace$ feature, as it appears in eight rules. As in our system, this feature must always be greater than a specific threshold. Here it must be greater

than 75 pixels, whereas it must be greater than the $medLineSpace$ feature in our system. The $precedingSpace$ feature seems a less important feature in the RIPPER rule-set than in ours, as it only appears in four rules. This suggests that the space following a block is more important than the one preceding it, as is the case for the annotation of Title blocks. Our rules relies mostly on the $blockType$, $precedingSpace$ and $followingSpace$ features to identify Title lines, whereas RIPPER uses these features alongside others, especially $stwCapital$ and $width$ which are both present in seven rules. $stwCapital$ is always True whereas $width$ is always smaller or equal than a specific threshold. This suggests Title lines should be shorter than most lines in the document.

RIPPER learned six rules to identify Header lines. In our system, a line is labelled as Header if it is contained in a Header block. The main feature used by RIPPER is $capitalProp$ which is present in five rules and is always greater or equal than 10 %. This suggests that Header lines should have more capital letters than common lines. The other main features used are $vpos$, $simHeader$ and $width$. $vpos$ is always below 985 pixels, suggesting that the Header line must be on the highest part of the document. Depending on the rules, the $simHeader$ feature must be greater than 55 % or 85 % percent, suggesting the importance of the header word set. The $width$ of the line is always set in a small range between 335 and 645 pixels. This suggests Header lines are smaller than most lines of text in the document.

RIPPER learned two rules to identify Firstlines, whereas our rule-based system contains five rules. As in our system, the main condition is that the $stwCapital$ feature is True. The first line of a paragraph is also often indented. Our rule-based system uses the $hpos$ feature to detect the indentation, whereas RIPPER uses the $width$ feature instead. Finally, RIPPER learned two rules to detect Text lines whereas this is the default annotation in our system. In RIPPER, the main condition is that the $stwCapital$ and the $stwDigit$ features are False or the $width$ of the line is an average value.

The use of the $blockType$ attribute is the main difference between our rule-set and RIPPER's. In our system, the $blockType$ attribute is an important feature, as any line contained in a Header or Title block inherits this label. RIPPER also uses this feature but to a lesser extent. It is only used in three rules for Title annotation and in one rule for Header annotation, and is always used alongside other features. This suggests that the categorisation of `TextLine` tags by the rule-learning approach relies less on the categorisation of `TextBlock` tags than our rule-based system.

12.2.3 Evaluation of the rule-learning approach

Table 12.6 shows the Precision, Recall and F1-scores obtained by the rule-learning approach on the `TextBlock` and `TextLine` classification tasks on the test set.

		Text			Title			Firstline			Header		
	Cat	P	R	F1	P	R	F1	P	R	F1	P	R	F1
TextBlock	1c	0.819	0.990	0.897	0.500	**0.625**	**0.550**				**0.857**	**0.333**	**0.480**
	2c	**0.851**	**1.000**	**0.920**	0.500	0.444	0.471				0.462	0.250	0.324
	3c+	0.852	0.999	**0.920**	**0.679**	0.297	0.413				0.000	0.000	0.000
	Mean	0.841	0.996	0.912	0.560	0.455	0.480				0.440	0.194	0.268
TextLine	1c	0.871	0.881	0.876	0.600	**0.290**	**0.391**	0.644	0.777	0.704	**0.667**	0.053	0.098
	2c	**0.923**	**0.947**	**0.935**	**0.710**	0.268	0.389	**0.754**	**0.877**	**0.811**	**0.667**	**0.065**	**0.118**
	3c+	0.876	0.908	0.892	0.610	0.207	0.309	0.614	0.709	0.658	0.000	0.000	0.000
	Mean	0.890	0.912	0.901	0.640	0.255	0.363	0.671	0.788	0.724	0.444	0.039	0.072

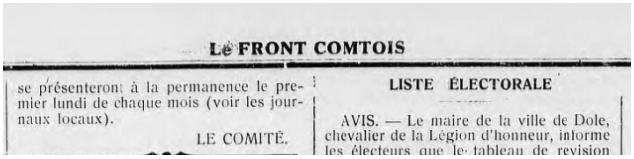
Table 12.6: Precision, Recall and F1 scores of the ensemble of RIPPER model on the `TextBlock` and `TextLine` classification task

The `TextBlock` classification rules perform the best on the 1c layout category. The classification of Text elements achieves the best results, with an F1 score of 0.912 on average. The classification rules for Titles and Headers achieve lower scores however, with F1 scores of 0.480 for Titles and 0.268 for Headers respectively.

The main source of error in the `TextBlock` classification task is a confusion between Header and Title blocks. Every Header block mislabelled as Title contains less than three lines, and vice versa. This suggests that more conditions are required to distinguish between small Header and Title blocks. Finally, many Header blocks are mislabelled as Text, suggesting a lack of rules to detect them.

The `TextLine` classification rules perform the best on the 2c layout category. The classification of Text elements achieves the best results with an F1 score of 0.901 on average. The classification of Firstline comes second with an average F1 score of 0.724. However, the classifications of Title and Header achieve lower F1 scores of 0.363 for Title and 0.072 for Header on average respectively.

Most errors in the `TextLine` classification task concern the Title and Header labels. Many Title lines are incorrectly annotated as Header. For instance in Figure 12.5, the article Title "Liste Electoral" (*Electoral List*) is incorrectly annotated as Header because of its proximity with the page's header. Moreover, most Title and Header lines were mislabelled as Text because of their width. This suggests either that the rules to detect these two labels are not precise enough, or that there are not enough rules in RIPPER's rule set.


Le FRONT COMTOIS

se présenteront à la permanence le premier lundi de chaque mois (voir les journaux locaux).

LE COMITÉ.

LISTE ÉLECTORALE

AVIS. — Le maire de la ville de Dole, chevalier de la Légion d'honneur, informe les électeurs que le tableau de revision

Figure 12.5: Example of incorrect annotation: the article Title "Liste Electorale" (*Electoral List*) is incorrectly annotated as Header, because of its proximity with the page's header

There is also a confusion between Text lines and Firstline. Many Text lines are incorrectly labelled as Firstline because they start with a capital letter. Figure 12.6 shows an example of such confusion, where the second line from the bottom, starting with the text *Petit Semeur*, is incorrectly annotated as Firstline instead of Text line because it starts with a capital letter. Similarly, many Firstlines are mislabelled as Text because of incorrect feature values. These incorrect feature values originate from the XML ALTO documents, and are caused by errors in the OCR transcription.

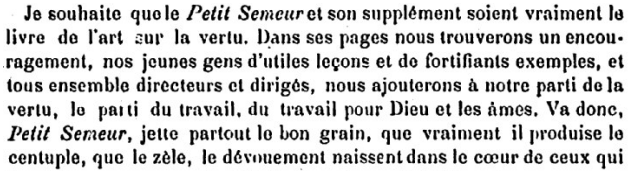

Je souhaite que le *Petit Semeur* et son supplément soient vraiment le livre de l'art sur la vertu. Dans ses pages nous trouverons un encouragement, nos jeunes gens d'utiles leçons et de fortifiants exemples, et tous ensemble directeurs et dirigés, nous ajouterons à notre parti de la vertu, le parti du travail, du travail pour Dieu et les âmes. Va donc, *Petit Semeur*, jette partout le bon grain, que vraiment il produise le centuple, que le zèle, le dévouement naissent dans le cœur de ceux qui

Figure 12.6: Example of incorrect annotation: the second line from the bottom is incorrectly annotated as Firstline because it starts with a capital letter

In conclusion, the RIPPER models we have trained on the LLA task seem to be able to identify lines of texts from XML ALTO documents. However, these models reach rather low Precision and Recall scores for the other logical categories, suggesting that the rule-sets they produce are not sufficient. On average, the RIPPER models reach higher Precision scores than Recall scores, suggesting their ability to learn rules that are precise rather than being general. Moreover, these models are easy and quick to train. Thus, they can help to quickly build an initial rule-set, upon which manual improvements can be added.

12.2.4 Machine-learning approach

In order to select the algorithm which is best suited for the `TextBlock` and `TextLine` classification tasks, we first compare the performances of the five following machine-learning algorithms: Support Vector Machine (SVM), Bagging, Random Forest, AdaBoost and Gradient Boosting. We use the *scikit-learn* implementation of these algorithms (Pedregosa et al., 2011).

We train these models on the train set of the LLA dataset described in Section 11.1. We train each model with their default hyper-parameters on the same feature set that was used to train the RIPPER algorithm.

Table 12.7 presents the macro Precision, Recall and F1 scores obtained by each model on the LLA test set. On the `TextBlock` classification task, Gradient Boosting achieves the highest macro scores, with a Precision score of 0.847, a Recall score of 0.719 and a F1-score of 0.783. Similarly, Gradient Boosting achieves the highest scores on the `TextLine` classification task, with a Precision score of 0.806, a Recall score of 0.687 and a F1-score of 0.746.

	`TextBlock`			`TextLine`		
Algorithm	P	R	F1	P	R	F1
SVM	0.750	0.644	0.697	0.717	0.578	0.647
Bagging	0.833	0.712	0.772	0.732	0.657	0.694
Random Forest	0.712	0.689	0.700	0.773	0.628	0.700
AdaBoost	0.775	0.647	0.711	0.720	0.672	0.696
Gradient Boosting	**0.847**	**0.719**	**0.783**	**0.806**	**0.687**	**0.746**

Table 12.7: Macro Precision, Recall and F1 scores of the Support Vector Machine (SVM), Bagging, Random Forest, AdaBoost and Gradient Boosting models on the `TextBlock` and `TextLine` classification tasks

This initial comparison shows that Gradient Boosting is the best algorithm for the `TextBlock` and `TextLine` classification tasks. We search for the optimal combination of hyper-parameter values among the values presented in Table 12.8 for the Gradient Boosting model for both tasks. The optimal values for each hyper-parameter for each task are shown in Table 12.9.

Hyper-parameter	Values
n_estimators	100, 150, 200
learning_rate	0.01, 0.005, 0.001
max_leaf_nodes	2, 3, 4, 5, 6, 7
min_samples_split	10, 20, 40, 60, 100
min_samples_leaf	1, 3, 5, 7, 9
max_features	2, 3, 4, 5, 6, 7
subsample	0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1

Table 12.8: Sets of hyper-parameter values for the Gradient Boosting model on the `TextBlock` and `TextLine` classification tasks

Hyper-parameter	TextBlock	TextLine
n_estimators	200	100
learning_rate	0.005	0.01
max_leaf_nodes	7	7
min_samples_split	40	100
min_samples_leaf	5	7
max_features	6	7
subsample	0.85	1

Table 12.9: Best hyper-parameter combinations for the Gradient Boosting model on the TextBlock and TextLine classification tasks

12.2.5 Evaluation of the machine-learning approach

We evaluate the machine-learning approach based on the Gradient Boosting model on the test set of the LLA dataset. Table 12.10 shows the results obtained by this approach in terms of Precision, Recall and F1 scores on the TextBlock and TextLine tasks.

		Text			Title			Firstline			Header		
	Cat	P	R	F1	P	R	F1	P	R	F1	P	R	F1
TextBlock	1c	0.863	0.962	0.910	0.500	0.375	0.429				**0.667**	**0.222**	**0.333**
	2c	0.891	**0.989**	**0.937**	0.706	**0.667**	**0.686**				0.625	0.208	0.312
	3c+	**0.967**	0.863	0.912	**0.789**	0.234	0.361				0.000	0.000	0.000
	Mean	0.907	0.938	0.920	0.665	0.425	0.492				0.431	0.144	0.215
TextLine	1c	0.778	**0.983**	0.868	0.667	0.194	0.300	0.641	0.207	0.312	**1.000**	**0.211**	**0.348**
	2c	**0.905**	0.949	**0.926**	0.758	**0.305**	**0.435**	**0.727**	**0.756**	**0.741**	0.875	0.113	0.200
	3c+	0.856	0.950	0.901	**0.941**	0.264	0.413	0.667	0.483	0.560	0.000	0.000	0.000
	Mean	0.846	0.961	0.898	0.788	0.254	0.383	0.678	0.482	0.538	0.625	0.108	0.183

Table 12.10: Precision, Recall and F1 scores of the Gradient Boosting models on the TextBlock and TextLine classification tasks

On the TextBlock classification task, the model performs the best on the 2c category. The classification of Text elements achieves the highest F1 score of 0.920 on average. The performance for Title classification is average with a mean F1 score of 0.429, but with a mean Precision of 0.665. Finally, the Header classification achieves the lowest scores, with a mean F1 score of 0.215 and an average Precision of 0.431. On the TextLine classification task, the model also performs the best on the 2c layout category, except for Header lines classification, where it performs the best on the 1c layout category. Text lines classification achieves the best scores, with a mean F1 score of 0.898

and a Recall of 0.961. Firstline classification achieves a mean F1 score of 0.538 and an average Precision score of 0.678. Both Header and Title classification achieve lower scores, with a mean F1 score of 0.183 and 0.383, but an average Precision score of 0.625 and 0.788 respectively.

Most errors in the `TextBlock` classification task are either Header blocks mislabelled as Title, or Text blocks mislabelled as Header. Most mislabelled Title and Header blocks are labelled as Text, which suggests the model has not been able to learn to recognise these labels correctly.

On the `TextLine` classification task, every mislabelled Text line is incorrectly labelled as Firstline, suggesting the model struggles to recognise the beginning of paragraphs. Moreover, most mislabelled Firstline, Title and Header lines are labelled as Text. Similarly, this suggests the model has not been able to learn to recognise these labels correctly.

In conclusion, the scores obtained by the Gradient Boosting model on the LLA task shows its ability to determine the logical label of `TextBlock` and `TextLine` tags from XML ALTO documents. In general, the model reaches higher Precision scores than Recall scores, suggesting it has been able to learn precise patterns from the data. The performances of the model could be improved by adding more samples of each logical category, as well as by training the model on more features.

12.3 Discussion

We compare the performances of our rule-based approach, of the rule-learning approach and of machine-learning approach to the LLA task. Table 12.11 shows the mean Precision, Recall and F1 scores of each approach on the `TextBlock` and `TextLine` classification tasks on the LLA test set.

		Text			Title			Firstline			Header		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
TextBlock	RB	**0.959**	0.966	**0.962**	0.600	**0.639**	**0.610**				**0.726**	**0.298**	**0.406**
	RP	0.841	**0.996**	0.912	0.560	0.455	0.480				0.440	0.194	0.268
	GB	0.907	0.938	0.920	**0.665**	0.425	0.492				0.431	0.144	0.215
TextLine	RB	**0.969**	**0.991**	**0.979**	0.595	**0.733**	**0.639**	**0.949**	**0.861**	**0.902**	**0.803**	**0.348**	**0.435**
	RP	0.890	0.912	0.901	0.640	0.255	0.363	0.671	0.788	0.724	0.444	0.039	0.072
	GB	0.846	0.961	0.898	**0.788**	0.254	0.383	0.678	0.482	0.538	0.625	0.108	0.183

Table 12.11: Mean Precision, Recall and F1 scores of the rule-based approach (RB), the ensemble of RIPPER models (RP) and the Gradient Boosting model (GB) on the `TextBlock` and `TextLine` classification tasks. The best scores for each metric for each label are shown in bold

All three approaches achieve relatively good results on the `TextBlock` classification task. Our rule-based approach has the best Precision and F1 score of 0.959 and 0.962 respectively, whereas RIPPER achieves the best Recall score, which is 0.996. The performances of the three approaches are between 0.425 and 0.665 on the Title block classification. Gradient Boosting has the best mean Precision score with 0.665, whereas the rule-based system has the best mean Recall and F1 score of 0.639 and 0.610 respectively. Finally, the performances for the Header block classification range from 0.144 to 0.726. Our approach achieves the best Precision, Recall and F1 scores, which are 0.726, 0.298 and 0.406 respectively. The Precision score is higher than the Recall score for every approach. This suggests that defining exhaustive rules to detect Header blocks is a difficult task.

Similarly for the `TextLine` classification task, all three approaches achieve high results. The performances for the classification of Title lines are variable. Gradient Boosting achieves a Precision score of 0.788. However, our approach achieves the best Recall and F1 scores, of 0.733 and 0.639, which are higher than those of the two other models. For Firstline classification, our approach obtains the best scores in every metric, with a Precision score of 0.949, a Recall score of 0.861 and a F1 score of 0.902. Gradient Boosting only achieves an average F1 score of 0.538, suggesting that paragraphs are easier to detect with simple rules. Finally, the performance for the classification of Header lines varies a lot. Our approach achieves the best Precision, Recall and F1 scores, which are 0.803, 0.348 and 0.435 respectively. Similarly, the Precision score is higher than the Recall score for every model.

In conclusion, our rule-based approach outperforms the two other approaches in nearly all evaluations. It obtains especially better Recall results, which indicates that it covers more types of every logical label than the other two models. When comparing Gradient Boosting and RIPPER, we

see that in general, Gradient Boosting achieves better Precision scores, whereas RIPPER achieves better Recall scores. Moreover, we notice that the RIPPER model achieves lower F1 scores on the `TextLine` classification task than on the `TextBlock` classification task. The `TextBlock` classification step is designed as an intermediary step that should facilitate the classification of `TextLine`, which is the final output of the algorithm. Unlike our approach however, RIPPER only uses sparsely the results of the `TextBlock` classification. This could explain its lower results on the `TextLine` classification task.

The results obtained by our approach show the relevance of our method, which we specifically conceived for historical documents. However, as shown in the comparative evaluation, the Gradient Boosting model can achieve very high Precision scores. Thus, it would be interesting to combine our approach with the machine-learning approach in an hybrid manner. Each model would be trained to detect specific labels, thus allowing to overcome the weaknesses of the other systems.

The RIPPER algorithm can learn very precise rules, which sometimes perform better than our manually crafted rules. Furthermore, RIPPER achieves better Precision scores than Recall scores on average. This suggests that RIPPER is better at producing fine-grained rules. As producing rules by hand is a time-consuming task, it would be interesting to use RIPPER in an exploratory manner and produce a base rule-set with high Precision scores. This rule set could then be manually updated in order to improve its Recall. Moreover, the layout of historical documents evolves rapidly, especially in newspapers. Thus, the RIPPER model could prove a useful tool to quickly build a rule set adapted to specific publication periods.

Part III

Semantic Annotation of historical documents: introducing the ELIJERE approach

The Semantic Annotation task consists in annotating the textual content of documents with semantic categories, such as sentiments (e.g. positive, negative), entity types (e.g. Person, Location, Time) or relations (e.g. place of birth, educated At) to name a few. Semantic annotations allow to structure the textual content of documents and can be exploited by search interfaces to assist in exploring a collection of documents. They can also be exploited to build augmented search interfaces, which allow a distant reading of the collection. We can divide the approaches to the Semantic Annotation task into two main categories: *rule-based approaches* and *machine-learning approaches*, among which we consider deep-learning approaches.

Rule-based approaches rely on sets of rules, which generally allow to achieve high Precision scores. However, these approaches do not generalise well, and are especially sensible to noisy input, such as errors in the OCR transcriptions. They usually achieve lower Recall scores. On the other hand, machine-learning approaches learn patterns from data. They are especially good at generalising from the data, and are less sensible to noisy input.

Both types of approaches require resources. Rule-based approaches require sets of rules which state the conditions that a textual content must meet in order to be annotated with a given semantic category. Adding and updating the rule-set allow the rule-based approach to detect and process new semantic categories. However, the more rules are stored in a rule-set, the harder it becomes to maintain and update. Moreover, building a rule-set is a time consuming task, which requires expertise of the application domain. Models used in machine-learning approaches must be trained on large annotated dataset. Building such annotated datasets remains a time-consuming and expensive task. Moreover, the model must be trained again on a sufficient amount of annotated data to learn to identify new semantic categories. Maintaining the quality and internal coherence of the annotations in large datasets can also be a challenge.

Recent deep-learning approaches, especially those based on the Transformer architecture, reduce the dependence on large annotated data, by allowing to fine-tune pre-trained models on smaller in-domain datasets. However, these approaches require important and expensive computational resources such as Graphical Processing Units (GPU) or Tensor Processing Units (TPU) for training and inference. Moreover, deep-learning approaches based on Neural Network and Transformer architectures are often considered "black boxes" (Barredo Arrieta et al., 2020) because of the difficulty to interpret their processes. On the other hand, rule-based approaches are interpretable, since they rely on explicitly stated rules, whereas traditional machine-learning approaches may be easier to interpret, depending on the complexity of the underlying architecture. Moreover, rule-based and machine-learning approaches require less computational resources and can run on CPUs.

In this part, we propose the Extensible, Lightweight and Interpretable Joint Extraction of Relations and Entities (ELIJERE) approach. Our approach relies on a set of two linguistic resources:

- a *Syntactic Index*, which describes what relation a syntactic pattern expresses, as well as the

kind of entities that are involved in the relation

- a *Lexical Index*, which describes how a relation is expressed lexically

We exploit these two resources to extract and categorise the mentions of entities involved in relations from sentences. The linguistic resources are built from lexico-syntactic patterns that are collected from a dataset of annotated sentences. These sentences are weakly annotated by applying the distant supervision method (Mintz et al., 2009). The distant supervision method assumes that, if two entities participate in a relation in a Knowledge Base, then at least one sentence which mentions both entities in a document expresses that relation (Riedel et al., 2010). By applying this method, we can quickly build an annotated dataset of sentences expressing any concept stored in a Knowledge Base, upon which we can build our linguistic resources. Thus, our approach is extensible and interpretable, since it relies on extensible and explicit linguistic resources. It is also lightweight as it requires low computing resources, and can run on CPUs. The main stages of our approach, which has been presented previously in Gutehrlé (2024a), are shown in Figure 12.7.

The rest of the part is organised as follows: In Chapter 13, we present the Distant Annotation of Relations and Entities in Sentences (DARES) method, to build a dataset of sentences annotated with the mentions of entities and relations. In Chapter 14, we describe our method to build the linguistic resources upon which the ELIJERE approach relies from the DARES dataset. Finally in Chapter 15, we describe our approach to the Joint Extraction of Relations and Entities task.

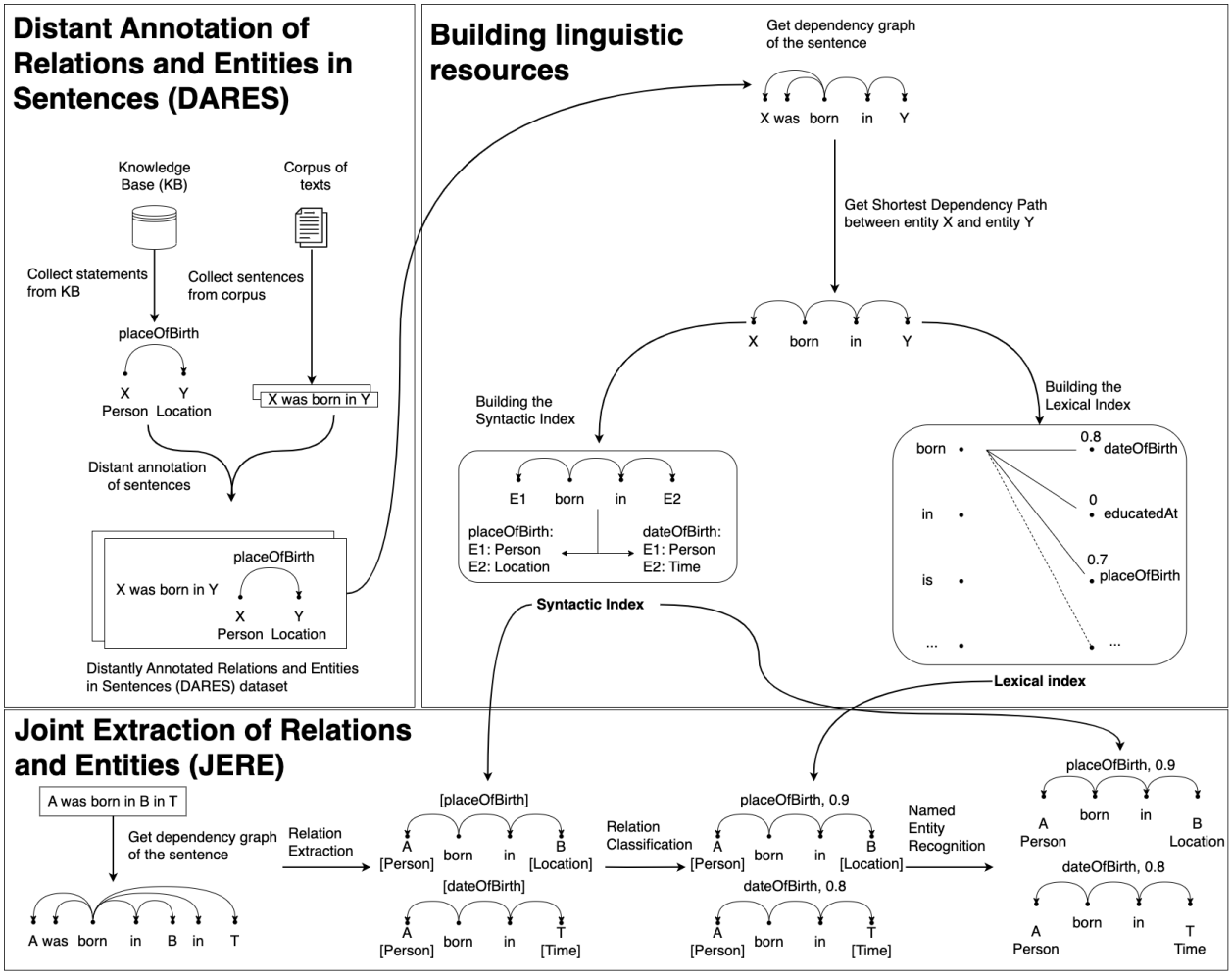


Figure 12.7: Main stages of the ELIJERE approach

Chapter 13

Distant Annotation of Relations and Entities in Sentences (DARES)

Table of contents

In this chapter, we present the Distant Annotated Relations and Entities in Sentences (DARES) method to build a dataset of sentences annotated with the mentions of entities and relations. This method relies on the the distant supervision method (Riedel et al., 2010) which allows to quickly build an annotated dataset of sentences expressing any concept stored in a Knowledge Base. Our dataset is built from a corpus of Wikipedia articles written in French about people and locations, as well as statements describing relations between entities, extracted from the Wikidata Knowledge Base.

The rest of this chapter is structured as follows: in Section 13.1, we describe the Distant Annotated Relations and Entities in Sentences (DARES) method to build a dataset of weakly annotated sentences. We describe the implementation to collect data from Wikipedia and Wikidata in Section 13.2. Finally, we describe the DARES dataset built with our method in Section 13.3.

13.1 Method

The linguistic resources upon which the ELIJERE approach relies are built from a dataset of sentences, where the mentions of relations and entities are annotated. To build this annotated dataset, we apply the following steps, which are also shown in Figure 13.1:

1. *statement collection*
2. *sentence collection*
3. *distant annotation*

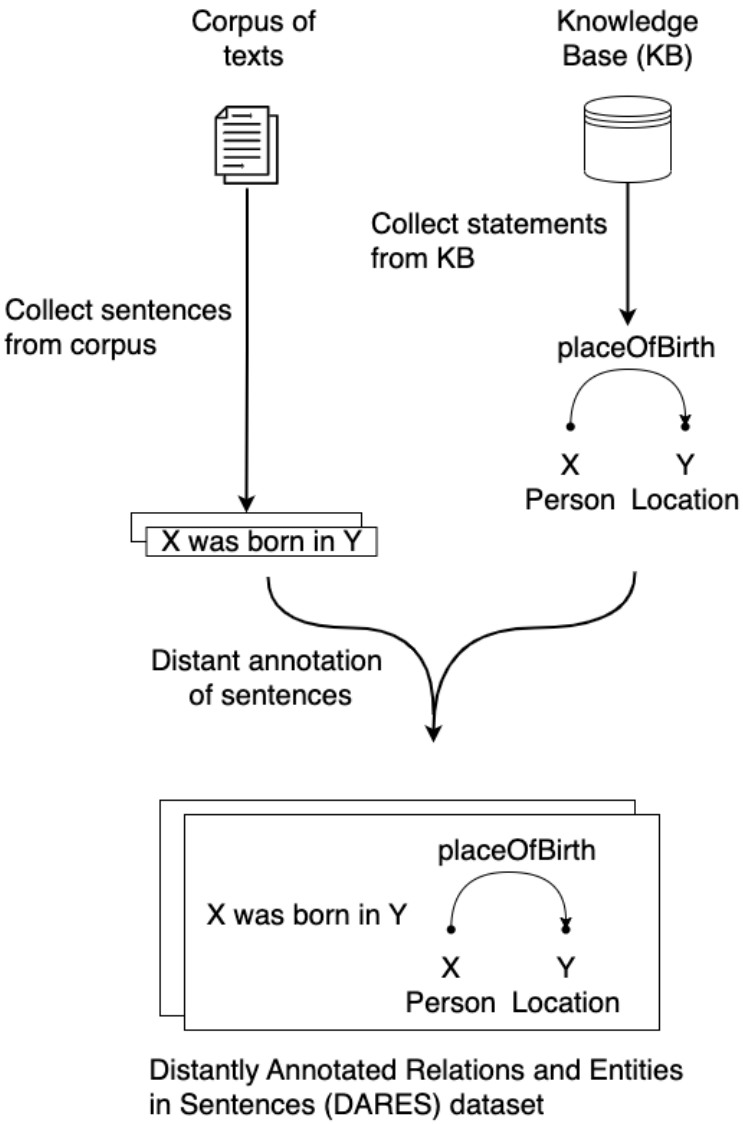


Figure 13.1: Main steps of the Distant Annotation of Relations and Entities in Sentences method

The *statement collection* step extracts triples from a Knowledge Base such as Wikidata or Freebase. A triple describes a property about an entity in a machine-readable format. It follows the form `(subject; predicate; object)`, where `subject` is the entity about which the statement is made, `predicate` is the relation or property describing the subject, and `object` is the value of the relation or property. For instance, given the subject "Douglas Adams", the object "1952" and the predicate "date of birth", the triple `(Douglas Adams; dateOfBirth; 1952)` states that "Douglas Adams" was born in 1952. Other examples of triples that can be extracted from the Wikidata Knowledge Base are:

1. `(Marie Curie; occupation; physicist)`
2. `(George Washington; placeOfDeath; Mount Vernon)`
3. `(France; sharesBorderWith; Belgium)`

The *sentence collection* step collects all sentences from a corpus of documents. In our experimentation, we collect all sentences from a set of Wikipedia articles. However, in a more general approach, the sentences can potentially be extracted from documents from various origins such as the World Wide Web. Given a corpus of articles collected from Wikipedia, we extract, for example, the following sentences:

1. Marie Curie was a physicist.
2. George Washington died in Mount Vernon.
3. France shares its border with Belgium to the north.

Finally, the *distant annotation* step applies the distant supervision method to associate the sentences with the statements. A sentence is associated with a statement if it contains mentions of both entities involved in the statement, i.e. mentions of the `subject` and the `object` of the triple. For instance, by applying the distant supervision method on the set of examples mentioned above, we would associate the preceding examples of sentences and statements as follows:

1. Marie Curie was a physicist. : `(Marie Curie; occupation; physicist)`
2. George Washington died in Mount Vernon. : `(George Washington; placeOfDeath; Mount Vernon)`
3. France shares its border with Belgium to the north. : `(France; sharesBorderWith; Belgium)`

13.2 Implementation

In this section, we describe our implementation of the DARES method. We first provide a brief description of the structure of a Wikidata page in Section 13.2.1, and describe in details the implementation of our method in Section 13.2.2. The code of this implementation is available on GitHub[1].

13.2.1 Structure of a Wikidata page

Wikidata contains Items, which describe any type of object, from a person, a place to a concept in the Knowledge Base. An Item stores links to related Wikimedia pages, such as Wikipedia, Wikiquote or Wikinews. For instance, the Item Q23 stores a link to the Wikipedia page related to "George Washington"[2]. Items are given a label, a description and a set of aliases. Labels, descriptions and aliases are usually available in multiple languages. For instance in English, the Item Q23 has the following label, description and aliases, which are also shown in Figure 13.2:

label : George Washington

description : president of the United States from 1789 to 1797

aliases : Father of the United States, The American Fabius, American Fabius

Information about an Item are recorded as `(Source;Property;Target)` triples named Statements, such as `(George Washington, date of birth, 22 February 1732)` or `(George Washington, place of death, Mount Vernon)`. Figure 13.3 shows an example of Statements in the Wikidata page of the Item Q23 ("George Washington"). The `Property` element in the Statement triple are special Items called Properties, which describe the relation between Items. Similarly, each Property has a label, a description and possible aliases. For instance in English, the Property P19 has the label "place of birth". The `Source` element of the Statement is the current Item I, for instance the Item Q23. The `Target` element of a Statement is usually another Item, as is the case for the *place of birth* or *educated at* Properties. It may be a date as is the case for the *date of birth* or *date of death* Properties. It may also be a quantitative or textual value, which is the case for Properties such as *number of children* or *height*. Finally, the value of the Target element of a Statement may be empty. A Target element of a Statement may record multiple values for the same Property.

[1] https://github.com/nicolasgutehrle/emontalproject
[2] https://en.wikipedia.org/wiki/George_Washington

George Washington (Q23)

president of the United States from 1789 to 1797

Father of the United States | The American Fabius | American Fabius

▾ In more languages

Configure

Language	Label	Description	Also known as
English	George Washington	president of the United States from 1789 to 1797	Father of the United States The American Fabius American Fabius
French	George Washington	révolutionnaire américain, premier président des États-Unis d'Amérique de 1789 à 1797	
Spanish	George Washington	1.er presidente de los Estados Unidos	
German	George Washington	erster Präsident der Vereinigten Staaten von Amerika (1732–1799)	Vater der Vereinigten Staaten

All entered languages

Figure 13.2: Example of the description, label and aliases of Item Q23 ("George Washington") in Wikidata

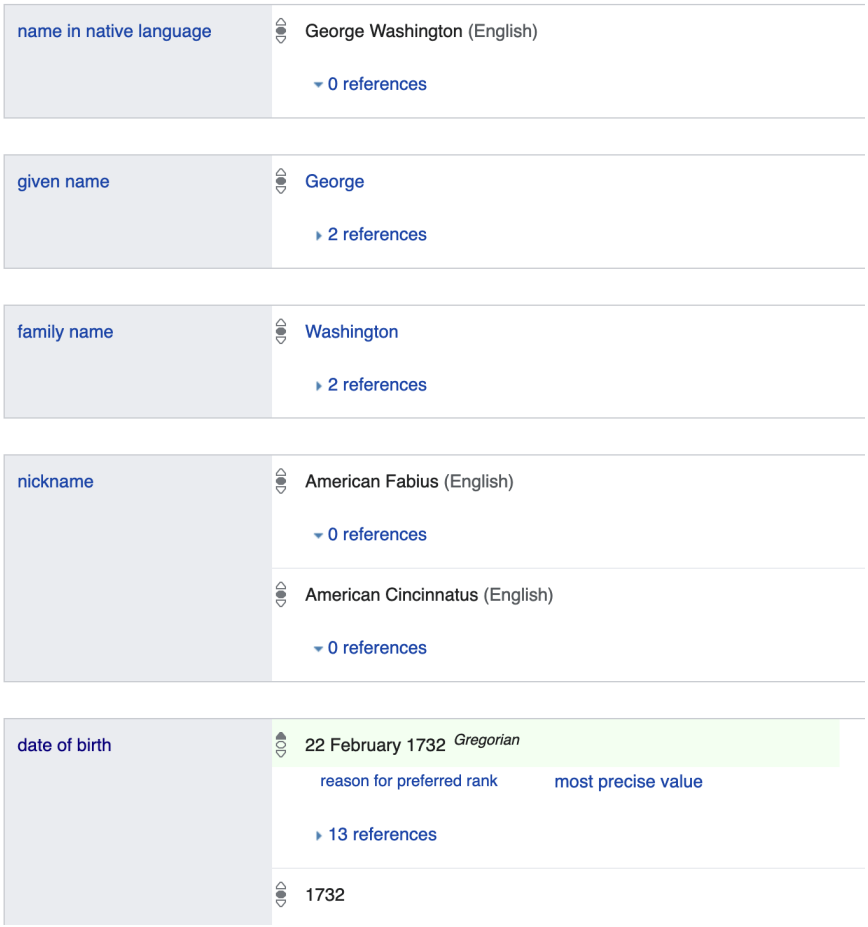


Figure 13.3: Example of Statements in the Wikidata page of the Item Q23 ("George Washington")

13.2.2 Implementation of the Distant Annotation of Relations and Entities in Sentences (DARES) method

We propose an implementation written in the Python programming language of the DARES method described in the previous Section. The main steps of the implementation are shown in Figure 13.4.

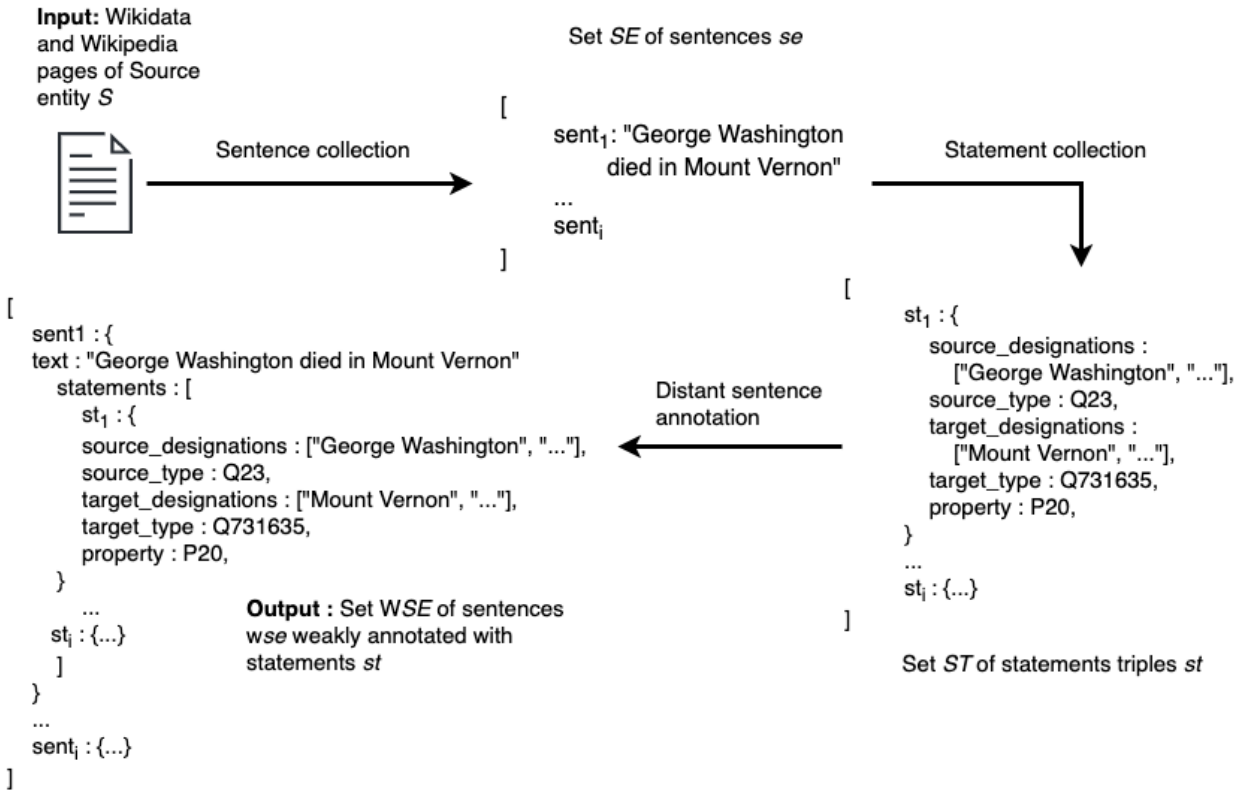


Figure 13.4: Main steps of the implementation of the Distant Annotation of Relations and Entities in Sentences (DARES) method

The *sentence collection* step extracts all sentences from the body of the Wikipedia page associated with the Source entity S. We use the *requests* library to collect Wikipedia and Wikidata pages, and the *BeautifulSoup* library to manipulate the HTML structure of the web pages. We collect the content of Wikipedia pages by extracting the content of *p* tags from their HTML pages. We segment the content of *p* tags into sentences by applying the method described in Section 11.4. To ensure the quality of further linguistic analysis, we clean the sentences of the elements described in Table 13.1. Figure 13.5 shows an example of an excerpt of the content of a Wikipedia page before and after applying the cleaning step.

The *statement collection* step extracts non-empty Statement triples from the Wikidata page associated with the Source entity S. We may extract all Statements or a selected subset. Figure 13.6 shows two examples of extracted statements, one for the Property P20 (*place of death*) and another

Rule	Regular Expression
Brackets and their content (often phonetic transcriptions)	`[/\[][^/\]]*[/\]]( ?Écouter)?`
Sentences starting with non-alphanumeric characters	`^[^a-zA-Z]+`
Isolated non alphanumeric symbols	`["«»\[\]]`

Trailing spaces	`^(,? |-|")`
	`( |")$`
	`’`
	`’ {2,}`

Table 13.1: Regular expressions to clean sentences from the content of Wikipedia pages

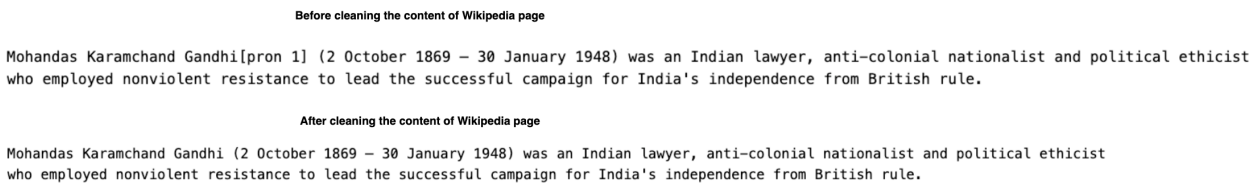


Figure 13.5: Excerpt of a Wikipedia page before and after applying the cleaning step

for the Property P569 (*date of birth*). We store the following data for each extracted statement:

- the property of the Statement, such as P20. It describes the type of relation between the Source and the Target entities
- the designations of the Source entity S. These are labels and aliases, e.g. "George Washington | Father of the United States | The American Fabius | American Fabius"
- the type of the Source entity, which is its Wikidata identifier, e.g. Q23
- the designations of the Target entity T
- the type of the Target entity

We apply the following process to extract the designations of the Target entity T: if T is an Item, we extract its labels and aliases from its corresponding Wikidata page. If T is a date, and we are working on another language than English, we automatically translate the date to the chosen language. For instance, "22 February 1732" in English would be translated into "*22 février 1732*" if we are working on the French language. Otherwise, if the value of T is numerical or textual, we take it as is. We use the *datetime* library to manipulate temporal data and the *babel* package to translate dates to the chosen language.

The type of the Target entity T is its Wikidata identifier if T is an Item. Otherwise, if T is a date, a temporal value, a quantitative value or a string, the type is respectively set to *date, time, quantity* or *string*. For instance, the type of the Target entity "Mount Vernon" will be Q731635, whereas the type of the Target entity "22 February 1732" will be "date".

Finally, the *distant sentence annotation* step applies the distant supervision method to find sentences which express a minimum of one relation. We determine that a sentence expresses a relation if it contains a designation of the Source and Target entities of a Statement. We choose to apply the fuzzy matching method to search for the presence of a designation of the Source and Target entities in a sentence. We select this method instead of an exact matching method in order to avoid missing the occurrence of a designation because of small orthographic differences, such as a plural forms. Thus, a sentence becomes associated with a Statement if it contains two tokens or two sequences of tokens that are at least N % similar to a designation of the Source and Target entities respectively. We use the *fuzzywuzzy* library to apply the fuzzy matching method, and set $N = 90\%$ as a similarity threshold. Figure 13.7 shows two sentences associated with a Statement. The first example shows a sentence associated with a Statement describing the *placeOfDeath* relation, whereas the second example shows a sentence associated with a Statement describing the *dateOfBirth* relation. The designations of the Source and Target entities which helped associating the sentences with the Statements are shown in bold.

```
st1 : {
   source_designations : [

      "George Washington",
      "Father of the United States",
      "The American Fabius",
      "American Fabius",
      ],

   source_type : Q23,

   target_designations : [
      "Mount Vernon",
      "Mount Vernon Estate",
      "Epswasson"
      ],

   target_type : Q731635,

   property : P20,
}
```

```
st2 : {
   source_designations : [

      "George Washington",
      "Father of the United States",
      "The American Fabius",
      "American Fabius",
      ],

   source_type : Q23,

   target_designations : [
      "22 February 1732",
      "22/02/1732"
      ],

   target_type : date,

   property : P569,
}
```

Figure 13.6: Examples of statements extracted for the P20 (*place of death*) and P569 (*date of birth*) Properties from the Wikidata page of the entity Q23

```
[
   sent1 : {
   text : "George Washington died
            in Mount Vernon"

      statements : [
         st1 : {
         source_designations : [
            "George Washington",
            "Father of the United States",
            "The American Fabius",
            "American Fabius"
      ],
         source_type : Q23,
         target_designations : [
            "Mount Vernon"
            "Mount Vernon Estate",
            "Epswasson"
      ],
         target_type : Q731635,
         property : P20,
      }

]
```

```
[
   sent2 : {
   text : "George Washington was born
            on February 22, 1732 "

      statements : [
         st1 : {
            source_designations : [
                  "George Washington",
                  "Father of the United States",
                  "The American Fabius",
                  "American Fabius"
            ],
            source_type : Q23,
            target_designations : [
                  "22 February 1732",
                  "22/02/1732"
         ],

         target_type : date,
         property : P569,
         }

]
```

Figure 13.7: Examples of sentences associated with at least one Statement. The designations of the Source and Target entities which helped associating the sentences with the Statements are shown in bold. The designation "22 February 1732" matches the mention "February 22, 1732" in the text, since we perform a fuzzy matching instead of an exact matching

13.3 Description of the DARES dataset

In this section, we present the Distantly Annotated Relations and Entities in Sentences (DARES) dataset, which is built by applying the implementation described above.

To build this dataset, we have collected sentences and statements from the Wikipedia and Wikidata pages related to 5,000 persons and 5,000 locations, which we have chosen in the following way: we have collected the Wikidata and Wikipedia pages related to persons by selecting the first 5,000 Items of type Q5 (*human*) from the What Links Here section on Wikidata[3]. Similarly, we have collected the Wikidata and Wikipedia pages of the first 5,000 Items having the Property P625 (*coordinate location*) to represent the *Location* type. We choose to collect Items having the *coordinate location* property to represent entities of type *Location*, since the data collected for Items with types such as Q2221906 (*geographic location*) or Q98929991 (*place*) did not prove useful.

The sentences in the DARES dataset are annotated with mentions of the relations described in Table 13.2. We selected 13 relations among the most common relations in the Items we have collected which we deemed interesting to model peoples and locations. Among these relations, six are related to persons, whereas seven are related to locations.

As explained in Section 13.2.2, the types of the Source and Target entities involved in a relation can be a Wikidata identifier *date, time, quantity* or *string*. In order to normalise the entity types involved in the relations in the DARES dataset, we have manually set the types of entities according to one of the four following entity labels:

- *Person*
- *Location*
- *Time*
- *Misc*

We have divided the DARES dataset into a train and a test set. We post-processed the train set to ensure the distribution of relations is balanced. Since this dataset has been labelled in a distant supervision manner, some annotations may not be correct. Thus, we have manually reviewed and corrected the annotations of the test set. Moreover, we added sentences to the test set which do not contain any of the selected properties, so as to ensure that the dataset contains negative samples. These sentences are annotated with the label *Other*. Table 13.3 shows the distribution of unique sentences in the train and test sets of the DARES dataset. Table 13.4 shows the distribution of relation labels, whereas Table 13.5 shows the distribution of entity types in the DARES dataset.

[3] https://www.wikidata.org/w/index.php?title=Special:WhatLinksHere/

```
{
    "id": "Q20",
    "content": [
        {
            "sent": "Après l'échec de l'instauration d'une un
            "sent_i": 5,
            "props": [
                {
                    "prop": "memberOf",
                    "sent": "Après l'échec de l'instauration
                    "source": "Norvège",
                    "source_type": "Location",
                    "target": "OTAN",
                    "target_type": "Location"
                }
            ]

        },
        {
            "sent": "En réaction, une Convention nationale se
            "sent_i": 67,
            "props": [
                {
                    "prop": "inception",
                    "sent": "En réaction, une Convention nati
                    "source": "Norvège",
                    "source_type": "Location",
                    "target": "17 mai 1814",
                    "target_type": "Time"
                }
            ]
        },
```


Figure 13.8: Extract of the DARES dataset, showing sentences and statements related to the entity *Norvège* (Norway)

Source type	Target type	Relation label	Description	Example
Person	Location	*placeOfBirth*	place where a person is born	Douglas Adams was born in Cambridge
Person	Time	*dateOfBirth*	date of birth of a person	Douglas Adams was born in in 1952
Person	Time	*dateOfDeath*	date of death of a person	Douglas Adams died in 2001
Person	Person	*spouse*	person with whom another person is married	Pierre Curie is the husband of Marie Curie
Person	Location	*educatedAt*	place where a person has studied	Marie Curie studied at the Flying University
Person	Misc	*occupation*	occupation of a person	Marie Curie was a physicist and chemist
Location	Location	*country*	country to which a location belongs	Lyon is a city in France
Location	Time	*inception*	date of foundation of a location	The Hôtel-Dieu in Paris was founded in 651
Location	Location	*capitalOf*	country to which a location is the capital of	Paris is the capital of France
Location	Person	*headOfGovernment*	location for which a person is the leading political figure	Anne Hidalgo is the mayor of Paris
Location	Location	*sharesBorderWith*	location with which another location shares its border with	France shares its border with Belgium
Location	Location	*memberOf*	location to which a location belongs	Dijon is part of the Bourgogne Franche-Comté region
Location	Location	*nextInBodyOfWater*	body of water to which a location is close to	Paris is located in a north-bending arc of the river Seine

Table 13.2: Description of entity types and relations in the DARES dataset

	Train set	Test set
Unique sentences	2,526	623

Table 13.3: Distribution of unique sentences in the train and test sets of the DARES dataset

	Train set		Test set	
Label	Count	Proportion	Count	Proportion
capitalOf	303	4.227 %	140	11.914 %
country	303	4.227 %	183	15.574 %
dateOfBirth	303	4.227 %	71	6.042 %
dateOfDeath	303	4.227 %	39	3.319 %
educatedAt	303	4.227 %	13	1.106 %
headOfGoverment	303	4.227 %	15	1.276 %
inception	303	4.227 %	12	1.021 %
memberOf	303	4.227 %	53	4.510 %
nextInBodyWater	303	4.227 %	47	4.000 %
occupation	303	4.227 %	125	10.638 %
Other	0	0 %	213	18.127 %
placeOfBirth	303	4.227 %	67	5.702 %
sharesBordersWith	303	4.227 %	174	14.808 %
spouse	303	4.227 %	23	1.957 %
Total	7,167	100 %	1,175	100 %

Table 13.4: Distribution of the relation labels in the train and test sets of the DARES dataset

	Train set		Test set	
Label	Count	Proportion	Count	Proportion
Person	2,424	30.769 %	391	19.628 %
Location	4,242	53.846 %	1,337	67.118 %
Time	909	11.538 %	128	6.425 %
Misc	303	3.846 %	136	6.827 %
Total	7,878	100 %	1,992	100 %

Table 13.5: Entity labels distribution in the train and test sets of the DARES dataset

Chapter 14

Building Linguistic Resources

Table of contents

In this chapter, we present the Syntactic and Lexical Indices, the main linguistic resources upon which our approach to extract and categorise relations and entities from a sentence relies. The Syntactic Index describes what relations a pattern expresses, as well as the kind of entities that are involved in the relation. On the other hand, the Lexical Index describes how a relation is expressed lexically. Both indices are built from lexico-syntactic patterns expressing a relation between entities that are collected from the DARES dataset. To collect the lexico-syntactic patterns, we rely on the assumption by R. Bunescu and Mooney (2005) that the relation between two entities lies in the Shortest Dependency Path between them. The main steps to build these resources from the DARES dataset are shown in Figure 14.1.

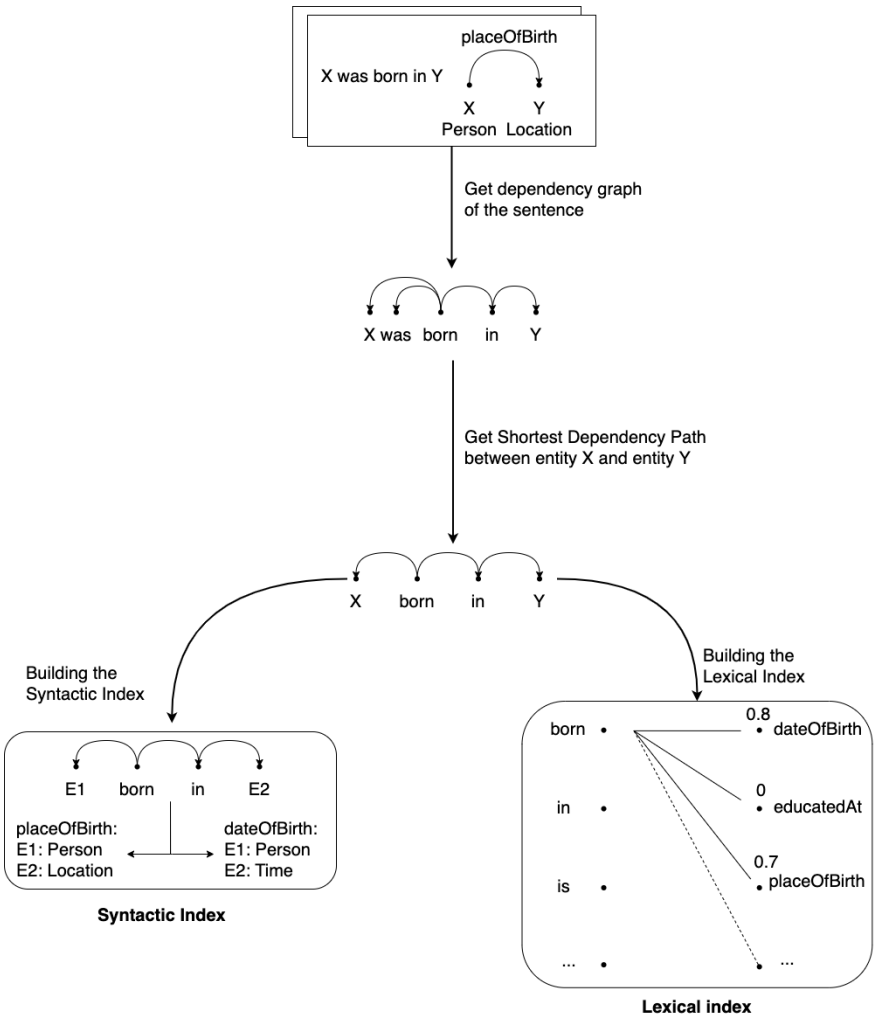


Figure 14.1: Main steps to build the linguistic resources required for the ELIJERE approach

14.1 Building the Syntactic Index

14.1.1 Method

We obtain the syntactic structures of the sentences as dependency graphs, such as the one shown in Figure 14.2. We choose to work at the syntactic level because it is easier to collect patterns expressing a relation between entities at this level than at the surface level of the sentence. The pattern to express a relation in a sentence may be either contiguous or non-contiguous. Contiguous patterns involve a sequence of words that follow each other, whereas non-contiguous patterns involve sequences of words that can be separated by other tokens in the sentence. For instance, the sentence "Douglas Adams was born in Cambridge in 1952" expresses the *placeOfBirth* and the *dateOfBirth* relations. The pattern to express the *placeOfBirth* relation involves the contiguous sequence of words "Douglas Adams was born in Cambridge", whereas the pattern to express the *dateOfBirth* relation involves the non-contiguous sequence of words "Douglas Adams was born [...] in 1952". Extracting such non-contiguous patterns requires working on the level of the syntactic structure of sentences and cannot be done by considering only the surface representation.

To collect the lexico-syntactic patterns from these graphs, we extract the Shortest Dependency Path (SDP) graphs between the entities. Two examples of such SDP graphs extracted from this sentence are shown in Figure 14.3.

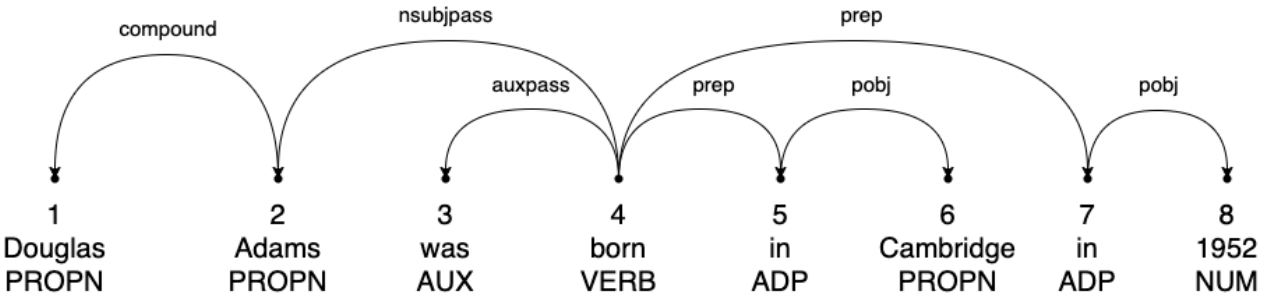


Figure 14.2: Dependency graph of the sentence "Douglas Adams was born in Cambridge in 1952"

The collected lexico-syntactic patterns may differ from each other by their vocabulary and the kind of entities involved in the expressed relation. However, they may share the same syntactic structure. Thus, we identify classes of patterns based on the following conditions: lexico-syntactic patterns belong to the same class if they share the same syntactic structure, the same part-of-speech (POS) tags and the same predicate. The predicate is the head of the pattern, i.e. the token which governs all the other tokens. The predicate is usually a verb, but can belong to other grammatical categories, such as noun. For instance, the two lexico-syntactic patterns in Figure 14.3 share the same syntactic structure and the verb "born" as predicate. However, the second pattern differs from the first one by its POS tags. Thus, the two patterns belong to different classes.

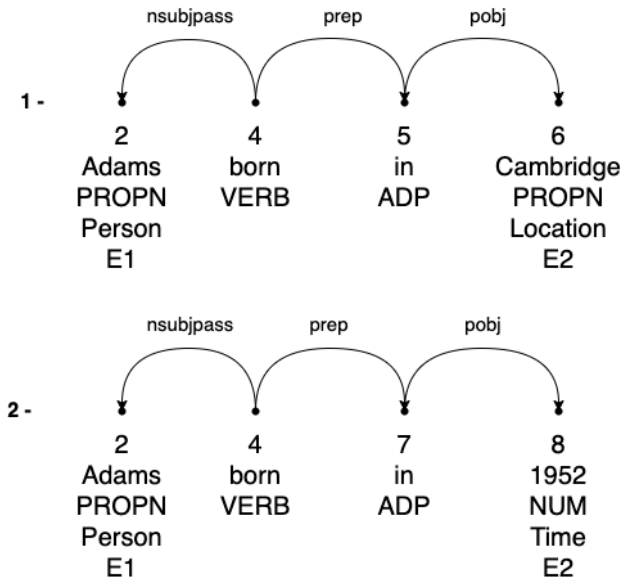


Figure 14.3: Shortest Dependency Path (SDP) graphs between the entities ("Adams", "Cambridge") and ("Adams", "1952"), extracted from the dependency graph shown in Figure 14.2

14.1.2 Implementation

Our pipeline to build the Syntactic Index takes as input the annotated sentences of the DARES dataset. It consists in two main steps, which are shown in Figure 14.4. Our implementation of the Syntactic Index has two main hyper-parameters:

predicate_type : the type of the surface form of the predicate of the syntactic pattern. This surface form is used as key in the Syntactic Index. It can be the text, lemma or the part-of-speech tag, or a combination of any of those. The **predicate_type** hyper-parameter is set to "text" by default

support_threshold : The minimum times a relation must appear in a pattern class. The **support_threshold** hyper-parameter is set to 0 by default

Since we are working on sentences written in French, we use the $fr\_dep\_news\_trf$ pipeline provided by the *spaCy* NLP framework to obtain the part-of-speech tag of words and dependency tree of the sentences. This pipeline uses a pre-trained CamemBERT model (Martin et al., 2020) to perform the linguistic analysis.

In order to extract the SDP graphs, we must find the nodes which correspond to the designations of the entities involved in a relation. For instance, node 1 and 2 of the dependency graph shown in Figure 14.2 correspond to the designation of the entity "Douglas Adams", whereas node 6 and 8 correspond to the "Cambridge" and "1952" entities respectively. As explained in the previous chapter, we apply the fuzzy matching method to find the designations of the entities in a sentence.

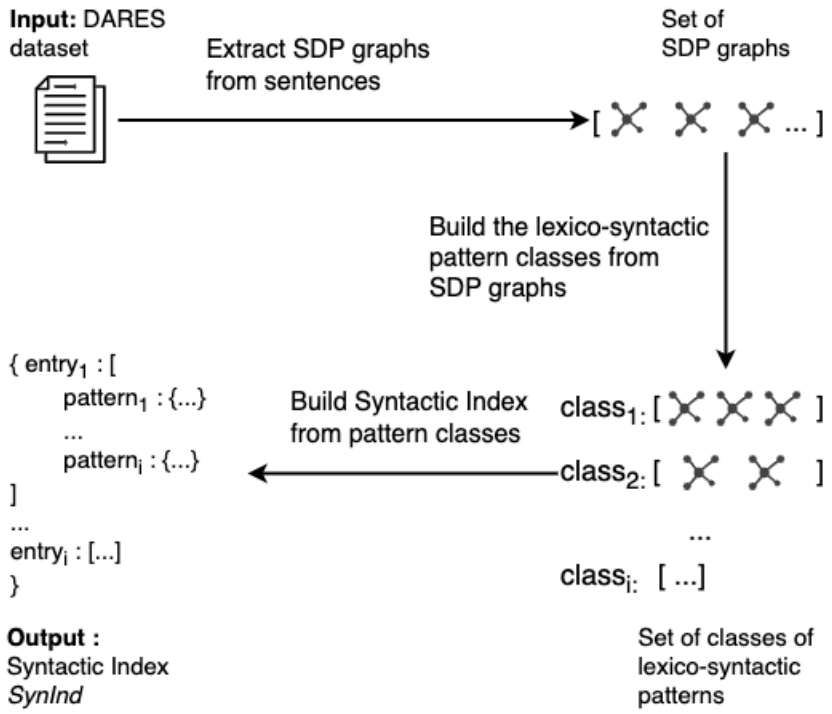


Figure 14.4: Main stages of the pipeline to build the Syntactic Index

Thus, we apply the fuzzy matching method again to find the node or group of nodes that is at least $N\%$ similar to the Source and Target entities designations. We use the *FuzzyMatcher* library[1] to perform the fuzzy matching step and set $N = 90\%$. After finding these groups of nodes, we extract the SDP graph between the root nodes of each group, i.e. the node with only out degrees and no in degrees. For instance, node 2 ("Adams") is the root node of the "Douglas Adams" group.

To build the classes of lexico-syntactic patterns, we perform an isomorphism test to determine if the SDP graphs are identical. Since the SDP graphs are labelled, directed and acyclic graphs, they must meet the following conditions to be considered as isomorphic:

- they must have the same amount of edges
- their edge connectivity must be identical
- the edge labels must be identical, i.e. the dependency role must be identical
- the part-of-speech tag of nodes must be identical
- the vocabulary, the part-of-speech tag and the dependency role of their predicate node must be identical. The predicate of a SDP graph is the node which has only out degrees and no in degree.

[1] https://github.com/RobinL/fuzzymatcher

We choose to store the Syntactic Index in the JSON format. Each pattern class is stored under a unique predicate entry. We store the pattern class with the elements described in Table 14.1. An example of the "born_VERB" entry in the Syntactic Index is shown in Figure 14.5. Our code is available on GitHub[2].

Element	Description
graph	a SDP graph, as representative of the class
size	the number of edges of the representative graph
relations	the set of unique relations the class can express
ambiguous	True if the class can express more than one relation, False otherwise
support	support score of each relation in the class
entities	the type of entities involved in the relations the class can express
source_node	the node which correspond to the Source entity in the graph
target_node	the node which correspond to the Target entity in the graph

Table 14.1: Elements of a pattern entry in the Syntactic Index

14.2 Building the Lexical Index

14.2.1 Method

The Lexical Index consists of a list of vocabulary where to each word is associated the possible relations in which this word can appear. The vocabulary is extracted from the same lexico-syntactic patterns upon which the Syntactic Index is built.

This index is built by measuring an association score between the vocabularies of the lexico-syntactic patterns and the relations. The frequency counts of the words appearing in the lexico-syntactic patterns are normalised into TF-IDF weights. For instance, the word "born" has a high association score with the *dateOfBirth* and *placeOfBirth* relations, since it essentially appears in patterns which express these relations. On the other hand, it has lower or null association scores with other relations in which it is rarely or never used, such as *educatedAt* or *placeOfDeath*. We can set a threshold so as to remove from the index any words whose association weight is too low.

The Lexical Index is inspired by the weighted inverted index of the Explicit Semantic Analysis (ESA) method proposed by Gabrilovich and Markovitch (2007). The weighted inverted index represents a Wikipedia concept, i.e. Wikipedia articles, as a vector of words occurring in the content of the article. Each word is assigned a TF-IDF weight, so as to indicate the strength of association between the word and the concepts. Unlike the weighted inverted index of the ESA

[2] https://github.com/nicolasgutehrle/emontalproject

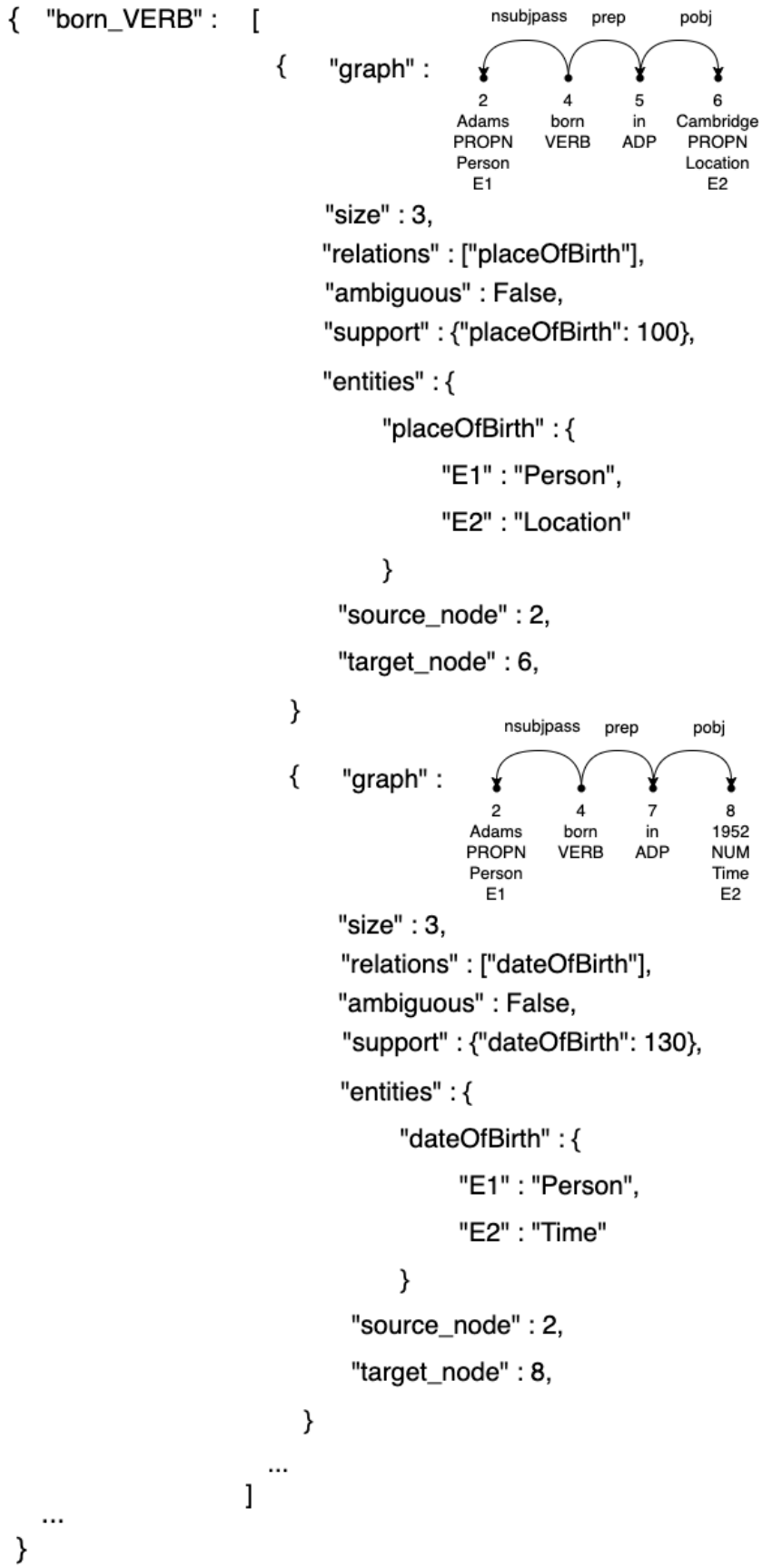


Figure 14.5: Example of the "born_VERB" entry in the Syntactic Index

method however, in our approach, we use the relations between entities as concepts, instead of Wikipedia articles.

14.2.2 Implementation

Our pipeline to build the Lexical Index takes as input the set of Shortest Dependency Path graphs, and is shown on Figure 14.6. Our implementation of the Lexical Index has three main hyper-parameters:

entry_type : the type of the surface form of the lexical units. It can be text, lemma or pos-tag, or a combination of any of those. The **entry_type** hyper-parameter is set to "text" by default

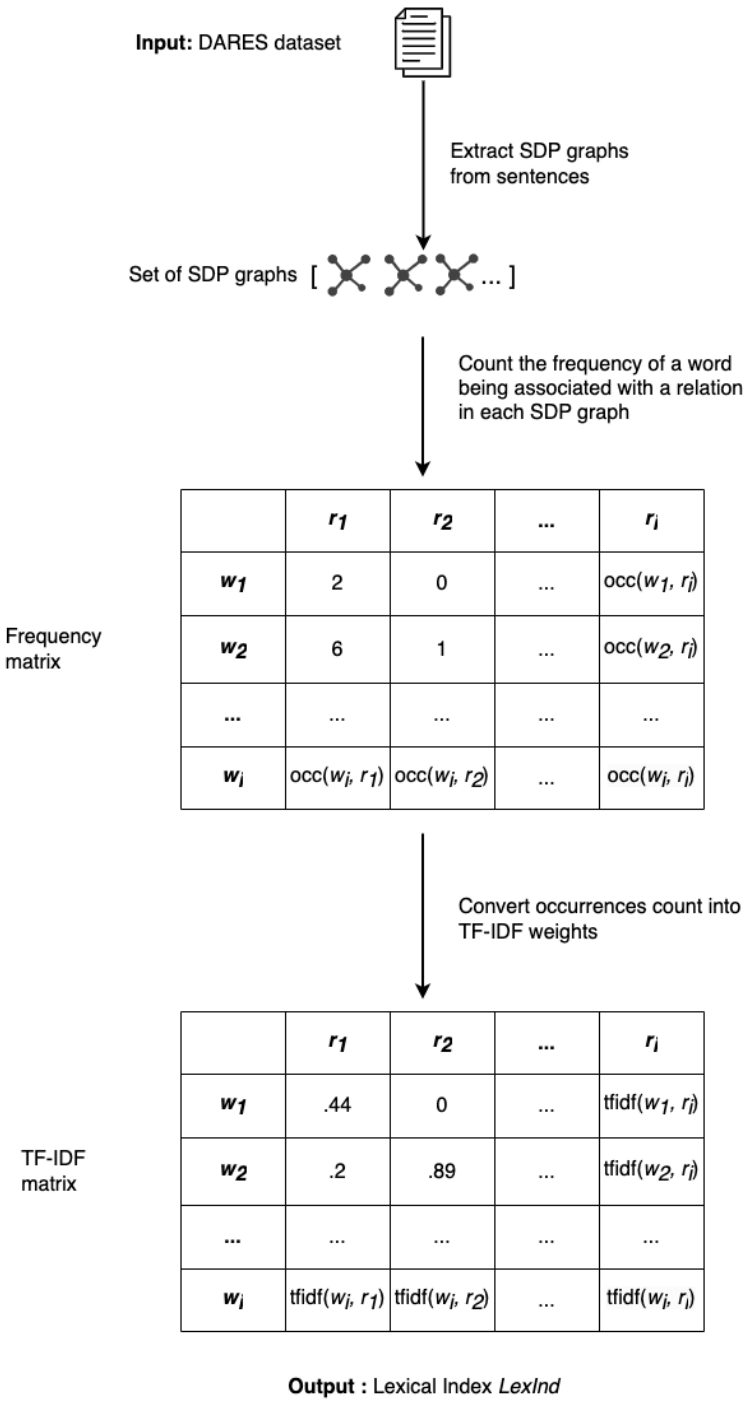
min_weight : sets a minimum threshold of the association score between the words and relations. The score of a word is automatically set to 0 if it is below this threshold. The **min_weight** hyper-parameter is set to 0 by default

pos-tag_filter : a list of part-of-speech tags to filter the lexical units. Any word of which the part-of-speech tag is in this list is removed from the index. By default, the **pos-tag_filter** is set to None, thus keeping every word, regardless of their part-of-speech tags

We store the frequency counts of words by relation in a $W \times R$ TF-IDF matrix, where R is the number of unique relations r express by the SDP graphs and W is the number of unique words w in the vocabulary of the SDP graphs. The words are stored in the matrix according to the surface form defined by the **entry_type** hyper-parameter.

We use the *TfidfTransformer* class from *scikit-learn* in Python to convert the word frequency matrix into a TF-IDF weight matrix. We choose to store the Lexical Index in the CSV format. Table 14.2 shows an example of the TF-IDF matrix obtained by the *tf-idf conversion* step. Words are stored as a concatenation of their lemma form and their part-of-speech tag. Our code is available on GitHub[3].

[3] https://github.com/nicolasgutehrle/emontalproject



	r_1	r_2	...	r_i
w_1	2	0	...	occ(w_1, r_i)
w_2	6	1	...	occ(w_2, r_i)
...	...	...	...	...
w_i	occ(w_i, r_1)	occ(w_i, r_2)	...	occ(w_i, r_i)

	r_1	r_2	...	r_i
w_1	.44	0	...	tfidf(w_1, r_i)
w_2	.2	.89	...	tfidf(w_2, r_i)
...	...	...	...	...
w_i	tfidf(w_i, r_1)	tfidf(w_i, r_2)	...	tfidf(w_i, r_i)

Figure 14.6: Main stages of the pipeline to build the Lexical Index

Words	occupation	placeOfBirth	nextInBodyWater	...	dateOfBirth	dateOfDeath	educatedAt
river_NOUN	0.000	0.000	1.000	...	0.000	0.000	0.000
university_NOUN	0.000	0.000	0.000	...	0.000	0.000	0.996
death_NOUN	0.219	0.000	0.000	...	0.000	0.902	0.000
...	...	...	...	...	...	...	
autor_NOUN	0.998	0.000	0.000	...	0.033	0.031	0.000
born_VERB	0.084	0.458	0.000	...	0.770	0.435	0.008

Table 14.2: Excerpt of a word-relations TF-IDF matrix. Words are stored as a concatenation of their lemma form and their part-of-speech tag. The **pos-tag_filter** and the **min_weight** hyper-parameters are set to None and 0 respectively

Chapter 15

Joint Extraction of Relations and Entities

Table of contents

In this chapter, we present the Joint Extraction of Relations and Entities pipeline which exploits the linguistic resources described in Chapter 14 to extract the mentions of relations and entities from sentences. Moreover, we propose an extension of the EMONTAL XML format originally presented in Section 11.4 in order to annotate the mentions of entities and relations in sentences.

The rest of this chapter is structured as follows: we first describe our Joint Extraction of Relations and Entities method in Section 15.1, before presenting its implementation in Section 15.2. Finally, we describe the extension of the EMONTAL XML format in Section 15.3.

15.1 Method

Our approach to the Joint Extraction of Relations and Entities task exploits the Syntactic and Lexical Indices that we have described above to extract and categorise entities and relations mentioned in a sentence. Our approach consists in the three following steps, which are shown in Figure 15.1:

1. *relation extraction*
2. *relation classification*
3. *named entity recognition*

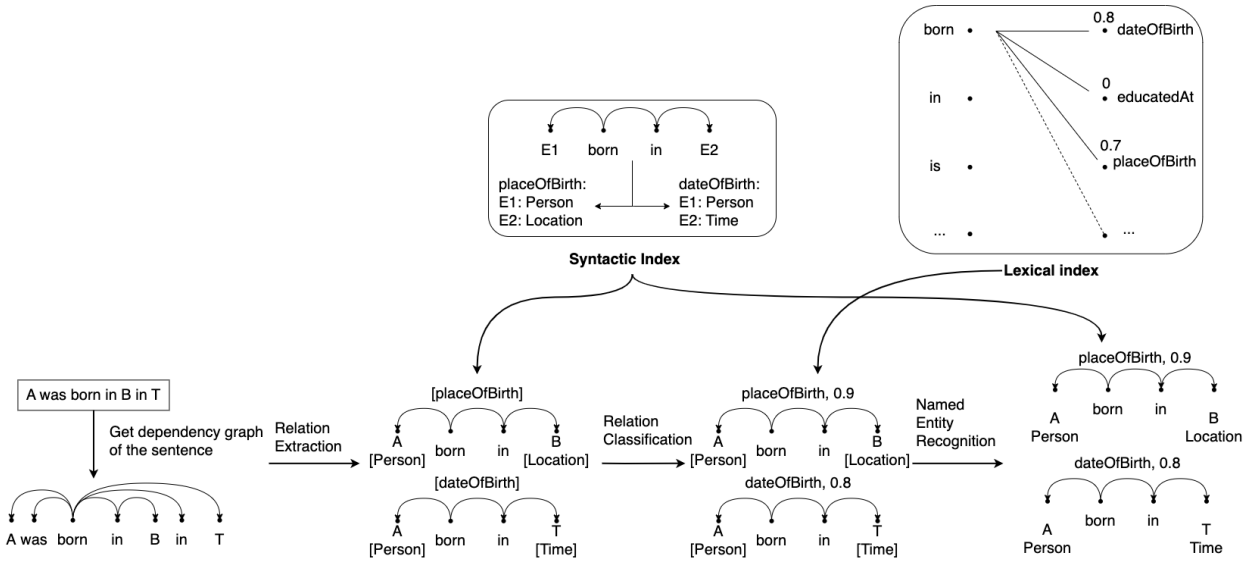


Figure 15.1: Main stages of our approach to the Joint Extraction of Relations and Entities task

The *relation extraction* step extracts the mentions of relations in the sentence. We apply the patterns stored in the Syntactic Index to find matching subgraphs in the sentence's dependency graph. Each candidate subgraph is associated with a set of possible relations, which are the relations the pattern can express according to the Syntactic Index. For instance, given the dependency graph of the sentence "The Republic of Guinea borders the Atlantic Ocean to the west and Senegal to the north" shown in Figure 15.2, we would find the candidate subgraphs shown in Figure 15.3 by applying the patterns stored in the Syntactic Index.

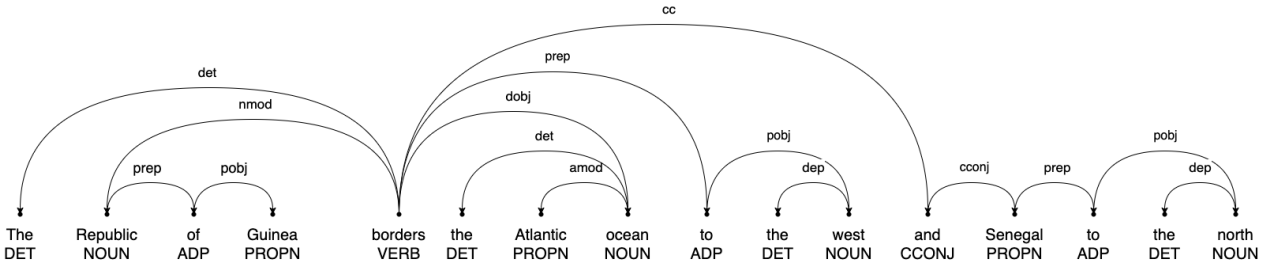


Figure 15.2: Dependency graph of the sentence "The Republic of Guinea borders the Atlantic Ocean to the west and Senegal to the north"

The *relation categorisation* step categorises the candidate subgraphs extracted in the *relation extraction* step. A candidate subgraph is categorised according to its vocabulary and by using the Lexical Index. Among the possible relations that are expressed by a candidate subgraph, we select the relation with which this subgraph has the highest association score. This score is calculated as the harmonic mean of the association scores of the vocabulary of the subgraph. We can set a threshold to ensure that the mean association score of the candidate subgraph with the predicted relation

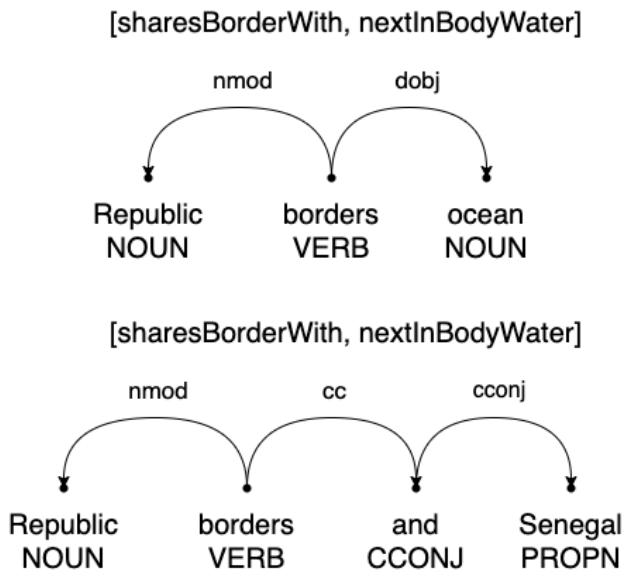


Figure 15.3: Candidate subgraphs extracted during the *relation extraction* step. Each subgraph is associated with a set of possible relations

is sufficiently high. For instance, given their vocabulary and their possible labels, the candidate graphs shown in Figure 15.3 are respectively categorised as *nextInBodyWater* and *sharesBorderWith* during the *relation categorisation* step, as shown in Figure 15.4.

Finally, the *named entity recognition* step determines the types of the entities involved in the relation. The entity types are determined by the lexico-syntactic pattern which matched the candidate subgraph in the *relation extraction* step, and the relation predicted in the *relation classification* step. For instance, as shown in Figure 15.5, every entity involved in the relations expressed by the two candidate graphs are of type *Location*.

15.2 Implementation

The main hyper-parameter of the our implementation of the JERE pipeline is the **semantic threshold** hyper-parameter. This hyper-parameter affects the output of *relation classification* step, which outputs the relation predicted for the candidate graph. The pipeline will output the predicted label if the mean association score is equal or higher than the semantic threshold. However, it will output the label *Other* if the semantic score is lower than the semantic threshold.

Our implementation takes as input the dependency graph of a sentence, from which it extracts the mentions of relations and the entities involved in them. The main steps of our implementation are shown in Figure 15.6.

The *relation extraction* step applies the patterns stored in the Syntactic Index to extract candi-

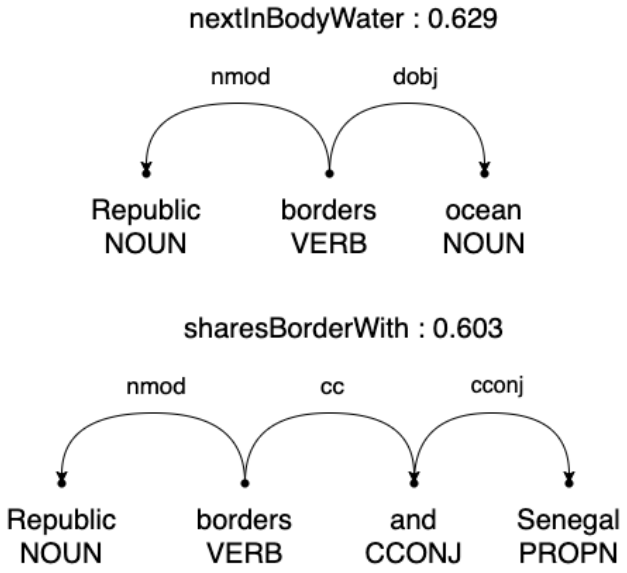


Figure 15.4: Candidate graphs categorised during the *relation classification* step

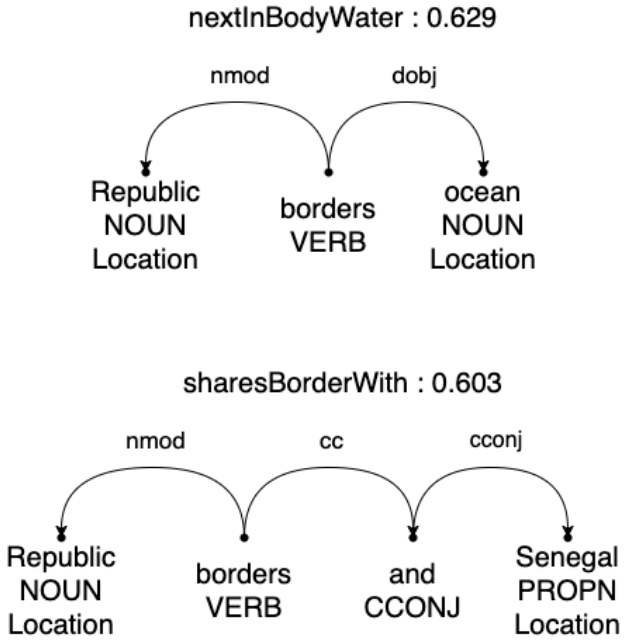


Figure 15.5: Entities involved in the relations categorised during the *named entity recognition* step

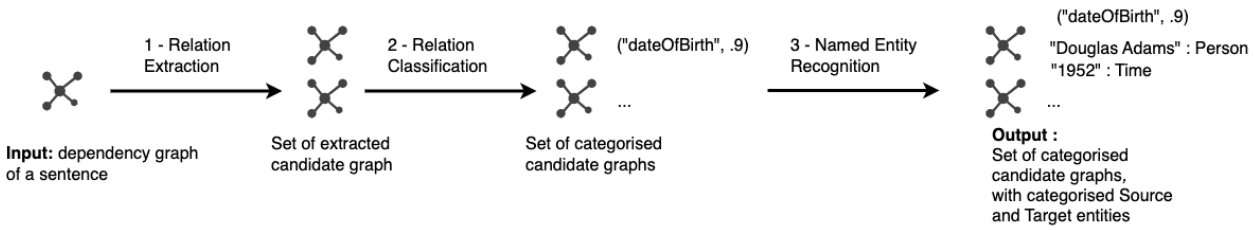


Figure 15.6: Joint Extraction of Relations and Entities pipeline

date graphs and the relations they can express from the dependency graph of a sentence. First, we search the Syntactic Index for an entry corresponding to a node in the sentence's dependency graph. Then, we apply the patterns stored in the matching entries to extract the candidate graph from the sentence's dependency graph. When extracting candidate graphs, we may extract candidates that are subgraphs of other candidates. Thus, we filter the set of candidate graphs, so as to only keep the longest possible candidates. We use the *networkx* library for any operation upon the graphs, such as the pattern matching step.

The *relation classification* step exploits the Lexical Index to categorise each candidate graph produced by the *relation extraction* stage. First, we produce a W x R matrix, where W is the number of words w from the candidate graph which has an entry in the Lexical Index, and R is the number of unique relation r in the set of possible relations. We measure the mean association score between the candidate graph and each possible relation by calculating the harmonic mean of each relation in the matrix. The candidate graph is assigned the relation label with which it has the highest mean association score. However, the candidate graph is assigned the label *Other* if the mean association score is below the value of the **semantic_threshold** hyper-parameter.

Figure 15.7 shows an example of the classification of the first candidate graph shown in Figure 15.4. The Lexical Index contains entries for the verb "borders" and the nouns "Republic" and "ocean". The candidate graph can express the *nextInBodyWater* and *sharesBorderWith* relations. Based on its vocabulary, the candidate graph is categorised with the *nextInBodyWater* relation, which obtains a mean association score of 0.629.

Finally, the *Named Entity Recognition* step determines the type and the boundaries of the Source and Target entities of the candidate graph. The types of the Source and Target entities are determined from the Syntactic Index, based on the pattern that matched in the *relation extraction* step and by the relation predicted in the *relation categorisation* step. The boundaries of an entity are based on the entity's root node in the sentence's dependency graph. The root node of an entity is determined by the pattern from the Syntactic Index which extracted the candidate graph. Moreover, we include in the entity's boundaries the nodes governed by the root node, which we filter according to their part-of-speech tags and dependency roles. We exclude every node of which part-of-speech tag is either a *punctuation*, a *pronoun*, a *coordinating conjunction*, an *adverb*, an *auxiliary* or a *determiner*. Similarly, we exclude every node of which dependency role is either *adnominal clause*, *apposition* or *conjunct*. We have defined the lists of filtering part-of-speech tags and dependency roles empirically from observation of the DARES dataset. We filter these nodes further by keeping the longest contiguous sequence of nodes which contains the entity's root node.

Figure 15.8 shows an example of the categorisation *Named Entity Recognition* step. Since the candidate graph shown in Figure 15.4 has been categorised as *nextInBodyWater* in the *relation classification* step, the type of the both entities is *Location*.

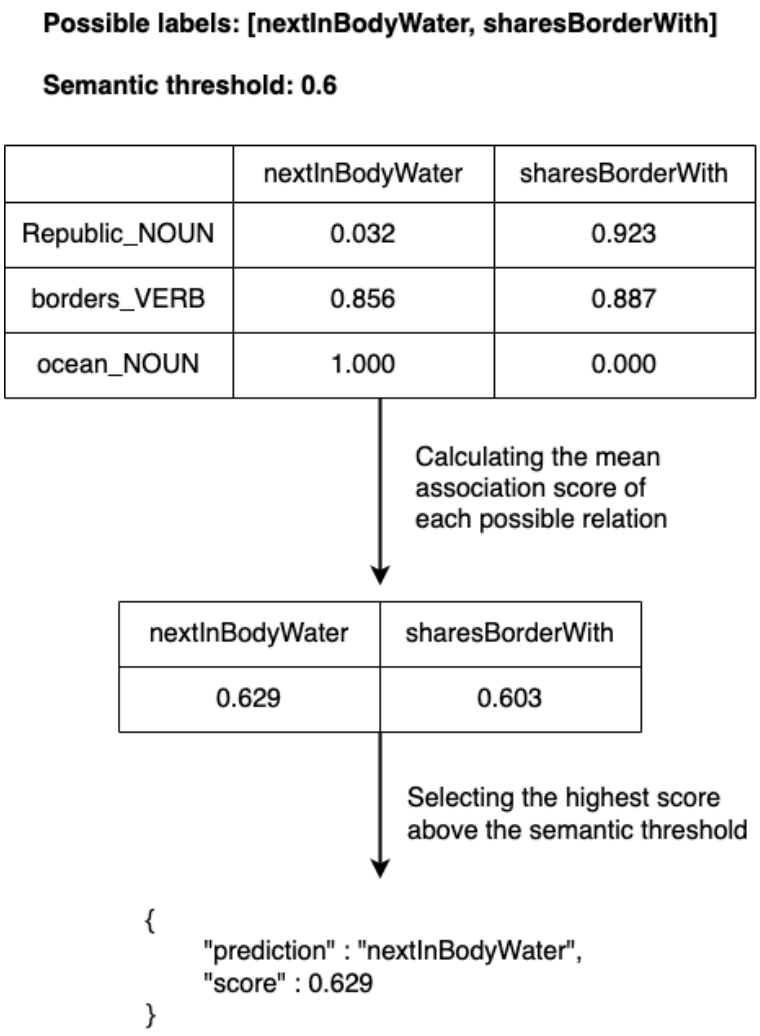


Figure 15.7: Example of the classification of the candidate graph shown in Figure 15.4

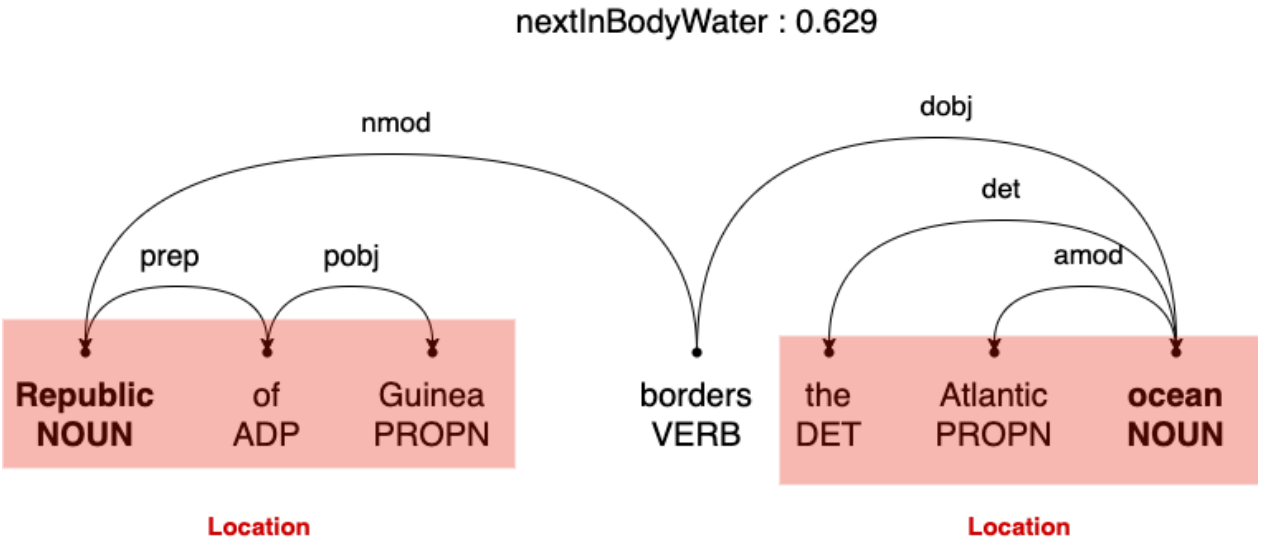


Figure 15.8: Example of the *Named Entity Recognition* step. The root node of each entity is shown in bold. The red boxes indicate the *Location* type

15.3 Annotating the mentions of relations and entities in the EMONTAL format

We propose an extension of the schema of the EMONTAL XML format originally presented in Section 11.4. This extension exploits the output of the JERE pipeline to annotate the mentions of entities and relations in the XML documents of the EMONTAL corpus. Table 15.1 describes the two tags added to the EMONTAL tagset described in Table 11.9. The complete XML Schema Definition (XSD) is shown in Annexe 19.4.

Tag name	Description
ent	mention of a Named Entity
rel	relation in which an entity is involved. This tag is nested inside a `ent` tag, and does not have a textual content

Table 15.1: Set of tags contained in the `content` tag of a document from our corpus in the EMONTAL XML format

`ent` tags have an *id* and a *type* attribute. The *id* attribute of `ent` tags is composed of their index number in the whole document, followed by their type, such as "pers_191". The *type* attribute indicates the type of the entity, such as *Person* or *Location*. The *id* attribute consists in the entity's type followed by a number. Figure 15.9 shows an example of `ent` and `rel` tags in an EMONTAL XML document from our corpus.

`rel` tags are nested within `ent` tags, and describe a relation in which an entity is involved.

The `rel` tag does not have a textual content. Each `rel` tag has an *id*, *type*, *score* and *anchor* attribute. The *id* attribute of `rel` tags is composed of the index number in the sentence, followed by the type of relation, such as *2_dateOfBirth*. The *type* attribute indicates the type of the relation, such as *educatedAt* or *spouse*. The *score* attribute is the association score predicted for this relation by the JERE pipeline. The *anchor* attribute is the anchor node of the pattern which extracted this relation, such as "born_VERB". The *id* attribute of `rel` tags connects `ent` tags that are involved in a relation. For instance, as shown in Figure 15.9, the `ent` tag with the "pers_191" *id* attribute describes an entity of type *Person*, whereas the `ent` tag with the "*loc_193*" *id* attribute describes an entity of type *Location*. These two tags share a `rel` tag, which has the *1_dateOfBirth id* attribute.

```
<sent id="sent_109">
 <ent id="pers_191" type="pers">
  <rel id="1_dateOfBirth" type="dateOfBirth" score="0.6991231441497803" rule="semantic" anchor="naître_VERB"/>
  <rel id="2_placeOfBirth" type="placeOfBirth" score="0.5120583176612854" rule="semantic" anchor="naître_VERB"/>
  Charles Geneviève Louis Auguste Thimolée d'Eon de Beaumont
  </ent>
  naquit à
  <ent id="loc_193" type="loc">
   <rel id="2_placeOfBirth" type="placeOfBirth" score="0.5120583176612854" rule="semantic" anchor="naître_VERB"/>
   Tonnerre
  </ent>
   en
  <ent id="time_194" type="time">
   <rel id="1_dateOfBirth" type="dateOfBirth" score="0.6991231441497803" rule="semantic" anchor="naître_VERB"/>
   1728
  </ent>
  ; il était fils du premier magistrat municipal de cette ville c'était, on le voit un très authentique
  <ent id="loc_195" type="loc">
  Bourguignon
  </ent>
</sent>
```

Figure 15.9: Examples of `ent` and `rel` tags from an EMONTAL XML document of our corpus (*Bulletin de la Société d'archéologie et d'histoire de Tonnerre*, June 1939)

Part IV

Evaluation and discussion of the ELIJERE approach

This part is dedicated to the evaluation and the discussion of the results obtained by our proposed ELIJERE approach on the Joint Extraction of Relations and Entities task. We evaluate two implementations of our approach: the first implementation, called *base ELIJERE model*, relies on the Syntactic Index to extract candidate graphs and on the Lexical Index to categorise these candidates based on their vocabulary. The second implementation, called *hybrid ELIJERE model*, also applies the syntactic patterns stored in the Syntactic Index to extract candidate graphs, but relies on a machine-learning based classifier to categorise these candidates based on their vocabulary. Our aim is to compare the classification of candidate graphs based on the association scores learned by the Lexical Index to existing approaches.

We first evaluate these two implementations on the DARES dataset, before evaluating them on the EMONTAL corpus. Our aim is to compare the results obtained by these models on the EMONTAL corpus with those obtained on the DARES dataset, and study how these scores are impacted by the errors produced by the OCR process, as well as by the differences in writing styles.

The rest of this part is structured as follows: in Chapter 16, we evaluate and discuss the two implementations of the ELIJERE approach on the DARES dataset. Finally in Chapter 17, we evaluate the two implementations of the ELIJERE approach on the EMONTAL corpus and propose a discussion of the results.

Chapter 16

Evaluation of the ELIJERE approach on the DARES dataset

Table of contents

In this chapter, we evaluate the ELIJERE approach on the Joint Extraction of Relations and Entities task. We evaluate two implementations of the ELIJERE approach: the first implementation relies on the Syntactic Index to extract candidate graphs and the Lexical Index to categorise these candidates based on their vocabulary. We call this first implementation the *base ELIJERE model*. The second implementation also applies the syntactic patterns stored in the Syntactic Index to extract candidate graphs, but relies on a machine-learning based classifier to categorise these candidates based on their vocabulary. We call this second implementation the *hybrid ELIJERE model*.

We evaluate each model on the Relation Extraction task and the Named Entity Recognition task separately. Our aim is to compare the classification of candidate graphs based on the association scores learned by the Lexical Index to existing approaches.

The rest of this chapter is organised as follows: we first describe the evaluation protocol and evaluation dataset in Section 16.1 and 16.2 respectively. We present the evaluation of the *base* and *hybrid ELIJERE model*s on the Relation Extraction and Named Entity Recognition task in Section 16.3 and Section 16.4 respectively. Finally, we compare and discuss the results of both models in Section 16.5. The structure of this chapter is presented in Figure 16.1.

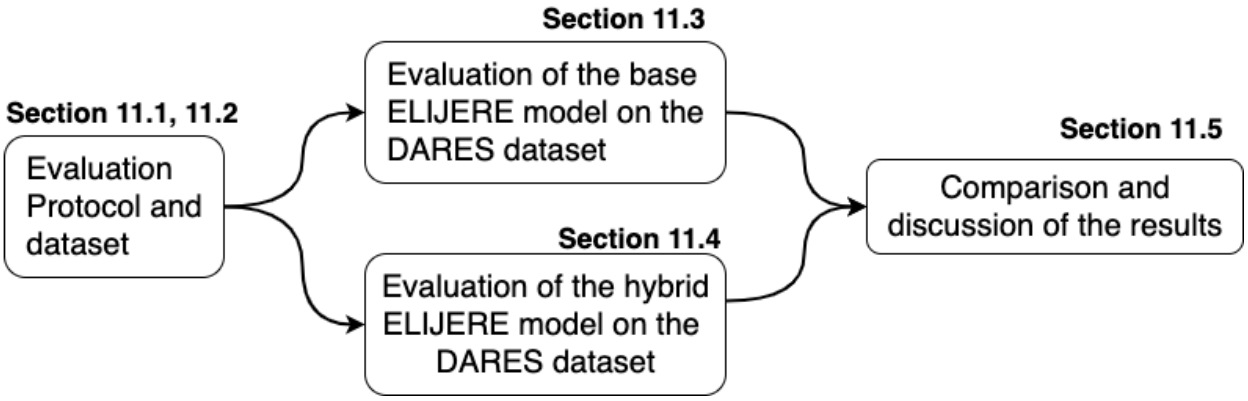


Figure 16.1: Structure of Chapter 16

16.1 Evaluation Protocol

We evaluate the ability of the models to categorise the relation expressed by a candidate graph in terms of Precision, Recall and F1 scores. We evaluate the ability of the model to determine the type and boundaries of the entities involved in the relation in terms of Precision, Recall and F1 scores, under the type, partial, strict and exact evaluation settings, as described in the SemEval evaluation scheme (Segura-Bedmar, Martínez, & Herrero-Zazo, 2013). We make ten evaluations of both models, for different values of the semantic threshold between 0 and 0.9. Our aim is to study the impact of the semantic threshold upon the performance of the models. When categorising the relation of the candidate graph, the model outputs the label *Other* if any of the following conditions is met:

- the model cannot find an entry in the Syntactic Index corresponding to the predicate node of the candidate graph
- the model cannot find a syntactic pattern in the Syntactic Index matching the candidate graph
- the semantic score predicted by the model is below the semantic threshold value

16.2 Description of the evaluation dataset

The distribution of relations and entities in the test set of the DARES dataset are respectively shown in Table 16.1 and Table 16.2. 18 % of the relations of the test set are labelled as *Other*. We ignore this label for the evaluation of the ELIJERE approach.

Labels	Count	Proportion
capitalOf	140	11.91 %
country	183	15.57 %
dateOfBirth	71	6.042 %
dateOfDeath	39	3.319 %
educatedAt	13	1.106 %
headOfGoverment	15	1.276 %
inception	12	1.021 %
memberOf	53	4.510 %
nextInBodyWater	47	4.000 %
occupation	125	10.638 %
Other	213	18.127 %
placeOfBirth	67	5.702 %
sharesBordersWith	174	14.808 %
spouse	23	1.957 %
Total	1,175	100 %

Table 16.1: Distribution of the relation labels in the test set of the DARES dataset

Label	Count	Proportion
Person	391	19.628 %
Location	1,337	67.118 %
Time	128	6.425 %
Misc	136	6.827 %
Total	1,992	100 %

Table 16.2: Distribution of the entity labels in the test set of the DARES dataset

16.3 Evaluation of the *base ELIJERE model*

The *base ELIJERE model* is an implementation of the ELIJERE approach presented in Part III. The model relies on the Syntactic Index to extract candidate graph and the Lexical Index to categorise these candidates based on their vocabulary. Both indices are built on the train set of the DARES dataset.

In the following sections, we first describe the implementation of the Syntactic and Lexical indices upon which the *base ELIJERE model* relies, before evaluating the model on the Relation Extraction task and the Named Entity Recognition task separately.

16.3.1 Description of the linguistic resources built on the DARES dataset

In this section, we present the Syntactic Index and the Lexical Index that we have built from the DARES dataset. We choose the following hyper-parameter values to build the Syntactic Index:

predicate_type : {lemma, pos}

support_threshold : 2

We set a minimum support threshold of 2 so as to filter out syntactic patterns which have only been annotated with a label once. Since the index is trained on a weakly annotated dataset, many wrongly labelled patterns can be removed in this way. We set the textual value of predicate nodes as a concatenation of their lemma and part-of-speech tag. This allows us to gather various forms of the same predicate node into the same entry. Table 16.3 shows the distribution of unique predicates, unique patterns, ambiguous patterns and mean number of patterns per predicate in the Syntactic Index.

Unique predicates	168
Unique patterns	322
Ambiguous patterns	4
Mean number of pattern per predicate	1

Table 16.3: Distribution of unique predicates, unique patterns, ambiguous patterns and mean number of patterns per predicate in the Syntactic Index

Figure 16.3 shows an excerpt of the entry "*naître_VERB*" (born_VERB) in the Syntactic Index, trained on the DARES dataset. Figure 16.2 shows the graph representation of the first extraction pattern in this entry. Each node is represented by its textual value, whereas edges are represented by their dependency label. This graph expresses the *dateOfBirth* relation.

We choose the following hyper-parameter values to build the Lexical Index:

entry_type : {lemma, pos}

min_weight : 0

pos-tag_filter : {PROPN}

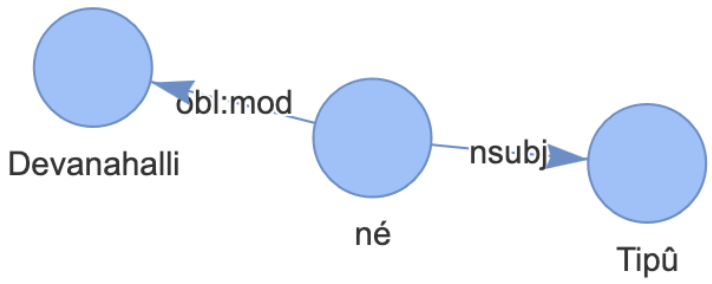


Figure 16.2: Graph representation of the first extraction pattern in the "*naître_VERB*" (born_VERB) entry in the Syntactic Index. This graph expresses the *dateOfBirth* relation

We set the textual value of the lexical units as a concatenation of their lemma and part-of-speech tag, so as to gather various forms of the same word into the same entry. We exclude proper nouns from the vocabulary, i.e. any word whose part-of-speech tag is *PROPN*, since these will usually indicate Named Entities. Figure 16.4 shows an excerpt of the Lexical Index built on the DARES dataset.

16.3.2 Evaluation on the Relation Extraction task

We evaluate the model on the Relation Extraction task in terms of Precision, Recall and F1 scores on the test set of the DARES dataset. The results of each evaluation are shown in Table 16.4. On average across all semantic thresholds, the model achieves a Precision score of 0.518, a Recall score of 0.120, and a F1 score of 0.189. The model achieves the best scores when the semantic threshold is set to 0, where it achieves a Precision score of 0.643, a Recall score of 0.148, and a F1 score of 0.233. It achieves the lowest scores when the semantic threshold is set to 0.9, where it achieves a Precision score of 0.357, a Recall score of 0.079, and a F1 score of 0.127.

The model achieves higher Precision scores than Recall scores. This suggests that the syntactic patterns stored in the Syntactic Index upon which the model relies capture a small proportion of all relations in the texts. The lower Recall scores also suggest that the Syntactic Index lacks the syntactic patterns required to categorise the candidate graphs of the test set. This lack of patterns can explain the average Precision scores, since the model outputs the label *Other* if it cannot find a syntactic pattern in the Syntactic Index matching the candidate graph. The model also outputs the label *Other* if the predicted semantic score is lower than the semantic threshold. This can explain

```
{"naître_VERB": [
    {
        "graph": {
            "directed": true,
            "multigraph": false,
            "graph": {},
            "nodes": [
                {
                    "text": "Laura",
                    "lemma": "Laura",
                    "pos": "PROPN",
                    "dep": "nsubj",
                    "char_idx": 0,
                    "id": 0
                },
                {
                    "text": "née",
                    "lemma": "naître",
                    "pos": "VERB",
                    "dep": "ROOT",
                    "char_idx": 12,
                    "id": 3
                },
                {
                    "text": "6",
                    "lemma": "6",
                    "pos": "NUM",
                    "dep": "obl:mod",
                    "char_idx": 19,
                    "id": 5
                }
            ],
            "links": [
                {
                    "dep": "nsubj",
                    "source": 3,
                    "target": 0
                },
                {
                    "dep": "obl:mod",
                    "source": 3,
                    "target": 5
                }
            ]
        },
        "size": 2,
        "props": [
            {
                "name": "dateOfBirth",
                "support": 71
            },
            {
                "name": "dateOfDeath",
                "support": 10
            }
        ],
        "ambiguous": 1,
        "source_types": [
            "Q5"
        ],
        "source_nodes": [
            0
        ],
        "target_types": [
            "time"
        ],
        "target_nodes": [
            5
        ],
        "ner_rules": {
            "dateOfBirth": {
                "source_type": "Q5",
                "target_type": "time"
            },
            "dateOfDeath": {
                "source_type": "Q5",
                "target_type": "time"
            }
        },
        "i": 1
    }
]
}
```


Figure 16.3: Excerpt of the entry "*naître_VERB*" (born_VERB) in the Syntactic Index built on the DARES dataset

	occupation	capitalOf	country	placeOfBirth
accueillir_VERB	0.0	0.7274244655538866	0.16690915905316975	0.0
un_PRON	0.12475522067989421	0.0	0.2645039652951811	0.0
bourse_NOUN	0.0	0.0	0.8133060482789962	0.0
sépare_VERB	0.0	0.0	0.0	0.0
rivière_NOUN	0.0	0.0	0.0	0.0
devenir_VERB	0.08664047322057213	0.2001437429243121	0.12246247009143181	0.04244568657461459
baie_NOUN	0.0	0.9467886518974421	0.0	0.0
ville_NOUN	0.0	0.5980367526688698	0.3659222959819651	0.2282926478109718
situe_VERB	0.0	0.0	0.927282363269959	0.0
nord_NOUN	0.0	0.2019673902594308	0.27805119729122774	0.0
mégalopole_NOUN	0.0	0.0	1.0	0.0
nord-est_NOUN	0.0	0.0	0.32212710299457653	0.22329953071643988
connaître_VERB	0.2743370966560816	0.6337321497304533	0.38776333098603744	0.1343993861945909
université_NOUN	0.0	0.04304051165628626	0.009875740981984774	0.0

Figure 16.4: Excerpt of the Lexical Index trained on the DARES dataset

the fact that the scores obtained by the model decrease as the semantic threshold increases, since the model will most likely output the label *Other*.

Threshold	P	R	F1
0.0	**0.643**	**0.148**	**0.233**
0.1	0.602	0.144	0.220
0.2	0.602	0.144	0.225
0.3	0.591	0.139	0.217
0.4	0.587	0.133	0.211
0.5	0.515	0.118	0.186
0.6	0.440	0.111	0.174
0.7	0.420	0.098	0.155
0.8	0.420	0.086	0.136
0.9	0.357	0.079	0.127
Mean	0.518	0.120	0.189

Table 16.4: Precision, Recall and F1 scores of *base ELIJERE model* on the test sets of the DARES dataset

Table 16.5 shows the Precision, Recall and F1 scores obtained by *base ELIJERE model* on the test set of the DARES dataset for each label when the semantic threshold is set to 0. We select the results obtained for this semantic threshold since it is the minimal threshold where the model achieves the best scores.

The model achieves high to very high Precision scores for multiple labels: it achieves a Pre-

cision score of 1.000 for the *dateOfBirth, dateOfDeath, occupation* and *placeOfBirth* labels. It also achieves high Precision scores for the *headOfGovernment, memberOf* and *spouse* label. The model performs the best on the *occupation* relation, with a Precision score of 1.000, a Recall score of 0.308 and a F1 score of 0.471. However, it achieves lower Recall and F1 scores. It especially fails to properly annotate any instance of the *sharesBorderWith* label. This suggests the patterns to express this label in the test set do not appear in the train set, upon which the Syntactic Index has been built. In general, the Precision scores being higher than the Recall scores suggests that the patterns stored by the Syntactic Index are precise, but lack diversity.

Label	P	R	F1
capitalOf	0.571	0.056	0.103
country	0.667	0.066	0.120
dateOfBirth	**1.000**	0.201	0.331
dateOfDeath	**1.000**	0.205	0.337
educatedAt	0.550	0.287	0.374
headOfGoverment	0.750	0.134	0.222
inception	0.500	0.071	0.125
memberOf	0.702	0.185	0.292
nextInBodyWater	0.393	0.079	0.131
occupation	**1.000**	**0.308**	**0.471**
placeOfBirth	**1.000**	0.205	0.339
sharesBordersWith	0.000	0.000	0.000
spouse	0.875	0.275	0.418

Table 16.5: Precision, Recall and F1 scores obtained by *base ELIJERE model* for each label when the semantic threshold is set to 0 on the test set of the DARES dataset

Table 16.6 shows the error types and distribution of samples wrongly labelled by *base ELIJERE model* on the test set of the DARES dataset. We identify four different types of error: *predicate not found, pattern not found, wrong label* and *semantic score too weak*.

The *predicate not found* error appears when the predicate node of the candidate graph is not present in the Syntactic Index. For instance, the *memberOf* relation in the sentence "*La Catalogne est une communauté de l'Espagne*" ("Catalonia is a community of Spain") is expressed by the pattern (`Catalogne; communauté; Espagne`) (`(Catalonia; community; Spain)`), here shown as a triple for simplicity . The predicate node of this pattern is the noun "*communauté*" (community). However, the Syntactic Index does not have an entry for the noun *communauté*. Thus, the model cannot predict a label for this candidate graph, and outputs the label *Other* instead. Depending on the value of the semantic threshold, 54.555 % to 58.505 % of errors produced by the model are of the *predicate not found* type. On average, 56.572 % of errors produced by the model are of this type.

	Predicate not found		Pattern not found		Wrong label		Semantic Score too weak		
Threshold	Count	Proportion	Count	Proportion	Count	Proportion	Count	Proportion	Total
0.0	509	58.505 %	326	37.471 %	32	3.678 %	3	0.344 %	870
0.1	509	57.972 %	326	37.129 %	26	2.961 %	17	1.936 %	878
0.2	509	57.972 %	326	37.129 %	26	2.961 %	17	1.936 %	878
0.3	509	57.191 %	326	36.629 %	20	2.247 %	35	3.932 %	890
0.4	509	57.062 %	326	36.547 %	20	2.242 %	37	4.147 %	892
0.5	509	56.057 %	326	35.903 %	20	2.202 %	53	5.837 %	908
0.6	509	55.872 %	326	35.784 %	17	1.866 %	59	6.476 %	911
0.7	509	55.628 %	326	35.628 %	9	0.983 %	71	7.759 %	915
0.8	509	54.908 %	326	35.167 %	8	0.862 %	84	9.061 %	927
0.9	509	54.555 %	326	34.941 %	0	0.000 %	98	10.503 %	933
Mean	509	56.572 %	326	36.232 %	17	2.000 %	47	5.193 %	900

Table 16.6: Error types and distribution of samples wrongly labelled by the *base ELIJERE model* on the test set of the DARES dataset

The *pattern not found* error appears when the predicate node is present in the Syntactic Index, but there is no syntactic pattern matching the candidate graph. For instance, the *nextInBodyWater* relation in the sentence "*Le Japon est un pays situé dans l'océan Pacifique*" ("Japan is a country located in the Pacific Ocean") is expressed by the pattern `(Japon; situé; océan) ((Japan; located; ocean))` in the predicate node of this pattern is the verb "*situé*" (located). The Syntactic Index contains an entry for this verb, however it does not store a matching extraction pattern. Thus, the model cannot predict a label for this candidate graph, and outputs the label *Other* instead. Depending on the value of the semantic threshold, 34.941 % to 37.471 % of errors produced by the model are of the *pattern not found* type. On average, 36.232 % of errors produced by the model are of this type.

Both *predicate not found* and *pattern not found* errors are caused by a lack of pattern in the Syntactic Index, and can be solved by increasing the diversity of syntactic patterns known by the Syntactic Index. A first solution to increase this diversity would be to build the Syntactic Index on a dataset of documents of various origins. Currently, the index is built on a dataset composed of only Wikipedia articles. Thus, the writing style of documents is identical. By including documents from other origins such as the World Wide Web or other encyclopedias, we could integrate various writing styles to the dataset, and thus increase the diversity of patterns known to the Syntactic Index.

To increase the diversity of the index, we could also use prompts to ask a generative Large Language Model such as ChatGPT to generate sentences which express specific relations, or ask it to rephrase known sentences. This process could be repeated until enough data have been collected. However, this process should be done manually, in order to ensure the quality of the generated sentences.

The amount of *predicate not found* and *pattern not found* errors may also be caused by the reliance of the model on the dependency structure of the candidate graph, as the model must find a syntactic pattern which exactly matches the structure of the candidate. A potential solution to this issue would be to build surface extraction patterns from the syntactic patterns stored in the Syntactic Index. For instance, given the sentence "*George Washington est né à Westmoreland County*" (George Washington was born in Westmoreland County), the system would learn an extraction pattern by extracting the Shortest Dependency Path (SDP) graph between the entities "George Washington" and "Westmoreland County". By taking into account the position of each token of the SDP graph in the original sentence, we can convert this syntactic pattern into the following surface pattern, expressed as a regular expression: `([A-Z]\w+ ?){1,2}.*(né).*([A-Z]\w+ ?){1,2}`. Each capturing group in the regular expression would correspond to a token or group of tokens from the Shortest Dependency Graph path, whereas anything outside of the capturing groups could be discarded. Such surface extraction patterns would allow to annotate the text in a non-contiguous manner, while removing the reliance of the system on the part-of-speech and dependency role analysis at extraction time.

The *wrong label* error appears when the model predicts an incorrect label. For instance, the pattern `(Louis XIV; devenu; roi) ((Louis XIV; became; king))` describes the *occupation* relation. However, the model incorrectly predicts the *educatedAt* relation with an association score of 0.415 when the semantic threshold is set to a minimum of 0. Depending on the value of the semantic threshold, 0.000 % to 3.678 % of errors produced by the model are of the *wrong label* type. On average, 2.000 % of errors produced by the model are of this type. Compared to the other error types, the *wrong label* error is the error less produced by the model. This shows the ability of the model to exploit the Lexical index to properly categorise the relation expressed by a candidate graph. The number of these errors decreases while the semantic threshold value increases. Thus, we can avoid such error by setting a higher semantic threshold.

The *wrong label* can also originate from incorrect syntactic patterns stored in the Syntactic Index or incorrect association scores from the Lexical index. The Syntactic and Lexical Indices are built from sentences annotated by applying the distant supervision method. Thus, some annotations may not be correct, leading to the Syntactic Index storing incorrect patterns and leading to the Lexical Index learning incorrect association scores between words and relations. Hence, we can reduce the amount of errors of type *wrong label* produced by the model by first correcting the incorrect annotations of the sentences upon which the Syntactic and Lexical indices are built. These corrections can be done manually or in an automatic way. For instance, we could measure the semantic similarity between sentences in order to cluster sentences expressing the same meaning together, while removing outliers and incorrect annotations. The similarity between sentences could be measured according to several methods, such as TF-IDF scores or word embeddings.

The *semantic score too weak* error appears when the classifier has not predicted a label because the semantic score is under the semantic threshold. For instance, the Shortest Dependency Graph `(Adams; né; Cambridge) ((Adams; born; Cambridge))` is labelled with the *placeOfBirth* relation. The model predicts the label *placeOfBirth* with a semantic score of 0.458. However, the model will output the *Other* label if the semantic threshold is at least set to 0.5. Depending on the value of the semantic threshold, 0.344 % to 10.503 % of errors produced by the model are of the *semantic score too weak* type. On average, 5.193 % of errors produced by the model are of this type. These errors increase alongside the increase of the semantic threshold value. Thus, we can avoid such error by setting a lower semantic threshold. We can also avoid this error by increasing the amount of annotated sentences from the training set, so that the Lexical Index learns better and more confident association scores between words and relations.

The *semantic score too weak* error may also be caused by words missing from the Lexical Index. As this index only stores association scores between words and relations for words it has seen in the training set. Thus, the Lexical Index suffers from the Out-Of-Vocabulary issue, since it is not able to produce a vector for words it does not know. A first possible solution to this issue would be to drastically increase the vocabulary of the Lexical Index. Another possible solution would be to learn association scores between subwords and relations, instead of learning association scores between words and relations. When producing the vector of a word, the Lexical Index would concatenate the vector representation of its subwords. Thus, the Lexical Index could produce a vector for any word. These subwords could be obtained by character n-gram of words, or other methods, as in the *fastText* word embeddings method.

In conclusion, the scores obtained by the *base ELIJERE model* on the DARES dataset demonstrate the validity of our method. The model achieves higher Precision scores than Recall scores, suggesting the Syntactic Index upon which the model relies stores precise patterns but lacks diversity. Moreover, the study of the types of errors produced by the model suggests that most errors are caused by syntactic patterns lacking from the Syntactic Index. Thus, the performance of the model can be improved by increasing the diversity of patterns stored in the Syntactic Index. Other errors produced by the model can be solved by first correcting the wrong annotations produced by the distant supervision method, thus improving the training set upon which the Syntactic and Lexical indices are built.

16.3.3 Evaluation on the Named Entity Recognition task

Table 16.7 shows the Precision, Recall and F1 scores obtained by the *base ELIJERE model* on the Named Entity Recognition task for each semantic threshold under each evaluation setting. The model achieves high to very high Precision scores, from 0.974 to 1.000 in the type setting, from

0.810 to 0.846 in the partial setting, from 0.603 to 0.684 in the strict setting and from 0.619 to 0.693 in the exact setting. On average, the model achieves a Precision score of 0.982 under the type setting, of 0.829 under the partial setting, of 0.649 under the strict setting, and 0.659 under the exact setting. The high scores under the type and partial settings indicate the ability of the model to properly determine the type of an entity, as well as to partially determine the boundaries of entities.

However, the model achieves lower Recall scores, from 0.063 to 0.156 in the type setting, from 0.051 to 0.134 in the partial setting, from 0.039 to 0.108 in the strict setting, and from 0.039 to 0.110 in the exact setting. On average, the model achieves a Recall score of 0.111 under the type setting, of 0.095 under the partial setting, of 0.074 under the strict setting, and of 0.076 under the exact setting

Consequently, the model achieves very low F1 scores, from 0.119 to 0.269 in the type setting, from 0.097 to 0.232 in the partial setting, from 0.074 to 0.187 in the strict setting, and from 0.074 to 0.189 in the exact setting. On average, the model achieves a F1 score of 0.199 under the type setting, of 0.169 under the partial setting, of 0.133 under the strict setting and of 0.135 under the exact setting.

The lower scores achieved under the exact and strict settings suggest that the model struggles to identify the exact boundaries of the entities. The boundaries of entities are determined by a set of rules which rely on the part-of-speech and dependency tags of words. These lower scores suggest that these rules are not sufficient and must be revised.

The lower Recall and F1-scores are correlated with the low Recall and F1-scores obtained by the model on the Relation Extraction task, as the prediction of the entity types depends on the predicted relation. For instance, during the Relation Extraction step, the model may have incorrectly predicted the *occupation* label instead of the *placeOfBirth* label for a candidate graph. Hence, the target entity is incorrectly labelled *Misc* instead of *Location*. Thus, we can improve the Recall and F1 scores obtained on the Named Entity Recognition task by first improving the Recall and F1 scores obtained on the Relation Extraction task.

Table 16.8 shows the F1 scores obtained by the *base ELIJERE model* for each entity type (*Person, Location, Time, Misc*) for each semantic threshold under each evaluation setting. The model performs the best under the type setting regardless of the semantic threshold. This shows the ability of the system to detect the correct entity types of the relation. However, the lower scores in the strict and exact settings indicate that the model struggles to correctly identify the boundaries of the entities. The null scores for the *Time* type under the strict and exact settings when the semantic threshold is set to 0.9 indicate the model fails to correctly identify the boundaries of this entity type, even partially.

The model identifies the *Misc* type the best, with a mean F1-scores, of 0.453 in the type setting. This is correlated with the fact that the model detected the *occupation* relation the best on the

Threshold	Setting	P	R	F1
0.0	type	**0.985**	0.156	0.269
	partial	**0.846**	0.134	0.232
	strict	**0.684**	0.108	0.187
	exact	**0.693**	0.110	0.189
0.1	type	**0.984**	0.142	0.248
	partial	**0.838**	0.121	0.212
	strict	**0.667**	0.096	0.168
	exact	**0.677**	0.098	0.171
0.2	type	**0.984**	0.142	0.248
	partial	**0.838**	0.121	0.212
	strict	**0.667**	0.096	0.168
	exact	**0.677**	0.098	0.171
0.3	type	**0.981**	0.124	0.220
	partial	**0.839**	0.106	0.188
	strict	**0.666**	0.084	0.149
	exact	**0.678**	0.085	0.152
0.4	type	**0.981**	0.122	0.217
	partial	**0.840**	0.105	0.186
	strict	**0.669**	0.083	0.148
	exact	**0.680**	0.084	0.151
0.5	type	**0.979**	0.106	0.191
	partial	**0.825**	0.089	0.161
	strict	**0.637**	0.069	0.125
	exact	**0.650**	0.070	0.127
0.6	type	**0.977**	0.100	0.181
	partial	**0.815**	0.083	0.151
	strict	**0.616**	0.063	0.115
	exact	**0.630**	0.064	0.117
0.7	type	**0.974**	0.087	0.161
	partial	**0.810**	0.073	0.134
	strict	**0.603**	0.054	0.100
	exact	**0.619**	0.056	0.103
0.8	type	**0.975**	0.075	0.139
	partial	**0.830**	0.064	0.119
	strict	**0.648**	0.050	0.093
	exact	**0.660**	0.051	0.095
0.9	type	**1.000**	0.063	0.119
	partial	**0.813**	0.051	0.097
	strict	**0.626**	0.039	0.074
	exact	**0.626**	0.039	0.074
Mean	type	**0.982**	0.111	0.199
	partial	**0.829**	0.095	0.169
	strict	**0.649**	0.074	0.133
	exact	**0.659**	0.076	0.135

Table 16.7: Precision, Recall and F1 scores obtained by the *base ELIJERE model* on the Named Entity Recognition task for each semantic threshold on the test set of the DARES dataset

Relation Extraction task, which is the only relation where the *Misc* type appears. The model struggles the most to identify the *Location* type, and achieves a mean F1 score of 0.105 under the type setting. This is correlated with the fact that the model struggles to correctly categorise relations such as *nextInBodyWater* or *sharesBorderWith* which involve entities of type *Location*, as shown in Table 16.5.

In conclusion, the high Precision scores obtained by the *base ELIJERE model* under the type setting suggests the ability of our approach to correctly categorise the entities involved in a relation, based on the relation predicted in the *relation classification* step. However, the lower scores obtained under the other settings suggest that the rules we have set to determine the boundaries of the entities are not sufficient and must be revised. The model achieves higher Precision scores than Recall scores, as in the Relation Extraction task. As the categorisation of Named Entities is based on the output of the Relation Extraction task, we must first improve the scores of the model on the Relation Task in order to improve its scores on the NER task.

16.4 Evaluation of the *hybrid ELIJERE model*

The *hybrid ELIJERE model* is a second implementation of the ELIJERE approach presented in Part III. The model relies on the Syntactic Index to extract candidate graphs and a machine-learning based classifier to categorise these candidates based on their vocabulary. The Syntactic Index upon which this model relies is the same index as the *base ELIJERE model*, whereas the machine-learning classifier is trained on the train set of the DARES dataset.

In the following sections, we first describe our process to select the machine-learning classifier most adapted to the JERE task, before evaluating the *hybrid ELIJERE model* on the Relation Extraction task and the Named Entity Recognition task separately on the DARES dataset.

16.4.1 Selecting a model

We compare the performances of the three following machine-learning algorithms: Support Vector Machine (SVM), Random Forest and XGBoost. We use the implementation of the Python library *scikit-learn* for the first two algorithms proposed, and the implementation from the *xgboost* Python library for the last one. Our aim is to determine which classifier is the best suited to categorise candidate graphs in the *hybrid ELIJERE model*. We train these models on the DARES dataset described in Section 13.3. We encode the textual content of the graphs in this dataset with the pre-trained ConceptNet Numberbatch word embeddings (Speer et al., 2018). Since these embeddings have been trained on the ConceptNet Knowledge Graph, they contain common-sense knowledge which makes them suited for our experiment.

Threshold	Setting	Per.	Loc.	Time	Misc
0.0	type	0.407	0.193	0.325	**0.466**
	partial	0.351	0.178	0.187	**0.375**
	strict	**0.296**	0.155	0.050	0.254
	exact	**0.296**	0.158	0.050	0.264
0.1	type	0.407	0.159	0.325	**0.466**
	partial	0.351	0.146	0.187	**0.375**
	strict	**0.296**	0.126	0.050	0.254
	exact	**0.296**	0.128	0.050	0.264
0.2	type	0.407	0.159	0.325	**0.466**
	partial	0.351	0.146	0.187	**0.375**
	strict	**0.296**	0.126	0.050	0.254
	exact	**0.296**	0.128	0.050	0.264
0.3	type	0.403	0.115	0.315	**0.466**
	partial	0.348	0.109	0.183	**0.375**
	strict	**0.292**	0.096	0.051	0.254
	exact	**0.292**	0.099	0.051	0.264
0.4	type	0.397	0.112	0.315	**0.466**
	partial	0.343	0.107	0.183	**0.375**
	strict	**0.290**	0.095	0.051	0.254
	exact	**0.290**	0.098	0.051	0.264
0.5	type	0.342	0.095	0.315	**0.440**
	partial	0.291	0.090	0.183	**0.354**
	strict	**0.240**	0.079	0.051	0.236
	exact	**0.240**	0.082	0.051	0.246
0.6	type	0.342	0.079	0.304	**0.440**
	partial	0.291	0.075	0.171	**0.354**
	strict	**0.240**	0.064	0.038	0.236
	exact	**0.240**	0.066	0.038	0.246
0.7	type	0.327	0.055	0.281	**0.440**
	partial	0.285	0.051	0.160	**0.354**
	strict	**0.242**	0.040	0.039	0.236
	exact	0.242	0.043	0.039	**0.246**
0.8	type	0.284	0.053	0.131	**0.440**
	partial	0.248	0.048	0.087	**0.354**
	strict	0.213	0.038	0.043	**0.236**
	exact	0.213	0.040	0.043	**0.246**
0.9	type	0.265	0.030	0.090	**0.445**
	partial	0.230	0.025	0.044	**0.342**
	strict	0.194	0.020	0.000	**0.238**
	exact	0.194	0.020	0.000	**0.238**
Mean	type	0.358	0.105	0.272	**0.453**
	partial	0.308	0.097	0.157	**0.363**
	strict	**0.259**	0.083	0.042	0.245
	exact	**0.259**	0.086	0.042	0.254

Table 16.8: F1 scores obtained by the *base ELIJERE model* for each entity type for each semantic threshold on the test set of the DARES dataset

During the *relation classification* step of the JERE pipeline described in Chapter 15.2, we rely on the machine-learning based classifier instead of the Lexical Index to categorise a candidate graph based on its vocabulary. Instead of outputing an association score between a graph and a relation, the model outputs a probability score for each relation the graph can express. The candidate graph is categorised with the relation for which the classifier has predicted the highest probability score.

Table 16.9 shows the evaluation of each model in terms of Precision, Recall and F1 scores. The Random Forest classifier achieves the highest scores on average, with a mean 0.666 Precision score, a mean 0.135 Recall score and a mean 0.215 F1 score. The XGBoost classifier achieves similar scores on average, with a mean 0.653 Precision score, a mean 0.134 Recall score and a mean 0.212 F1 score. However, the SVM model achieves the lowest scores, with a mean 0.561 Precision score, a mean 0.127 Recall score and a mean 0.199 F1 score. Random Forest achieves the best Precision score of 0.703 when the semantic threshold is set between 0.5 and 0.7. Every model achieves a maximum Recall score of 0.149 when the semantic threshold is set between 0 and 0.3 on average. Finally, Random Forest and XGBoost achieve a maximum F1-score of 0.238 when the semantic threshold is set to between 0.2.

We perform a grid search with the sets of hyper-parameters shown in Table 16.10 to tune the Random Forest and XGBoost classifiers. The best value for each hyper-parameter is shown in bold in the same Table.

Table 16.11 shows the Precision, Recall and F1-scores obtained by the Random Forest and XGBoost models, trained with the best sets of hyper-parameters values, as shown in Table 16.10. On average across threshold, the XGBoost classifier achieves the best results, with a mean Precision score of 0.675, a mean Recall score of 0.137 and a mean F1 score of 0.218. XGBoost achieves the best Precision score of 0.699 when the semantic threshold is set to 0.5. It achieves the best Recall score of 0.149 and best F1 scores of 0.238 when the threshold is set between 0.2 and 0.4. Moreover, compared to Random Forest, the scores achieved by XGBoost are more stable across thresholds. Thus, we select the XGBoost model trained with the best selection of hyper-parameter values, as the classifier of the *hybrid ELIJERE model*.

16.4.2 Evaluation on the Relation Extraction task

We evaluate the *hybrid ELIJERE model* on the Relation Extraction task in terms of Precision, Recall and F1 scores on the test set of the DARES dataset. The results of each evaluation are shown in Table 16.12. On average across semantic threshold, the model achieves a Precision score of 0.675, a Recall score of 0.137, and a F1 score of 0.218. The model performs the best when the semantic threshold is set between 0.2 and 0.4, where it achieves a Recall score of 0.149 and a F1 score of 0.238. However, the model achieves the best Precision score of 0.699 when the semantic threshold

	Threshold	P	R	F1
SVM	0.0	0.635	**0.149**	0.232
	0.1	0.635	**0.149**	0.232
	0.2	0.637	**0.149**	0.233
	0.3	**0.665**	**0.149**	**0.237**
	0.4	0.664	0.136	0.215
	0.5	0.623	0.131	0.206
	0.6	0.576	0.128	0.200
	0.7	0.433	0.112	0.173
	0.8	0.443	0.110	0.172
	0.9	0.303	0.058	0.093
	Mean	0.561	0.127	0.199
Random Forest	0.0	0.653	**0.149**	0.233
	0.1	0.666	**0.149**	0.237
	0.2	0.683	**0.149**	**0.238**
	0.3	0.683	**0.149**	**0.238**
	0.4	0.683	**0.149**	**0.238**
	0.5	**0.703**	0.135	0.216
	0.6	**0.703**	0.124	0.199
	0.7	**0.703**	0.123	0.198
	0.8	0.632	0.114	0.184
	0.9	0.550	0.102	0.165
	Mean	0.666	0.135	0.215
XGBoost	0.0	0.647	**0.149**	0.232
	0.1	0.647	**0.149**	0.232
	0.2	0.660	**0.149**	0.236
	0.3	0.677	**0.149**	**0.238**
	0.4	0.676	0.147	0.234
	0.5	**0.698**	0.134	0.213
	0.6	**0.698**	0.123	0.197
	0.7	**0.698**	0.123	0.197
	0.8	0.657	0.116	0.185
	0.9	0.477	0.097	0.152
	Mean	0.653	0.134	0.212

Table 16.9: Precision, Recall and F1 scores of the SVM, Random Forest and XGBoost on the DARES dataset

XGBoost	n_estimators	50, 100, 150, 200, **350**, 500
	max_depth	1, 2, 3, **4**, 5, 6, 7, 8, 9, 10
	learning_rate	0.0001, 0.001, 0.01, **0.1**, 1.0
	subsample	.25, **.5**, .75, 1
Random Forest	n_estimators	100, **150**, 200, 250, 300
	criterion	**gini**, entropy , log_loss
	max_depth	0, 3, 6, 9, **12**
	max_features	**sqrt**, log2
	bootstrap	**True**, False
	min_samples_split	2, **5**, 10
	min_samples_leaf	1, 2, **4**

Table 16.10: Sets of hyper-parameter values for the XGBoost and Random Forest classifiers. The best value for each hyper-parameter is shown in bold

	Threshold	P	R	F1
Random Forest	0.0	0.647	**0.149**	0.232
	0.1	0.647	**0.149**	0.232
	0.2	0.677	**0.149**	**0.238**
	0.3	**0.676**	0.147	0.234
	0.4	0.666	0.144	0.229
	0.5	0.651	0.128	0.203
	0.6	0.450	0.094	0.149
	0.7	0.485	0.092	0.149
	0.8	0.446	0.074	0.122
	0.9	0.375	0.063	0.103
	Mean	0.572	0.119	0.189
XGBoost	0.0	0.647	**0.149**	0.232
	0.1	0.660	**0.149**	0.236
	0.2	0.677	**0.149**	**0.238**
	0.3	0.677	**0.149**	**0.238**
	0.4	0.677	**0.149**	**0.238**
	0.5	**0.699**	0.138	0.221
	0.6	0.698	0.125	0.200
	0.7	0.698	0.123	0.197
	0.8	0.657	0.119	0.189
	0.9	0.657	0.119	0.189
	Mean	0.675	0.137	0.218

Table 16.11: Precision, Recall and F1 scores on the test set of the DARES dataset of the Random Forest and XGBoost classifiers trained with the best values of hyper-parameters

is set to 0.5. The model achieves the lowest scores when the semantic threshold is set between 0.8 and 0.9, where it achieves a Precision score of 0.657, a Recall score of 0.137 and a F1 score of 0.189. The scores obtained by the hybrid approach are stable across the semantic threshold. This suggests the ability of the XGBoost classifier to properly categorise a candidate graph, regardless of the semantic threshold.

Threshold	P	R	F1
0.0	0.647	**0.149**	0.232
0.1	0.660	**0.149**	0.236
0.2	0.677	**0.149**	**0.238**
0.3	0.677	**0.149**	**0.238**
0.4	0.677	**0.149**	**0.238**
0.5	**0.699**	0.138	0.221
0.6	0.698	0.125	0.200
0.7	0.698	0.123	0.197
0.8	0.657	0.119	0.189
0.9	0.657	0.119	0.189
Mean	0.675	0.137	0.218

Table 16.12: Precision, Recall and F1 scores on the test set of the DARES dataset of the Random Forest and XGBoost classifiers trained with the best values of the hyper-parameters

Table 16.13 shows the Precision, Recall and F1 scores obtained by the *hybrid ELIJERE model* on the test set of the DARES dataset for each label when the semantic threshold is set to 0.2. We select the results when the semantic threshold is set to 0.2 since it is the minimal threshold where the model performs the best.

The model achieves high to very high Precision scores for multiple labels: it achieves a Precision score of 1.000 for the *dateOfBirth, dateOfDeath, memberOf, occupation, placeOfBirth*. It also achieves high Precision scores for the *educatedAt, headOfGovernment* and *spouse* relations. The model performs the best on the *occupation* relation, and achieves a Precision score of 1.000, a Recall score of 0.308 and a F1 score of 0.471. However, the model achieves lower Recall and F1 scores for each relation label. It also fails to properly categorise any instance of the *shares-BordersWith* relation. As for the evaluation of the *base ELIJERE model* on the DARES dataset, the Precision scores being higher than the Recall scores suggests that the patterns stored by the Syntactic Index are precise, but lack diversity.

Table 16.14 shows the error type and distribution of samples wrongly labelled by the *hybrid ELIJERE model* on the test set of the DARES dataset. We identify the same error types as in the evaluation of the *base ELIJERE model* on the DARES dataset, i.e the *predicate not found, pattern not found, wrong label* and *semantic score too weak* error types.

Label	P	R	F1
capitalOf	0.571	0.056	0.103
country	0.667	0.066	0.120
dateOfBirth	**1.000**	0.201	0.331
dateOfDeath	**1.000**	0.205	0.337
educatedAt	0.800	0.287	0.397
headOfGoverment	0.750	0.134	0.222
inception	0.500	0.071	0.125
memberOf	**1.000**	0.185	0.308
nextInBodyWater	0.625	0.100	0.173
occupation	**1.000**	**0.308**	**0.471**
placeOfBirth	**1.000**	0.044	0.083
sharesBordersWith	0.000	0.000	0.000
spouse	0.875	0.275	0.418

Table 16.13: Precision, Recall and F1 scores obtained by the *hybrid ELIJERE model* on the test set of the DARES dataset for each label when the semantic threshold is set to 0.2

The *predicate not found* error appears when the predicate node of the candidate graph is not present in the Syntactic Index. Depending on the value of the semantic threshold, 54.555 % to 58.505 % of errors produced by the model are of the *predicate not found* type. Similarly, the *pattern not found* error appears when the predicate node is present in the Syntactic Index, but there is no syntactic pattern matching the candidate graph. Depending on the value of the semantic threshold, 34.941 % to 37.471 % of errors produced by the model are of the *pattern not found* type. Most error of these types can can be solved by increasing the diversity of syntactic patterns known by the Syntactic Index or by relying on surface patterns instead.

The *wrong label* error appears when the model predicts an incorrect label. Depending on the value of the semantic threshold, 0.000 % to 3.678 % of errors produced by the model are of the *wrong label* type. The *semantic score too weak* error appears when the classifier has not predicted a label because the semantic score is under the semantic threshold. Depending on the value of the semantic threshold, 0.344 % to 10.503 % of errors produced by the model are of the *semantic score too weak* type. As for the *base ELIJERE model*, most errors of this type can either be solved by changing the value of the semantic threshold, or by first correcting the incorrect annotations of the sentences upon which the XGBoost classifier is trained.

In conclusion, the high Precision scores, as well as the low amount of *wrong label* and *semantic score too weak* error types suggest that the *hybrid ELIJERE model* is able to properly categorise the relations when relying on the XGBoost classifier. Most of the remaining errors are caused by the lack of diversity of patterns known by the Syntactic Index. Thus, the performances of the hybrid model can mainly be increased by first increasing the diversity of patterns stored in the Syntactic

	Predicate not found		Pattern not found		Wrong label		Semantic Score too weak		
Threshold	Count	Proportion	Count	Proportion	Count	Proportion	Count	Proportion	Total
0.0	509	58.573 %	326	37.514 %	34	3.912 %	0	0 %	869
0.1	509	58.573 %	326	37.514 %	31	3.567 %	3	0.345 %	869
0.2	509	58.573 %	326	37.514 %	30	3.452 %	4	0.460 %	869
0.3	509	58.573 %	326	37.514 %	30	3.452 %	4	0.460 %	869
0.4	509	58.573 %	326	37.514 %	30	3.452 %	4	0.460 %	869
0.5	509	57.840 %	326	37.045 %	26	2.954 %	19	2.159 %	880
0.6	509	56.744 %	326	36.343 %	23	2.564 %	39	4.347 %	897
0.7	509	56.618 %	326	36.262 %	23	2.558 %	41	4.560 %	899
0.8	509	56.119 %	326	35.942 %	17	1.874 %	55	6.063 %	907
0.9	509	56.119 %	326	35.942 %	17	1.874 %	55	6.063 %	907
Mean	509	57.630 %	326	36.910 %	26	2.965 %	22	2.491 %	883

Table 16.14: Error type and distribution of samples wrongly labelled by the *hybrid ELIJERE model* on the test set of the DARES dataset

Index.

16.4.3 Evaluation on the Named Entity Recognition task

Table 16.15 shows the Precision, Recall and F1 scores obtained by the *hybrid ELIJERE model* on the Named Entity Recognition task for each semantic threshold under each evaluation setting. The model achieves high to very high Precision scores, from 0.966 to 0.985 in the type setting, from 0.836 to 0.851 in the partial setting, from 0.662 to 0.688 in the strict setting and from 0.672 to 0.703 in the exact setting. On average, the model achieves a Precision score of 0.980 under the type setting, of 0.693 under the exact setting, of 0.846 under the partial setting and of 0.649 under the strict setting.

However, the model achieves very low Recall scores, from 0.104 to 0.156 in the type setting, from 0.089 to 0.137 in the partial setting, from 0.071 to 0.108 in the strict setting, and from 0.072 to 0.112 in the exact setting. On average, the model achieves a Recall score of 0.982 under the type setting, of 0.659 under the exact setting, of 0.829 under the partial setting and of 0.649 under the strict setting.

Consequently, the model achieves very low F1 scores, from 0.187 to 0.268 in the type setting, from 0.161 to 0.235 in the partial setting, from 0.128 to 0.186 in the strict setting, and from 0.131 to 0.192 in the exact setting. On average, the model achieves a F1-score of 0.982 under the type setting, of 0.659 under the exact setting, of 0.829 under the partial setting and of 0.649 under the strict setting.

As for the evaluation of the *base ELIJERE model* on the DARES dataset, the higher scores under the type and partial settings indicate the ability of the model to properly determine the type

of an entity, as well as partially determining the boundaries of entities. The low Recall and F1-scores are correlated with the low Recall and F1-scores obtained by the model on the Relation Extraction task. Moreover, errors in the predicted entity type are also correlated with the predicted relation. Thus, such errors can be corrected by first improving the scores on the Relation Extraction task.

Table 16.16 shows the F1 scores obtained by the *hybrid ELIJERE model* for each entity type for each semantic threshold under each evaluation setting. The model identifies the *Misc* type the best, with an average F1-score of 0.456 in the type setting. As with *base ELIJERE model*, this is correlated with the fact that the *hybrid ELIJERE model* detects the *occupation* relation the best on the Relation Extraction task. The model achieves comparable results on the *Person* and *Time* types, with mean F1 scores of 0.366 and 0.263 respectively. This is correlated with the fact the model achieved high Precision scores for the *dateOfBirth, dateOfDeath, educatedAt, headOfGovernment, occupation, placeOfBirth* and *spouse* relation. However, the model achieves the lowest scores on the *Location* type, with a mean F1 scores of to 0.162. Similarly, this is correlated with the fact the model achieved low Precision scores for the *capitalOf, country, inception, nextInBodyWater* and *sharesBorderWith* relations.

In conclusion, as in the evaluation of the *base ELIJERE model* on the DARES dataset, the high Precision scores obtained by the hybrid model under the type setting suggest that our approach is able to correctly categorise the entities involved in a relation. On the other hand, the lower scores obtained under the other settings suggests the rules we have set to determine the boundaries of the entities are not sufficient and must be revised. Moreover, since the categorisation of Named Entities is based on the output of the Relation Extraction task, we must first improve the scores of the model on the Relation Task in order to improve its scores on the NER task.

16.5 Discussion

In this section, we compare the performances of the *base* and *hybrid ELIJERE model*s on the test set of the DARES dataset. Table 16.17 shows the mean score obtained by the *base* and *hybrid ELIJERE model*s on the Relation Extraction task. The *hybrid ELIJERE model* achieves the best scores, with a mean Precision score of 0.675, a mean Recall score of 0.137 and a mean F1 score of 0.218. Table 16.18 compares the mean distribution of error types produced by the *base* and *hybrid ELIJERE model*s on the DARES dataset. The hybrid model produces 883 errors whereas the *base* model produces 900 errors. The hybrid model produces less *semantic score too weak* errors, suggesting the model outputs more confident scores than the *base* model. From these results, the *hybrid ELIJERE model* seems to be more adapted to the Relation Extraction task than the *base* model.

Threshold	Setting	P	R	F1
0.0	type	**0.966**	0.156	0.268
	partial	**0.846**	0.137	0.235
	strict	**0.671**	0.108	0.186
	exact	**0.692**	0.112	0.192
0.1	type	**0.985**	0.156	0.269
	partial	**0.846**	0.134	0.232
	strict	**0.684**	0.108	0.187
	exact	**0.693**	0.110	0.189
0.2	type	**0.985**	0.155	0.268
	partial	**0.845**	0.133	0.230
	strict	**0.682**	0.107	0.185
	exact	**0.691**	0.109	0.188
0.3	type	**0.985**	0.155	0.268
	partial	**0.845**	0.133	0.230
	strict	**0.682**	0.107	0.185
	exact	**0.691**	0.109	0.188
0.4	type	**0.985**	0.155	0.268
	partial	**0.845**	0.133	0.230
	strict	**0.682**	0.107	0.185
	exact	**0.691**	0.109	0.188
0.5	type	**0.984**	0.140	0.245
	partial	**0.836**	0.119	0.208
	strict	**0.662**	0.094	0.165
	exact	**0.672**	0.096	0.167
0.6	type	**0.981**	0.120	0.214
	partial	**0.850**	0.104	0.185
	strict	**0.688**	0.084	0.150
	exact	**0.699**	0.085	0.153
0.7	type	**0.980**	0.118	0.210
	partial	**0.851**	0.102	0.183
	strict	**0.691**	0.083	0.148
	exact	**0.703**	0.084	0.151
0.8	type	**0.978**	0.104	0.187
	partial	**0.841**	0.089	0.161
	strict	**0.668**	0.071	0.128
	exact	**0.682**	0.072	0.131
0.9	type	**0.978**	0.104	0.187
	partial	**0.841**	0.089	0.161
	strict	**0.668**	0.071	0.128
	exact	**0.682**	0.072	0.131
Mean	type	**0.980**	0.129	0.226
	partial	**0.846**	0.111	0.195
	strict	**0.681**	0.090	0.157
	exact	**0.693**	0.091	0.160

Table 16.15: Precision, Recall and F1 scores obtained by the *hybrid ELIJERE model* on the Named Entity Recognition task for each semantic threshold on the test set of the DARES dataset

Threshold	Setting	Per.	Loc.	Time	Misc
0.0	type	0.407	0.193	0.325	**0.466**
	partial	0.351	0.184	0.187	**0.375**
	strict	**0.296**	0.155	0.050	0.254
	exact	**0.296**	0.163	0.050	0.264
0.1	type	0.407	0.193	0.325	**0.466**
	partial	0.351	0.178	0.187	**0.375**
	strict	**0.296**	0.155	0.050	0.254
	exact	**0.296**	0.158	0.050	0.264
0.2	type	0.407	0.191	0.325	**0.466**
	partial	0.351	0.175	0.187	**0.375**
	strict	**0.296**	0.153	0.050	0.254
	exact	**0.296**	0.156	0.050	0.264
0.3	type	0.407	0.191	0.325	**0.466**
	partial	0.351	0.175	0.187	**0.375**
	strict	**0.296**	0.153	0.050	0.254
	exact	**0.296**	0.156	0.050	0.264
0.4	type	0.407	0.191	0.325	**0.466**
	partial	0.351	0.175	0.187	**0.375**
	strict	**0.296**	0.153	0.050	0.254
	exact	**0.296**	0.156	0.050	0.264
0.5	type	0.369	0.167	0.325	**0.466**
	partial	0.317	0.153	0.187	**0.375**
	strict	**0.264**	0.132	0.050	0.254
	exact	**0.264**	0.135	0.050	**0.264**
0.6	type	0.331	0.145	0.197	**0.466**
	partial	0.284	0.134	0.126	**0.375**
	strict	0.237	0.117	0.055	**0.254**
	exact	0.237	0.119	0.055	**0.264**
0.7	type	0.324	0.144	0.184	**0.466**
	partial	0.279	0.133	0.120	**0.375**
	strict	0.234	0.116	0.055	**0.254**
	exact	0.234	0.118	0.055	**0.264**
0.8	type	0.324	0.106	0.184	**0.466**
	partial	0.279	0.099	0.120	**0.375**
	strict	0.234	0.084	0.055	**0.254**
	exact	0.234	0.087	0.055	**0.264**
0.9	type	0.324	0.106	0.184	**0.466**
	partial	0.279	0.099	0.120	**0.375**
	strict	0.234	0.084	0.055	**0.254**
	exact	0.234	0.087	0.055	**0.264**
Mean	type	0.366	0.162	0.263	**0.456**
	partial	0.314	0.149	0.154	**0.362**
	strict	**0.268**	0.130	0.052	0.255
	exact	**0.244**	0.124	0.046	0.237

Table 16.16: F1 scores obtained by the *hybrid ELIJERE model* for each entity type for each semantic threshold on the test set of the DARES dataset

	P	R	F1
base ELIJERE model	0.518	0.120	0.189
hybrid ELIJERE model	**0.675**	**0.137**	**0.218**

Table 16.17: Mean scores across threshold obtained by the *base* and *hybrid ELIJERE models* on the test set of the DARES dataset

	Predicate not found		Pattern not found		Wrong label		Semantic Score too weak		
Threshold	Count	Proportion	Count	Proportion	Count	Proportion	Count	Proportion	Total
Base ELIJERE model	**509**	**56.572** %	**326**	**36.232** %	**17**	**2.000** %	47	5.193 %	900
Hybrid ELIJERE model	**509**	57.630 %	**326**	36.910 %	26	2.965 %	**22**	**2.491** %	**883**

Table 16.18: Mean error types and distribution of samples wrongly labelled by the base and *hybrid ELIJERE model* on the test set of the DARES dataset

Table 16.19 compares the mean scores across threshold in each evaluation setting obtained by the two models on the Named Entity Recognition task. The *hybrid ELIJERE model* achieves the highest scores, although both models achieve similar scores across evaluation settings. The model achieves Precision scores of 0.980 in the type setting, 0.846 in the partial setting, 0.681 in the strict setting and 0.693 in the exact setting. The *base ELIJERE model* achieves a mean Precision score 0.982 under the type setting, which is slightly higher than the score obtained by the *hybrid* model.

The *hybrid* model achieves lower Recall scores, with a score of 0.129 in the type setting, a score of 0.111 in the partial setting, a score of 0.090 in the strict setting and 0.091 in the exact setting. Consequently, it achieves lower F1 scores, with 0.226 in the type setting, 0.195 in the partial setting and 0.157 in the strict setting and 0.160 in the exact setting. From these results, the hybrid model seem slightly more adapted to the Named Entity Recognition task than the *base* model.

Threshold	Setting	P	R	F1
base ELIJERE model	type	**0.982**	0.111	0.199
	partial	**0.829**	0.095	0.169
	strict	**0.649**	0.074	0.133
	exact	**0.659**	0.076	0.135
hybrid ELIJERE model	type	**0.980**	0.129	0.226
	partial	**0.846**	0.111	0.195
	strict	**0.681**	0.090	0.157
	exact	**0.693**	0.091	0.160

Table 16.19: Mean Precision, Recall and F1 scores obtained by the *base* and *hybrid ELIJERE models* on the Named Entity Recognition task on the test set of the DARES dataset

The scores obtained by the *hybrid ELIJERE model*, compared to the *base* model, suggest that the hybrid model is more suited for the Joint Extraction of Relations and Entities task than the *base* model. Its scores are especially much higher on the Relation Extraction task. However, they are similar on the Named Entity Recognition task to those obtained by the *base ELIJERE model*.

The high Precision scores obtained by both models on the Named Entity Recognition task under the type setting shows the ability of the system to properly detect entity types. However, the low scores under the strict, exact and partial settings indicate the models struggle to correctly identify the boundaries of the entities. Currently, the boundaries of an entity are detected with hand-crafted rules, which rely on the part-of-speech and dependency roles of tokens. In order to improve the detection of boundaries, we could first upgrade these hand-crafted rules. For instance, these boundaries could be learned during the process of building the Syntactic Index from the SDP graphs.

As indicated in Section 16.3.2, the incorrect categorisation of a candidate graph may originate from incorrect annotations in the train set upon which the linguistic resources are built. These incorrect annotations are caused by the distant supervision method. Thus, we must add a post-processing step to correct these annotations in order to improve the performances of our approach. This post-processing step could be done manually or automatically, for instance by relying on semantic clustering methods.

The scores obtained by the *base ELIJERE model* are lower than those obtained by the hybrid model classifier on average. However, the output of the *base* model might be easier to interpret, since each word is associated with explicit relations. Thus, we can determine how each word contributed to the categorisation of a candidate graph. On the other, the classification of relations with machine-learning models such as XGBoost may be less interpretable, depending on the complexity of the model. As explained in Section 16.3.2, the Lexical Index suffers from the Out-Of-Vocabulary issue. A potential solution to this issue would be to learn association scores between subwords and relations, instead of learning association scores between words and relations. Thus, the Lexical Index could produce a vector for any word based on its subwords.

In conclusion, the evaluation of the ELIJERE approach on the DARES dataset confirms that our approach to the Joint Extraction of Relations and Entities task is extensible, lightweight and interpretable. By annotating the dataset in a distant supervision manner, our method is able to quickly collect patterns for the extraction of entities involved in a relation. This method allows our approach to collect extraction patterns for any relation and any entity in any language available on Wikidata and Wikipedia. Moreover, the output and process of our approach are interpretable, as they rely on explicit syntactic extraction patterns.

The low Recall and F1 scores are the main issue of the ELIJERE approach, which are caused by the low variation of extraction patterns in the dataset upon which the Syntactic and Lexical indices

are built. As suggested in Section 16.3.2, we consider several ways to increase the diversity of the indices: we could build the indices on a dataset composed of documents of various origins, instead of a unique origin. We could also ask a Large Language Model such as ChatGPT to rephrase the known sentences or even suggest new ones. Finally, we could build surface extraction patterns from the syntactic patterns stored in the Syntactic Index, in order to limit the reliance of the model on the dependency analysis of a candidate graph.

We have compared the performances of two implementations of our JERE system. However, we should also compare their performances to other models, trained on the Joint Extraction of Relation and Entities task. More specifically, we should compare our method with Graphical Neural Networks, trained on the dependency graph representation of sentences. To train this model, we could cast the Named Entity Recognition task as a node classification task, where the model would have to determine if a node is part of a Named Entity or not. Similarly, we could cast the Relation Extraction task as a graph classification task, where the model would have to determine if the path between two detected entities in the graph describe a relation or not.

Chapter 17

Evaluation of the ELIJERE approach on the EMONTAL corpus

Table of contents

In this chapter, we evaluate the ELIJERE approach on the Joint Extraction of Relations and Entities task on the EMONTAL corpus. We evaluate the same two implementations of the ELIJERE approach as in Chapter 16, i.e. the *base ELIJERE model* and the *hybrid ELIJERE model*. We evaluate each model on the Relation Extraction task and the Named Entity Recognition task separately. Our aim is to compare the results obtained by these models on the EMONTAL corpus with those obtained on the DARES dataset, and study how these scores are impacted by the errors produced by the OCR process, as well as by the differences in writing styles.

The rest of this chapter is organised as follows: we first describe the evaluation protocol and evaluation dataset in Section 17.1 and 17.2 respectively. We present the evaluation of the *base* and

*hybrid ELIJERE model*s on the Relation Extraction and Named Entity Recognition task in Section 17.3 and Section 17.4 respectively. Finally, we compare and discuss the results of both models in Section 17.5. The structure of this chapter is presented in Figure 17.1.

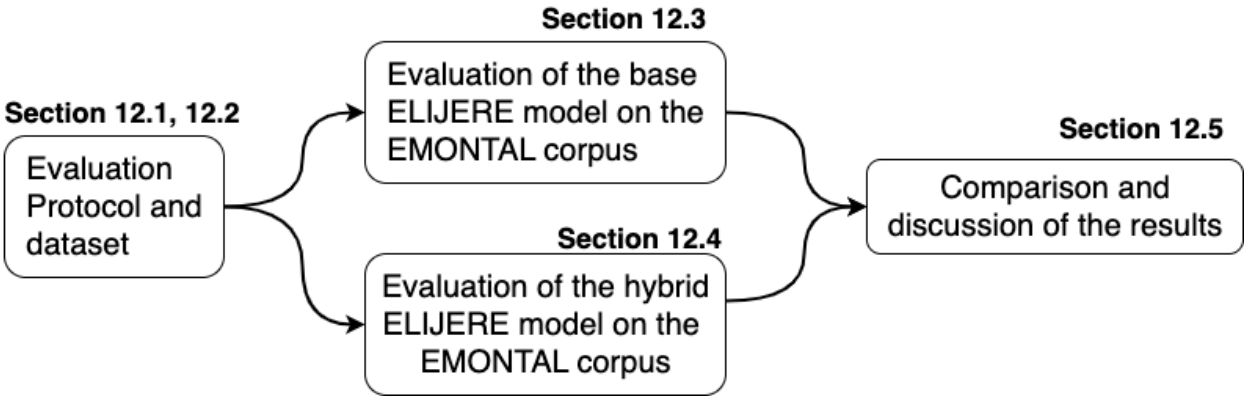


Figure 17.1: Structure of Chapter 17

17.1 Evaluation Protocol

The evaluation protocol of the ELIJERE approach on the EMONTAL corpus is identical to the evaluation protocol applied on the DARES dataset: we evaluate the ability of the models to categorise the relation expressed by a candidate graph in terms of Precision, Recall and F1 scores. We evaluate the ability of the model to determine the type and boundaries of the entities involved in the relation in terms of Precision, Recall and F1 scores, under the type, partial, strict and exact evaluation settings, as described in the SemEval evaluation scheme (Segura-Bedmar, Martínez, & Herrero-Zazo, 2013). We make ten evaluation of both models, each time by setting a semantic threshold between 0 and 0.9. Our aim is to study the impact of the semantic threshold upon the performances of the model. When categorising the relation of the candidate graph, the model outputs the label *Other* if any of the following conditions is met:

- the model cannot find an entry in the Syntactic Index corresponding to the predicate node of the candidate graph
- the model cannot find a syntactic pattern in the Syntactic Index matching the candidate graph
- the semantic score predicted by the model is below the semantic threshold value

17.2 Description of the evaluation dataset

In order to evaluate how the ELIJERE approach performs on historical documents, we have built a dataset of annotated sentences containing the mentions of relations and entities from the EMONTAL corpus. This dataset follows the same structure as the DARES dataset.

We produced this dataset in a semi-automatic way: first, we applied the ELIJERE method to collect sentences from the EMONTAL corpus, and automatically annotate the mentions of entities and relations in them. We manually reviewed and corrected the annotations of these sentences, as well as added any relations and entities the model may have missed. Moreover, we manually added negative samples in the dataset by adding graphs which did not express any relations. We categorised these negative samples with the label *Other*. Finally, we randomly selected 299 sentences from the set of corrected sentences. Table 17.1 shows the distribution of relation labels, whereas Table 17.2 shows the distribution of entity labels in this dataset.

Label	Count	Proportion
capitalOf	3	0.711%
country	3	0.711%
dateOfBirth	40	9.503%
dateOfDeath	36	8.550%
educatedAt	15	3.563%
inception	18	4.276%
memberOf	17	4.035%
occupation	63	14.964%
Other	112	26.600%
placeOfBirth	55	13.067%
spouse	33	7.837%
sharesBordersWith	26	6.174 %
Total	421	100 %

Table 17.1: Distribution of relation labels in the EMONTAL Joint Extraction of Relations and Entities dataset

17.3 Evaluation on the *base ELIJERE model*

In this section, we evaluate the *base ELIJERE model* described in Section 16.3 on the EMONTAL corpus. Our aim is to evaluate how the noisy input produced by the OCR process and the variations in writing styles impact the performances of the model. In the following sections, we first evaluate the model on the Relation Extraction task, before evaluating it on the Named Entity Recognition task separately on the EMONTAL corpus.

Label	Count	Proportion
Person	551	60.152 %
Location	195	21.288 %
Time	102	11.135 %
Misc	68	7.423 %
Total	916	100 %

Table 17.2: Distribution of entity labels in the EMONTAL Joint Extraction of Relations and Entities dataset

17.3.1 Evaluation on the Relation Extraction task

Table 17.3 shows the Precision, Recall and F1 scores of *base ELIJERE model* on the EMONTAL corpus on the Relation Extraction task. On average across all semantic thresholds, the model achieves a Precision score of 0.352, a Recall score of 0.154, and a F1 score of 0.199. The model achieves the best scores when the semantic threshold is set between 0 and 0.5, where it achieves a Precision score of 0.394, a Recall score of 0.187, and a F1 score of 0.238. On the other hand, the model achieves the lowest scores when the semantic threshold is set to 0.9, where it achieves a Precision score of 0.266, a Recall score of 0.0082, and a F1 score of 0.114.

On average, the model achieves higher Precision scores than Recall scores. The model achieves lower scores on the EMONTAL corpus than on DARES dataset. This can be explained by the artefacts and incorrect transcriptions of the textual content of the EMONTAL corpus produced by the OCR process. These incorrect transcriptions have an impact on the dependency analysis of sentences, which yields incorrect dependency graphs. Thus, the model cannot find syntactic patterns matching the candidate graph. Similarly, words incorrectly transcribed cannot be found in the Lexical Index. Thus, the model cannot rely on the vocabulary of the candidate graph to categorise it.

The lower Recall scores can also be explained by the difference in writing styles between the EMONTAL corpus and Wikipedia. The syntactic patterns from the Syntactic Index were built from sentences collected from Wikipedia. However, the writing styles in the EMONTAL corpus differ from the writing style of Wikipedia articles. Thus, the relations in the EMONTAL corpus may be expressed in ways that are not found in Wikipedia articles. For instance, the occupation of a person in the EMONTAL corpus is often expressed according to the pattern "`Name, occupation`", such as "*Gérard Lenfant, pharmacien*: ("Gérard Lenfant, pharmacist"). Such patterns may not appear in Wikipedia articles.

One possible solution to this issue would be to apply a bootstrapping method similar to the SnowBall system (Agichtein & Gravano, 2000), in order to fine-tune our model to the target corpus. First, we could apply our approach on the EMONTAL corpus to extract the mentions of entities

and relations from sentences. Then, we could exploit the extracted information to annotate the EMONTAL corpus in a distant supervision manner, and thus, collect new SDP graphs from the corpus. Thus, we could increase and adapt the Syntactic and Lexical Indices with patterns and vocabulary that are adapted to the target corpus. This process could be repeated until the linguistic resources can no longer be augmented, or after another given condition is met. Since our objective was to first observe the impact of the OCR errors and various writing styles in the EMONTAL corpus on the ELIJERE approach, we did not implement this bootstraping approach in this work.

Threshold	P	R	F1
0.0	**0.394**	**0.187**	**0.238**
0.1	**0.394**	**0.187**	**0.238**
0.2	**0.394**	**0.187**	**0.238**
0.3	**0.394**	**0.187**	**0.238**
0.4	**0.394**	**0.187**	**0.238**
0.5	**0.394**	**0.187**	**0.238**
0.6	0.353	0.162	0.204
0.7	0.266	0.091	0.127
0.8	0.266	0.084	0.117
0.9	0.266	0.082	0.114
Mean	0.352	0.154	0.199

Table 17.3: Precision, Recall and F1 scores of *base ELIJERE model* on the EMONTAL corpus

Table 17.4 shows the Precision, Recall and F1 scores obtained by *base ELIJERE model* on EMONTAL corpus for each label when the semantic threshold is set to 0. We select the results when the semantic threshold is set to 0 since it is the minimal threshold where the model performs the best. The model achieves high to very high Precision scores for multiple labels: it achieves a Precision score of 1.000 for the *dateOfBirth, dateOfDeath* and *inception* labels. It also achieves high Precision scores for the *placeOfBirth* and *memberOf* labels. The model performs the best on the *inception* relation, with a Precision score of 1.000, a Recall score of 0.833 and a F1 score of 0.909. However, it achieves lower Recall and F1 scores. The model especially fails to categorise any mention of the *capitalOf, country, educatedAt, sharesBordersWith* and *spouse* relations. As in the evaluation on the DARES dataset, this suggests that these labels are expressed in ways that are unknown to the model. This fact may be increased by the errors in the OCR transcriptions as well as the different writing styles of the EMONTAL corpus.

Table 17.5 shows the error type and distribution of samples wrongly labelled by *base ELIJERE model* on the EMONTAL corpus. We identify the same error types as in the evaluation of the *base ELIJERE model* on the DARES dataset, i.e the *predicate not found, pattern not found, wrong label* and *semantic score too weak* error types.

Label	P	R	F1
capitalOf	0.000	0.000	0.000
country	0.000	0.000	0.000
dateOfBirth	**1.000**	0.614	0.760
dateOfDeath	**1.000**	0.242	0.385
educatedAt	0.000	0.000	0.000
inception	**1.000**	**0.833**	**0.909**
memberOf	0.750	0.127	0.216
occupation	0.050	0.016	0.024
placeOfBirth	0.732	0.305	0.431
sharesBordersWith	0.000	0.000	0.000
spouse	0.000	0.000	0.000

Table 17.4: Precision, Recall and F1 scores obtained by *base ELIJERE model* on the EMONTAL corpus when the semantic threshold is set to 0

The *predicate not found* error appears when the predicate node of the candidate graph is not present in the Syntactic Index. Depending on the value of the semantic threshold, 47.761 % to 57.657 % of errors produced by the model are of the *predicate not found* type. On average, 52.909 % of errors produced by the model are of this type. Most errors of this type produced by the model can be corrected by increasing the diversity of syntactic patterns known by the Syntactic Index. However, this error type is further increased by the errors produced by the OCR process, as the model may miss the entry corresponding to a word in the Syntactic Index because of an incorrect transcription. A possible solution to this issue would be to apply the fuzzy matching method, and search for approximate matches instead of exact matches in the entries of the Syntactic Index during the *relation extraction* step.

The *pattern not found* error appears when the predicate node is present in the Syntactic Index, but there is no syntactic pattern matching the candidate graph. Depending on the value of the semantic threshold, 31.343 % to 37.837 % of errors produced by the model are of the *pattern not found* type. On average, 34.722 % of errors produced by the model are of this type. Most errors of this type produced by the model can also be corrected by increasing the diversity of syntactic patterns known by the Syntactic Index. Similarly, this error type is further increased by the errors produced by the OCR process, which produces incorrect analysis of the dependency structure of a sentence, and prevents the model from finding matching patterns.

A possible solution to this issue would also be to search for approximate matches instead of exact matches among the syntactic patterns stored in the Syntactic Index during the *relation extraction* step. For instance, a syntactic pattern could match a candidate graph if they share the same graph structure, regardless of the part-of-speech tags and dependency roles of words. Thus, a syntactic

pattern could match, regardless of the errors in the dependency analysis of the sentence caused by the OCR process. However, this approach may reduce the general Precision of the model.

The *wrong label* error appears when the model predicts an incorrect label. Depending on the value of the semantic threshold, 0.000 % to 4.504 % of errors produced by the model are of the *wrong label* type. On average, 2.746 % of errors produced by the model are of this type. The amount of these errors decrease while the semantic threshold value increases. Thus, as for the evaluation on the DARES dataset, we can avoid such error by setting a higher semantic threshold, or by first correcting the incorrect annotations of the sentences upon which the Syntactic and Lexical indices are built. Because the categorisation of a candidate graph is based on its vocabulary, we may also apply an approximate matching to find entries from the Lexical Index, despite the incorrect transcriptions of the OCR process.

The *semantic score too weak* error appears when the classifier has not predicted a label because the association score is under the semantic threshold. Depending on the value of the semantic threshold, 0.000 % to 20.895 % of errors produced by the model are of the *semantic score too weak* type. On average, 9.622 % of errors produced by the model are of this type. These errors increase alongside the increase of the semantic threshold value. Thus, as for the evaluation on the DARES dataset, we can avoid such error by setting a lower semantic threshold. We can also avoid this error by increasing the amount of annotated sentences from the training set, so that the Lexical Index learns better association score between words and relations.

	Predicate not found		Pattern not found		Wrong label		Semantic Score too weak		
Threshold	Count	Proportion	Count	Proportion	Count	Proportion	Count	Proportion	Total
0.0	128	57.657 %	84	37.837 %	10	4.504 %	0	0.000 %	222
0.1	128	57.657 %	84	37.837 %	10	4.504 %	0	0.000 %	222
0.2	128	57.657 %	84	37.837 %	10	4.504 %	0	0.000 %	222
0.3	128	54.008 %	84	35.443 %	10	4.219 %	15	6.329 %	237
0.4	128	54.008 %	84	35.443 %	10	4.219 %	15	6.329 %	237
0.5	128	50.996 %	84	33.466 %	4	1.593 %	35	13.944 %	251
0.6	128	50.793 %	84	33.333 %	4	1.587 %	36	14.285 %	252
0.7	128	50.793 %	84	33.333 %	4	1.587 %	36	14.285 %	252
0.8	128	47.761 %	84	31.343 %	2	0.746 %	54	20.149 %	268
0.9	128	47.761 %	84	31.343 %	0	0.000 %	56	20.895 %	268
Mean	128	52.909 %	84	34.722 %	6	2.746 %	24.	9.622 %	243

Table 17.5: Distribution of error types and samples wrongly labelled by the *base ELIJERE model* on the EMONTAL corpus

In conclusion, the scores obtained by the ELIJERE approach on the EMONTAL corpus for the Relation Extraction task show that our approach is highly impacted by the artefacts and incorrect transcriptions produced by the OCR process, as well as by the diverse writing styles of the historical documents in the EMONTAL corpus. A potential solution to these issues would be to rely on

approximate matching methods to extract and categorise the candidate graphs, instead of relying on exact matching methods.

17.3.2 Evaluation on the Named Entity Recognition task

Table 17.6 shows the Precision, Recall and F1 scores obtained by the *base ELIJERE model* on the Named Entity Recognition task for each semantic threshold. On the development set, the model achieves high to very high Precision scores, from 0.727 to 1.000 in the type setting, from 0.783 to 0.813 in the partial setting, from 0.425 to 0.566 in the strict setting and from 0.566 to 0.626 in the exact setting. On average, the model achieves a Precision score of 0.856 under the type setting, of 0.800 under the partial setting, of 0.488 under the strict setting, and 0.600 under the exact setting.

However, the model achieves very low Recall scores, from 0.053 to 0.195 in the type setting, from 0.041 to 0.204 in the partial setting, from 0.030 to 0.116 in the strict setting, and from 0.030 to 0.157 in the exact setting. On average, the model achieves a Recall score of 0.133 under the type setting, of 0.132 under the partial setting, of 0.076 under the strict setting, and 0.099 under the exact setting.

Consequently, the model achieves very low F1 scores, from 0.101 to 0.312 in the type setting, from 0.079 to 0.326 in the partial setting, from 0.057 to 0.185 in the strict setting, and from 0.057 to 0.251 in the exact setting. On average, the model achieves a F1 score of 0.225 under the type setting, of 0.220 under the partial setting, of 0.129 under the strict setting, and 0.167 under the exact setting.

As in the evaluation on the DARES dataset, the errors and low Recall and F1-scores obtained by the model on the Named Entity Recognition task are correlated with the errors on the Relation Extraction task. Thus, we can improve the performances of the model on the NER task by first improving its performances on the RE task. The lower scores obtained under the partial, strict and exact settings suggests the rules we have conceived to determine the boundaries of entities are not sufficient. These errors are further increased by the errors produced by the OCR process and the different writing styles of the corpus.

Table 17.7 shows the F1 scores obtained by the model for each entity type for each semantic threshold. The model performs the best under the type setting for the *Person* and *Time* types. This is correlated with the fact the model achieves the best scores for relations involving these two entities types on the Relation Extraction task, such as the *dateOfBirth, dateOfDeath, inception* and *spouse* relations. The model identifies the *Time* type the best, with a mean F1-score up to 0.457 under the type setting.

However, the model struggles the most to categorise the *Location* and *Misc* entity types. It achieves a mean F1-score of 0.180 for the *Location* type and a mean F1-score of 0.134 for the *Misc*

Threshold	Setting	P	R	F1
0.0	type	**0.779**	0.195	0.312
	partial	**0.813**	0.204	0.326
	strict	**0.462**	0.116	0.185
	exact	**0.626**	0.157	0.251
0.1	type	**0.779**	0.195	0.312
	partial	**0.813**	0.204	0.326
	strict	**0.462**	0.116	0.185
	exact	**0.626**	0.157	0.251
0.2	type	**0.779**	0.195	0.312
	partial	**0.813**	0.204	0.326
	strict	**0.462**	0.116	0.185
	exact	**0.626**	0.157	0.251
0.3	type	**0.727**	0.143	0.239
	partial	**0.813**	0.161	0.268
	strict	**0.425**	0.083	0.139
	exact	**0.626**	0.124	0.207
0.4	type	**0.727**	0.143	0.239
	partial	**0.813**	0.161	0.268
	strict	**0.425**	0.083	0.139
	exact	**0.626**	0.124	0.207
0.5	type	**0.944**	0.120	0.213
	partial	**0.784**	0.099	0.177
	strict	**0.513**	0.065	0.115
	exact	**0.569**	0.072	0.128
0.6	type	**0.942**	0.117	0.208
	partial	**0.779**	0.096	0.171
	strict	**0.501**	0.062	0.110
	exact	**0.559**	0.068	0.122
0.7	type	**0.942**	0.117	0.208
	partial	**0.779**	0.096	0.171
	strict	**0.501**	0.062	0.110
	exact	**0.559**	0.068	0.122
0.8	type	**0.941**	0.056	0.107
	partial	**0.809**	0.048	0.092
	strict	**0.559**	0.034	0.063
	exact	**0.618**	0.037	0.070
0.9	type	**1.000**	0.053	0.101
	partial	**0.783**	0.041	0.079
	strict	**0.566**	0.030	0.057
	exact	**0.566**	0.030	0.057
Mean	type	**0.856**	0.133	0.225
	partial	**0.800**	0.132	0.220
	strict	**0.488**	0.076	0.129
	exact	**0.600**	0.099	0.167

Table 17.6: Precision, Recall and F1 scores obtained by the *base ELIJERE model* on the Named Entity Recognition task for each semantic threshold on the EMONTAL corpus

type under the partial setting. This is correlated with the fact that the model struggles to correctly identify relations involving the *Location* and *Misc* type, such as the *capitalOf, country, educatedAt, memberOf, sharesBordersWith* and *occupation*. The model achieves a constant F1 score of 0.031 in each setting for the *Misc* type when the threshold is set at a minimum of 0.9. This is correlated with the fact that the model struggles to identify the *occupation* relation when the semantic threshold is set to 0.9.

In conclusion, the scores obtained by the *base ELIJERE model* on the EMONTAL corpus for the Named Entity Recognition task show that our approach is impacted by the artefacts and incorrect transcriptions produced by the OCR process, as well as by the diverse writing styles of the historical documents in the EMONTAL corpus. The OCR issues especially impact our rules to determine the boundaries of entities, which exploit the part-of-speech tags and dependency roles of words. Thus, we must update our rules, so that they adapt to the noisy input from the historical documents of the EMONTAL corpus. Moreover, since the categorisation of entities depends on the relation predicted, we must first improve the performance of the model on the Relation Extraction task to improve its performance on the NER task.

17.4 Evaluation of the *hybrid ELIJERE model*

In this section, we evaluate the *hybrid ELIJERE model* described in Section 16.4 on the EMONTAL corpus. Our aim is to evaluate how the noisy input produced by the OCR process and the variations in writing styles impact the performance of the model. In the following sections, we first evaluate the model on the Relation Extraction task, before evaluating it on the Named Entity Recognition task separately on the EMONTAL corpus.

17.4.1 Evaluation on the Relation Extraction task

Table 17.8 shows the Precision, Recall and F1 scores of the *hybrid ELIJERE model* on the EMONTAL corpus on the Relation Extraction task. On average across semantic threshold, the model achieves a Precision score of 0.469, a Recall score of 0.166, and a F1 score of 0.218. The model performs the best when the semantic threshold is set between 0 and 0.4, where it achieves a Precision score of 0.541, Recall score of 0.206 and a F1 score of 0.271. The model achieves the lowest scores when the semantic threshold is set between to 0.9, where it achieves a Precision score of 0.371, a Recall score of 0.107 and a F1 score of 0.145.

Similarly to the *base ELIJERE model*, the hybrid model achieves lower scores on the EMONTAL corpus than on the DARES dataset. This suggests the hybrid model is also impacted by the noisy input produced by the OCR process and the divergence of the writing styles in the EMON-

Threshold	Setting	Per.	Loc.	Time	Misc
0.0	type	0.267	0.296	**0.627**	0.028
	partial	0.265	0.383	**0.484**	0.158
	strict	0.151	0.197	**0.342**	0.028
	exact	0.191	0.323	**0.342**	0.144
0.1	type	0.267	0.296	**0.627**	0.028
	partial	0.265	0.383	**0.484**	0.158
	strict	0.151	0.197	**0.342**	0.028
	exact	0.191	0.323	**0.342**	0.144
0.2	type	0.267	0.296	**0.627**	0.028
	partial	0.265	0.383	**0.484**	0.158
	strict	0.151	0.197	**0.342**	0.028
	exact	0.191	0.323	**0.342**	0.144
0.3	type	0.267	0.136	**0.467**	0.028
	partial	0.265	0.275	**0.336**	0.158
	strict	0.151	0.117	**0.205**	0.028
	exact	0.191	0.254	**0.205**	0.144
0.4	type	0.267	0.136	**0.467**	0.028
	partial	0.265	0.275	**0.336**	0.158
	strict	0.151	0.117	**0.205**	0.028
	exact	0.191	0.254	**0.205**	0.144
0.5	type	0.253	0.048	**0.467**	0.029
	partial	0.198	0.043	**0.336**	0.146
	strict	0.144	0.037	**0.205**	0.029
	exact	0.144	0.037	**0.205**	0.146
0.6	type	0.253	0.022	**0.467**	0.029
	partial	0.198	0.017	**0.336**	0.146
	strict	0.144	0.011	**0.205**	0.029
	exact	0.144	0.011	**0.205**	0.146
0.7	type	0.253	0.022	**0.467**	0.029
	partial	0.198	0.017	**0.336**	0.146
	strict	0.144	0.011	**0.205**	0.029
	exact	0.144	0.011	**0.205**	0.146
0.8	type	0.140	0.022	**0.200**	0.030
	partial	0.112	0.017	**0.155**	0.090
	strict	0.084	0.011	**0.110**	0.030
	exact	0.084	0.011	**0.110**	0.090
0.9	type	0.127	0.022	**0.200**	0.031
	partial	0.099	0.017	**0.155**	0.031
	strict	0.070	0.011	**0.110**	0.031
	exact	0.070	0.011	**0.110**	0.031
Mean	type	0.233	0.129	**0.457**	0.028
	partial	0.210	0.180	**0.339**	0.134
	strict	0.134	0.090	**0.227**	0.028
	exact	0.147	0.154	**0.216**	0.124

Table 17.7: F1 scores obtained by the *base ELIJERE model* for each entity type for each semantic threshold on the EMONTAL corpus

TAL corpus.

Threshold	P	R	F1
0.0	**0.541**	**0.206**	**0.271**
0.1	**0.541**	**0.206**	**0.271**
0.2	**0.541**	**0.206**	**0.271**
0.3	**0.541**	**0.206**	**0.271**
0.4	**0.541**	**0.206**	**0.271**
0.5	0.500	0.181	0.236
0.6	0.371	0.115	0.150
0.7	0.371	0.115	0.150
0.8	0.371	0.115	0.150
0.9	0.371	0.107	0.145
Mean	0.469	0.166	0.218

Table 17.8: Precision, Recall and F1 scores of the *hybrid ELIJERE model* on the EMONTAL corpus

Table 17.9 shows the Precision, Recall and F1 scores obtained by the *hybrid ELIJERE model* on the EMONTAL corpus for each label when the semantic threshold is set to 0. We select the results when the semantic threshold is set to 0 since it is the minimal threshold where the model performs the best. The model achieves high to very high Precision scores for multiple labels: it achieves a Precision score of 1.000 for the *dateOfBirth, dateOfDeath* and *inception* relations. The model performs the best on the *inception* relation, where it achieves a Precision score of 1.000, a Recall score of 0.833, and a F1 score of 0.909 F1 score. However, the model achieves lower Recall and F1 scores, and fails to properly categorise the *capitalOf, educatedAt* and *memberOf* relations. However, unlike the *base ELIJERE model*, the hybrid model is able to correctly categorise the *sharesBordersWith* and *spouse* relations.

Table 17.10 shows the error type and distribution of samples wrongly labelled by the hybrid model on the EMONTAL corpus. We identify the same error types as in the evaluation of the *base ELIJERE model* on the EMONTAL corpus, i.e the *predicate not found, pattern not found, wrong label* and *semantic score too weak* error types.

Depending on the value of the semantic threshold, 50.592 % to 57.657 % of errors produced by the model are of the *predicate not found* type. Similarly, depending on the value of the semantic threshold, 33.201 % to 37.837 % of errors produced by the model are of the *pattern not found* type. Most errors of these types can can be solved by increasing the diversity of syntactic patterns stored in the Syntactic Index, or by relying on approximate matching methods.

Depending on the value of the semantic threshold, 2.371 % to 4.504 % of errors produced by the model are of the *wrong label* type. Similarly, 0 % to 13.833 % of errors produced by the model are of the *semantic score too weak* type. Moreover, the model produces errors of this type when the

Label	P	R	F1
capitalOf	0.000	0.000	0.000
country	0.500	0.333	0.400
dateOfBirth	**1.000**	0.441	0.612
dateOfDeath	**1.000**	0.271	0.426
educatedAt	0.000	0.000	0.000
inception	**1.000**	**0.833**	**0.909**
memberOf	0.000	0.000	0.000
occupation	0.500	0.016	0.031
placeOfBirth	0.732	0.305	0.431
sharesBordersWith	0.500	0.050	0.091
spouse	0.500	0.087	0.148

Table 17.9: Precision, Recall and F1 scores obtained by the Lexical Index and the *hybrid ELIJERE model* on the EMONTAL corpus when the semantic threshold is set to 0

semantic threshold is set to 0.5 at minimum. This suggests the model outputs probability scores that are lower than 0.4 in general. Thus, as for the *base ELIJERE model*, most errors of this type can either be solved by changing the value of the semantic threshold, or by first correcting the incorrect annotations of the sentences upon which the XGBoost classifier is trained.

	Predicate not found		Pattern not found		Wrong label		Semantic Score too weak		
Threshold	Count	Proportion	Count	Proportion	Count	Proportion	Count	Proportion	Total
0.0	128	57.657 %	84	37.837 %	10	4.504 %	0	0 %	222
0.1	128	57.657 %	84	37.837 %	10	4.504 %	0	0 %	222
0.2	128	57.657 %	84	37.837 %	10	4.504 %	0	0 %	222
0.3	128	57.657 %	84	37.837 %	10	4.504 %	0	0 %	222
0.4	128	57.657 %	84	37.837 %	10	4.504 %	0	0 %	222
0.5	128	54.468 %	84	35.744 %	6	2.553 %	17	7.234 %	235
0.6	128	50.793 %	84	33.333 %	6	2.380 %	34	13.492 %	252
0.7	128	50.793 %	84	33.333 %	6	2.380 %	34	13.492 %	252
0.8	128	50.793 %	84	33.333 %	6	2.380 %	34	13.492 %	252
0.9	128	50.592 %	84	33.201 %	6	2.371 %	35	13.833 %	253
Mean	128	54.572 %	84	35.813 %	8	3.458 %	15	6.154 %	235

Table 17.10: Error type and distribution of samples wrongly labelled by the *hybrid ELIJERE model* on the EMONTAL corpus

In conclusion, as for the evaluation of the *base ELIJERE model* on the EMONTAL corpus, the scores obtained by the *hybrid ELIJERE approach* on this corpus for the Relation Extraction task show that our approach is highly impacted by the artefacts and incorrect transcriptions produced by the OCR process, as well as by the diverse writing styles of the historical documents in the EMONTAL corpus. However, the categorisation of candidate graphs by the XGBoost classifier

seems less impacted by these issues than the Lexical Index, as shown by the higher Precision, Recall and F1 scores obtained by the model. Most errors produced by the model can be improved by improving the diversity of patterns known by the Syntactic Index, or by relying on approximate matching methods.

17.4.2 Evaluation on the Named Entity Recognition task

Table 17.11 shows the Precision, Recall and F1 scores obtained by the *hybrid ELIJERE model* on the Named Entity Recognition task for each semantic threshold. The model achieves high to very high Precision scores, from 0.779 to 0.866 in the type setting, from 0.794 to 0.813 in the partial setting, from 0.462 to 0.569 in the strict setting and from 0.587 to 0.667 in the exact setting. On average, the model achieves a Precision score of 0.827 under the type setting, of 0.818 under the partial setting, of 0.506 under the strict setting, and 0.637 under the exact setting.

However, the model achieves very low Recall scores, from 0.110 to 0.195 in the type setting, from 0.105 to 0.204 in the partial setting, from 0.071 to 0.116 in the strict setting, and from 0.084 to 0.157 in the exact setting. On average, the model achieves a Recall score of 0.160 under the type setting, of 0.160 under the partial setting, of 0.097 under the strict setting, and 0.124 under the exact setting.

Consequently, the model achieves very low F1 scores, from 0.195 to 0.312 in the type setting, from 0.187 to 0.326 in the partial setting, from 0.127 to 0.185 in the strict setting, and from 0.149 to 0.251 in the exact setting. On average, the model achieves a Precision score of 0.265 under the type setting, of 0.264 under the partial setting, of 0.161 under the strict setting, and 0.205 under the exact setting.

The model achieves better scores under the partial setting when the semantic threshold is set below 0.5. This is correlated with the fact that the model outputs association scores that are lower than 0.4, as explained in the previous section. This suggests the model is able to correctly identify the type and partially identify the boundaries of entities when the semantic threshold is set to a value below 0.5.

As in the evaluation on the DARES dataset, the errors and low Recall and F1-scores obtained by the hybrid model on the Named Entity Recognition task are correlated with the errors on the Relation Extraction task. Thus, we can improve the performances of the model on the NER task by first improving its performances on the RE task. The lower scores obtained under the partial, strict and exact settings suggests the rules we have conceived to determine the boundaries of entities are not sufficient. These errors are further increased by the errors produced by the OCR process and the different writing styles of the corpus.

Table 16.16 shows the F1 scores obtained by the *hybrid ELIJERE model* for each entity type for

Threshold	Setting	P	R	F1
0.0	type	**0.779**	0.195	0.312
	partial	**0.813**	0.204	0.326
	strict	**0.462**	0.116	0.185
	exact	**0.626**	0.157	0.251
0.1	type	**0.779**	0.195	0.312
	partial	**0.813**	0.204	0.326
	strict	**0.462**	0.116	0.185
	exact	**0.626**	0.157	0.251
0.2	type	**0.779**	0.195	0.312
	partial	**0.813**	0.204	0.326
	strict	**0.462**	0.116	0.185
	exact	**0.626**	0.157	0.251
0.3	type	**0.779**	0.195	0.312
	partial	**0.813**	0.204	0.326
	strict	**0.462**	0.116	0.185
	exact	**0.626**	0.157	0.251
0.4	type	**0.779**	0.195	0.312
	partial	**0.813**	0.204	0.326
	strict	**0.462**	0.116	0.185
	exact	**0.626**	0.157	0.251
0.5	type	**0.910**	0.174	0.292
	partial	**0.794**	0.151	0.254
	strict	**0.519**	0.099	0.167
	exact	**0.587**	0.112	0.188
0.6	type	**0.866**	0.114	0.201
	partial	**0.830**	0.108	0.191
	strict	**0.559**	0.073	0.13
	exact	**0.661**	0.086	0.152
0.7	type	**0.866**	0.114	0.201
	partial	**0.830**	0.108	0.191
	strict	**0.559**	0.073	0.130
	exact	**0.661**	0.086	0.152
0.8	type	**0.866**	0.114	0.201
	partial	**0.830**	0.108	0.191
	strict	**0.559**	0.073	0.130
	exact	**0.661**	0.086	0.152
0.9	type	**0.859**	0.110	0.195
	partial	**0.833**	0.105	0.187
	strict	**0.559**	0.071	0.127
	exact	**0.667**	0.084	0.149
Mean	type	**0.827**	0.160	0.265
	partial	**0.818**	0.160	0.264
	strict	**0.506**	0.097	0.161
	exact	**0.637**	0.124	0.205

Table 17.11: Precision, Recall and F1 scores obtained by the the *hybrid ELIJERE model* on the Named Entity Recognition task for each semantic threshold on the EMONTAL corpus

each semantic threshold under each evaluation setting. The model identifies the *Time* type the best, with a mean F1-scores of 0.538 under the type setting. This is correlated with the fact that the *hybrid ELIJERE model* detects the *dateOfBirth, dateOfDeath* and *inception* relations the best during on the Relation Extraction task. The model achieves comparable results on the *Person* and *Location* types, with mean F1 scores up to 0.219 and 0.280 under the type and partial settings respectively. This is correlated with the fact the model achieved high Precision scores for the *capitalOf, country, educatedAt, memberOf, sharesBordersWith* and *spouse* relation. However, the model achieves the lowest scores on the *Misc* type, with a mean F1 scores of 0.145. Similarly, this is correlated with the fact the model achieved low Precision scores for the *occupation* relation.

In conclusion, as the evaluation of the *base ELIJERE model* on the EMONTAL corpus, the scores obtained by the *hybrid ELIJERE approach* on this corpus for the Named Entity Recognition task show that our approach is impacted by the artefacts and incorrect transcriptions produced by the OCR process, as well as by the diverse writing styles of the historical documents in the EMONTAL corpus. Since the hybrid model performed slightly better on the Relation Extraction task, its scores are also slightly better on the Named Entity Recognition task. In order to improve the performances of the model, we must improve the performances of the model on the Relation Extraction task, as well as update our rules to determine the boundaries of entities, so that they adapt to the noisy input from the historical documents of the EMONTAL corpus.

17.5 Discussion

In this section, we compare and discuss the performances of the *base* and *hybrid ELIJERE models* on the test set of the EMONTAL corpus. Table 17.13 shows the mean score obtained by the base and *hybrid ELIJERE models* on the Relation Extraction task. The *hybrid ELIJERE model* achieves the best scores, with a mean Precision score of 0.469, a mean Recall score of 0.166 and a mean F1 score of 0.218.

Table 17.14 compares the mean distribution of error types produced by the *base* and *hybrid ELIJERE models* on the EMONTAL corpus. The *hybrid* model produces 235 errors whereas the *base* model produces 243 errors. The *hybrid* model produces less *semantic score too weak* errors, suggesting the model outputs more confident scores than the *base* model, and may be less impacted by the OCR errors. From these results, the *hybrid ELIJERE model* seems more adapted to the Relation Extraction task on the EMONTAL corpus than the *base* model.

Table 17.15 compares the mean scores across all thresholds in each evaluation setting obtained by the two models on the Named Entity Recognition task. The *hybrid ELIJERE model* achieves the highest scores, although both models achieve similar scores across evaluation settings. The model achieves Precision scores of 0.827 in the type setting, 0.818 in the partial setting, 0.506 in the strict

Threshold	Setting	Per.	Loc.	Time	Misc
0.0	type	0.267	0.296	**0.627**	0.028
	partial	0.265	0.383	**0.484**	0.158
	strict	0.151	0.197	**0.342**	0.028
	exact	0.191	0.323	**0.342**	0.144
0.1	type	0.267	0.296	**0.627**	0.028
	partial	0.265	0.383	**0.484**	0.158
	strict	0.151	0.197	**0.342**	0.028
	exact	0.191	0.323	**0.342**	0.144
0.2	type	0.267	0.296	**0.627**	0.028
	partial	0.265	0.383	**0.484**	0.158
	strict	0.151	0.197	**0.342**	0.028
	exact	0.191	0.323	**0.342**	0.144
0.3	type	0.267	0.296	**0.627**	0.028
	partial	0.265	0.383	**0.484**	0.158
	strict	0.151	0.197	**0.342**	0.028
	exact	0.191	0.323	**0.342**	0.144
0.4	type	0.267	0.296	**0.627**	0.028
	partial	0.265	0.383	**0.484**	0.158
	strict	0.151	0.197	**0.342**	0.028
	exact	0.191	0.323	**0.342**	0.144
0.5	type	0.251	0.242	**0.627**	0.028
	partial	0.208	0.213	**0.484**	0.158
	strict	0.143	0.143	**0.342**	0.028
	exact	0.150	0.164	**0.342**	0.144
0.6	type	0.158	0.204	**0.428**	0.028
	partial	0.139	0.185	**0.360**	0.158
	strict	0.097	0.124	**0.291**	0.028
	exact	0.104	0.145	**0.291**	0.144
0.7	type	0.158	0.204	**0.428**	0.028
	partial	0.139	0.185	**0.360**	0.158
	strict	0.097	0.124	**0.291**	0.028
	exact	0.104	0.145	**0.291**	0.144
0.8	type	0.158	0.204	**0.428**	0.028
	partial	0.139	0.185	**0.360**	0.158
	strict	0.097	0.124	**0.291**	0.028
	exact	0.104	0.145	**0.291**	0.144
0.9	type	0.158	0.193	**0.410**	0.028
	partial	0.139	0.180	**0.341**	0.158
	strict	0.097	0.124	**0.271**	0.028
	exact	0.104	0.146	**0.271**	0.144
Mean	type	0.219	0.251	**0.538**	0.041
	partial	0.204	0.280	**0.425**	0.145
	strict	0.129	0.164	**0.319**	0.039
	exact	0.141	0.221	**0.292**	0.129

Table 17.12: F1 scores obtained by the *hybrid ELIJERE model* for each entity type for each semantic threshold on the EMONTAL corpus

	P	R	F1
base ELIJERE model	0.352	0.154	0.199
hybrid ELIJERE model	**0.469**	**0.166**	**0.218**

Table 17.13: Mean scores across all thresholds obtained by the *base* and *hybrid ELIJERE model*s on the test set of the EMONTAL corpus

	Predicate not found		Pattern not found		Wrong label		Semantic Score too weak		
Threshold	Count	Proportion	Count	Proportion	Count	Proportion	Count	Proportion	Total
base ELIJERE model	128	**52.909** %	84	**34.722** %	6	**2.746** %	24	9.622 %	243
hybrid ELIJERE model	128	54.572 %	84	35.813 %	8	3.458 %	**15**	**6.154** %	**235**

Table 17.14: Mean error types and distribution of samples wrongly labelled by the *base* and *hybrid ELIJERE models* on the test set of the EMONTAL corpus

setting and 0.637 in the exact setting. The *base ELIJERE model* achieves a mean Precision score 0.856 under the type setting, which is slightly higher than the score obtained by the *hybrid* model.

However, the *hybrid* model achieves lower Recall scores, with a score of 0.160 in the type and partial settings, a score of 0.097 in the strict setting and 0.124 in the exact setting. Consequently, it achieves lower F1 scores, with 0.265 in the type setting, 0.264 in the partial setting and 0.161 in the strict setting and 0.205 in the exact setting. From these results, the *hybrid* model seems to be slightly more adapted to the Named Entity Recognition task on the EMONTAL corpus than the *base* model.

Threshold	Setting	P	R	F1
base ELIJERE model	type	**0.856**	0.133	0.225
	partial	**0.800**	0.132	0.220
	strict	**0.488**	0.076	0.129
	exact	**0.600**	0.099	0.167
hybrid ELIJERE model	type	**0.827**	0.160	0.265
	partial	**0.818**	0.160	0.264
	strict	**0.506**	0.097	0.161
	exact	**0.637**	0.124	0.205

Table 17.15: Mean Precision, Recall and F1 scores obtained by the *base* and *hybrid ELIJERE model*s on the Named Entity Recognition task on the test set of the EMONTAL corpus

The *hybrid ELIJERE model* seems to be the most suited for the Joint Extraction of Relations and Entities task. Its scores are especially much higher on the Relation Extraction task. The scores obtained by both models on the Named Entity Recognition task are similar, although the hybrid

model performs slightly better than the base model.

In conclusion, the evaluation of the ELIJERE approach on the EMONTAL corpus shows that our approach is impacted by the artefacts and incorrect transcriptions produced by the OCR process. A better correction of these errors is thus necessary to first improve the performances of our approach on historical documents. As suggested, our approach could also rely on approximate matching methods such as fuzzy matching instead of exact matching methods so that it becomes flexible and adaptable to the noisy input from historical documents. Making our approach more flexible would increase its Recall scores, at the cost of lowering its Precision scores however.

Our approach is also impacted by the writing styles of the documents in the EMONTAL corpus, which are different from the Wikipedia articles from which the linguistic resources of the ELIJERE approach are built. The bootstraping method inspired by the SnowBall approach suggested in Section 17.3.1 may help to fine-tune our approach to a target dataset such as our corpus.

Part V

Augmented search interfaces

In addition to their role in structuring the textual content of documents, semantic annotations can be exploited by search engines to index documents. Thus, search engines may allow the users to search documents using keyword-based queries and filter the results according to the semantic annotations of the documents. Moreover, these semantic annotations can be exploited to build augmented search interfaces, such as maps, timelines, or networks of entities to name a few, which would complement traditional search engines (Gutehrlé, 2024b; Gutehrlé & Atanassova, 2023; Jatowt, 2021). Such interfaces would allow a distant reading (Moretti, 2013) of the documents, and may help reveal hidden patterns in the data and lead to new research questions, while keeping access to the original documents in order to allow close reading.

In this part, we focus on augmented search interfaces that are based on maps and timelines. First, we propose a framework for building an interface to study how spatial imaginaries (Watkins, 2015) are represented in a dataset of historical newspapers published during WWI. This interface provides several modules, such as a map or a concordance table, that allow close and distant readings of the underlying dataset. Secondly, we propose a conceptual framework to automatically generate informative, readable and interpretable timelines from collections of historical documents. We address the challenges related to generating timelines from such documents, before suggesting several methods to implement this framework, and discussing potential extensions and applications. These frameworks have not been applied to the EMONTAL corpus, as they are the result of external projects on the processing of historical documents that were realised in the context of this PhD thesis. However, they show how the semantic annotations obtained by our approaches can be further exploited to build semantically rich interfaces dedicated to archival documents.

The rest of this part is structured as follows: we describe our framework to build an interface for the study of spatial imaginaries in Chapter 18, before presenting our conceptual framework to automatically generate timelines in Chapter 19.

Chapter 18

Maps

Table of contents

In this chapter we propose a framework to build an interface to study the expression of spatial imaginaries in historical newspapers. Spatial imaginaries are "stories and ways of talking about places and spaces that transcend language as embodied performances by people in the material world" (Watkins, 2015), that are shared by large groups of people, or a society as a whole (Davoudi et al., 2018). In order to study these spatial imaginaries, the interface provides a distant reading tool where locations that have been automatically extracted from a dataset of historical newspapers are projected to a map. Moreover, the interface provides close reading tools such as a concordance table and links to the original documents, thus allowing to study the mentions of locations in their textual context.

The web interface as well as our code are publicly available[1][2]. This work has been presented in Gutehrlé et al. (2021), and was originally initiated by the SpaceWars team during the Helsinki Digital Humanities Hackathon 2021[3]. The results obtained by the SpaceWars team during this hackathon are presented in a blog post, available on the NewsEye website[4].

[1] http://spacewars.newseye.eu/

[2] https://github.com/dhh21/SpaceWars

[3] https://www.helsinki.fi/en/digital-humanities/dhh-hackathon/helsinki-digital-humanities-hackathon-2021-dhh21

[4] https://www.newseye.eu/blog/news/where-did-it-happen-spatial-imaginaries-of-world-war-i/

The rest of this chapter is structured as follows: we first provide an overview of digital mappings projects applied in a Digital Humanities context. Then, we describe the dataset underlying this interface in Section 18.2 before describing the interface itself in Section 18.3. Finally, we present a use case of this interface, as well as some of its limitations in Section 18.4.

18.1 Related works

There has been extensive theoretical studies in the field of spatial digital humanities and on the methods of geographical text analysis and digital mapping (Bodenhamer et al., 2010; Cooper et al., 2016; Eide, 2015). Several works in the field of Digital Humanities have focused on mapping the locations mentioned in literary works. For instance, *A Literary Atlas of Europe*[5] proposes interactive maps and other visualisations tools to study how fictions make use of real or fictional spaces (Piatti et al., 2011). Similarly, *A Map of Paradise Lost*[6] proposes an exploratory map to visualise the locations mentioned in John Milton's *Paradise Lost*. Each location on the map is completed with contextual information such as editorial notes or extracts from the original text where the location is mentioned (El Khatib & Currell, 2018). An example of this interface is shown in Figure 18.1.

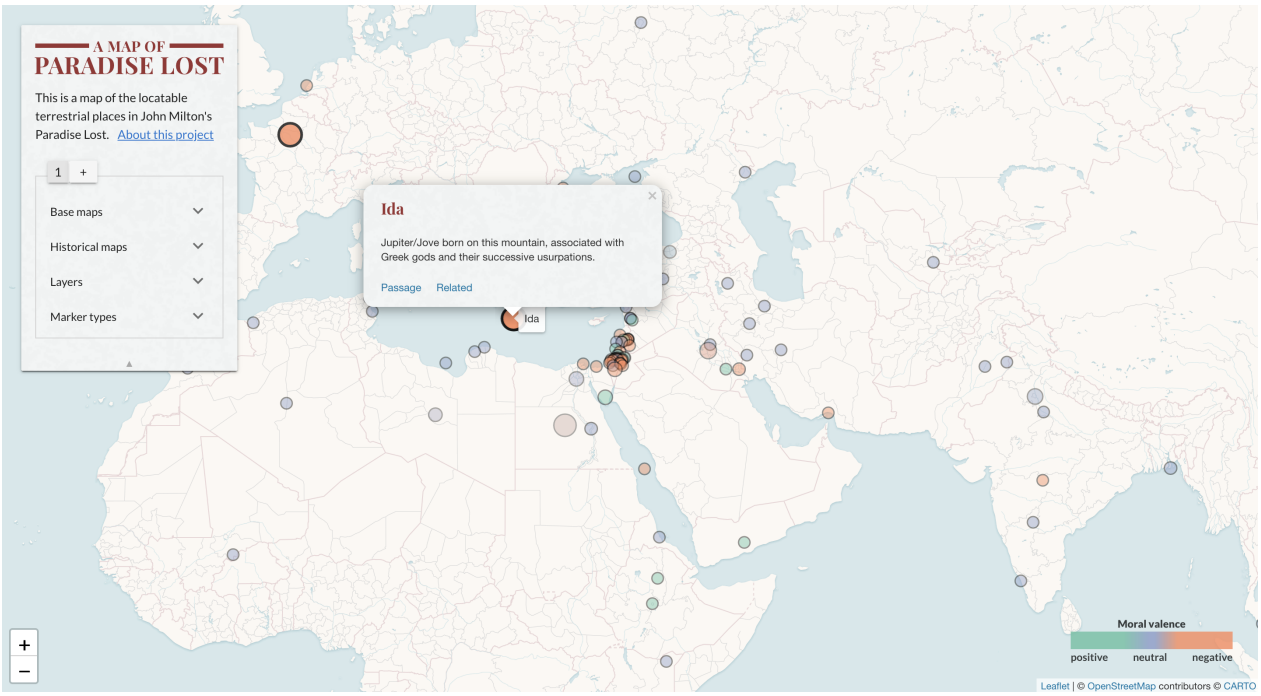


Figure 18.1: Extract from the interface of the *A Map of Paradise Lost* project

[5] http://www.literaturatlas.eu/en/index.html
[6] https://olvidalo.github.io/paradise-lost/

Similar projects have been realised on documents other than literary works. For instance, Moncla (2015) proposes an approach for the automatic geocoding of itineraries described in the textual content of documents. This approach is applied to a corpus of hiking descriptions, thus allowing to automatically generate the itineraries taken by hikers. E. Klein et al. (2014) propose a map to visualise the exchange of commodities in the world as described in a corpus of documents from the 19th century. The locations in the document have been identified by applying Named Entity Recognition methods, and associated with geographical coordinates with the Edinburgh geoparser tool (Alex et al., 2015).

Moreover, several projects have focused on mapping ancient places. For instance, the *al-Ṯurayyā* Project[7] is a gazetteer storing coordinates for over 2,000 toponyms and route sections of the early Islamic world, based on Georgette Cornu's *Atlas du monde arabo-islamique à l'époque classique: IXe-Xe siècles* (Cornu, 1983). This gazetteer can be accessed and searched via an interactive map that is shown in Figure 18.2. The *ORBIS: The Stanford Geospatial Network Model of the Roman World* [8] project proposes an interactive map to reconstruct the financial cost as well as the time required to travel during the Roman empire era. (Scheidel, 2015). Similarly, the *Running Reality*[9] project proposes a spatio-temporal system to model the evolution of human civilisations. The interface can render any day of any year as a map that allows users to interact with and gain information about the displayed spatial entities and objects. An example of the *ORBIS* and *RunningReality* interfaces are respectively shown in Figure 18.3 and 18.4.

As explained in Section 6.1, the automatic generation of maps from documents requires to perform a geocoding step to associate the mentions of locations to geographical coordinates. This step requires external resources known as gazetteers, such as Geonames[10] or Nominatim[11] for contemporary locations or Pleiades for known existing ancient places (Bagnall et al., 2016). When generating such maps in a fully automatic manners, the mentions of locations in the documents are usually obtained by applying Named Entity Recognition methods, as in E. Klein et al. (2014) or Govind and Spaniol (2017) for instance.

The geocoding step remains however a challenging task (Melo & Martins, 2017), as what appears as a reference to a location may actually refer to another kind of entity. For instance, the mention "Lancaster" may refer to the city in the United Kingdom, or to a kind of airplane. Moreover, a mention of a location may be ambiguous, and refer to multiple real locations. For instance, the mention "Paris" may refer to the city of Paris in France, or the city of Paris in the United States. Thus, a post-processing step of *toponym resolution* to solve these entities is required (Moncla,

[7] https://althurayya.github.io/
[8] https://orbis.stanford.edu/
[9] https://www.runningreality.org/
[10] https://www.geonames.org/
[11] https://nominatim.org/release-docs/develop/develop/Ranking/

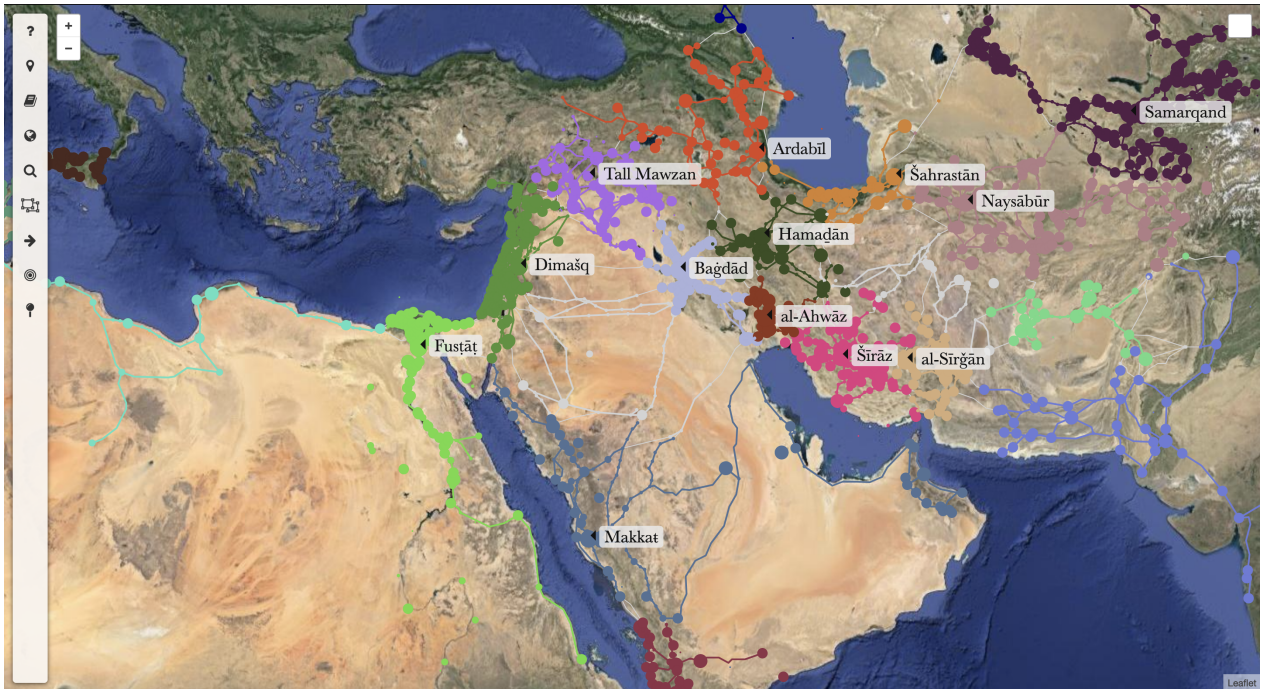


Figure 18.2: Extract from the interface of the *al-Ṯurayyā* project

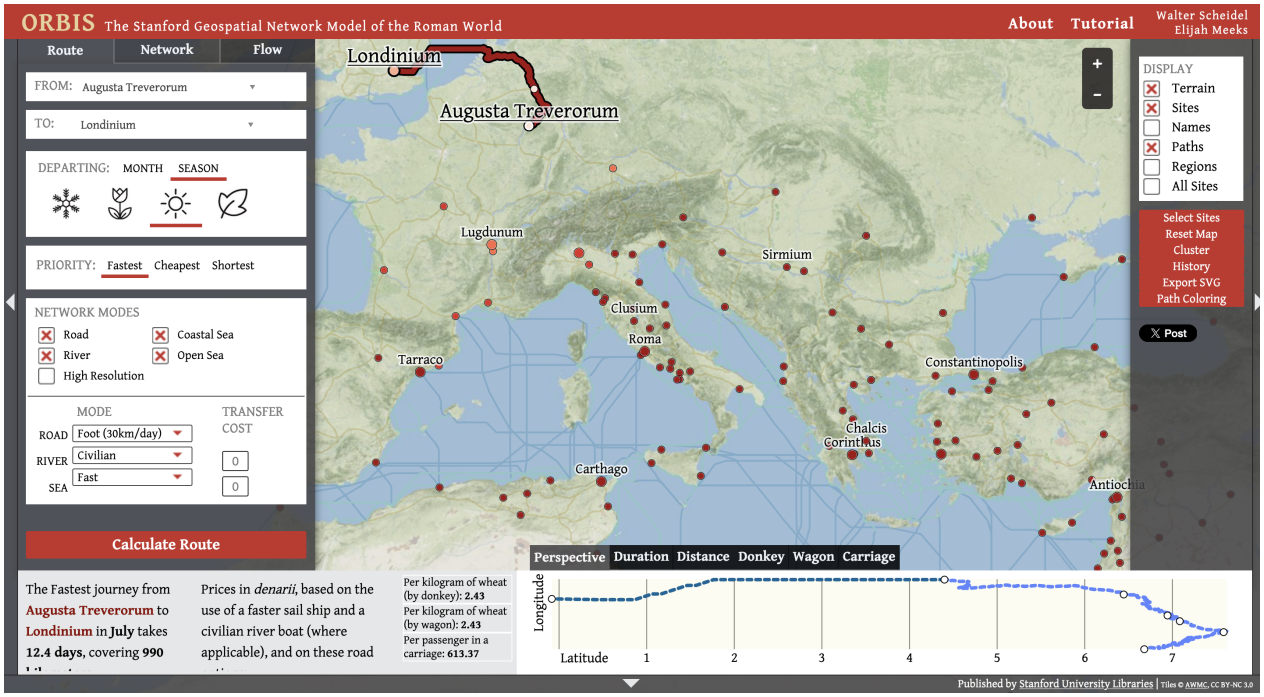


Figure 18.3: Extract from the interface of the *ORBIS* project

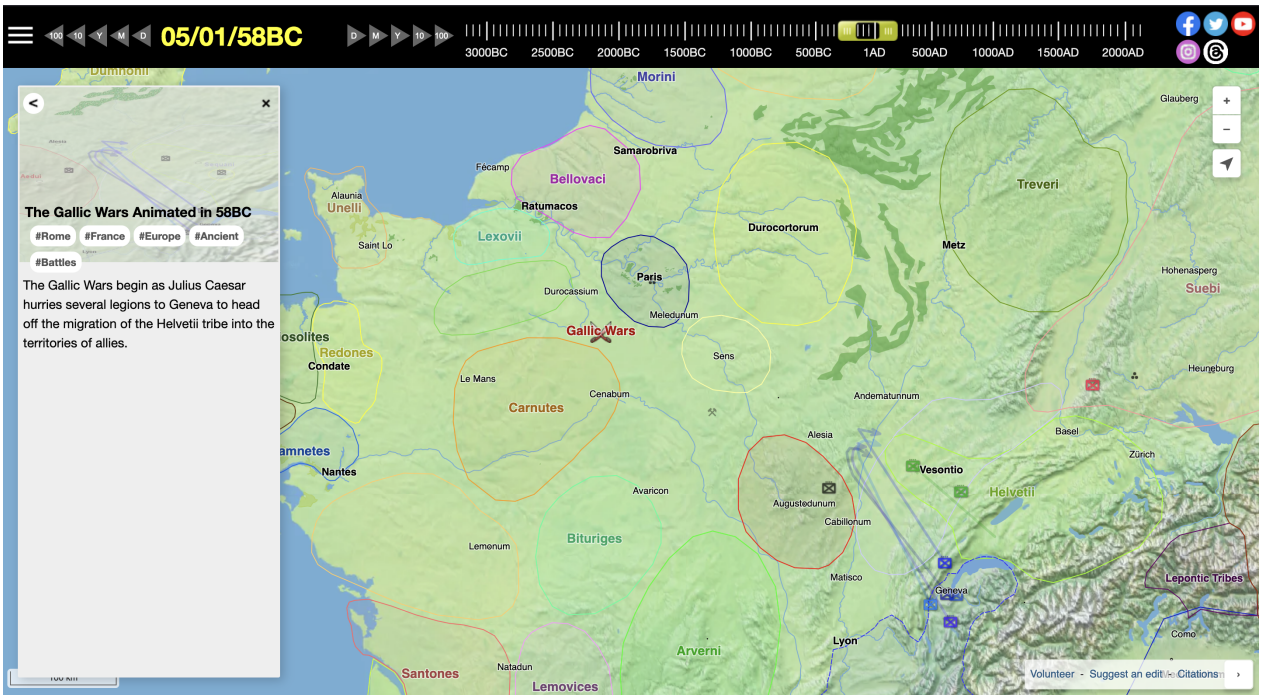


Figure 18.4: Extract from the interface of the *Running Reality* project

2015). Furthermore, the names and coordinates of a location may change over time. For instance, the borders of France in 1914 were different than the contemporary borders, whereas the Prussian state, which entailed several contemporary countries, does not exist anymore. Thus, the geocoding step requires gazetteers that are adapted to the period described by the dataset, in order to avoid introducing anachronisms on the map. Although some resources such as Pleiades or *al-Ṯurayyā* exist today, such resources remain rare.

18.2 Dataset

Our interface relies on a dataset of newspapers published between 1913 and 1920 in three different languages: *Arbeiter Zeitung*, *Illustrierte Kronenzeitung* (German), *Le Matin*, *L'Oeuvre* (French), and *Helsingin Sanomat* (Finnish). This dataset has been provided by courtesy of the *NewsEye* project[12] (Doucet et al., 2020). The transcriptions of the textual content of the newspapers have been obtained by OCR methods, whereas the mentions of locations have automatically been annotated with Named Entity Recognition methods. Whenever possible, the extracted entities have been automatically linked to a Wikidata resource by applying Named Entity Resolution methods (Boroş et al., 2020; Linhares Pontes et al., 2019b). Table 18.1 shows the distribution of entities of type *Location* by newspapers for each year in the dataset.

[12] https://www.newseye.eu/

Newspaper	1913	1914	1915	1916	1917	1918	1919	1920
Arbeiter Zeitung	42,559	39,892	34,630	33,160	31,972	33,961	34,470	34,349
Helsingin Sanomat	23,287	24,184	26,014	24,997	22,700	20,389	0	0
Illustrierte Kronen Zeitung	46,350	50,403	48,857	55,531	51,694	48,334	44,724	42,052
Le Matin	13,3395	10,7850	59,117	47,921	44,371	34,353	46,061	58,041
L'Oeuvre	0	0	3,578	9,660	9,282	10,479	15,668	15,316
Total	245,591	222,329	172,196	171,269	160,019	147,516	140,923	149,758

Table 18.1: Distribution of entities of type *Location* by newspapers for each year


Der Krieg zwiſchen England und Deutſchland.
Nachrichten über die Stimmungen gegen den Krieg.
Wien, 11. Auguſt.

Figure 18.5: Excerpt from the *Neue Freie Presse*, published on the 11th of August, 1914

The years 1913 and 1914 have on average more entity mentions than other years in the dataset. This may be caused by errors in the OCR process, which produced more name variants. Moreover, some documents are missing for some years: our dataset lacks documents for *L'Oeuvre* for 1913 and 1914, and *Helsingin Sanomat* for 1919 and 1920. This reflects a peculiarity of our dataset rather than a historical fact, as these newspapers were actually published in these years. Moreover, the quality of the OCR and NER varies across newspaper issues and languages, which creates major challenges for assessing the accuracy of the tokens recognised as locations. For example, Figure 18.5 shows an excerpt of an article from the Austrian newspaper *Neue Freie Presse*[13], which announces a war between Britain and Germany. However, the NER system recognised only one location in this text, namely Vienna, which is the publication location of the newspaper rather than the location of the event. Such problems are common when processing historical datasets (Dumais, 2001; Kantor & Voorhees, 2000) and explains why automatic processing is usually verified by close reading. In order to limit the variations caused by the OCR process, we have normalised the entity mentions by applying a partial-matching based approach Dumais, 2001; Jin et al., 2003. For a detailed presentation of this method, see Gutehrlé et al. (2021).

We have applied a geocoding step to associate each location mentions to geographical coordinates and project them on a map. For this step, we rely on the *Nominatim* open-source georeferencing service. As explained in the previous Section, contemporary gazetteers such as Geonames or Nominatim do not account for historical geographical entities, such as *Austria-Hungary*

[13] Although not present in our dataset, the *Neue Freie Presse* was present in the dataset used by the SpaceWars team during the Helsinki Digital Humanities Hackathon 2021

or *Ottoman Empire*. Thus, they are not entirely suited to encode locations mentioned in historical documents with geographical coordinates. However, historical gazetteers such as Pleiades or *al-Ṯurayyā* could not be used, as the locations they describe are not suited for the historical context of our dataset. Moreover, dedicated historical map services such as *OpenHistoricalMaps*[14] or *RunningReality* could not be used because of a lack of Application Programming Interface (API).

Finally, we completed our dataset with a list of capital cities from this period, as well as data related to battles that occurred during WWI. We exploit these external data to compare the locations mentioned in our dataset with known locations and events. We collected these data from Wikidata, which is, to the best of our knowledge, the only available Knowledge Base which provides access to such historical information in a machine-readable format. We extracted battles from Wikidata using SPARQL with the *WWI* keyword as a query. Most battles in the Wikidata ontology are linked to a specific front: some are linked to WWI directly, whereas others are linked to other conflicts that occurred during the war, such as the Finnish Civil War. Thus, we ensured to extract all the battles that are directly or indirectly related to WWI. For each battle, we extracted its name and coordinates, start and end dates, the war fronts it belongs to (e.g. Western front, Finnish Civil War), the country where it took place, and its duration in days.

18.3 Interface

Our interface is composed of a map module and a concordance table module, which allow to easily switch between a distant and a close reading of the dataset. These modules allow to observe how the spatial imaginaries are expressed in the whole dataset or in a selected subset. The dataset can be filtered according to the language, the title or the publication date of the newspapers.

The map module shows the location mentions present in the selected dataset, where the more frequent a location is, the bigger it appears on the map. The base map shows the borders and the capitals of countries between 1913 and 1920. This base map relies on data provided by Schvitz et al. (2021), which combines multiple sources to represent country borders from 1886 to 2019. Each country on the map is associated with a colour, where the darker the colour, the more frequently that country is mentioned in the selected dataset. Moreover, the map shows contextual information such as capital cities and battles that occurred in the selected time period. The data related to capital cities and battles come from the data we have extracted from Wikidata. As with location mentions, the longer a battle lasted, the bigger it appears on the map. Hovering or clicking on an entity shows the metadata related to it, such as its frequency, the newspaper it appears in or the link to the Wikidata resource associated with that entity mention, provided that link has been found[15].

[14] *OpenHistoricalMap* https://openhistoricalmap.org

[15] More details on the interface can be seen in our tutorial video: https://youtu.be/iIpEvM9IFaM

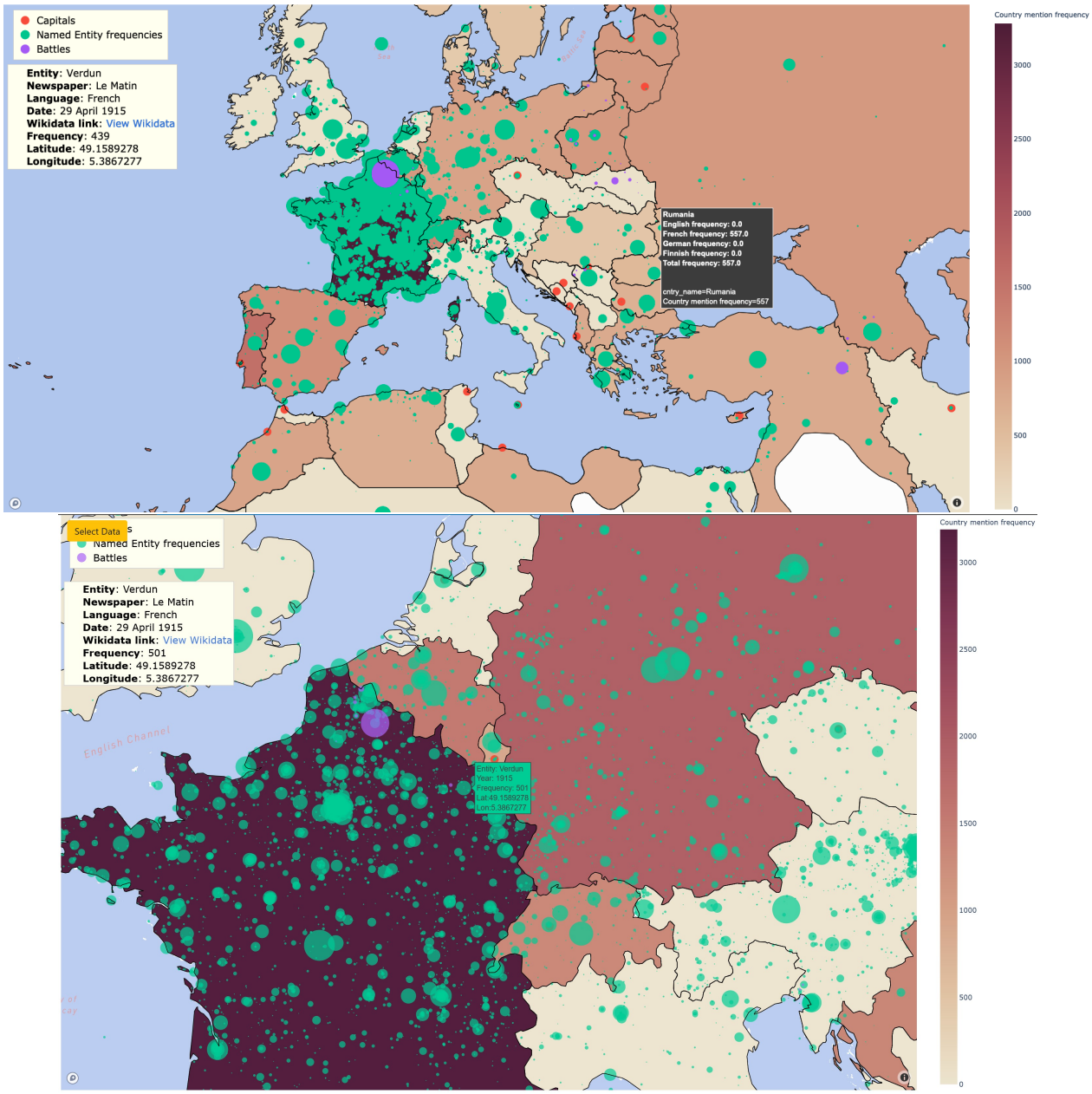


Figure 18.6: Example of the map module. Every location mentions in in *Le Matin* (1913-1915) are visualised. The legend of the map appears on the top-left corner of the interface. The data related to select location "Verdun" are also shown on the box on the left corner of the interface.

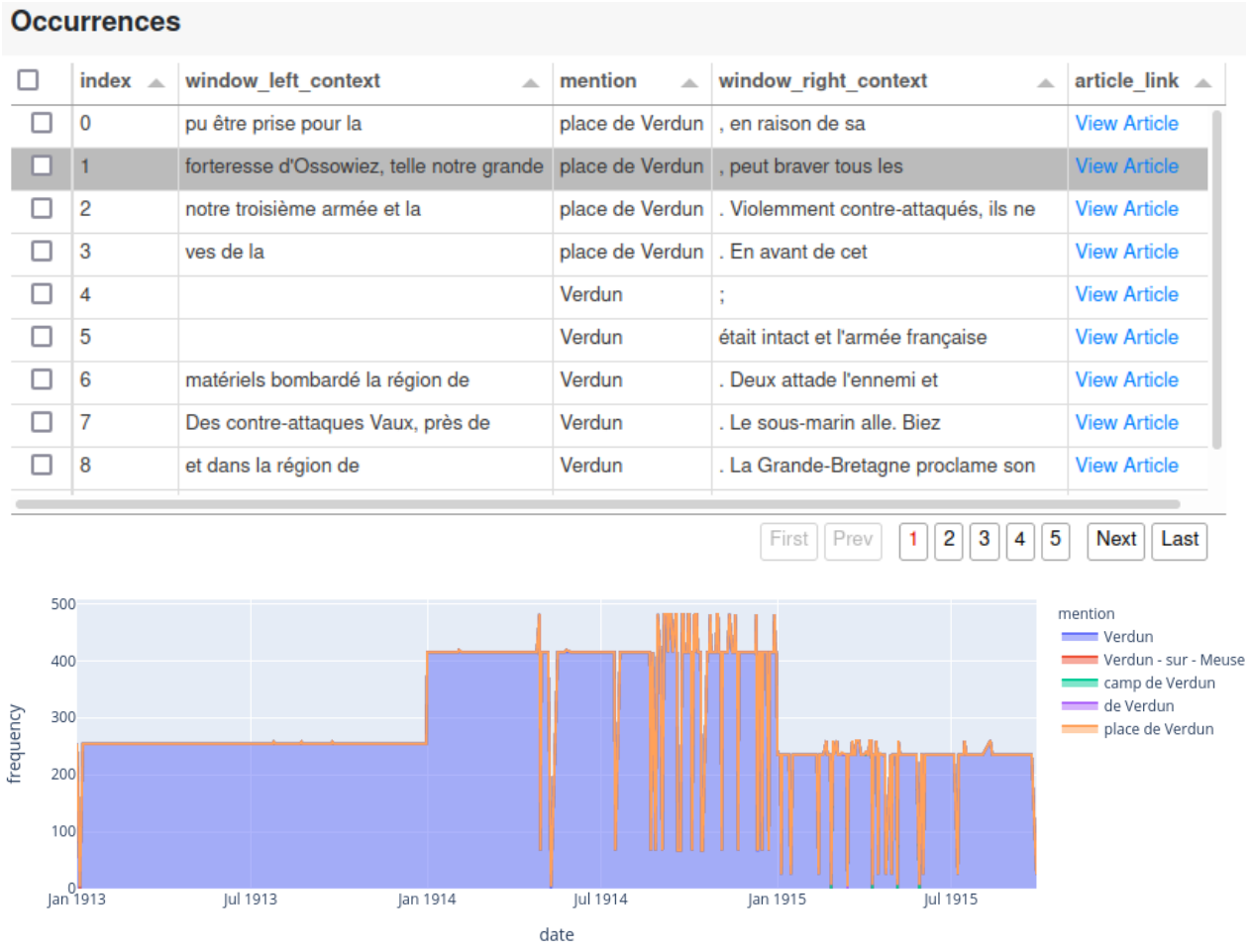


Figure 18.7: Occurrence of "Verdun" over time and in context in *Le Matin* (1913-1915) in the concordance table module

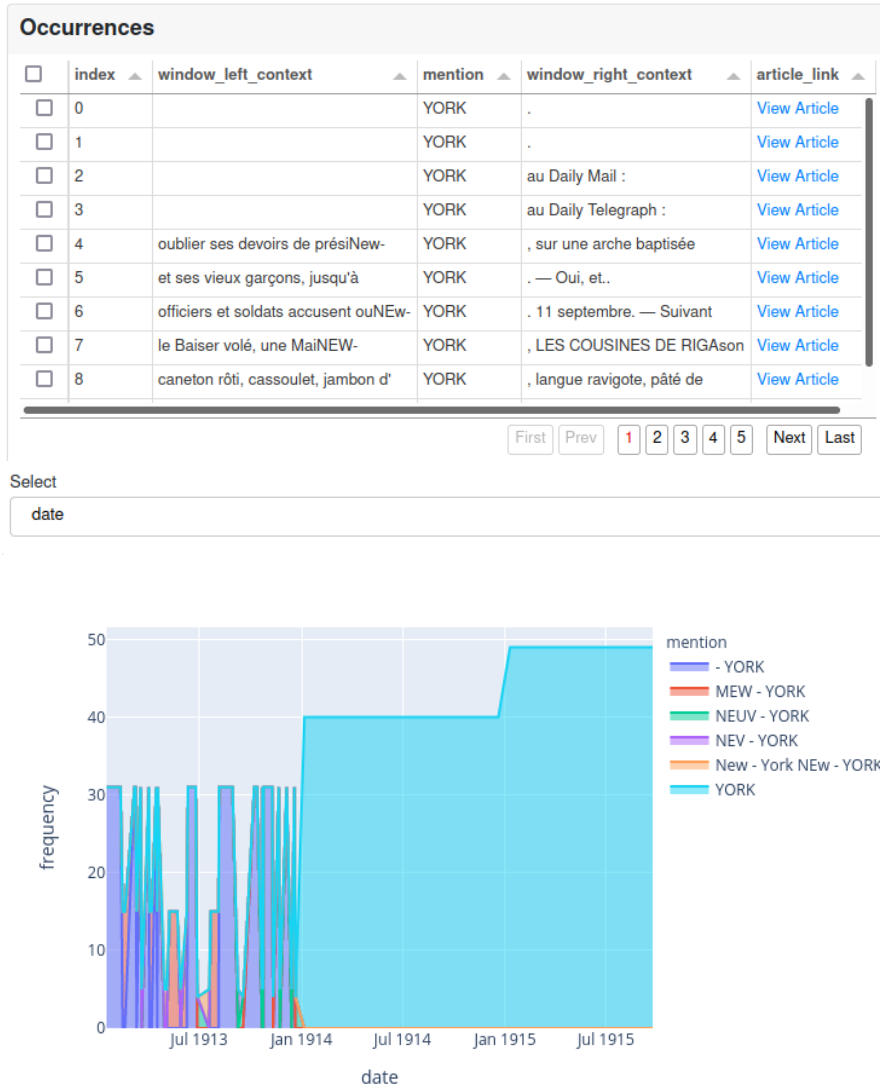


Figure 18.8: Occurrence of "York" over time and in context in *Le Matin* (1913-1915) in the concordance tablemodule

Figure 18.6 shows the results for *Le Matin* between 1913 and 1915 on the map module. Every location mentioned in this subset appear as a green dot, whereas the battles that occurred in that period appear as purple dot. Similarly, the capital of each country appear as an orange dot. The legend of the map appears on the top-left corner of the interface. The data related to a selected location, in this case "Verdun", are shown on the box on the left corner of the interface.

The concordance table module provides a concordance table where every mention of a location from the original document is surrounded by its left and right context. By default, these contexts are limited to 5 words before and after the mention. Each mention in the table is associated with a link that redirects the user to the NewsEye platform, which stores the original scans of the documents. This feature allows the user to extend their research by returning to the original document. Moreover, the concordance table is linked to a plot that shows the distribution of entity mentions across time, by language or by newspaper. Two examples of the concordance table and the plot are shown in Figure 18.7 and Figure 18.8. These figures respectively show the occurrences of "Verdun" and "York" over time and in context in *Le Matin* between 1913 and 1915.

18.4 Discussion

We have tested our interface by observing how the spatial imaginaries are expressed in the issues of *Le Matin* published between 1913 and 1915. This case study has revealed that most locations mentioned in this subset are located in France, Germany, Switzerland and Belgium, as shown by the map module in Figure 18.6. Those countries also appear more frequently than other surrounding countries, as shown by their darker filling colour. This suggests that *Le Matin* had a Eurocentric view of the war and focused on the Western front, even though the fighting spread also to Africa and Asia. This is an example of biases that are relevant for historical research and could be easily found by using our interface.

Moreover, we have chosen to focus on the the spatial imaginaries related to the city of Verdun. Because of its proximity with the German border, Verdun was a key position in the conflict. This is confirmed by the mentions of this location in *Le Matin* between 1913 and 1915, which nearly exclusively appears in war reports or articles related to the conflict, as shown in Figure 18.7. Interestingly, other location mentions related to Verdun reinforce the relation of the city with the topic of war. For instance, Verdun is sometimes mentioned as "*place de Verdun*" (seat of Verdun). "*place de*" usually indicates a square in a city, however in this case, it always refers to the city of *Verdun* itself. There are also mentions of the military camp in Verdun ("*camp de Verdun*"), but only in articles published in 1913 referring to the Franco-Prussian war. Both these mentions insist on the importance of the city as a defensive position, even before the beginning of WWI. The only exception to that is the mention of "*Verdun-sur-Meuse*", which was the official name of the city

between 1801 and 1970, and which is only mentioned to indicate the birthplace of a person. These are examples of elements that have been identified thanks to a distant reading of the dataset, and that have been confirmed by a close reading of the document.

These case studies show the validity and usefulness of such an interface. The map module we propose allows a distant reading of the corpus, which may help reveal hidden patterns in the data, whereas the concordance table module allows to read the dataset in a close manner, and to investigate elements that have been identified thanks to the map module. However, as we have highlighted earlier, the conception of such interfaces dedicated to the analysis of historical datasets remains challenged by the lack of open machine-readable representations of historical knowledge, even for exhaustively studied periods such as WWI. This lack of dedicated resources implies that NLP methods should rely on other types of resources, such as Wikidata or Geonames, that are suited for the processing of contemporary data, and not entirely suited for the processing of historical documents.

Chapter 19

Timelines

Table of contents

In this chapter we propose a conceptual framework for the Archive TimeLine Summarisation (ATLS) task, which aims to automatically generate informative, readable and interpretable timelines from archival collections of historical documents. These timelines could serve as a distant reading tool and as a first step in exploring a dataset by providing an overview of its key events. Moreover, these timelines could be combined with search engines and used as an interface to search results returned by issuing queries over large datasets, as suggested in Alonso et al. (2021) and Swan and Allan (2000). From there, the user could zoom into the documents in order to proceed to close reading. Furthermore, these summaries would be presented in chronological order, thus preserving the link between events, and could also be contextualised by adding data from external knowledge bases as in Ceroni et al. (2014).

The ATLS framework is an extension of the TimeLine Summarisation task, and addresses the challenges that standard TLS methods face when applied to archive collections, such as the sparsity of data, OCR problems, context shifts and linguistic changes over time in order to generate timelines based on these datasets. This work has been previously presented in Gutehrlé et al. (2022), and is the result of a three months research stay between April and July 2022 at the Digital Sci-

ence Center (DiSC) of the University of Innsbruck (Austria). This research stay was supervised by Professor Adam Jatowt (Department of Computer Science & DiSC, Deputy Head of Digital Science Center, University of Innsbruck, Austria) and Professor Antoine Doucet (L3i laboratory, La Rochelle University, France).

The rest of this chapter is structured as follows: we first present an overview of existing TLS methods in Section 19.1. We then describe the type of expected dataset the ATLS can process in Section 19.2, before presenting the framework in details in Section 19.3. Finally, we propose a discussion of the application of the ATLS framework in Section 19.4.

19.1 Related works

The TimeLine Summarisation task (TLS), which is a subfield of the Multi-Document Summarisation task (MDS), consists in summarising multiple documents by generating a timeline where important events detected in the dataset are associated with a time unit, such as a day. For instance, Swan and Allan (2000) generate clusters of Named Entities and noun chunks that best describe major news topics covered in a subset of the TDT-2 dataset (Allan et al., 1998), which contains text transcripts of broadcast news spanning from January 1, 1998, to June 30, 1998, in English. K.-H. Nguyen et al. (2014) generate timelines by detecting events that are the most relevant to a user query. They apply their methodology on a dataset of newswire texts in English covering the 2004-2011 period provided by the AFP French news agency. Duan et al. (2017) extend these methods to summarise the common history of similar entities such as Japanese Cities or French scientists. Examples of timelines generated by such methods are shown in Figure 19.1.

Most approaches to the TLS task generate timelines by applying the two following steps: a *date selection* step which identifies and ranks the key dates in the documents, and a *date summarisation* step which generates a summary of an event occurring at a specific date by picking important sentences in the documents published on that date. To identify important dates in the dataset, Gholipour Ghalandari and Ifrim (2020) select the most frequent date mentions, G. Tran, Herder, and Markert (2015) rely on a graph-ranking model, whereas Kessler et al. (2012) combine a clustering model and a supervised classifier. For the second step, La Quatra et al. (2021) apply state-of-the-art methods for Text Summarisation (TS) such as TextRank (Mihalcea & Tarau, 2004) whereas Martschat and Markert (2018) adapt methods from the Multi-Document Summarisation (MDS) field.

TLS methods are generally extractive, i.e. the summary is created by extracting textual elements such as titles, sentences or paragraphs from the input data (G. Tran, Alrifai, & Herder, 2015). Other works are abstractive, i.e. the summary is a completely new text generated by the system (Steen & Markert, 2019). TLS methods in general tend to be applied to summarise datasets describing

Date	Summary
2011-01-25	Thousands of protesters spilled into the streets of Egypt on Tuesday , an unprecedented display of anti-government rage inspired in part by the tumult in the nearby North African nation of Tunisia.
2011-01-26	Twitter says its site is being blocked in Egypt Egyptians took to the streets in what could be a sequel to the recent revolution in Tunisia witter , Facebook and YouTube were widely used in Tunisia 's uprising and in Iran last year -LRB-.
2011-01-28	With parts of his capital ablaze , Mubarak said he was asking his government to resign and would soon announce a new one , pledging to address the concerns of thousand of Egyptians protesting in Cairo 's streets . Amre Moussa , the Arab League 's secretary-general and a veteran Egyptian diplomat , joined protesters in Cairo 's Tahrir Square on Friday , state-run Nile TV reported .

From	To	Top headlines
5/2010	7/2011	Syrian officials launch tear gas against protesters Security forces shoot at protesters New York Times journalist with the Pulitzer died of an Asthma attack in Syria
8/2011	3/2012	Assad promises elections in February in Syria US withdraws ambassador from Syria for security reasons NATO says goodbye to Libya and the world turns to Syria
7/2012	12/2012	Meeting of senior officials in Geneva failed agreement to end violence in Syria Russia delivers three war helicopters to Syria Red Cross says Syria is in civil war
7/2016	11/2016	Maternity unit among hospitals bombed in Idlib air strikes Russian helicopter shot down in Syria. Turkish army enters Syria

Figure 19.1: Examples of generated timelines by Y. Yu et al. (2021) (top) and Campos et al. (2018) (bottom), summarising a set of documents about respectively Egyptian protests and the Syrian War. The top timeline outputs a summary on a day-to-day basis, whereas the bottom timeline lists events using uneven periods of time

large events, such as the Egyptian protests or the Syrian War (Martschat & Markert, 2018; G. Tran, Herder, & Markert, 2015). These methods require that the dataset covers a constrained period of time and is homogeneous, i.e. that the documents cover the same topic. Standard TLS methods are thus not suited to summarise heterogeneous or longitudinal datasets, such as archival collections of historical documents.

There exist a few datasets available for the TLS task, such as 17 Timelines (T17) (G. B. Tran et al., 2013), CRISIS (G. Tran, Alrifai, & Herder, 2015), ENTITIES (Gholipour Ghalandari & Ifrim, 2020), CovidTLS (La Quatra et al., 2021) or TLS-Covid19 (Pasquali et al., 2021) which are constructed from contemporary news articles. However, datasets are often lacking in most projects. It is therefore necessary to create a dataset from scratch, as in Bedi et al. (2017), Ge et al. (2015), Minard et al. (2015), and K.-H. Nguyen et al. (2014) or extend existing ones as in Y. Yu et al. (2021). Due to this lack of datasets, evaluating TLS systems is a difficult task. The date selection step can be evaluated with the F1-measure (Gholipour Ghalandari & Ifrim, 2020; La Quatra et al., 2021) or with the Mean Average Precision (MAP) metric (K.-H. Nguyen et al., 2014). The date summary is often evaluated with one of the ROUGE metrics (C.-Y. Lin, 2004) to compare a ground-truth timeline and a generated one (Duan et al., 2020; Gholipour Ghalandari & Ifrim, 2020; K.-H. Nguyen et al., 2014; Y. Yu et al., 2021). Methods relying on event detection such as Bedi et al. (2017), Ge et al. (2015), and Minard et al. (2015) often evaluate their system in terms of Precision, Recall and F1-measure. However, most projects often lack datasets and must then resort to human evaluation as in Duan et al. (2017), Swan and Allan (2000), and G. Tran, Alrifai, and Herder (2015).

In order to generalise the application of TLS, Y. Yu et al. (2021) propose a Multiple TimeLine Summarisation (MTLS) system, which generates a timeline for each story found in the dataset. To do so, it first detects events mentioned in the dataset and measures their saliency and consistency. An event linking step determines the link between these events in order to generate each timeline. Similarly, Duan et al. (2020) propose the Comparative TimeLine Summarisation (CTLS) task, which generates a comparative timeline highlighting the contrast between two timestamped timeline documents (e.g. biographies, historical sections, ...) by computing local and global importance of events.

Some works such as Chieu and Lee (2004), Kessler et al. (2012), K.-H. Nguyen et al. (2014), and Pasquali et al. (2019b) can be described as Query-based TimeLine Summarisation (QTLS), as they apply TLS on documents related to a user query such as documents returned by a search engine. QTLS generally consists in the two following steps: *event detection* and *event ranking*. To detect events, Chieu and Lee (2004) select any sentence where the terms of the query appear, K.-H. Nguyen et al. (2014) cluster by a common date every sentence returned by the query, and Pasquali et al. (2019b) detect the most frequent dates in the time span covered by the documents. Other

works train a classifier to detect important events (Chasin, 2010) or rank events by their importance with a Learning-to-Rank model (Ge et al., 2015). However, these classifiers need training data, which are difficult to create since defining what is important is a subjective matter. This can lead to disappointing results, as shown in Chasin (2010). To determine the importance of events, K.-H. Nguyen et al. (2014) first score them according to their relevancy and saliency to the query, then rerank them to ensure a diverse timeline. Chieu and Lee (2004) rank the importance of a sentence according to their "interest" and "burstiness", then remove duplicate sentences to ensure diversity. Pasquali et al. (2019b) use the keyword extractor YAKE! (Campos et al., 2018) to weight the terms in the event description. Duplicate event descriptions are detected with the Levenshtein similarity measure and removed. Those methods finally select the top most important events to generate the timeline.

19.2 Expected Dataset

The Archive TimeLine Summarisation framework takes as input a longitudinal dataset composed of timestamped documents, such as news articles from a historical newspaper collection. This dataset can be standalone or made of documents returned by a search engine for a given query $\mathcal{Q}$. The dataset could be in a raw format or may have been pre-processed. We would suggest at least the two following pre-processing steps: first, we recommend to clean the dataset if it has been processed with OCR, either manually or semi-automatically, since the OCR quality will impact further steps (T. T. H. Nguyen et al., 2021b). Secondly, we recommend to detect temporal expressions, as they are a good indicator of event mentions. Temporal expressions are either explicit (e.g. "February 17, 1995") or implicit (e.g. "yesterday", "next month"). One can use tools such as HeidelTime (Strötgen & Gertz, 2010) or SUTime (Chang & Manning, 2012) to detect temporal expressions in text and resolve them to an absolute date format, simplifying their use in the TLS process. However, we must keep in mind that the detection of temporal expressions, especially implicit ones, is still a challenging task. Moreover, available tools such as these were mainly conceived for contemporary data, and thus may not work as properly on historical data.

The input dataset could be pre-processed further by applying NLP methods such as Name Entity Recognition (NER), Topic Modelling (TM), Event Extraction (EE), Relation Extraction (RE), Keyword Extraction (KE), or Keyword Generation (KG). Such annotations could be used to index the dataset and allow the user to query documents about a specific Named Entity or topic, as in the impresso[1] or the NewsEye[2] platforms.

[1] https://impresso-project.ch/app/
[2] https://www.newseye.eu/

19.3 Framework

The ATLS framework consists of a *Timeline Generation* and a *Timeline Presentation* step, which are shown in Figure 19.2. The first step extracts textual elements describing an event from the dataset, and attributes them an importance score. The first step has to run only once over the processed dataset, since it aims to detect the elements composing the timeline to be generated. The second step generates the timeline by filtering events and selecting their description. In contrast, the second step can be run multiple times to update the timeline.

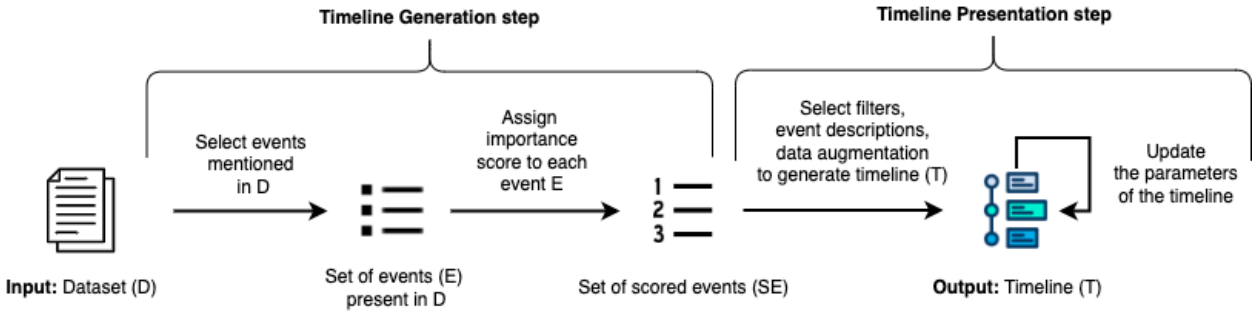


Figure 19.2: Conceptual pipeline for building the ATLS system

19.3.1 Timeline Generation

The *Timeline Generation* step extracts mentions of events from the dataset, and attributes them an importance score. Although events can be defined in many ways, a commonly accepted definition is "something that is *happening* or that is holding true in a given circumstance", as stated in the TimeML guidelines (Saurí et al., 2006).

Events can be detected in multiple ways: one could detect them through statistical analysis of the corpus. For instance, Chieu and Lee (2004) measure the occurrences of similar sentences associated with the same date, whereas Pasquali et al. (2019b) measure the occurrences of articles in atomic time intervals to later aggregate them and determine the bursty time periods. These statistical methods are especially suited for homogeneous datasets, but may not work as well on heterogeneous or fragmentary datasets. One could also train a Learning-to-Rank model on summaries created by experts in order to detect important sentences as in G. B. Tran et al. (2013). This would, however, require training data which remain rare.

Alternatively, one could use an Event Detection model to detect and annotate events in the dataset, as in Chasin (2010). Event Detection is another task that has been extensively studied in the NLP field, and has even been applied in a Digital Humanities context (N. K. Nguyen et al., 2020). However, training such model requires annotated resources that are often lacking, especially

for historical data. Moreover, the output of these models are impacted by the OCR quality of documents.

Finally, we could select as event any sentence containing at least one time expression, either explicit or implicit, as in Duan et al. (2019) and K.-H. Nguyen et al. (2014). This selection could be refined by taking sentences that also contain a Named Entity, as in Abujabal and Berberich (2015) and Bedi et al. (2017). One could then apply algorithms such as Affinity Propagation (Frey & Dueck, 2007) or Chinese Whispers (Biemann, 2006) to gather sentences describing the same event, as in Rusu et al. (2014), Steen and Markert (2019), and Y. Yu et al. (2021).

Regardless of the method used to detect them, events should all be associated with a temporal value. These could be the temporal expressions occurring with the event mentions, or the Document Creation Date (DCD) if no time expressions are present in the content of the document. Alternatively, approaches for estimating the focus time of text, in absence of any temporal expressions can be applied to associate event-related sentences with particular points of time (Jatowt et al., 2015).

As mentioned in Section 19.1, the importance of an event can be measured in a supervised or semi-supervised manner with a classifier (Chasin, 2010; Ge et al., 2015). This method, however, requires training data that are difficult to obtain or produce. Furthermore, the process leading a classifier to a prediction is generally not explained. Since the goal of this framework is to assist in the study of longitudinal datasets, it is necessary that the process of generating a timeline is interpretable. Thus, we would suggest to measure the importance score in an unsupervised manner by extracting features from the dataset, as in Campos et al. (2018), Chieu and Lee (2004), and K.-H. Nguyen et al. (2014). Some of the features that we think could help measure this importance score are listed below, with suggestions on how to compute them:

Redundancy : The more frequently an event is mentioned, the more important it should be. One can then simply count the occurrences of events, or as an alternative, assign them importance weights by calculating their TF-IDF scores over all the time units. However, as the data might be fragmentary in archive datasets, this feature should rather not be used alone

Contemporary references : an event may be important at a given time if other events occurring around the same period of time refer to it. Thus, to evaluate this feature, we could count how often an event is referred to from the descriptions of other events in a given short period of time around that event

Retrospective references : Similarly, an event is likely to be important if documents keep mentioning it some time after it occurred. To assess this kind of across-time reference to the event, one could count how often (and perhaps for how long) an event is mentioned by other events that occurred after a given period of time. Other solutions may rely on computing

random walks over graphs composed of timestamped events and/or entities to measure the amount of signal propagation from the past towards "the recent times" (Jatowt et al., 2016)

Causality : an event is likely to be important if it is the cause of other events that occurred after it. To evaluate the causality of an event, one could use *date reference graphs* as in G. Tran, Herder, and Markert (2015), which measure the frequency of references, the topical influence and temporal influence between two events to determine a causal link. It is also possible to use Causal Relation Extraction (CRE) methods as presented by Gao et al. (2019) for instance. However, the CRE task is far from solved and may require to pre-process the dataset further

Common sense : some events are clearly more important than other, e.g. the birth of a child or marrying a partner are usually more important events in a family history than repainting a house. To represent that kind of common sense knowledge and compute this feature, it may be necessary to create a dataset of events that are deemed important to train a 1-class classifier (1CC) as in Duan et al. (2019) or a Learning-to-Rank model as in Ge et al. (2015). Note that while important events can be collected from historical textbooks or history-related content, gathering unimportant events may be less easy and more problematic; hence the solution could be to rely on a 1CC task.

Using these features, a straightforward formula to calculate the importance of an event could be:

$$\alpha \cdot F1 + \beta \cdot F2 + \gamma \cdot F3 + \delta \cdot F4 + \epsilon \cdot F5$$

where $F1, F2, F3, F4, F5$ are the scaled values of the features described above and $\alpha, \beta, \gamma, \delta, \epsilon$ are hyper-parameters of which value is defined by the user or document archive custodians. Similarly to event detection, the user could be asked to select any of these features to compute this score.

Some periods may contain much more documents than others. For instance, fewer documents may be available during a war time because of censorship or paper restriction. This lack of documents may lead to events that are far more or far less mentioned than others, and bias frequency-based features such as *redundancy*, *contemporary* and *retrospective references*. Thus, these features should be normalised before being incorporated. Furthermore, we suggest these features since they are easy to compute, but we also acknowledge that they may not be sufficient to measure the importance of an event from the perspective of an expert such as a historian. Because the formula to compute the importance score is modular, one could incorporate more features in collaboration with experts.

19.3.2 Timeline Presentation

The *Timeline Presentation* step generates the timeline from events that have been detected and scored in the previous step. We present sets of filters to select which events should appear on the timeline and how they should be presented. We also describe an optional step of timeline augmentation using external data.

A dataset may contain hundreds or thousands of mentioned events. Thus, it is necessary to select those that will be added to the timeline. To do so, we can use filters such as described below. The weight of these filters could be changed on the user interface, thus allowing users to instantly update the timeline.

Top N : top N most important events are retained

Importance Threshold (IT) : only events of which the importance score is superior to a pre-fixed threshold IT are taken. Individual thresholds for the features described in Section 19.3.1 that make up the importance score can also be set

Topical Diversity Threshold ($TopDT$) : removes redundant event mentions and ensures the timeline is topically diverse. Topical diversity can be simply measured using Maximal Marginal Relevance (MMR) (Goldstein-Stewart & Carbonell, 1998) or the *n*-gram blocking metric as in Liu (2019)

Temporal Diversity Threshold ($TempDT$) : ensures every time unit on the generated timeline is evenly represented by setting a minimum and maximum number of events that can appear at each time unit

There are multiple ways to represent an event on a timeline. One could select a sentence that describes the event. If this sentence is too long, one could use sentence compression methods (Filippova & Strube, 2008) to only keep its most important part. As mentioned earlier, an event might be represented by a cluster of sentences. The user can thus select one sentence among this cluster or generate a cloud of terms of all sentences contained in it, as in Duan et al. (2019). One could also use headlines if the target documents are articles as in Pasquali et al. (2019b) and G. Tran, Alrifai, and Herder (2015).

Finally, we could also use a Natural Language Generation (NLG) system as in Steen and Markert (2019), as these generated texts are often easier to understand than text extracted from the documents. However, abstractive methods such as these may suffer from inaccuracies or hallucinations, i.e. generate information that is not present in the original documents. Thus, abstractive methods might generate improper event descriptions and lose the connection with the original documents. On the other hand, a common drawback of purely extractive methods is that selected sentences may

require some context or at least post-processing for users to be able to properly understand them (e.g. pronouns may need to be resolved or we may need to add definitions or descriptions of some entities or events).

To properly understand them, some events may require contextual knowledge that is missing from the processed dataset. This can especially happen if the user is not a domain expert. Such contextual knowledge may be found in knowledge bases such as Wikidata or Wikipedia Year pages (see for example N. K. Tran et al., 2015). Thus, timelines generated by an ATLS system could be augmented with contextual data provided by external Knowledge Bases as in Ceroni et al., 2014. These augmented timelines could help in explaining a dataset by summarising it and providing the user with the necessary knowledge to understand it. Unfortunately, most resources created by experts are not in a machine-readable format (Gutehrlé et al., 2021). Hence, this step may require more effort.

19.4 Discussion

Most TLS methods generate timelines through statistical analysis of the input dataset. They also often require that the input corpus contains documents of a similar type and similar content. However, an archive collection may be heterogeneous and contain documents of different authors, genres, topics and periods. It may also be fragmentary and not as complete as a contemporary dataset. In contrast, the ATLS framework that we propose allows to deal with the issues of such heterogeneous, longitudinal and fragmentary datasets. We present the main contributions of the ATLS framework compared to standard TLS methods in Table 19.1.

	TLS	ATLS
Covered period	Shorter	Longer
Input Data Size	Small / Medium	Large
Documents type	Timestamped documents (e.g. news articles)	
Document Format	Usually born digital	Often digitized
Data Integrity	Usually complete	Can be fragmentary
Presence of noise	Less likely	Depends on OCR quality
Semantic evolution	Less common	Possible (esp. over long time)
Need for query-based filtering	Optional (depends on data size and heterogeneity)	
Need for contextualization	Less likely	More likely (esp. over long time)
Need for interpretable output	Yes	Yes

Table 19.1: Comparison of the TLS and ATLS tasks

Search engines augmented with timelines would be especially useful in a Digital Humanities context such as for facilitating the study of historical datasets, as they would provide the necessary

context to understand past events and to structure the event landscape. They could also help the user to understand the history of a particular entity such as a person or a location, or even a group of such entities through providing a bird's-eye view of the relevant data. A good example of such search engine augmented with TLS is the Conta-me Histórias (Tell me stories) platform[3], where the user can query news articles from the Portuguese web archive. The user-friendly interface allows a distant reading of the documents returned by the query through a timeline that summarises them, but also allows close reading by preserving the link to the original documents. An example of the Conta-me Histórias is shown in Figure 19.3. To the best of our knowledge, works on applying TLS methods to provide a distant reading of archival collections remain however rare.

As mentioned in Section 19.1, the evaluation of a TLS system is a difficult task because of the lack of evaluation datasets and the inherent subjectivity of the task. In order to evaluate the output of an ATLS system, we would suggest to manually assess the produced timelines, either by following some evaluation criteria as in Duan et al. (2017), or by comparing them with resources created by experts such as timelines derived from history books as in Bedi et al. (2017). One could also use this system to bootstrap an evaluation dataset specific to the given corpus, towards an automatic evaluation.

Timelines are usually represented linearly, where each time unit is of the same size (usually a day or a year). However, the optimal granularity of temporal units might vary when generating a timeline over a long period of time. For example, when referring to a distant past, humans tend to often describe entire decades or years rather than discussing each day or month which is more common for the recent past. Furthermore, events mentioned in historical documents might not always be recorded with the same temporal precision (e.g., some events may have missing dates, the dates can be imprecise or difficult to be inferred). A possible solution would be to generate logarithmic timelines, where the granularity of the time unit changes over time, as suggested in Jatowt and Au Yeung (2011).

If the documents in the datasets are annotated with Named Entities, one could generate entity-based timelines. This could help understand the history of a specific entity such as a person or a location as in Duan et al. (2019). This idea could be extended by generating aggregate timelines for multiple entities at the same time. These timelines could be agglomerative or contrastive and respectively show the similarities and differences between the history of multiple entities of the same type (e.g., cities in the same region or country, scientists of the same area). Similar to Duan et al. (2020), such comparative timelines would allow to study the history of entities of the same or similar type, e.g. Berlin vs. Paris, or even entities of different types, e.g. Paris and the writer Victor Hugo.

[3] https://contamehistorias.pt/arquivopt

Figure 19.3: Timeline of events related to the query "world war 2" in the last 10 years produced by the Conta-me Histórias interface

Discussion and Limitations

In this part we discuss some of the limitations of the research presented in this manuscript.

Firstly, the results and evaluations presented in this PhD thesis focus only on the main contributions of our work, which are our proposed approach to the Logical Layout Analysis task (LLA), the DARES dataset and the ELIJERE approach. In parallel, we have made the choice to present a full processing pipeline for the historical documents, from images processed by OCR to obtain their textual structure in the XML ALTO format, to building the augmented search interfaces. To do this, and to be able to complete this work within the time frame of a PhD degree, some of the implementations and methods have not been evaluated in this manuscript. In particular, we have not evaluated the proposed rules for post-processing the OCR transcriptions of the EMONTAL corpus. As explained in Chapter 10, our aim was not to propose a complete correction of the OCR transcriptions, which would have required a much more complex post-processing pipeline. Instead, our aim was to propose a simple method to limit the impact of non-word errors on further processes applied to the corpus, such as our approach to the LLA task or the ELIJERE approach. At the same time, the evaluation of the ELIJERE approach on the EMONTAL corpus has clearly shown that the remaining errors in the OCR transcriptions have a strong impact on the its performance. Thus, an evaluation of our OCR post-processing rules might be necessary in order to evaluate their efficiency and determine if other post-processing methods would prove to be better suited to clean the transcriptions of the EMONTAL corpus.

Furthermore, as explained in Chapter 11, we have built a dataset in order to evaluate our approach to the LLA task. This dataset was built by manually selecting documents from the EMONTAL corpus with various layouts and with a satisfying OCR transcription quality. The logical label corresponding to each blocks and lines of text in the dataset were then manually added by a single annotator. However, this manual annotation process is not the focus of our work and thus we do not provide here the annotation guidelines that were used or the detailed protocol. These could be object to future publications for the sake of reproducibility of the construction of such datasets. Moreover, this manual construction of the dataset may have introduced a bias which we did not take into consideration for the evaluation of our approach to the LLA task. Thus, in order to limit the impact of this bias, it would be necessary to evaluate our proposed approach to the LLA task on other datasets composed of randomly selected historical documents.

As discussed in Chapter 16 and Chapter 17, the ELIJERE approach that we propose in this work suffers from several limitations. Firstly, the lexico-syntactic patterns that are learned lack diversity and may be incorrectly annotated by the distant supervision process. They directly depend on the diversity of expressions and richness of the DARES dataset, upon which the lexico-syntactic patterns are collected. Secondly, the ELIJERE approach struggles to adapt to the writing styles of the target corpus and to noisy inputs such as incorrect OCR transcriptions. The disparity of styles between those of the DARES dataset and those of the target corpus is a major factor for

the lower scores obtained during the evaluation on the EMONTAL corpus. Moreover, the DARES dataset and the lexico-syntactic patterns upon which the ELIJERE approach relies are built from a limited Knowledge Base and a related corpus. Although the ELIJERE approach could theoretically work with any Knowledge Bases and corpora, it currently relies heavily on the structure of the Wikidata Knowledge Base and the Wikipedia corpus. Thus, we must investigate the adaptability of the ELIJERE approach to other Knowledge Bases and corpora, and evaluate how this change impacts its performance.

The evaluation of our approach could also be improved by adapting our method to take into consideration other existing datasets. For the moment, we have evaluated the ELIJERE approach on custom datasets built from Wikipedia and from the EMONTAL corpus. Standard datasets such as DocRed for evaluating approaches on the Relation Extraction task could also be considered. This issue will be addressed in future works in order to properly evaluate the efficiency of our approach, and to compare it with other approaches to the Relation Extraction and the Joint Extraction of Relations and Entities tasks.

Finally, in Part V, we have proposed two conceptual frameworks to build semantically rich interfaces based on maps and timelines dedicated to the exploration of archival collections. We have described their motivations, implementations and given examples based on other datasets and projects. However, these frameworks are the result of external projects on the processing of historical documents that we realised in the context of this PhD thesis, and they were not directly applied to the EMONTAL corpus. We plan to implement and apply these frameworks to the EMONTAL corpus, which would allow us to show the relevance and efficiency of such interfaces for exploring archival collections. The development of such web-based interfaces would also contribute to open the access to the EMONTAL corpus to a wider audience, such as scholars in Humanities and to the general public, and raise interest to the importance of the exploration of archival collections and their content. More specifically, the usefulness of such interface could be evaluated by allowing the users to perform historical research related to specific topics, or to conduct historical studies, such as genealogical or prosopographic works. Such works would allow, for example, to identify the events, places and moments related to an individual, in order to construct a map of their movements, a timeline of the most important events that occurred in their lives or establish their social network.

Conclusion

In this thesis, we have proposed new NLP approaches to the problem of exploring and exploiting archival collections of historical documents. Thanks to the digitisation campaigns undertaken by archives and libraries in recent years, archival collections of historical documents have become accessible to a wider audience. However, searching for information in these collections remains a challenge due to the lack of structure of their textual content. In this context, we have proposed several methods to structure these documents by annotating their textual content with semantic annotations.

As we have highlighted in this thesis, the pre-processing of documents in order to prepare them for the application of NLP methods is an important stage. In this work, we have focused on three pre-processing steps. Firstly, we have proposed an approach to clean the transcriptions of the documents obtained by OCR methods. This approach is based on heuristic rules to remove artefacts generated by the hyphenation process and to clean erroneous transcriptions. Our priority with this approach was to limit the impact of incorrect transcriptions on subsequent processing.

Secondly, we have proposed an approach to the task of Logical Layout Analysis (LLA) to determine the logical structure of documents. Our approach is based on geometrical, morphological and semantic features extracted from documents in the XML ALTO format. We have compared the performance of three implementations of this approach: a rule-based model based on hand-crafted rules, a rule-learning implementation based on the RIPPER algorithm, and a machine-learning implementation based on the Gradient Boosting algorithm. The evaluation of these implementations shows that the rules we have designed specifically for our corpus outperform the other two methods. However, this evaluation also suggests that combining methods in a hybrid way can achieve even better results. It also suggests that rule-learning methods can help in the construction of initial rule sets, or in the construction of rules dedicated to specific layouts.

Thirdly, we applied our approach to the LLA task to convert the documents in the XML ALTO format from our corpus to the EMONTAL format. This new format is inspired by the Dublin Core and XML DocBook formats, and describes the logical structure of documents. This makes it suitable for the application of NLP methods. We have proposed a set of tags to integrate the semantic annotations related to the task of Joint Extraction of Relations and Entities. However, this format could be extended to integrate semantic annotations related to other NLP methods such as Event Extraction, Sentiment Analysis, etc.

Considering the semantic processing of historical documents, a major contribution of this thesis is the Extensible, Lightweight and Interpretable Joint Extraction of Relations and Entities (ELIJERE) approach, which is a novel approach that we have proposed to the Joint Extraction of Relations and Entities (JERE) task. Unlike most other approaches to the JERE task, which rely on deep-learning architectures, our approach relies on two linguistic resources, called Syntactic Index and Lexical Index, to extract and categorise the relations and entities mentioned in sentences. These

resources are built from lexico-syntactic patterns that express relationships between entities. The Syntactic Index describes which relations a lexico-syntactic pattern expresses, as well as the kind of entities involved in the relations. The Lexical Index describes how a relation is expressed lexically. To collect these patterns, we have introduced the DARES dataset, which consists of weakly annotated sentences collected from Wikipedia articles. These sentences are weakly annotated by applying the distant supervision method (Mintz et al., 2009), while the lexico-syntactic patterns are collected by extracting the Shortest Dependency Path (R. Bunescu & Mooney, 2005) between the entities involved in a relation. By applying this method, we can quickly build an annotated dataset of sentences expressing any concept stored in a Knowledge Base, upon which we can build our linguistic resources. Thus, our approach is extensible and interpretable as it is based on extensible and explicit linguistic resources. It is also lightweight, as it requires little computational resources and can run on CPUs. We have evaluated a base implementation as well as a hybrid implementation of the ELIJERE approach on the DARES dataset and the EMONTAL corpus. The evaluations on both datasets show that the hybrid implementation largely outperforms the base implementation. In future works, we intend to evaluate the ELIJERE approach on standard datasets such as the New York Times dataset or TACRED, in order to compare it with other approaches to the JERE task, in particular deep-learning approaches.

We have evaluated our method in several steps: we have first evaluated the ELIJERE approach on the DARES dataset, before evaluating it on the EMONTAL corpus, in order to study the impact of OCR errors and of the various writing styles of our corpus on our approach. For each dataset, we have evaluated our approach on the Relation Extraction and Named Entity Recognition tasks separately. The results of the evaluation on the DARES dataset show that our approach needs improvement in several directions. Since the annotation obtained by the distant supervision method may be incorrect, a post-processing step is required to clean these errors, and improve the performance of the system. We have made several suggestions for post-processing these annotations, such as applying clustering methods. Although our approach is able to collect precise patterns, these patterns lack are however lacking in diversity. We have made several suggestions to collect more diverse patterns, such as relying on surface patterns instead of lexico-syntactic patterns, collecting the patterns from a corpus not only composed of Wikipedia articles, or ask a Large Language Model, such as ChatGPT, to generate new sentences from which we can collect patterns. The evaluation also shows that our approach is able to categorise the entities involved in a relation, but struggles to properly determine their boundaries. Similarly, we have proposed several solutions to this issue, such as learning the entities boundaries alongside the lexico-syntactic patterns from the DARES dataset, instead of relying on handcrafted rules.

The results of the evaluation on the EMONTAL corpus show that our approach is impacted by the errors produced by the OCR process, and struggles to adapt to the various writing styles of the

corpus. We have made several suggestions to adapt the linguistic resources to the target corpus, for instance by applying the bootstrapping method. Similarly, we have made several suggestions to make the patterns more robust to the noise introduced by incorrect OCR transcriptions, for instance by relying on fuzzy matching methods instead of exact matching methods. Moreover, the approach that we used to clean the errors produced by the OCR process only takes into account the forms of tokens and does not consider the contexts of occurrences. Thus, it may not clean some of the errors that would require to take the context into account. Although we have not evaluated our approach to clean the OCR transcriptions, the impact that these errors have on the ELIJERE approach suggests that the cleaning method is not sufficient, and should be improved.

Finally, we have proposed two conceptual frameworks that exploit the semantic annotations of the documents to build augmented search interfaces with close and distant reading functionalities. First, we proposed a framework for building an interface that allows to study the expression of spatial imaginaries in documents. This interface consists of a map module and a concordance table module that provide different perspectives on the underlying corpus. We described the necessary steps to build this interface, before testing it on a dataset of European newspapers published during the period of the First World War. Secondly, we proposed the Archive TimeLine Summarisation (ATLS) framework for the automatic generation of timelines over longitudinal, heterogeneous and fragmented document collections, such as archival collections. We have made several proposals to implement this framework and generate informative, readable and interpretable timelines, before discussing its potential applications and extensions.

Such semantically rich interfaces, combined with search engines, may prove to be useful tools for providing an overview of historical collections, as well as serving as a novel means of accessing information in archives. Although we have not applied these frameworks to our corpus, the semantic annotation methods we have proposed solve some of the main problems associated with the implementation of such interfaces. Our future work will consist in applying these frameworks to our corpus and building semantically rich interfaces dedicated to the *Bourgogne* and *Franche-Comté* collections. We also intend to ask humanities scholars (historians, archivists, etc.) to evaluate the quality of the maps and timelines generated and to assess the effectiveness of these frameworks for the study of archival collections.

To conclude, by experimenting with a corpus of historical printed periodicals of various origins published in the 19th and 20th centuries in the *Bourgogne* and *Franche-Comté* regions of France, we have developed a pipeline for the processing of such historical documents. The approaches developed in our work have numerous applications in the field of Digital Humanities. Firstly, search engines can be improved by exploiting the semantic annotations added to the textual data by our methods. Secondly, richer interfaces, based on visual modules such as timelines and maps, can be created to allow distant reading access to the collections. However, the difficulty of searching

for information in documents also concerns corpora of other origins and time periods, such as web pages, financial documents or scientific publications. In this context, the methods proposed in this thesis have been designed so that they can be applied to corpora of other origins than archival collections.

As we have highlighted in this work, the semantic annotation of documents requires resources such as rule sets or annotated datasets, that still remain rare today. For this reason, we have focused on methods such as rule-learning or distant supervision to accelerate the construction of such resources. In future works, we intend to focus on methods such as these that support the creation of resources required for rule-based and machine learning approaches. Furthermore, the evaluation of our approach to the LLA task, as well as the evaluation of the ELIJERE approach, suggest that hybrid approaches that combine rule-based and machine-learning methods can achieve better results than approaches that rely on a single method. We intend to delve deeper in the conception of such hybrid approaches in future works.

Annexes

Annexe A: Document collections in the EMONTAL corpus

Regional fund: Franche-Comté

1	L'Avenir de l'Est : organe du Front national de lutte pour la libération et l'indépendance de la France, de Meurthe-et-Moselle, Vosges, Haute-Marne, Haute-Saône, Doubs et du Territoire de Belfort
2	Bulletin de la Société amicale des anciens combattants du 4e RAC
3	Bulletin de la Société d'agriculture, sciences et arts de Poligny (Jura)
4	Circulaire économique et financière de Bourgogne et de Franche-Comté : publication mensuelle ["ou" publication bi-mensuelle]
5	Délibérations du Conseil général du territoire de Belfort
6	Le Démocrate : Organe des Républicains radicaux et radicaux-socialistes du pays de Montbéliard
7	Le Franc-comtois. Ed. par les régions du Doubs, Haute-Saône, Jura et Territoire de Belfort du Parti communiste français. Ed. spéciale
8	Le Franc-Comtois de Paris : journal d'information des départements Doubs, Jura, Haute-Saône, Haut-Rhin
9	La Franc-comtoise. Journal des Comités féminins de Franche-Comté. Ed. de Belfort et du Territoire
10	La Franche-Comté à Paris : journal d'informations des départements du Doubs, Jura, Haute-Saône, Haut-Rhin ["puis" organe des intérêts des Franc-Comtois à Paris, et d'informations des départements du Doubs, du Jura, de la Haute-Saône et du Territoire de Belfort]
11	La Franche-Comté libre. Organe régional du Front national pour la liberté et l'indépendance de la France. Doubs et territoire de Belfort
12	La Haute-Saône libre : organe départemental du Front national
13	Maîtrise de la Métropole / Association des anciens professeurs et élèves

Table 19.2: Documents belonging to the *Fond régional : Franche-Comté* in the EMONTAL corpus

14	Mémoires et documents inédits pour servir à l'histoire de la Franche-Comté / publiés par l'Académie de Besançon
15	L'Écho paroissial de Censeau : paraissant le 1er de chaque mois
16	Le Petit écho de Sainte-Madeleine : Besançon
17	Le Petit écho du 21e Régiment d'infanterie
18	Le Petit semeur : bulletin mensuel du Patronage central de Besançon
19	Le Semeur : journal communiste : Territoire de Belfort, Doubs, Jura, Haute-Saône
20	Le Semeur. Organe régional du Parti communiste. [Région de Montbéliard]
21	Le Socialiste comtois : organe hebdomadaire de la Fédération du Doubs (Parti socialiste, Section française de l'Internationale ouvrière) ["puis" organe hebdomadaire des fédérations socialistes du Doubs, de la Haute-Saône et du Haut-Rhin]
22	Les Sports comtois : organe sportif de Franche-Comté, paraissant le 1er de chaque mois et s'intéressant à tous les sports
23	Vers l'avenir : organe mensuel de la Jeunesse catholique de Franche-Comté et du Territoire-de-Belfort
24	La Vie meilleure : revue sociologique et littéraire de l'Est, paraissant le 1er et le 15 de chaque mois / [gérant : Alfred Jacquin]
25	Vive labeur : organe du Sillon de l'Est : paraît le 10 du mois
26	Bulletin de la Société belfortaine d'émulation
27	Supplément illustré du Petit comtois
28	[Franche-Comté Libération]
29	Tableaux de l'économie franc-comtoise / Institut national de la statistique et des études économiques, Service régional de Franche-Comté
30	Frontière : journal de marche du Groupement Frontière

Table 19.3: Documents belonging to the *Fond régional : Franche-Comté* in the EMONTAL corpus

31	La Libre Comté : organe régional du Mouvement de la libération nationale
32	Rapports et procès-verbaux des délibérations - Conseil général du Jura / Conseil général du Jura
33	Rapports et délibérations / Conseil général du Doubs
34	Rapports et délibérations / Conseil général de la Haute-Saône
35	Le Front comtois : organe du Front populaire de Dole et de la région ["puis" hebdomadaire du Front populaire]
36	Annuaire : programme des cours et des conférences / Université de Besançon
37	Les Sports de l'Est : hebdomadaire illustré de tous les sports, Lorraine, Franche-Comté, Bourgogne, Champagne
38	Bulletin paroissial d'Arc-lès-Gray
39	Bulletin paroissial de la région de Granges
40	Bulletin paroissial de Baume-les-Dames : paraissant le 1er dimanche du mois
41	Écho paroissial de Saint-Jean
42	Bulletin mensuel de la paroisse de Champlitte
43	Bulletin paroissial de Saint-Vit
44	Bulletin paroissial de Vesoul : [écho de N. D. de la Motte]
45	Bulletin paroissial de Belleherbe
46	Le Patriote comtois : organe des comités du Front national de Franche-Comté

Table 19.4: Documents belonging to the *Fond régional : Franche-Comté* in the EMONTAL corpus

Regional fund: Bourgogne

1	Les Annales chalonnaises et bourguignonnes : recueil bi-mensuel d'articles et de documents d'histoire locale publié par souscription / sous le patronage de la Société d'histoire et d'archéologie de Chalon-sur-Saône
2	Annales d'Igé en Mâconnais : recueil de documents et de matériaux pour servir à l'histoire de cette commune
3	Annuaire administratif, commercial et historique du Département de Saône-et-Loire...
4	L'Aurore nouvelle de Saône-et-Loire : Journal ["puis" Organe] socialiste indépendant
5	Le Bien du terrien : viticole, agricole et littéraire : supplément mensuel indépendant au Cri des terriens : organe d'éducation rurale
6	La Bourgogne littéraire et scientifique : revue régionaliste
7	La Bourgogne rurale : revue mensuelle d'agriculture, de viticulture et d'horticulture ["puis" fondé en 1903 par Jean Guicherd, inspecteur général de l'agriculture : organe général des groupements agricoles, viticoles et horticoles...]
8	La Bresse louhannaise : bulletin mensuel de la Société d'agriculture et d'horticulture de l'arrondissement de Louhans
9	Bulletin bi-mensuel de l'Ecole Saint-François de Sales de Dijon
10	Bulletin d'histoire et d'archéologie religieuses du Diocèse de Dijon
11	Bulletin de l'Union agricole et viticole : société d'encouragement à l'agriculture de l'arrondissement de Chalon-sur-Saône
12	Bulletin de la Société bourguignonne d'apiculture : organe de propagande d'apiculture méthodiste et rationnelle, spéciale pour la région de la Basse-Bourgogne
13	Bulletin de la Société centrale d'horticulture de l'Yonne
14	Bulletin de la Société d'agriculture de l'arrondissement de Charolles : organe du Syndicat des éleveurs, du Syndicat des emboucheurs du charollais et du brionnais et du Syndicat agricole
15	Bulletin de la Société d'archéologie et d'histoire de Tonnerre / président Léon Lacroix
16	Bulletin de la Société départementale d'agriculture de la Nièvre : ["puis" paraissant tous les mois "puis plus de sous-titre"]
17	Bulletin de la Société vigneronne de l'arrondissement de Beaune (Côte-d'Or)
18	Bulletin du Comité d'agriculture de Beaune
19	Bulletin du Photo-club sénonais : revue trimestrielle illustrée
20	Bulletin du Syndicat agricole du département de l'Yonne

Table 19.5: Documents belonging to the *Fond régional : Bourgogne* in the EMONTAL corpus

21	Bulletin du Syndicat viticole de la côte dijonnaise : ["puis" fondé en 1890]
22	Bulletins de la Société des sciences naturelles de Saône-et-Loire
23	Circulaire économique et financière de Bourgogne et de Franche-Comté : publication mensuelle ["ou" publication bi-mensuelle]
24	Le Collectionneur : bulletin mensuel de publicité, philatélie, cartophilie...
25	Le Combattant républicain : journal de la FNCR de l'Yonne
26	Le Courrier de la Nièvre : ancien Petit Nivernais : journal hebdomadaire, politique et littéraire, paraissant le dimanche
27	Le Cri des terriens : Côte d'Or, Saône-et-Loire, Yonne, Nièvre, Ain : organe de défense paysanne et du bloc des viticulteurs et cultivateurs bourguignons
28	L'Eduen : feuille de liaison périodique de la Société éduenne des lettres, sciences et arts et de la Société d'histoire naturelle d'Autun
29	L'Émancipateur : hebdomadaire communiste de la région nivernaise
30	Enquêtes sur la Révolution en Côte-d'Or / Comité départemental de la Côte-d'Or ; avec une note de E. Champeaux
31	L'Essor : revue mensuelle bourguignonne, artistique, littéraire ["puis" revue bourguignonne d'art et de littérature "puis" bulletin trimestriel des artistes et artisans d'art bourguignons]
32	L'Étincelle : organe mensuel de Saône-et-Loire
33	La Fédération agricole de l'Yonne : organe bi-mensuel ["ou" organe tri-mensuel] de la Fédération des associations agricoles, de l'Union des syndicats professionnels agricoles, de l'Union départementale des sociétés d'assurance mutuelle contre la mortalité du bétail...
34	La France féminine : revue mensuelle pendant la guerre : organe de défense des droits de la femme française
35	Jeunesse du Maquis
36	Journal des ouvriers du Creusot
37	Le Libre-penseur de Saône-et-Loire : Organe et propriété de la Fédération départementale des sociétés de libre-pensée de Saône-et-Loire. Mensuel
38	Mémoires de la Commission des antiquités du département de la Côte-d'Or
39	Le Mutilé : organe des mutilés, réformés et victimes de la guerre
40	Le Patriote de Saône-et-Loire : journal des mouvements unis de résistance

Table 19.6: Documents belonging to the *Fond régional : Bourgogne* in the EMONTAL corpus

41	Recueil des instructions relatives à la défense nationale / Préfecture de la Côte-d'Or
42	Revue champenoise et bourguignonne : (Aube, Côte-d'Or, Haute-Marne, Marne, Seine-et-Marne et Yonne) : archéologie, beaux-arts, bibliographie, biographie, documents inédits, géologie, histoire : revue régionale d'histoire et d'érudition paraissant tous les deux mois / publiée sous le patronage de MM. Ernest Babelon et Auguste Longnon,... ; [Eugène Maury, secrétaire de la rédaction]
43	La Revue du Centre : Nivernais, Berry, Bourbonnais, Bourgogne, Orléanais ["puis plus de sous-titre, puis" revue mensuelle illustrée "puis" revue trimestrielle illustrée]
44	Russie d'aujourd'hui. Organe mensuel des Amis de l'Union soviétique. Région de Saône-et-Loire
45	La Sardine : revue mensuelle : conserves alimentaires : sardines, maquereaux, thon, sprats, etc., salaisons / directeur Paul-Louis Tregan
46	Mémoires / Société d'archéologie de Beaune. Histoire, lettres, sciences et arts
47	Mémoires / Société d'histoire, d'archéologie et de littérature de l'arrondissement de Beaune
48	Bulletin... / Société d'histoire naturelle d'Autun
49	Assemblée générale du..., rapport sur l'exercice... / Syndicat d'initiative d'Avalon et du Morvan...
50	La Terre de Bourgogne : La Bourgogne agricole et La Bourgogne rurale réunies : organe de la Fédération des associations agricoles et viticoles de la Côte-d'Or, du Syndicat de défense paysanne, de l'Office agricole départemental et des services agricoles
51	Union populaire. [Chalon-sur-Saône]
52	Vaincre ! Organe de Saône-et-Loire du Front national de lutte pour la liberté et l'indépendance de la France
53	La Vigne américaine : sa culture, son avenir en Europe
54	La Voix de la Bourgogne. Edition nouvelle éd. par la Région bourguignonne du Parti communiste S.F.I.C. ...
55	La Voix de la Bourgogne. Organe de la Région communiste de Saône-et-Loire
56	La Voix des femmes de Bourgogne
57	La Voix des femmes de la Côte-d'Or
58	La Voix des femmes de Saône-et-Loire
59	La Voix du mutilé de Mâcon et de la région : organe mensuel de l'Association des mutilés et veuves de guerre
60	L'Yonne. Organe régional du rassemblement national

Table 19.7: Documents belonging to the *Fond régional : Bourgogne* in the EMONTAL corpus

61	La Revue de Bourgogne
62	Alauda : études et notes ornithologiques / recueil publié par P. Paris,...
63	Mémoires de la Société académique du Nivernais
64	Bulletin de la Société des sciences historiques et naturelles de l'Yonne
65	Bulletin de la Société scientifique artistique de Clamecy
66	Bulletin / Association bourguignonne des sociétés savantes
67	Pro Alesia : revue mensuelle des fouilles d'Alise et des questions relatives à Alesia / publiée sous le patronage de la Société des sciences de Semur par M. Louis Matruchot,...
68	Mémoires de la Société pour l'histoire du droit et des institutions des anciens pays bourguignons, comtois et romands
69	Société des amis des arts, sciences, archéologie et histoire locale de la Bresse louhannaise
70	Société d'histoire naturelle du Creusot / gérant A. Dulac
71	Mémoires de l'Académie des sciences, arts et belles-lettres de Dijon
72	Mémoires de la Société éduenne
73	Bulletin de la Société des sciences historiques et naturelles de Semur (Côte-d'Or)
74	Revue périodique de vulgarisation des sciences naturelles et préhistoriques de la Physiophile / Société d'études d'histoire naturelle de Montceau-les-Mines
75	Mémoires de la Société bourguignonne de géographie et d'histoire
76	Bulletin de la Société archéologique et biographique du canton de Montbard
77	La Mère éducatrice : revue mensuelle d'éducation populaire
78	Bulletin de la Société d'études d'Avallon
79	La Revue du Charolais : histoire régionale, lettres, sciences, actualités : Charolais, Brionnais et Bourbonnais
80	Mémoires de la Société d'histoire et d'archéologie de Chalon-sur-Saône

Table 19.8: Documents belonging to the *Fond régional : Bourgogne* in the EMONTAL corpus

81	Bulletin de la Société archéologique de Sens
82	Bulletin de la Société d'études du Brionnais
83	Rapports du Préfet, procès-verbaux des délibérations / Conseil général de la Nièvre
84	Rapports et délibérations / Conseil général du Département de la Côte-d'Or
85	Rapports et délibérations / Conseil général, Saône-et-Loire
86	La Revue de la Nièvre et du Centre
87	[Le Corps et l'esprit]
88	Rapports et délibérations / Département de l'Yonne, Conseil général
89	Les Sports de l'Est : hebdomadaire illustré de tous les sports, Lorraine, Franche-Comté, Bourgogne, Champagne
90	Bulletin de la Société d'horticulture et viticulture de la Côte-d'Or : agrégée à la Société d'acclimatation
91	Bulletin paroissial de Saint-Bénigne de Dijon
92	Bulletin paroissial mensuel de Saint-Apollinaire
93	Bulletin paroissial de Notre-Dame, Dijon
94	Écho paroissial de Cruzy-le-Châtel : diocèse de Sens, Yonne
95	Bulletin paroissial de Saint-Vincent de Châlon-s-S. : paraissant tous les mois
96	Bulletin paroissial de Saint-Pierre et du Sacré-Coeur, Châlon-sur-Saône : paraissant tous les mois
97	Bulletin paroissial de Brassy
98	Annales du doyenné de Pontailler-sur-Saône : diocèse de Dijon (Côte-d'Or) : paraissant le premier dimanche de chaque mois ["puis" bulletin mensuel]
99	Bulletin paroissial de Tilchâtel et d'Échevannes : Diocèse de Dijon, Côte d'Or
100	Bulletin paroissial de Saint-Pierre de Dijon
101	Bulletin paroissial de Saint-Michel
102	Notre clocher : bulletin paroissial mensuel de Saint-Paul de Dijon
103	L'Écho de Saint-Jean : bulletin mensuel paroissial
104	La Bourgogne combattante / édité par le Front national de lutte pour la liberté et l'indépendance de la France
105	Bulletin régional / PCF Nivernais, Berry, Touraine, Beauce
106	En avant ! : organe clandestin du Front national pour l'Ouest Côte d'Orient
107	Le Jeune patriote de Bourgogne / édité par les comités bourguignons du Front patriotique de la jeunesse
108	Le Patriote nivernais : organe départemental du Front national de lutte pour l'indépendance de la France
109	La Terre / édité par la Région de l'Yonne du Parti communiste français
110	La Vie ouvrière
111	Le Travailleur : organe de la région Côte d'Orienne du Parti communiste français
112	Le Travailleur de l'Yonne : organe régional du Parti communiste français
113	L'Yonne libre : organe régional du Front national

Table 19.9: Documents belonging to the *Fond régional : Bourgogne* in the EMONTAL corpus

Annexe B: Thematic categories of the documents in the EMONTAL corpus

Agriculture

Bulletin de la Société d'agriculture, sciences et arts de Poligny (Jura)
La Terre de Bourgogne : La Bourgogne agricole et La Bourgogne rurale réunies : organe de la Fédération des associations agricoles et viticoles de la Côte-d'Or, du Syndicat de défense paysanne, de l'Office agricole départemental et des services agricoles
Bulletin de la Société d'horticulture et viticulture de la Côte-d'Or : agrégée à la Société d'acclimatation
La Bresse louhannaise : bulletin mensuel de la Société d'agriculture et d'horticulture de l'arrondissement de Louhans
La Bourgogne rurale : revue mensuelle d'agriculture, de viticulture & d'horticulture (puis fondé en 1903 par Jean Guicherd, inspecteur général de l'agriculture : organe général des groupements agricoles, viticoles et horticoles...)
La Vigne américaine : sa culture, son avenir en Europe
La Fédération agricole de l'Yonne : organe bi-mensuel (ou organe tri-mensuel) de la Fédération des associations agricoles, de l'Union des syndicats professionnels agricoles, de l'Union départementale des sociétés d'assurance mutuelle contre la mortalité du bétail...
Bulletin du Syndicat agricole du département de l'Yonne
Bulletin de la Société vigneronne de l'arrondissement de Beaune (Côte-dOr)
Bulletin de la Société d'études d'Avallon

Table 19.10: Document collections in the EMONTAL corpus belonging to the *Agriculture* thematic category

Bulletin de la Société départementale d'agriculture de la Nièvre : (puis paraissant tous les mois puis plus de sous-titre)
Bulletin de l'Union agricole et viticole : société d'encouragement à l'agriculture de l'arrondissement de Chalon-sur-Saône
La Revue de Bourgogne
Le Bien du terrien : viticole, agricole et littéraire : supplément mensuel indépendant au Cri des terriens : organe déducation rurale
Bulletin du Comité d'agriculture de Beaune
Bulletin de la Société d'agriculture de l'arrondissement de Charolles : organe du Syndicat des éleveurs, du Syndicat des emboucheurs du charollais et du brionnais et du Syndicat agricole
Bulletin du Syndicat viticole de la côte dijonnaise : (puis fondé en 1890)
Bulletin de la Société centrale d'horticulture de l'Yonne
Bulletin de la Société bourguignonne d'apiculture : organe de propagande d'apiculture méthodiste et rationnelle, spéciale pour la région de la Basse-Bourgogne

Table 19.11: Document collections in the EMONTAL corpus belonging to the *Agriculture* thematic category

Fighters and patriotism

Le Petit écho du 21e Régiment d'infanterie
La Libre Comté : organe régional du Mouvement de la libération nationale
Franche-Comté Libération
Bulletin de la Société amicale des anciens combattants du 4e RAC
Recueil des instructions relatives à la défense nationale / Préfecture de la Côte-dOr
Le Mutilé : organe des mutilés, réformés et victimes de la guerre
Vaincre ! Organe de Saône-et-Loire du Front national de lutte pour la liberté et l'indépendance de la France
Jeunesse du Maquis
En avant ! : organe clandestin du Front national pour l'Ouest Côte d'Orient
Le Jeune patriote de Bourgogne / édité par les comités bourguignons du Front patriotique de la jeunesse
L'Aurore nouvelle de Saône-et-Loire : Journal (puis Organe) socialiste indépendant
La Revue de la Nièvre et du Centre
L'Yonne. Organe régional du rassemblement national
L'Yonne libre : organe régional du Front national
Le Libre-penseur de Saône-et-Loire : Organe et propriété de la Fédération départementale des sociétés de libre-pensée de Saône-et-Loire. Mensuel
La Voix de la Bourgogne. Organe de la Région communiste de Saône-et-Loire
Russie d'aujourd'hui. Organe mensuel des Amis de l'Union soviétique. Région de Saône-et-Loire
Le Courrier de la Nièvre : ancien Petit Nivernais : journal hebdomadaire, politique et littéraire, paraissant le dimanche
Le Combattant républicain : journal de la FNCR de l'Yonne
Le Patriote nivernais : organe départemental du Front national de lutte pour l'indépendance de la France
Le Patriote de Saône-et-Loire : journal des mouvements unis de résistance
La Bourgogne combattante / édité par le Front national de lutte pour la liberté et l'indépendance de la France
La Voix du mutilé de Mâcon et de la région : organe mensuel de l'Association des mutilés et veuves de guerre

Table 19.12: Document collections in the EMONTAL corpus belonging to the *Fighters and patriotism* thematic category

Generalist and partisan newspapers

Le Semeur. Organe régional du Parti communiste. (Région de Montbéliard)
La Franche-Comté libre. Organe régional du Front national pour la liberté et l'indépendance de la France. Doubs et territoire de Belfort
L'Avenir de lEst : organe du Front national de lutte pour la libération et l'indépendance de la France, de Meurthe-et-Moselle, Vosges, Haute-Marne, Haute-Saône, Doubs et du Territoire de Belfort
Le Démocrate : Organe des Républicains radicaux et radicaux-socialistes du pays de Montbéliard
Le Franc-comtois. Ed. par les régions du Doubs, Haute-Saône, Jura et Territoire de Belfort du Parti communiste français. Ed. spéciale
La Haute-Saône libre : organe départemental du Front national
Frontière : journal de marche du Groupement Frontière
Le Patriote comtois : organe des comités du Front national de Franche-Comté
Le Semeur : journal communiste : Territoire de Belfort, Doubs, Jura, Haute-Saône
Vive labeur : organe du Sillon de lEst : paraît le 10 du mois
Le Front comtois : organe du Front populaire de Dole et de la région (puis hebdomadaire du Front populaire)
Le Socialiste comtois : organe hebdomadaire de la Fédération du Doubs (Parti socialiste, Section française de l'Internationale ouvrière) (puis organe hebdomadaire des fédérations socialistes du Doubs, de la Haute-Saône et du Haut-Rhin)
L'Émancipateur : hebdomadaire communiste de la région nivernaise
Union populaire. (Chalon-sur-Saône)
Le Travailleur de lYonne : organe régional du Parti communiste français
Bulletin régional / PCF Nivernais, Berry, Touraine, Beauce

Table 19.13: Document collections in the EMONTAL corpus belonging to the *Generalist and partisan newspapers* thematic category

La Vie ouvrière
La Voix de la Bourgogne. Edition nouvelle éd. par la Région bourguignonne du Parti communiste S.F.I.C. ...
Le Travailleur : organe de la région Côte dOrienne du Parti communiste français
L'Étincelle : organe mensuel de Saône-et-Loire
Journal des ouvriers du Creusot
La Terre / édité par la Région de l'Yonne du Parti communiste français
Le Cri des terriens : Côte d'Or, Saône-et-Loire, Yonne, Nièvre, Ain : organe de défense paysanne et du bloc des viticulteurs et cultivateurs bourguignons
La Revue du Centre : Nivernais, Berry, Bourbonnais, Bourgogne, Orléanais (puis plus de sous-titre, puis revue mensuelle illustrée puis revue trimestrielle illustrée)
La Franche-Comté à Paris : journal d'informations des départements du Doubs, Jura, Haute-Saône, Haut-Rhin (puis organe des intérêts des Franc-Comtois à Paris, et d'informations des départements du Doubs, du Jura, de la Haute-Saône et du Territoire de Belfort)
Le Franc-Comtois de Paris : journal d'information des départements Doubs, Jura, Haute-Saône, Haut-Rhin
Supplément illustré du Petit comtois

Table 19.14: Document collections in the EMONTAL corpus belonging to the *Generalist and partisan newspapers* thematic category

Leisure

Les Sports comtois : organe sportif de Franche-Comté, paraissant le 1er de chaque mois et sintéressant à tous les sports
Maîtrise de la Métropole / Association des anciens professeurs et élèves
Les Sports de lEst : hebdomadaire illustré de tous les sports, Lorraine, Franche-Comté Bourgogne, Champagne
Le Collectionneur : bulletin mensuel de publicité, philatélie, cartophilie...
Bulletin du Photo-club sénonais : revue trimestrielle illustrée

Table 19.15: Document collections in the EMONTAL corpus belonging to the *Leisure* thematic category

Local powers and economy

Assemblée générale du..., rapport sur l'exercice... / Syndicat d'initiative d'Avalon et du Morvan...
Annuaire administratif, commercial et historique du Département de Saône-et-Loire...
Tableaux de l'économie bourguignonne
Circulaire économique et financière de Bourgogne et de Franche-Comté : publication mensuelle (ou publication bi-mensuelle)
Tableaux de l'économie franc-comtoise / Institut national de la statistique et des études économiques, Service régional de Franche-Comté
Circulaire économique et financière de Bourgogne et de Franche-Comté : publication mensuelle (ou publication bi-mensuelle)
Rapports et délibérations / Conseil général du Département de la Côte-d'Or
Rapports du Préfet, procès-verbaux des délibérations / Conseil général de la Nièvre
Rapports et délibérations / Conseil général, Saône-et-Loire
Rapports et délibérations / Département de l'Yonne, Conseil général
Rapports et délibérations / Conseil général du Doubs
Rapports et délibérations / Conseil général de la Haute-Saône
Délibérations du Conseil général du territoire de Belfort
Rapports et procès-verbaux des délibérations - Conseil général du Jura / Conseil général du Jura
Bulletin de la Société d'études du Brionnais
La Sardine : revue mensuelle : conserves alimentaires : sardines, maquereaux, thon, sprats, etc., salaisons / directeur Paul-Louis Tregan

Table 19.16: Document collections in the EMONTAL corpus belonging to the *Local powers and economy* thematic category

Religious

Le Petit écho de Sainte-Madeleine : Besançon
Bulletin paroissial de Vesoul : (écho de N. D. de la Motte)
Écho paroissial de Saint-Jean
L'Écho paroissial de Chenaux : paraissant le 1er de chaque mois
Bulletin paroissial de la région de Granges
Vers l'avenir : organe mensuel de la Jeunesse catholique de Franche-Comté et du Territoire-de-Belfort
Le Petit semeur : bulletin mensuel du Patronage central de Besançon
Bulletin paroissial de Baume-les-Dames : paraissant le 1er dimanche du mois
Bulletin paroissial de Saint-Vit
Bulletin paroissial de Belleherbe
Bulletin mensuel de la paroisse de Champlitte
Bulletin paroissial d'Arc-lès-Gray
Bulletin paroissial de Brassy
Bulletin paroissial de Notre-Dame, Dijon
Bulletin bi-mensuel de l'Ecole Saint-François de Sales de Dijon
Bulletin paroissial de Saint-Bénigne de Dijon
L'Écho de Saint-Jean : bulletin mensuel paroissial
Bulletin paroissial mensuel de Saint-Apollinaire
Bulletin paroissial de Saint-Vincent de Châlon-s-S. : paraissant tous les mois
Notre clocher : bulletin paroissial mensuel de Saint-Paul de Dijon
Bulletin paroissial de Saint-Pierre et du Sacré-Coeur, Châlon-sur-Saône : paraissant tous les mois
Bulletin paroissial de Tilchâtel et d'Échevannes : Diocèse de Dijon, Côte dOr
Annales du doyenné de Pontailler-sur-Saône : diocèse de Dijon (Côte-d'Or) : paraissant le premier dimanche de chaque mois (puis bulletin mensuel)
Bulletin paroissial de Saint-Michel
Écho paroissial de Cruzy-le-Châtel : diocèse de Sens, Yonne
Bulletin paroissial de Saint-Pierre de Dijon

Table 19.17: Document collections in the EMONTAL corpus belonging to the *Religious* thematic category

Science and culture

Mémoires de la Société bourguignonne de géographie et d'histoire
La Revue du Charolais : histoire régionale, lettres, sciences, actualités : Charolais, Brionnais et Bourbonnais
La Revue du Charolais : histoire régionale, lettres, sciences, actualités : Charolais, Brionnais et Bourbonnais
Mémoires et documents inédits pour servir à l'histoire de la Franche-Comté / publiés par l'Académie de Besançon
Mémoires et documents inédits pour servir à l'histoire de la Franche-Comté / publiés par l'Académie de Besançon
Bulletin d'histoire et d'archéologie religieuses du Diocèse de Dijon
Revue champenoise et bourguignonne : (Aube, Côte-d'Or, Haute-Marne, Marne, Seine-et-Marne et Yonne) : archéologie, beaux-arts, bibliographie, biographie, documents inédits, géologie, histoire : revue régionale d'histoire et d'érudition paraissant tous les deux mois / publiée sous le patronage de MM. Ernest Babelon et Auguste Longnon,... ; (Eugène Maury, secrétaire de la rédaction)
Bulletin de la Société archéologique et biographique du canton de Montbard
Mémoires de la Commission des antiquités du département de la Côte-d'Or
Mémoires / Société d'histoire, d'archéologie et de littérature de l'arrondissement de Beaune
Mémoires de la Société d'histoire et d'archéologie de Chalon-sur-Saône
Pro Alesia : revue mensuelle des fouilles d'Alise et des questions relatives à Alesia / publiée sous le patronage de la Société des sciences de Semur par M. Louis Matruchot,...
Mémoires de la Société pour l'histoire du droit et des institutions des anciens pays bourguignons, comtois et romands
Mémoires / Société d'archéologie de Beaune. Histoire, lettres, sciences et arts
Société des amis des arts, sciences, archéologie et histoire locale de la Bresse louhannaise
Bulletin de la Société d'archéologie et d'histoire de Tonnerre / président Léon Lacroix

Table 19.18: Document collections in the EMONTAL corpus belonging to the *Science and culture* thematic category

Les Annales chalonnaises et bourguignonnes : recueil bi-mensuel d'articles et de documents d'histoire locale publié par souscription / sous le patronage de la Société d'histoire et d'archéologie de Chalon-sur-Saône
Bulletin de la Société des sciences historiques et naturelles de l'Yonne
Bulletin de la Société des sciences historiques et naturelles de Semur (Côte-d'Or)
Société d'histoire naturelle du Creusot / gérant A. Dulac
Annales d'Igé en Mâconnais : recueil de documents et de matériaux pour servir à l'histoire de cette commune
Enquêtes sur la Révolution en Côte-d'Or / Comité départemental de la Côte-d'Or ; avec une note de E. Champeaux
Bulletin de la Société belfortaine d'émulation
Revue périodique de vulgarisation des sciences naturelles et préhistoriques de la Physiophile / Société d'études d'histoire naturelle de Montceau-les-Mines
Mémoires de la Société académique du Nivernais
Société des amis des arts et des sciences de Tournus
Bulletins de la Société des sciences naturelles de Saône-et-Loire
Bulletin... / Société d'histoire naturelle d'Autun
L'Eduen : feuille de liaison périodique de la Société éduenne des lettres, sciences et arts et de la Société d'histoire naturelle d'Autun
Alauda : études et notes ornithologiques / recueil publié par P. Paris,...
Bulletin de la Société archéologique de Sens
Bulletin trimestriel de la Société d'histoire naturelle et des amis du Muséum d'Autun
Bulletin / Association bourguignonne des sociétés savantes
Mémoires de la Société éduenne

Table 19.19: Document collections in the EMONTAL corpus belonging to the *Science and culture* thematic category

Annuaire : programme des cours et des conférences / Université de Besançon
La Vie meilleure : revue sociologique et littéraire de l'Est, paraissant le 1er et le 15 de chaque mois / (gérant : Alfred Jacquin)
La Vie meilleure : revue sociologique et littéraire de l'Est, paraissant le 1er et le 15 de chaque mois / (gérant : Alfred Jacquin)
Mémoires de l'Académie des sciences, arts et belles-lettres de Dijon
Mémoires de l'Académie des sciences, arts et belles-lettres de Dijon
Annales de l'Académie de Mâcon : société des arts, sciences, belles-lettres et d'agriculture
(Le Corps et l'esprit)
La Bourgogne littéraire et scientifique : revue régionaliste
Bulletin de la Société nivernaise des lettres, sciences et arts
LÉssor : revue mensuelle bourguignonne, artistique, littéraire ("puis" revue bourguignonne dárt et de littérature "puis" bulletin trimestriel des artistes et artisans dárt bourguignons)
Bulletin de la Société scientifique artistique de Clamecy

Table 19.20: Document collections in the EMONTAL corpus belonging to the *Science and culture* thematic category

Women in society

La Voix des femmes de la Côte-d'Or
La Voix des femmes de Saône-et-Loire
La Voix des femmes de Bourgogne
La Mère éducatrice : revue mensuelle déducation populaire
La France féminine : revue mensuelle pendant la guerre : organe de défense des droits de la femme française
La Franc-comtoise. Journal des Comités féminins de Franche-Comté. Ed. de Belfort et du Territoire

Table 19.21: Document collections in the EMONTAL corpus belonging to the *Women in society* thematic category

Annexe C: XSD schema of the EMONTAL XML format

Listing 19.1: XSD schema of the EMONTAL XML format

```
<xs:schema attributeFormDefault="unqualified" elementFormDefault="qualified"
    xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="ark" type="xs:string"/>
  <xs:element name="identifier" type="xs:string"/>
  <xs:element name="date" type="xs:float"/>
  <xs:element name="title">
    <xs:complexType>
      <xs:simpleContent>
        <xs:extension base="xs:string">
          <xs:attribute type="xs:string" name="id" use="optional"/>
        </xs:extension>
      </xs:simpleContent>
    </xs:complexType>
  </xs:element>
  <xs:element name="contributor" type="xs:string"/>
  <xs:element name="publisher" type="xs:string"/>
  <xs:element name="languague" type="xs:string"/>
  <xs:element name="creator" type="xs:string"/>
  <xs:element name="source" type="xs:string"/>
  <xs:element name="typedoc" type="xs:string"/>
  <xs:element name="nqamoyen" type="xs:float"/>
  <xs:element name="dewey" type="xs:string"/>
  <xs:element name="image_url" type="xs:string"/>
  <xs:element name="header">
    <xs:complexType>
      <xs:simpleContent>
        <xs:extension base="xs:string">
          <xs:attribute type="xs:string" name="block_id" use="optional"/>
          <xs:attribute type="xs:string" name="id" use="optional"/>
        </xs:extension>
```

```
      </xs:simpleContent>
    </xs:complexType>
  </xs:element>
  <xs:element name="articles">
    <xs:complexType mixed="true">
      <xs:choice maxOccurs="unbounded" minOccurs="0">
        <xs:element ref="article"/>
      </xs:choice>
    </xs:complexType>
  </xs:element>
  <xs:element name="ent">
    <xs:complexType>
      <xs:simpleContent>
        <xs:extension base="xs:string">
          <xs:attribute type="xs:string" name="id" use="optional"/>
          <xs:attribute type="xs:string" name="type" use="optional"/>
        </xs:extension>
      </xs:simpleContent>
    </xs:complexType>
  </xs:element>
  <xs:element name="sent">
    <xs:complexType mixed="true">
      <xs:sequence>
        <xs:element ref="ent" maxOccurs="unbounded" minOccurs="0"/>
      </xs:sequence>
      <xs:attribute type="xs:string" name="id" use="optional"/>
    </xs:complexType>
  </xs:element>
  <xs:element name="para">
    <xs:complexType mixed="true">
      <xs:sequence>
        <xs:element ref="sent" maxOccurs="unbounded" minOccurs="0"/>
      </xs:sequence>
      <xs:attribute type="xs:string" name="block_id" use="optional"/>
      <xs:attribute type="xs:string" name="id" use="optional"/>
    </xs:complexType>
  </xs:element>
  <xs:element name="other">
    <xs:complexType>
      <xs:simpleContent>
        <xs:extension base="xs:string">
          <xs:attribute type="xs:string" name="block_id" use="optional"/>
          <xs:attribute type="xs:string" name="id" use="optional"/>
        </xs:extension>
```

```
74        </xs:simpleContent>
75      </xs:complexType>
76    </xs:element>
77    <xs:element name="text">
78      <xs:complexType mixed="true">
79        <xs:sequence>
80          <xs:element ref="para" minOccurs="0"/>
81          <xs:element ref="other" minOccurs="0"/>
82        </xs:sequence>
83        <xs:attribute type="xs:string" name="id" use="optional"/>
84      </xs:complexType>
85    </xs:element>
86    <xs:element name="article">
87      <xs:complexType>
88        <xs:sequence>
89          <xs:element ref="title"/>
90          <xs:element ref="text"/>
91        </xs:sequence>
92        <xs:attribute type="xs:string" name="id" use="optional"/>
93      </xs:complexType>
94    </xs:element>
95    <xs:element name="page">
96      <xs:complexType>
97        <xs:sequence>
98          <xs:element ref="header"/>
99          <xs:element ref="articles"/>
100       </xs:sequence>
101       <xs:attribute type="xs:byte" name="id" use="optional"/>
102     </xs:complexType>
103   </xs:element>
104   <xs:element name="metadata">
105     <xs:complexType>
106       <xs:sequence>
107         <xs:element ref="ark"/>
108         <xs:element ref="identifier"/>
109         <xs:element ref="date"/>
110         <xs:element ref="title"/>
111         <xs:element ref="contributor"/>
112         <xs:element ref="publisher"/>
113         <xs:element ref="languague"/>
114         <xs:element ref="creator"/>
115         <xs:element ref="source"/>
116         <xs:element ref="typedoc"/>
117         <xs:element ref="nqamoyen"/>
```

```
118         <xs:element ref="dewey"/>
119         <xs:element ref="image_url"/>
120       </xs:sequence>
121     </xs:complexType>
122   </xs:element>
123   <xs:element name="content">
124     <xs:complexType>
125       <xs:sequence>
126         <xs:element ref="page" maxOccurs="unbounded" minOccurs="0"/>
127       </xs:sequence>
128     </xs:complexType>
129   </xs:element>
130   <xs:element name="document">
131     <xs:complexType>
132       <xs:sequence>
133         <xs:element ref="metadata"/>
134         <xs:element ref="content"/>
135       </xs:sequence>
136     </xs:complexType>
137   </xs:element>
138 </xs:schema>
```

Annexe D: Publications related to the Ph.D thesis

Publications related to the thesis

Gutehrlé, N. (2024a). Semantic search in archive collections through interpretable and adaptable relation extraction about person and places. In N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, & I. Ounis (Eds.), *Advances in information retrieval* (pp. 315–318). Springer Nature Switzerland.

Gutehrlé, N. (2024b). Comprendre les archives : explorer et valoriser les documents historiques grâce à l'annotation sémantique. *Printemps de la Donnée 2024*. https://doi.org/10.5281/zenodo.11609500

Gutehrlé, N., & Atanassova, I. (2021a, October). *Dataset for Logical-layout analysis on French historical newspapers* (Version 1.0). Zenodo. https://doi.org/10.5281/zenodo.5752440

Gutehrlé, N., & Atanassova, I. (2021b). Logical layout analysis applied to historical newspapers. *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, 85–94. https://aclanthology.org/2021.nlp4dh-1.10

Gutehrlé, N., & Atanassova, I. (2022). Processing the structure of documents: Logical Layout Analysis of historical newspapers in French. *Journal of Data Mining & Digital Humanities*, *NLP4DH*. https://doi.org/10.46298/jdmdh.9093

Gutehrlé, N., & Atanassova, I. (2023). Comprendre les archives : vers de nouvelles interfaces de recherche reposant sur l'annotation sémantique des documents Understanding Archives : Towards New Research Interfaces Relying on the Semantic Annotation of Documents. *CiDE.23 : Document et archivage : pratiques formelles et informelles*. https://hal.science/hal-04523110

Gutehrlé, N., Doucet, A., & Jatowt, A. (2022). Archive TimeLine summarization (ATLS): Conceptual framework for timeline generation over historical document collections. *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, So-*

cial Sciences, Humanities and Literature, 13–23. https://aclanthology.org/2022.latechclfl-1.3

Gutehrlé, N., Harlamov, O., Karimi, F., Wei, H., Jean-Caurant, A., & Pivovarova, L. (2021). Spacewars: A web interface for exploring the spatio-temporal dimensions of wwi newspaper reporting. *HistoInformatics 2021 – 6th International Workshop on Computational History*.

Participation to popular science events

- *Nuit des Chercheurs* - Besançon (2022, 2023)
- *Fête de la Science* - Besançon (2023)
- *OVNI* - Morteau, Arbois (2023)
- *Printemps de la donnée* - Online (2024)

Blog and news articles presenting the project

- *Where did it happen ?: Spatial imaginaries of World War I*, NewsEye Blog, 2021[4]
- *Donner un nouveau sens aux documents historiques*, journal *En Direct*, n° 312, 2024[5]

[4] https://www.newseye.eu/fr/le-blog/news/where-did-it-happen-spatial-imaginaries-of-world-war-i/

[5] https://flowpaper.com/online-pdf-viewer/?pdf=https://endirect.univ-fcomte.fr/content/uploads/2024/04/NÂř312-mai-juin-2024-web.pdf#page=4

List of Figures

List of Tables

Bibliography

Abujabal, A., & Berberich, K. (2015). Important events in the past, present, and future, 1315–1320. https://doi.org/10.1145/2740908.2741692

Afli, H., Barrault, L., & Schwenk, H. (2016). Ocr error correction using statistical machine translation. *Int. J. Comput. Linguistics Appl.*, *7*(1), 175–191.

Agerri, R., & Rigau, G. (2016). Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, *238*, 63–82.

Agichtein, E., & Gravano, L. (2000). Snowball: Extracting relations from large plain-text collections. *Proceedings of the Fifth ACM Conference on Digital Libraries*, 85–94. https://doi.org/10.1145/336597.336644

Aguilar, S. T., Tannier, X., & Chastang, P. (2016). Named entity recognition applied on a data base of medieval latin charters. the case of chartae burgundiae. *3rd International Workshop on Computational History (HistoInformatics 2016)*.

Ahnert, R., Griffin, E., Ridge, M., & Tolfo, G. (2023). *Collaborative historical research in the age of big data: Lessons from an interdisciplinary project*. Cambridge University Press.

Akbik, A., & Löser, A. (2012). KrakeN: N-ary facts in open information extraction. *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, 52–56. https://aclanthology.org/W12-3010

Akl, H. A., Gupta, A., & Mariko, D. (2019). FinTOC-2019 shared task: Finding title in text blocks. *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)*, 58–62. https://www.aclweb.org/anthology/W19-6408

Al Azawi, M., Liwicki, M., & Breuel, T. M. (2015). Combination of multiple aligned recognition outputs using wfst and lstm. *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 31–35.

Albertin, F., Astolfo, A., Stampanoni, M., Peccenini, E., Hwu, Y., Kaplan, F., & Margaritondo, G. (2015a). Ancient administrative handwritten documents: X-ray analysis and imaging. *Journal of synchrotron radiation*, *22*(2), 446–451.

Albertin, F., Astolfo, A., Stampanoni, M., Peccenini, E., Hwu, Y., Kaplan, F., & Margaritondo, G. (2015b). X-ray spectrometry and imaging for ancient administrative handwritten documents. *X-Ray Spectrometry*, *44*(3), 93–98.

Alex, B., Byrne, K., Grover, C., & Tobin, R. (2015). Adapting the edinburgh geoparser for historical georeferencing. *International Journal of Humanities and Arts Computing*, *9*(1), 15–35.

Alfonseca, E., Filippova, K., Delort, J.-Y., & Garrido, G. (2012). Pattern learning for relation extraction with a hierarchical topic model. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 54–59. https://aclanthology.org/P12-2011

Ali, M., Saleem, M., & Ngomo, A.-C. N. (2021). Unsupervised relation extraction using sentence encoding. *The Semantic Web: ESWC 2021 Satellite Events: Virtual Event, June 6–10, 2021, Revised Selected Papers 18*, 136–140.

Allan, J., Papka, R., & Lavrenko, V. (1998). On-line new event detection and tracking. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 37–45. https://doi.org/10.1145/290941.290954

Alonso, O., Marchesin, S., Najork, M., & Silvello, G. (Eds.). (2021). *Proceedings of the second international conference on design of experimental search & information retrieval systems, padova, italy, september 15-18, 2021* (Vol. 2950). CEUR-WS.org. http://ceur-ws.org/Vol-2950

Amrhein, C., & Clematide, S. (2018). Supervised ocr error detection and correction using statistical and neural machine translation methods. *Journal for Language Technology and Computational Linguistics (JLCL)*, *33*(1), 49–76.

Antonacopoulos, A., & Bridson, D. (2007). Performance analysis framework for layout analysis methods. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, *2*, 1258–1262.

Antonacopoulos, A., Bridson, D., Papadopoulos, C., & Pletschacher, S. (2009). A realistic dataset for performance evaluation of document layout analysis. *2009 10th International Conference on Document Analysis and Recognition*, 296–300.

Antonacopoulos, A., Clausner, C., Papadopoulos, C., & Pletschacher, S. (2011). Historical document layout analysis competition. *2011 International Conference on Document Analysis and Recognition*, 1516–1520.

Antonacopoulos, A., & Ritchings, R. (1995). Representation and classification of complex-shaped printed regions using white tiles. *Proceedings of 3rd international conference on document analysis and recognition*, *2*, 1132–1135.

Appelt, D. E., Hobbs, J. R., Bear, J., Israel, D., & Tyson, M. (1993). Fastus: A finite-state processor for information extraction from real-world text. *IJCAI*, *93*, 1172–1178.

Ardanuy, M. C., Hosseini, K., McDonough, K., Krause, A., van Strien, D., & Nanni, F. (2020). A deep learning approach to geographical candidate selection through toponym matching. *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, 385–388.

Ares Oliveira, S., di Lenardo, I., Tourenc, B., & Kaplan, F. (2019). A deep learning approach to cadastral computing. *Digital humanities conference*, (CONF).

Arnold, P., & Rahm, E. (2014). Extracting semantic concept relations from wikipedia. *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*. https://doi.org/10.1145/2611040.2611079

Asahara, M., & Matsumoto, Y. (2003). Japanese named entity extraction with redundant morphological analysis. *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, 8–15.

Atanassova, I. (2012). *Exploitation informatique des annotations sémantiques automatiques d'excom pour la recherche d'informations et la navigation* [Doctoral dissertation, Paris 4].

Augenstein, I., Das, M., Riedel, S., Vikraman, L., & McCallum, A. (2017). SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 546–555. https://doi.org/10.18653/v1/S17-2091

Bagnall, R., Talbert, R. J., Bond, S., Becker, J., Elliott, T., Gillies, S., Horne, R., McCormick, M., Rabinowitz, A., et al. (2016). Pleiades: A gazetteer of past places. *Database. Pleiades: A Gazetteer of Past Places. Institute for the Study of the Ancient World, New York University*.

Baldwin, T., De Marneffe, M.-C., Han, B., Kim, Y.-B., Ritter, A., & Xu, W. (2015). Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. *Proceedings of the workshop on noisy user-generated text*, 126–135.

Barlas, P., Adam, S., Chatelain, C., & Paquet, T. (2014). A typed and handwritten text block segmentation system for heterogeneous and complex documents. *2014 11th IAPR International Workshop on Document Analysis Systems*, 46–50.

Barman, R., Ehrmann, M., Clematide, S., Oliveira, S., & Kaplan, F. (2020). Combining visual and textual features for semantic segmentation of historical newspapers. *ArXiv*, *abs/2002.06144*.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*, 82–115. https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.012

Bassil, Y., & Alwani, M. (2012). Ocr post-processing error correction algorithm using google online spelling suggestion. *arXiv preprint arXiv:1204.0191*.

Bedi, H., Patil, S., Hingmire, S., & Palshikar, G. (2017). Event timeline generation from history textbooks. *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, 69–77. https://aclanthology.org/W17-5912

Beelen, K., Nanni, F., Coll Ardanuy, M., Hosseini, K., Tolfo, G., & McGillivray, B. (2021, August). When time makes sense: A historically-aware approach to targeted sense disambiguation. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the association for computational linguistics: Acl-ijcnlp 2021* (pp. 2751–2761). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.243

Bekoulis, G., Deleu, J., Demeester, T., & Develder, C. (2018). Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, *114*, 34–45. https://doi.org/10.1016/j.eswa.2018.07.032

Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. *Advances in neural information processing systems*, *13*.

Benikova, D., Biemann, C., & Reznicek, M. (2014). Nosta-d named entity annotation for german: Guidelines and dataset. *LREC*, 2524–2531.

Bernard, G., Suire, C., Faucher, C., & Doucet, A. (2021). A comprehensive extraction of relevant real-world-event qualifiers for semantic search engines. *Linking Theory and Practice of Digital Libraries: 25th International Conference on Theory and Practice of Digital Libraries, TPDL 2021, Virtual Event, September 13–17, 2021, Proceedings 25*, 153–164.

Bertin, M., Kauppinen, T., & Atanassova, I. (2015). Exploitation de données spatiales provenant d'articles scientifiques pour le suivi des maladies tropicales. *Gestion et Analyse des données Spatiales et Temporelles (GAST'2015), 15ème conférence internationale sur l'extraction et la gestion des connaissances (EGC-2015)*, 21–32.

Bhutani, N., Jagadish, H. V., & Radev, D. (2016). Nested propositions in open information extraction. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 55–64. https://doi.org/10.18653/v1/D16-1006

Biemann, C. (2006). Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems. *Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing*, 73–80. https://aclanthology.org/W06-3812

Bikel, D. M., Miller, S., Schwartz, R., & Weischedel, R. (1998). Nymble: A high-performance learning name-finder. *arXiv preprint cmp-lg/9803003*.

Binmakhashen, G. M., & Mahmoud, S. A. (2019). Document layout analysis: A comprehensive survey. *ACM Computing Surveys (CSUR)*, *52*(6), 1–36.

Blevins, C. (2014). Space, nation, and the triumph of region: A view of the world from houston. *The Journal of American History*, *101*(1), 122–147.

Bodenhamer, D. J., Corrigan, J., & Harris, T. M. (Eds.). (2010). *The Spatial Humanities: GIS and the Future of Humanities Scholarship*. Indiana University Press.

Bogdanov, S., Constantin, A., Bernard, T., Crabbé, B., & Bernard, E. (2024). Nuner: Entity recognition encoder pre-training via llm-annotated data. https://arxiv.org/abs/2402.15343

Borin, L., Kokkinakis, D., & Olsson, L.-J. (2007). Naming the past: Named entity and animacy recognition in 19th century swedish literature. *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007).*, 1–8.

Boros, E., Hamdi, A., Linhares Pontes, E., Cabrera-Diego, L. A., Moreno, J. G., Sidere, N., & Doucet, A. (2020, November). Alleviating digitization errors in named entity recognition for historical documents. In R. Fernández & T. Linzen (Eds.), *Proceedings of the 24th conference on computational natural language learning* (pp. 431–441). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.conll-1.35

Boros, E., Moreno, J. G., & Doucet, A. (2021). Event detection with entity markers. *European Conference on Information Retrieval*, 233–240.

Boros, E., Moreno, J. G., & Doucet, A. (2022). Exploring entities in event detection as question answering. *European Conference on Information Retrieval*, 65–79.

Boros, E., Nguyen, N. K., Lejeune, G., & Doucet, A. (2022). Assessing the impact of ocr noise on multilingual event detection over digitised documents. *International Journal on Digital Libraries*, *23*(3), 241–266.

Boros, E., Pontes, E. L., Cabrera-Diego, L. A., Hamdi, A., Moreno, J. G., Sidère, N., & Doucet, A. (2020). Robust named entity recognition and linking on historical multilingual documents. *Conference and Labs of the Evaluation Forum (CLEF 2020)*, *2696*(Paper 171), 1–17.

Boroş, E., Hamdi, A., Pontes, E. L., Cabrera-Diego, L.-A., Moreno, J. G., Sidere, N., & Doucet, A. (2020). Alleviating digitization errors in named entity recognition for historical documents. *Proceedings of the 24th Conference on Computational Natural Language Learning*, 431–441.

Borovikov, E., Zavorin, I., & Turner, M. (2004). A filter based post-ocr accuracy boost system. *Proceedings of the 1st ACM workshop on Hardcopy document processing*, 23–28.

Borthwick, A., Sterling, J., Agichtein, E., & Grishman, R. (1998). Description of the mene named entity system as used in muc-7. *Proceedings of the Seventh Message Understanding Conference (MUC-7), Fairfax, Virginia, April 29-May 1, 1998*.

Bossy, R., Golik, W., Ratkovic, Z., Bessieres, P., & Nédellec, C. (2013). Bionlp shared task 2013–an overview of the bacteria biotope task. *Proceedings of the BioNLP shared task 2013 workshop*, 161–169.

Brin, S. (1998). Extracting patterns and relations from the world wide web. *International Workshop on the Web and Databases*.

Broux, Y., & Depauw, M. (2015). Developing Onomastic Gazetteers and Prosopographies for the Ancient World Through Named Entity Recognition and Graph Visualization: Some Examples from Trismegistos People. In L. M. Aiello & D. McFarland (Eds.), *Social Informatics* (pp. 304–313). Springer International Publishing.

Bukhari, S. S., Shafait, F., & Breuel, T. M. (2011). Improved document image segmentation algorithm using multiresolution morphology. *Document recognition and retrieval XVIII*, *7874*, 109–116.

Bunescu, R., & Mooney, R. (2005). A shortest path dependency kernel for relation extraction. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 724–731. https://aclanthology.org/H05-1091

Bunescu, R. C., & Mooney, R. J. (2005). Subsequence kernels for relation extraction. *Proceedings of the 18th International Conference on Neural Information Processing Systems*, 171–178.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2018). Yake! collection-independent automatic keyword extractor. In G. Pasi, B. Piwowarski, L. Azzopardi, & A. Hanbury (Eds.), *Advances in information retrieval* (pp. 806–810). Springer International Publishing.

Ceroni, A., Tran, N. K., Kanhabua, N., & Niederée, C. (2014). Bridging temporal context gaps using time-aware re-contextualization. *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1127–1130. https://doi.org/10.1145/2600428.2609526

Chan, Y. S., & Roth, D. (2011). Exploiting syntactico-semantic structures for relation extraction. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 551–560. https://aclanthology.org/P11-1056

Chang, A. X., & Manning, C. (2012). SUTime: A library for recognizing and normalizing time expressions. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3735–3740. http://www.lrec-conf.org/proceedings/lrec2012/pdf/284_Paper.pdf

Chasin, R. (2010). Event and temporal information extraction towards timelines of wikipedia articles.

Chieu, H. L., & Lee, Y. K. (2004). Query based event extraction along a timeline. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 425–432. https://doi.org/10.1145/1008992.1009065

Chiron, G., Doucet, A., Coustaty, M., & Moreux, J.-P. (2017). ICDAR2017 Competition on Post-OCR Text Correction. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 1423–1428. https://doi.org/10.1109/icdar.2017.232

Chiu, J. P., & Nichols, E. (2016). Named entity recognition with bidirectional lstm-cnns. *Transactions of the association for computational linguistics*, *4*, 357–370.

Christopoulou, F., Miwa, M., & Ananiadou, S. (2018). A walk-based model on entity graphs for relation extraction. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 81–88. https://doi.org/10.18653/v1/P18-2014

Chrons, O., & Sundell, S. (2011). Digitalkoot: Making old archives accessible using crowdsourcing. *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Clausner, C., Papadopoulos, C., Pletschacher, S., & Antonacopoulos, A. (2015). The enp image and ground truth dataset of historical newspapers. *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 931–935. https://doi.org/10.1109/ICDAR.2015.7333898

Clematide, S., Furrer, L., & Volk, M. (2016). Crowdsourcing an ocr gold standard for a german and french heritage corpus. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 975–982.

Cohen, W. W. (1995). Repeated incremental pruning to produce error reduction. *Machine Learning Proceedings of the Twelfth International Conference ML95*.

Colavizza, G., Ehrmann, M., & Bortoluzzi, F. (2019). Index-driven digitization and indexation of historical archives. *Frontiers in Digital Humanities*, *6*, 4.

Colavizza, G., Romanello, M., & Kaplan, F. (2018). The references of references: A method to enrich humanities library catalogs with citation data. *International Journal on Digital Libraries*, *19*(2-3), 151–161.

Collier, N., & Kim, J.-D. (2004). Introduction to the bio-entity recognition task at jnlpba. *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, 73–78.

Collins, M., & Duffy, N. (2001). Convolution kernels for natural language. *Advances in neural information processing systems*, *14*.

Collins, M., & Singer, Y. (1999). Unsupervised models for named entity classification. *1999 Joint SIGDAT conference on empirical methods in natural language processing and very large corpora*.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, *12*(ARTICLE), 2493–2537.

Cooper, D., Donaldson, C., & Murrieta-Flores, P. (Eds.). (2016). *Literary mapping in the digital age* (First published). Routledge, Taylor & Francis Group.

Cordell, R., & Smith, D. (2017). Viral texts: Mapping networks of reprinting in 19th-century newspapers and magazines.

Cornu, G. (1983). *Atlas du monde arabo-islamique à l'époque classique: Ix.-x. siècles*. Brill.

Councill, I. G., Giles, C. L., & Kan, M.-Y. (2008). Parscit: An open-source crf reference string parsing package. *LREC*, *8*, 661–667.

Crane, G., & Jones, A. (2006). The challenge of virginia banks: An evaluation of named entity analysis in a 19th-century newspaper collection. *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, 31–40.

Cui, L., Wei, F., & Zhou, M. (2018). Neural open information extraction.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, *7*(3), 171–176.

Davoudi, S., Crawford, J., Raynor, R., Reid, B., Sykes, O., & Shaw, D. (2018). Spatial imaginaries: Tyrannies or transformations? *Town Planning Review*.

De Paiva, V., Oliveira, D. A. B., Higuchi, S., Rademaker, A., & De Melo, G. (2014). Exploratory information extraction from a historical dictionary. *2014 IEEE 10th International Conference on e-Science*, *2*, 11–18. https://doi.org/10.1109/eScience.2014.50

De Toni, F., Akiki, C., De La Rosa, J., Fourrier, C., Manjavacas, E., Schweter, S., & Van Strien, D. (2022). Entities, dates, and languages: Zero-shot on historical texts with t0. *arXiv preprint arXiv:2204.05211*.

Del Corro, L., & Gemulla, R. (2013). Clausie: Clause-based open information extraction. *Proceedings of the 22nd international conference on World Wide Web*, 355–366.

Deléger, L., Bossy, R., Chaix, E., Ba, M., Ferré, A., Bessieres, P., & Nédellec, C. (2016). Overview of the bacteria biotope task at bionlp shared task 2016. *Proceedings of the 4th BioNLP shared task workshop*, 12–22.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

D'hondt, E., Grouin, C., & Grau, B. (2016). Low-resource ocr error detection and correction in french clinical texts. *Proceedings of the seventh international workshop on health text mining and information analysis*, 61–68.

di Lenardo, I. (2022). The replica project: Co-designing a discovery engine for digital art history. *Multimodal Technologies and Interaction*, *6*(11), 100.

di Lenardo, I., Seguin, B. L. A., & Kaplan, F. (2016). *Visual patterns discovery in large databases of paintings* (tech. rep.).

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., & Weischedel, R. (2004). The automatic content extraction (ACE) program – tasks, data, and evaluation. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf

Dominguès, C., Jolivet, L., Brando, C., & Cargill, M. (2019). Place and Sentiment-based Life story Analysis. *Revue française des sciences de l'information et de la communication*, (17). https://doi.org/10.4000/rfsic.7228

Doucet, A., Gasteiner, M., Granroth-Wilding, M., Kaiser, M., Kaukonen, M., Labahn, R., Moreux, J.-P., Muehlberger, G., Pfanzelter, E., Therenty, M.-E., et al. (2020). Newseye: A digital investigator for historical newspapers. *15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020*.

Duan, Y., Jatowt, A., & Tanaka, K. (2017). Discovering typical histories of entities by multi-timeline summarization. *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, 105–114. https://doi.org/10.1145/3078714.3078725

Duan, Y., Jatowt, A., & Tanaka, K. (2019). History-driven entity categorization. In J. Shao, M. L. Yiu, M. Toyoda, D. Zhang, W. Wang, & B. Cui (Eds.), *Web and big data* (pp. 349–364). Springer International Publishing.

Duan, Y., Jatowt, A., & Yoshikawa, M. (2020). Comparative timeline summarization via dynamic affinity-preserving random walk. *ECAI*.

Dumais, S. (2001). Improved String Matching Under Noisy Channel Conditions (Proceedings of CIKM 01). *Proceedings of CIKM 01*, 357–364. https://www.microsoft.com/en-us/research/publication/improved-string-matching-under-noisy-channel-conditions/

Duong, Q., Hämäläinen, M., & Hengchen, S. (2020). An unsupervised method for ocr post-correction and spelling normalisation for finnish. *arXiv preprint arXiv:2011.03502*.

Düring, M., Romanello, M., Ehrmann, M., Beelen, K., Guido, D., Deseure, B., Bunout, E., Keck, J., & Apostolopoulos, P. (2023). Impresso text reuse at scale. an interface for the exploration of text reuse data in semantically enriched historical newspapers. *Frontiers in big Data*, *6*.

Eberts, M., & Ulges, A. (2019). Span-based joint entity and relation extraction with transformer pre-training. *arXiv preprint arXiv:1909.07755*.

Efremova, J., García, A. M., Zhang, J., & Calders, T. (2015). Towards population reconstruction: Extraction of family relationships from historical documents. *First International Workshop on Population Informatics for Big Data*, 1–9.

Ehrmann, M. (2008). *Les entitées nommées, de la linguistique au tal: Statut théorique et méthodes de désambiguısation* [Doctoral dissertation, Paris Diderot University].

Ehrmann, M., Colavizza, G., Rochat, Y., & Kaplan, F. (2016). Diachronic evaluation of ner systems on old newspapers. *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, 97–107.

Ehrmann, M., Colavizza, G., Topalov, O., Cella, R., Drago, D., Erboso, A., Zugno, F., Bellavitis, A., Sapienza, V., & Kaplan, F. (2016). *From documents to structured data: First milestones of the garzoni project* (tech. rep.).

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2021a). Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*.

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2021b). Named entity recognition and classification on historical documents: A survey. https://doi.org/10.48550/ARXIV.2109.11406

Ehrmann, M., Romanello, M., Clematide, S., Ströbel, P., & Barman, R. (2020). Language resources for historical newspapers: The impresso collection.

Ehrmann, M., Romanello, M., Flückiger, A., & Clematide, S. (2020). Extended overview of clef hipe 2020: Named entity processing on historical newspapers. *CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum*, *2696*(CONF).

Ehrmann, M., Romanello, M., Najem-Meyer, S., Doucet, A., Clematide, S., Faggioli, G., Ferro, N., Hanbury, A., & Potthast, M. (2022). Extended overview of hipe-2022: Named entity recognition and linking in multilingual historical documents. *CEUR Workshop Proceedings*, (3180), 1038–1063.

Ehrmann, Watter, Romanello, Clematide, & Flückiger. (2020, January). *Impresso named entity annotation guidelines*. https://doi.org/10.5281/zenodo.3604227

Eide, Ø. (2015). *Media Boundaries and Conceptual Modelling: Between Texts and Maps*. Palgrave Macmillan UK. https://doi.org/10.1057/9781137544582

El Khatib, R., & Currell, D. (2018). Mapping the moralized geography of paradise lost. *Digital Milton*, 129–152.

Elsahar, H., Demidova, E., Gottschalk, S., Gravier, C., & Laforest, F. (2017). Unsupervised open relation extraction. In *Lecture notes in computer science* (pp. 12–16). Springer International Publishing. https://doi.org/10.1007/978-3-319-70407-4_3

Erdmann, A., Brown, C., Joseph, B., Janse, M., Ajaka, P., Elsner, M., & de Marneffe, M.-C. (2016). Challenges and solutions for latin named entity recognition. *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 85–93.

Estrella, P., & Paliza, P. (2014). Ocr correction of documents generated during argentina's national reorganization process. *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 119–123.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., & Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, *165*(1), 91–134.

Evershed, J., & Fitch, K. (2014). Correcting noisy ocr: Context beats confusion. *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 45–51.

Exchanges, O. (2017). Tracing global information networks in historical newspaper repositories, 1840-1914. *Ed. by Oceanic Exchanges Project Team. Boston, MA*.

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. *Proceedings of the 2011 conference on empirical methods in natural language processing*, 1535–1545.

Farmakiotou, D., Karkaletsis, V., Koutsias, J., Sigletos, G., Spyropoulos, C. D., & Stamatopoulos, P. (2000). Rule-based named entity recognition for greek financial texts. *Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000)*, 75–78.

Feng, W., He, W., Yin, F., Zhang, X.-Y., & Liu, C.-L. (2019). Textdragon: An end-to-end framework for arbitrary shaped text spotting. *Proceedings of the IEEE/CVF international conference on computer vision*, 9076–9085.

Filippova, K., & Strube, M. (2008). Dependency tree based sentence compression. *INLG*.

Fischer, A., Frinken, V., Fornés, A., & Bunke, H. (2011). Transcription alignment of latin manuscripts using hidden markov models. *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, 29–36.

Fonseca Cacho, J. R., & Taghva, K. (2020). Ocr post processing using support vector machines. *Intelligent Computing: Proceedings of the 2020 Computing Conference, Volume 2*, 694–713.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*(5814), 972–976. https://doi.org/10.1126/science.1136800

Fu, T.-J., Li, P.-H., & Ma, W.-Y. (2019, July). GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1409–1418). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1136

Gábor, K., Buscaldi, D., Schumann, A.-K., QasemiZadeh, B., Zargayouna, H., & Charnois, T. (2018). SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. *Proceedings of the 12th International Workshop on Semantic Evaluation*, 679–688. https://doi.org/10.18653/v1/S18-1111

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. *International Joint Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:5291693

Gao, L., Choubey, P. K., & Huang, R. (2019). Modeling document-level causal structures for event causal relation identification. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1808–1817.

Gatos, B., Mantzaris, S., Chandrinos, K., Tsigris, A., & Perantonis, S. J. (1999). Integrated algorithms for newspaper page decomposition and article tracking. *Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR'99 (Cat. No. PR00318)*, 559–562.

Ge, T., Pei, W., Ji, H., Li, S., Chang, B., & Sui, Z. (2015). Bring you to the past: Automatic generation of topically relevant event chronicles. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 575–585. https://doi.org/10.3115/v1/P15-1056

Généreux, M., & Spano, D. (2015). Nlp challenges in dealing with ocr-ed documents of derogated quality. *Workshop on Replicability and Reproducibility in Natural Language Processing: adaptive methods, resources and software at IJCAI*, 25–27.

Gholipour Ghalandari, D., & Ifrim, G. (2020). Examining the state-of-the-art in news timeline summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1322–1334. https://doi.org/10.18653/v1/2020.acl-main.122

Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., & Yuret, D. (2007). SemEval-2007 task 04: Classification of semantic relations between nominals. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 13–18. https://aclanthology.org/S07-1003

Göhring, A., & Volk, M. (2011). The text+ berg corpus: An alpine french-german parallel resource.

Goldstein-Stewart, J., & Carbonell, J. G. (1998). Summarization: (1) using MMR for Diversity-Based Reranking and (2) Evaluating Summaries. *TIPSTER*.

Govind & Spaniol, M. (2017). Elevate: A framework for entity-level event diffusion prediction into foreign language communities. *Proceedings of the 2017 ACM on Web Science Conference*.

Grishman, R., & Sundheim, B. M. (1996). Message understanding conference-6: A brief history. *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Grover, C., Givon, S., Tobin, R., & Ball, J. (2008). Named entity recognition for digitised historical texts. *LREC*.

Gutehrlé, N. (2024a). Semantic search in archive collections through interpretable and adaptable relation extraction about person and places. In N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, & I. Ounis (Eds.), *Advances in information retrieval* (pp. 315–318). Springer Nature Switzerland.

Gutehrlé, N. (2024b). Comprendre les archives : explorer et valoriser les documents historiques grâce à l'annotation sémantique. *Printemps de la Donnée 2024*. https://doi.org/10.5281/zenodo.11609500

Gutehrlé, N., & Atanassova, I. (2021a). Dataset for logical-layout analysis on french historical newspapers.

Gutehrlé, N., & Atanassova, I. (2021b). Logical layout analysis applied to historical newspapers. *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, 85–94. https://aclanthology.org/2021.nlp4dh-1.10

Gutehrlé, N., & Atanassova, I. (2022). Processing the structure of documents: Logical layout analysis of historical newspapers in french. *Journal of Data Mining & Digital Humanities*.

Gutehrlé, N., & Atanassova, I. (2023). Comprendre les archives : vers de nouvelles interfaces de recherche reposant sur l'annotation sémantique des documents Understanding Archives : Towards New Research Interfaces Relying on the Semantic Annotation of Documents. *CiDE.23 : Document et archivage : pratiques formelles et informelles*. https://hal.science/hal-04523110

Gutehrlé, N., Doucet, A., & Jatowt, A. (2022). Archive timeline summarization (atls): Conceptual framework for timeline generation over historical document collections. *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 13–23.

Gutehrlé, N., Harlamov, O., Karimi, F., Wei, H., Jean-Caurant, A., & Pivovarova, L. (2021). Spacewars: A web interface for exploring the spatio-temporal dimensions of wwi newspaper reporting. *CEUR Workshop Proceedings*.

Hakala, K., & Pyysalo, S. (2019). Biomedical named entity recognition with multilingual bert. *Proceedings of the 5th workshop on BioNLP open shared tasks*, 56–61.

Hamdi, A., Linhares Pontes, E., Boros, E., Nguyen, T. T. H., Hackl, G., Moreno, J. G., & Doucet, A. (2021). A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2328–2334.

Hammarström, H., Virk, S. M., & Forsberg, M. (2017). Poor man's ocr post-correction: Unsupervised recognition of variant spelling applied to a multilingual document collection. *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, 71–75.

Han, X., Gao, T., Lin, Y., Peng, H., Yang, Y., Xiao, C., Liu, Z., Li, P., Sun, M., & Zhou, J. (2020). More data, more relations, more context and more openness: A review and outlook for relation extraction.

Han, X., Zhu, H., Yu, P., Wang, Z., Yao, Y., Liu, Z., & Sun, M. (2018a). Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. https://doi.org/10.48550/ARXIV.1810.10147

Han, X., Zhu, H., Yu, P., Wang, Z., Yao, Y., Liu, Z., & Sun, M. (2018b). FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4803–4809. https://doi.org/10.18653/v1/D18-1514

Hasegawa, T., Sekine, S., & Grishman, R. (2004). Discovering relations among named entities from large corpora. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 415–422. https://doi.org/10.3115/1218955.1219008

He, H., & Sun, X. (2017). A unified model for cross-domain and semi-supervised named entity recognition in chinese social media. *Proceedings of the AAAI conference on artificial intelligence*, *31*(1).

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*. https://aclanthology.org/C92-2082

Hébert, D., Palfray, T., Nicolas, S., Tranouez, P., & Paquet, T. (2014). Automatic article extraction in old newspapers digitized collections. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/2595188.2595195

Hedderich, M. A., Lange, L., & Klakow, D. (2021). Anea: Distant supervision for low-resource named entity recognition. *arXiv preprint arXiv:2102.13129*.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., & Szpakowicz, S. (2010). SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. *Proceedings of the 5th International Workshop on Semantic Evaluation*, 33–38. https://aclanthology.org/S10-1006

Hládek, D., Staš, J., Ondáš, S., Juhár, J., & Kovács, L. (2017). Learning string distance with smoothing for ocr spelling correction. *Multimedia Tools and Applications*, *76*, 24549–24567.

Holley, R. (2008). THE AUSTRALIAN NEWSPAPERS DIGITISATION PROGRAM: Helping communities access and explore their newspaper heritage.

Hosseini, K., Wilson, D. C., Beelen, K., & McDonough, K. (2022). Mapreader: A computer vision pipeline for the semantic exploration of maps at scale. *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Geospatial Humanities*, 8–19.

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Hubková, H. (2019). Named-entity recognition in czech historical texts: Using a cnn-bilstm neural network model.

Huiyu Sun, R. G. (2022). Lexicalized dependency paths based supervised learning for relation extraction. *Computer Systems Science and Engineering*, *43*(3), 861–870. https://doi.org/10.32604/csse.2022.030759

Huynh, V.-N., Hamdi, A., & Doucet, A. (2020, November). When to Use OCR Post-correction for Named Entity Recognition? In *22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020* (pp. 33–42). https://doi.org/10.1007/978-3-030-64452-9_3

Ide, N., & Woolner, D. (2004). Exploiting semantic web technologies for intelligent access to historical documents. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. http://www.lrec-conf.org/proceedings/lrec2004/pdf/248.pdf

Jatowt, A. (2021). Timeline as information retrieval and ranking unit in news search. *DESIRES*.

Jatowt, A., & Au Yeung, C.-m. (2011). Extracting collective expectations about the future from large text collections. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, 1259–1264. https://doi.org/10.1145/2063576.2063759

Jatowt, A., Kawai, D., & Tanaka, K. (2016). Predicting importance of historical persons using wikipedia. *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, 1909–1912.

Jatowt, A., Yeung, C. A., & Tanaka, K. (2015). Generic method for detecting focus time of documents. *Inf. Process. Manag.*, *51*(6), 851–868.

Jean-Caurant, A., Tamani, N., Courboulay, V., & Burie, J.-C. (2017). Lexicographical-based order for post-ocr correction of named entities. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, *1*, 1192–1197.

Ji, B., Liu, R., Li, S., Yu, J., Wu, Q., Tan, Y., & Wu, J. (2019). A hybrid approach for named entity recognition in chinese electronic medical record. *BMC medical informatics and decision making*, *19*(2), 149–158.

Jin, R., Zhai, C., & Hauptmann, A. (2003). Information retrieval for ocr documents: A content-based probabilistic correction model [Document Recognition and Retrieval X ; Conference date: 22-01-2003 Through 24-01-2003]. *Proceedings of SPIE - The International Society for Optical Engineering*, *5010*, 128–135. https://doi.org/10.1117/12.472838

Jung, S. J., Kim, H., & Jang, K. S. (2024). Llm based biological named entity recognition from scientific literature. *2024 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 433–435.

Jurafsky, D., & Martin, J. (2008, February). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (Vol. 2).

Kahle, P., Colutto, S., Hackl, G., & Mühlberger, G. (2017). Transkribus-a service platform for transcription, recognition and retrieval of historical documents. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, *4*, 19–24.

Kambhatla, N. (2004). Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 178–181. https://aclanthology.org/P04-3022

Kantor, P., & Voorhees, E. (2000). The trec-5 confusion track: Comparing retrieval methods for scanned text. *Information Retrieval*, *2*, 165–176. https://doi.org/10.1023/A:1009902609570

Kaplan, F. (2015). The venice time machine. *Proceedings of the 2015 ACM Symposium on Document Engineering*, 73–73.

Kaplan, F. (2020). Big data of the past, from venice to europe. *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, 1–1.

Kaplan, F., & di Lenardo, I. (2017). Big data of the past. *Frontiers in Digital Humanities*, *4*, 12. https://doi.org/10.3389/fdigh.2017.00012

Kassner, N., Dufter, P., & Schütze, H. (2021). Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 3250–3258. https://doi.org/10.18653/v1/2021.eacl-main.284

Kate, R. J., & Mooney, R. (2010, July). Joint entity and relation extraction using card-pyramid parsing. In M. Lapata & A. Sarkar (Eds.), *Proceedings of the fourteenth conference on computational natural language learning* (pp. 203–212). Association for Computational Linguistics. https://aclanthology.org/W10-2924

Kessler, R., Tannier, X., Hagège, C., Moriceau, V., & Bittar, A. (2012). Finding salient dates for building thematic timelines. *ACL*.

Kew, T., Shaitarova, A., Meraner, I., Goldzycher, J., Clematide, S., & Volk, M. (2019). Geotagging a diachronic corpus of alpine texts: Comparing distinct approaches to toponym recognition. *Proceedings of the Workshop on Language Technology for Digital Historical Archives*, 11–18.

Khirbat, G. (2017). Ocr post-processing text correction using simulated annealing (opteca). *Proceedings of the Australasian Language Technology Association Workshop 2017*, 119–123.

Kiessling, B., Tissot, R., Stokes, P., & Ezra, D. S. B. (2019). Escriptorium: An open source platform for historical document analysis. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, *2*, 19–19.

Kim, J.-H., & Woodland, P. C. (2000). A rule-based named entity recognition system for speech input. *Sixth International Conference on Spoken Language Processing*.

Kim, Y., Jernite, Y., Sontag, D., & Rush, A. (2016). Character-aware neural language models. *Proceedings of the AAAI conference on artificial intelligence*, *30*(1).

Klampfl, S., & Kern, R. (2013). An unsupervised machine learning approach to body text and table of contents extraction from digital scientific articles. *TPDL*.

Klein, E., Alex, B., Grover, C., Tobin, R., Coates, C., Clifford, J., Quigley, A., Hinrichs, U., Reid, J., Osborne, N., et al. (2014). Digging into data white paper: Trading consequences.

Klein, G., Kim, Y., Deng, Y., Nguyen, V., Senellart, J., & Rush, A. M. (2018). Opennmt: Neural machine translation toolkit.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, 177–180.

Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D. M., et al. (2015). The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, *7*(1), 1–17.

Kristanti, T., & Romary, L. (2020). Delft and entity-fishing: Tools for clef hipe 2020 shared task. *CLEF 2020-Conference and Labs of the Evaluation Forum*, *2696*.

Kuru, O., Can, O. A., & Yuret, D. (2016). Charner: Character-level named entity recognition. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 911–921.

La Quatra, M., Cagliero, L., Baralis, E., Messina, A., & Montagnuolo, M. (2021). Summarize dates first: A paradigm shift in timeline summarization. In *Proceedings of the 44th international acm sigir conference on research and development in information retrieval* (pp. 418–427). Association for Computing Machinery. https://doi.org/10.1145/3404835.3462954

Labusch, K., Kulturbesitz, P., Neudecker, C., & Zellhöfer, D. (2019). Bert for named entity recognition in contemporary and historical german. *Proceedings of the 15th conference on natural language processing*, 9–11.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Lavrenko, V., Rath, T. M., & Manmatha, R. (2004). Holistic word recognition for handwritten historical documents. *First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings.*, 278–287.

Leppänen, L., & Toivonen, H. (2021). A baseline document planning method for automated journalism. *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 101–111.

Li, H., Wang, P., & Shen, C. (2017). Towards end-to-end text spotting with convolutional recurrent neural networks. *Proceedings of the IEEE international conference on computer vision*, 5238–5246.

Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 50–70.

Li, Q., & Ji, H. (2014, June). Incremental joint extraction of entity mentions and relations. In K. Toutanova & H. Wu (Eds.), *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 402–412). Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-1038

Li, X., Chen, K., Long, Y., & Zhang, M. (2024). Llm with relation classifier for document-level relation extraction. https://arxiv.org/abs/2408.13889

Liang, C., Yu, Y., Jiang, H., Er, S., Wang, R., Zhao, T., & Zhang, C. (2020). Bond: Bert-assisted open-domain named entity recognition with distant supervision. *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 1054–1064.

Liao, W., & Veeramachaneni, S. (2009). A simple semi-supervised algorithm for named entity recognition. *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, 58–65.

Limsopatham, N., & Collier, N. H. (2016). Bidirectional lstm for named entity recognition in twitter messages.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 74–81. https://aclanthology.org/W04-1013

Lin, D., & Pantel, P. (2001). Dirt – discovery of inference rules from text.

Linhares Pontes, E., Cabrera-Diego, L. A., Moreno, J. G., Boros, E., Hamdi, A., Doucet, A., Sidere, N., & Coustaty, M. (2022). Melhissa: A multilingual entity linking architecture for historical press articles. *International Journal on Digital Libraries*, 1–28.

Linhares Pontes, E., Hamdi, A., Sidere, N., & Doucet, A. (2019a). Impact of ocr quality on named entity linking. *Digital Libraries at the Crossroads of Digital Information for the Future: 21st International Conference on Asia-Pacific Digital Libraries, ICADL 2019, Kuala Lumpur, Malaysia, November 4–7, 2019, Proceedings 21*, 102–115.

Linhares Pontes, E., Hamdi, A., Sidere, N., & Doucet, A. (2019b). Impact of ocr quality on named entity linking. *Digital Libraries at the Crossroads of Digital Information for the Future*, 102–115. https://doi.org/10.5281/zenodo.3529180

Liu, Y. (2019). Fine-tune BERT for extractive summarization. https://doi.org/10.48550/ARXIV.1903.10318

Llobet, R., Cerdan-Navarro, J.-R., Perez-Cortes, J.-C., & Arlandis, J. (2010). Ocr post-processing using weighted finite-state transducers. *2010 20th International Conference on Pattern Recognition*, 2021–2024.

Lopez de Lacalle, O., & Lapata, M. (2013). Unsupervised relation extraction with general domain knowledge. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 415–425. https://aclanthology.org/D13-1040

Luan, Y., He, L., Ostendorf, M., & Hajishirzi, H. (2018). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3219–3232. https://doi.org/10.18653/v1/D18-1360

Lund, W. B., Kennard, D. J., & Ringger, E. K. (2013). Combining multiple thresholding binarization values to improve ocr output. *Document Recognition and Retrieval XX*, *8658*, 254–264.

Luong, M.-T., Nguyen, T. D., & Kan, M.-Y. (2012). Logical structure recovery in scholarly articles with rich document features. In *Multimedia storage and retrieval innovations for digital library systems* (pp. 270–292). IGI Global.

Ma, X., & Hovy, E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Makarov, P., & Clematide, S. (2018a, August). Neural transition-based string transduction for limited-resource setting in morphology. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th international conference on computational linguistics* (pp. 83–93). Association for Computational Linguistics. https://aclanthology.org/C18-1008

Makarov, P., & Clematide, S. (2018b, November). Imitation learning for neural morphological string transduction. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2877–2882). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1314

Mao, S., Rosenfeld, A., & Kanungo, T. (2003). Document structure analysis algorithms: A literature survey. *Document recognition and retrieval X*, *5010*, 197–207.

Marjanen, J., Pivovarova, L., Zosa, E., & Kurunmäki, J. (2019). Clustering ideological terms in historical newspaper data with diachronic word embeddings. *5th International Workshop on Computational History, HistoInformatics 2019*.

Marjanen, J., Zosa, E., Hengchen, S., Pivovarova, L., & Tolonen, M. (2020a). Topic modelling discourse dynamics in historical newspapers. https://doi.org/10.48550/ARXIV.2011.10428

Marjanen, J., Zosa, E., Hengchen, S., Pivovarova, L., & Tolonen, M. (2020b). Topic modelling discourse dynamics in historical newspapers. *arXiv preprint arXiv:2011.10428*.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., & Sagot, B. (2020). CamemBERT: A tasty French language model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7203–7219. https://www.aclweb.org/anthology/2020.acl-main.645

Martschat, S., & Markert, K. (2018). A temporally sensitive submodularity framework for timeline summarization. *Proceedings of the 22nd Conference on Computational Natural Language Learning*, 230–240. https://doi.org/10.18653/v1/K18-1023

Mausam, Schmitz, M., Soderland, S., Bart, R., & Etzioni, O. (2012). Open language learning for information extraction. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 523–534. https://aclanthology.org/D12-1048

McCallum, A., & Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons.

McDonough, K., Moncla, L., & Van de Camp, M. (2019). Named entity recognition goes to old regime france: Geographic text analysis for early modern french corpora. *International Journal of Geographical Information Science*, *33*(12), 2498–2522.

Melo, F., & Martins, B. (2017). Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS*, *21*(1), 3–38.

Michael, J., Labahn, R., Grüning, T., & Zöllner, J. (2019). Evaluating sequence-to-sequence models for handwritten text recognition. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1286–1293.

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404–411. https://aclanthology.org/W04-3252

Milanova, I., Silc, J., Serucnik, M., Eftimov, T., & Gjoreski, H. (2019). Locale: A rule-based location named-entity recognition method for latin text. *HistoInformatics@ TPDL*, 13–20.

Minard, A.-L., Speranza, M., Agirre, E., Aldabe, I., van Erp, M., Magnini, B., Rigau, G., & Urizar, R. (2015). SemEval-2015 task 4: TimeLine: Cross-document event ordering. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 778–786. https://doi.org/10.18653/v1/S15-2132

Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. *Proceedings of the Joint Conference of the 47th Annual Meeting of*

the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 1003–1011. https://aclanthology.org/P09-1113

Mishra, S., & Diesner, J. (2016). Semi-supervised named entity recognition in noisy-text. *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, 203–212.

Miwa, M., & Bansal, M. (2016). End-to-end relation extraction using lstms on sequences and tree structures. *arXiv preprint arXiv:1601.00770*.

Miwa, M., & Sasaki, Y. (2014, October). Modeling joint entity and relation extraction with table representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1858–1869). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1200

Moncla, L. (2015). *Automatic reconstruction of itineraries from descriptive texts* [Doctoral dissertation, Pau].

Moncla, L., Gaio, M., Joliveau, T., Le Lay, Y.-F., Boeglin, N., & Mazagol, P.-O. (2019). Mapping urban fingerprints of odonyms automatically extracted from French novels. *International Journal of Geographical Information Science*, *33*(12), 2477–2497. https://doi.org/10.1080/13658816.2019.1584804

Moretti, F. (2013). *Distant reading*. Verso Books.

Mutuvi, S., Doucet, A., Odeo, M., & Jatowt, A. (2018). Evaluating the Impact of OCR Errors on Topic Modeling. In *Maturity and Innovation in Digital Libraries. 20th International Conference on Asia-Pacific Digital Libraries, ICADL 2018, Hamilton, New Zealand, November 19-22, 2018, Proceedings* (pp. 3–14). https://doi.org/10.1007/978-3-030-04257-8_1

Nadeau, D., Turney, P. D., & Matwin, S. (2006). Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. *Advances in Artificial Intelligence: 19th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2006, Québec City, Québec, Canada, June 7-9, 2006. Proceedings 19*, 266–277.

Nafchi, H. Z., Ayatollahi, S. M., Moghaddam, R. F., & Cheriet, M. (2013). An efficient ground truthing tool for binarization of historical manuscripts. *2013 12th International Conference on Document Analysis and Recognition*, 807–811.

Namboodiri, A., & Jain, A. (2007, March). Document structure and layout analysis. https://doi.org/10.1007/978-1-84628-726-8_2

Nebhi, K. (2013). A rule-based relation extraction system using dbpedia and syntactic parsing. *Proceedings of the NLP-DBPEDIA-2013 Workshop co-located with the 12th International Semantic Web Conference (ISWC 2013)*.

Neudecker, C. (2016). An open corpus for named entity recognition in historic newspapers. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4348–4352.

Neudecker, C., Wilms, L., Faber, W. J., & van Veen, T. (2014). Large-scale refinement of digital historic newspapers with named entity recognition. *Proc IFLA Newspapers/GENLOC Pre-Conference Satellite Meeting*, 232–246.

Nguyen, D. P., Matsuo, Y., & Ishizuka, M. (2007). Subtree mining for relation extraction from Wikipedia. *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, 125–128. https://aclanthology.org/N07-2032

Nguyen, K.-H., Tannier, X., & Moriceau, V. (2014). Ranking multidocument event descriptions for building thematic timelines. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1208–1217. https://aclanthology.org/C14-1114

Nguyen, N. K., Boros, E., Lejeune, G., & Doucet, A. (2020). Impact analysis of document digitization on event extraction. In P. Basile, V. Basile, D. Croce, & E. Cabrio (Eds.), *Proceedings of the 4th workshop on natural language for artificial intelligence (NL4AI 2020) co-located with the 19th international conference of the italian association for artificial intelligence (ai*ia 2020), anywhere, november 25th-27th, 2020* (pp. 17–28, Vol. 2735). CEUR-WS.org. http://ceur-ws.org/Vol-2735/paper28.pdf

Nguyen, T. T. H., Jatowt, A., Coustaty, M., & Doucet, A. (2021a). Survey of post-ocr processing approaches. *ACM Comput. Surv.*, *54*(6). https://doi.org/10.1145/3453476

Nguyen, T. T. H., Jatowt, A., Coustaty, M., & Doucet, A. (2021b). Survey of Post-OCR processing approaches. *ACM Comput. Surv.*, *54*(6), 124:1–124:37.

Nguyen, T. T. H., Jatowt, A., Nguyen, N.-V., Coustaty, M., & Doucet, A. (2020). Neural machine translation with bert for post-ocr error detection and correction. *Proceedings of the ACM/IEEE joint conference on digital libraries in 2020*, 333–336.

Nguyen, T.-T.-H., Coustaty, M., Doucet, A., Jatowt, A., & Nguyen, N.-V. (2018). Adaptive edit-distance and regression approach for post-ocr text correction. *Maturity and Innovation in Digital Libraries: 20th International Conference on Asia-Pacific Digital Libraries, ICADL 2018, Hamilton, New Zealand, November 19-22, 2018, Proceedings 20*, 278–289.

Niklaus, C., Cetto, M., Freitas, A., & Handschuh, S. (2018). A survey on open information extraction. *Proceedings of the 27th International Conference on Computational Linguistics*, 3866–3878. https://aclanthology.org/C18-1326

Nissim, M., Matheson, C., Reid, J., et al. (2004). Recognising geographical entities in scottish historical documents. *Proceedings of the Workshop on Geographic Information Retrieval at SIGIR 2004*.

Niyogi, D., & Srihari, S. (1995). Knowledge-based derivation of document logical structure. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 472–475 vol.1. https://doi.org/10.1109/ICDAR.1995.599038

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19–51.

O'Gorman, L. (1993). The document spectrum for page layout analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, *15*, 1162–1173.

Ohta, T., Tateisi, Y., Kim, J.-D., Mima, H., & Tsujii, J. (2002). The genia corpus: An annotated research abstract corpus in molecular biology domain. *Proceedings of the human language technology conference*, 73–77.

Oliveira, S. A., & Kaplan, F. (2018). Comparing human and machine performances in transcribing 18th century handwritten venetian script. *ADHO/EHD 2018-Mexico City*.

Oliveira, S. A., Seguin, B., & Kaplan, F. (2018). Dhsegment: A generic deep-learning approach for document segmentation. *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 7–12.

Packer, T. L., Lutes, J. F., Stewart, A. P., Embley, D. W., Ringger, E. K., Seppi, K. D., & Jensen, L. S. (2010). Extracting person names from diverse and noisy ocr text. *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, 19–26.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Pasquali, A., Campos, R., Ribeiro, A., Santana, B. S., Jorge, A. M., & Jatowt, A. (2021). Tls-covid19: A new annotated corpus for timeline summarization. *ECIR*.

Pasquali, A., Mangaravite, V., Campos, R., Jorge, A. M., & Jatowt, A. (2019a). Interactive system for automatically generating temporal narratives. In L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, & D. Hiemstra (Eds.), *Advances in information retrieval* (pp. 251–255). Springer International Publishing.

Pasquali, A., Mangaravite, V., Campos, R., Jorge, A. M., & Jatowt, A. (2019b). Interactive system for automatically generating temporal narratives. In L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, & D. Hiemstra (Eds.), *Advances in information retrieval* (pp. 251–255). Springer International Publishing.

Pawar, S., Bhattacharya, P., & Palshikar, G. K. (2017). End-to-end relation extraction using markov logic networks.

Pawar, S., Bhattacharyya, P., & Palshikar, G. K. (2021). Techniques for jointly extracting entities and relations: A survey.

Pawar, S., Palshikar, G. K., & Bhattacharyya, P. (2017). Relation extraction : A survey.

Pedrazzini, N., & McGillivray, B. (2022). Machines in the media: Semantic change in the lexicon of mechanization in 19th-century british newspapers.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Perez-Cortes, J. C., Amengual, J.-C., Arlandis, J., & Llobet, R. (2000). Stochastic error-correcting parsing for ocr post-processing. *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, *4*, 405–408.

Phi, V.-T., Santoso, J., Shimbo, M., & Matsumoto, Y. (2018). Ranking-based automatic seed selection and noise reduction for weakly supervised relation extraction. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 89–95. https://doi.org/10.18653/v1/P18-2015

Piatti, B., Reuschel, A.-K., & Hurni, L. (2011). A literary atlas of europe-analysing the geography of fiction with an interactive mapping and visualisation system. *Proceedings of the 25th international cartographic conference*, 3–8.

Piotrowski, M. (2012, September). *Natural language processing for historical texts* (Vol. 5). https://doi.org/10.2200/S00436ED1V01Y201207HLT017

Piskorski, J., Pivovarova, L., Šnajder, J., Steinberger, J., & Yangarber, R. (2017). The first cross-lingual challenge on recognition, normalization and matching of named entities in slavic languages. *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*.

Pivovarova, L., Jean-Caurant, A., Avikainen, J., Alnajjar, K., Granroth-Wilding, M., Leppänen, L., Zosa, E., & Toivonen, H. (2020). Personal research assistant for online exploration of historical news. *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, 481–485.

Platas, M. L. D., Muñoz, S. R., González-Blanco, E., Fabo, P. R., & Mellado, E. Á. (2020). Medieval spanish (12th–15th centuries) named entity recognition and attribute annotation system based on contextual information. *Journal of the Association for Information Science and Technology*, *72*, 224–238. https://api.semanticscholar.org/CorpusID:225430990

Plum, A., Ranasinghe, T., Jones, S., Orasan, C., & Mitkov, R. (2022). Biographical semi-supervised relation extraction dataset. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3121–3130.

Pontes, E. L., Cabrera-Diego, L. A., Moreno, J. G., Boros, E., Hamdi, A., Sidère, N., Coustaty, M., & Doucet, A. (2020). Entity linking for historical documents: Challenges and solutions. *Digital Libraries at Times of Massive Societal Transition: 22nd International Conference*

on Asia-Pacific Digital Libraries, ICADL 2020, Kyoto, Japan, November 30–December 1, 2020, Proceedings 22, 215–231.

Prasad, A., Kaur, M., & Kan, M.-Y. (2018). Neural parscit: A deep learning based reference string parser. *International Journal on Digital Libraries*, *19*, 323–337. https://link.springer.com/article/10.1007/s00799-018-0242-1

Pratikakis, I., Zagoris, K., Barlas, G., & Gatos, B. (2017). Icdar2017 competition on document image binarization (dibco 2017). *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, *1*, 1395–1403.

Quaresma, P., & Finatto, M. J. B. (2020). Information extraction from historical texts: A case study. *DHandNLP@PROPOR*. https://api.semanticscholar.org/CorpusID:218910145

Quirk, C., & Poon, H. (2017). Distant supervision for relation extraction beyond the sentence boundary. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1171–1182. https://aclanthology.org/E17-1110

Ramakrishnan, C., Patnia, A., Hovy, E., & Burns, G. (2012). Layout-aware text extraction from full-text pdf of scientific articles. *Source code for biology and medicine*, *7*, 7. https://doi.org/10.1186/1751-0473-7-7

Rangoni, Y., Belaıd, A., & Vajda, S. (2012). Labelling logical structures of document images using a dynamic perceptive neural network. *International Journal on Document Analysis and Recognition (IJDAR)*, *15*(1), 45–55.

Reul, C., Springmann, U., Wick, C., & Puppe, F. (2018). Improving ocr accuracy on early printed books by utilizing cross fold training and voting. *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, 423–428.

Richter, C., Wickes, M., Beser, D., & Marcus, M. (2018). Low-resource post processing of noisy ocr output for historical corpus digitisation. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Riedel, S., Yao, L., & McCallum, A. (2010). Modeling relations and their mentions without labeled text. *ECML/PKDD*.

Riedl, M., Betz, D., & Padó, S. (2019). Clustering-based article identification in historical newspapers. *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 12–17. https://doi.org/10.18653/v1/W19-2502

Riedl, M., & Padó, S. (2018). A named entity recognition shootout for german. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 120–125.

Rigaud, C., Doucet, A., Coustaty, M., & Moreux, J.-P. (2019). Icdar 2019 competition on post-ocr text correction. *2019 international conference on document analysis and recognition (ICDAR)*, 1588–1593.

Ringlstetter, C., Hadersbeck, M., Schulz, K. U., & Mihov, S. (2007). Text correction using domain dependent bigram models from web crawls. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-2007) Workshop on Analytics for Noisy Unstructured Text Data*, 47–54.

Rink, B., & Harabagiu, S. (2010). UTD: Classifying semantic relations by combining lexical and semantic resources. *Proceedings of the 5th International Workshop on Semantic Evaluation*, 256–259. https://aclanthology.org/S10-1057

Ritter, A., Clark, S., Etzioni, O., et al. (2011). Named entity recognition in tweets: An experimental study. *Proceedings of the 2011 conference on empirical methods in natural language processing*, 1524–1534.

Ritze, D., Zirn, C., Greenstreet, C., Eckert, K., & Ponzetto, S. P. (2014). Named entities in court: The marinelives corpus. *Language Resources and Technologies for Processing and Linking Historical Documents and Archives-Deploying Linked Open Data in Cultural Heritage–LRT4HDA Workshop Programme*, 26.

Ro, Y., Lee, Y., & Kang, P. (2020). Multi
2OIE: Multilingual open information extraction based on multi-head attention with BERT. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1107–1117. https://doi.org/10.18653/v1/2020.findings-emnlp.99

Rocktäschel, T., Weidlich, M., & Leser, U. (2012). Chemspot: A hybrid system for chemical named entity recognition. *Bioinformatics*, *28*(12), 1633–1640.

Rodrigues Alves, D., Colavizza, G., & Kaplan, F. (2018). Deep reference mining from scholarly literature in the arts and humanities. *Frontiers in Research Metrics and Analytics*, 21.

Romanello, M. (2018). Detecting text reuse in newspapers data with passim. *Hacking the News, co-located with DHNordic 2018*.

Romanello, M., Ehrmann, M., Clematide, S., & Guido, D. (2020). The impresso system architecture in a nutshell.

Roth, D., & Yih, W.-t. (2004). A linear programming formulation for global inference in natural language tasks. *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, 1–8. https://aclanthology.org/W04-2401

Rouhou, A. C., Dhiaf, M., Kessentini, Y., & Salem, S. B. (2022). Transformer-based approach for joint handwriting and named entity recognition in historical document. *Pattern Recognition Letters*, *155*, 128–134.

Ruokolainen, T., & Kettunen, K. (2018). À la recherche du nom perdu–searching for named entities with stanford ner in a finnish historical newspaper and journal collection. *13th IAPR International Workshop on Document Analysis Systems*, 1–2.

Rusu, D., Hodson, J., & Kimball, A. (2014). Unsupervised techniques for extracting and clustering complex events in news. *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, 26–34. https://doi.org/10.3115/v1/W14-2905

Saad, R. S., Elanwar, R. I., Kader, N. A., Mashali, S., & Betke, M. (2016). Bce-arabic-v1 dataset: Towards interpreting arabic document images for people with visual impairments. *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, 1–8.

Sammut, C., & Webb, G. I. (Eds.). (2017). *Encyclopedia of machine learning and data mining*. Springer. https://doi.org/10.1007/978-1-4899-7687-1

Sang, E. F., & De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Saurí, R., Moszkowicz, J., Knippen, B., Gaizauskas, R., Setzer, A., & Pustejovsky, J. (2006). Timeml annotation guidelines version 1.2.1.

Schaefer, R., & Neudecker, C. (2020). A two-step approach for automatic ocr post-correction. *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 52–57.

Scheidel, W. (2015). Orbis: The Stanford Geospatial Network Model of the Roman World. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2609654

Schelb, J., Ehrmann, M., Romanello, M., & Spitz, A. (2022). Ecce: Entity-centric corpus exploration using contextual implicit networks. *Companion Proceedings of the Web Conference 2022*, 278–281. https://doi.org/10.1145/3487553.3524237

Schneider, R., Oberhauser, T., Klatt, T., Gers, F. A., & Löser, A. (2017). Analysing errors of open information extraction systems. *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, 11–18. https://doi.org/10.18653/v1/W17-5402

Schroeder, C. T., & Zeldes, A. (2016). Raiders of the lost corpus. *Digital Humanities Quarterly*, *10*(2).

Schulz, S., & Kuhn, J. (2017). Multi-modular domain-tailored ocr post-correction. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2716–2726.

Schvitz, G., Rüegger, S., Girardin, L., Cederman, L.-E., Weidmann, N., & Gleditsch, K. S. (2021). Mapping The International System, 1886-2017: The CShapes 2.0 Dataset. *Journal of Conflict Resolution*. https://doi.org/10.1177/00220027211013563

Seganti, A., Firląg, K., Skowronska, H., Satława, M., & Andruszkiewicz, P. (2021). Multilingual entity and relation extraction dataset and model. *Proceedings of the 16th Conference of the*

European Chapter of the Association for Computational Linguistics: Main Volume, 1946–1955. https://doi.org/10.18653/v1/2021.eacl-main.166

Segura-Bedmar, I., Martínez, P., & Herrero-Zazo, M. (2013). SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 341–350. https://aclanthology.org/S13-2056

Segura-Bedmar, I., Martínez Fernández, P., & Herrero Zazo, M. (2013). Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013).

Sekine, S. (1998). Nyu: Description of the japanese ne system used for met-2. *http://www. muc. saic. com/*.

Sekine, S., & Nobata, C. (2004). Definition, dictionaries and tagger for extended named entity hierarchy. *LREC*, 1977–1980.

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Miceli Barone, A. V., Mokry, J., & Nădejde, M. (2017, April). Nematus: A toolkit for neural machine translation. In A. Martins & A. Peñas (Eds.), *Proceedings of the software demonstrations of the 15th conference of the European chapter of the association for computational linguistics* (pp. 65–68). Association for Computational Linguistics. https://aclanthology.org/E17-3017

Shaalan, K. (2014). A survey of arabic named entity recognition and classification. *Computational Linguistics*, *40*(2), 469–510.

Shaalan, K., & Oudah, M. (2014). A hybrid approach to arabic named entity recognition. *Journal of Information Science*, *40*(1), 67–87.

Shafii, M., & Sid-Ahmed, M. (2015). Skew detection and correction based on an axes-parallel bounding box. *International Journal on Document Analysis and Recognition (IJDAR)*, *18*(1), 59–71.

Shang, Y.-M., Huang, H., & Mao, X.-L. (2022). Onerel:joint entity and relation extraction with one module in one step.

Sheikhshab, G., Birol, I., & Sarkar, A. (2018). In-domain context-aware token embeddings improve biomedical named entity recognition. *Proceedings of the ninth international workshop on health text mining and information analysis*, 160–164.

Shen, Z., Zhang, K., & Dell, M. (2020). A large dataset of historical japanese documents with complex layouts. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2336–2343.

Shih, F. Y., & Chen, S.-S. (1996). Adaptive document block segmentation and classification. *IEEE transactions on systems, man, and cybernetics, part B (cybernetics)*, *26*(5), 797–802.

Simistira, F., Bouillon, M., Seuret, M., Würsch, M., Alberti, M., Ingold, R., & Liwicki, M. (2017). Icdar2017 competition on layout analysis for challenging medieval manuscripts. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, *1*, 1361–1370.

Simistira, F., Seuret, M., Eichenberger, N., Garz, A., Liwicki, M., & Ingold, R. (2016). Diva-hisdb: A precisely annotated large dataset of challenging medieval manuscripts. *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 471–476.

Singh, A. K. (2008). Named entity recognition for south and south East Asian languages: Taking stock. *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*. https://aclanthology.org/I08-5003

Smith, D. A. (2002). Detecting and browsing events in unstructured text. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 73–80.

Souza, F., Nogueira, R., & Lotufo, R. (2019). Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*.

Speer, R., Chin, J., & Havasi, C. (2018). Conceptnet 5.5: An open multilingual graph of general knowledge.

Springmann, U., Reul, C., Dipper, S., & Baiter, J. (2018). Ground truth for training ocr engines on historical documents in german fraktur and early modern latin. *arXiv preprint arXiv:1809.05501*.

Sprugnoli, R., Tonelli, S., Marchetti, A., & Moretti, G. (2016). Towards sentiment analysis for historical texts. *Digit. Scholarsh. Humanit.*, *31*, 762–772.

Sprugnoli, R., et al. (2018). Arretium or arezzo? a neural approach to the identification of place names in historical texts. *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, 360–365.

Srivastava, S., Sanglikar, M., & Kothari, D. (2011). Named entity recognition system for hindi language: A hybrid approach. *International Journal of Computational Linguistics (IJCL)*, *2*(1), 10–23.

Stanovsky, G., & Dagan, I. (2016). Creating a large benchmark for open information extraction. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2300–2305. https://doi.org/10.18653/v1/D16-1252

Stanovsky, G., Michael, J., Zettlemoyer, L., & Dagan, I. (2018). Supervised open information extraction. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 885–895. https://doi.org/10.18653/v1/N18-1081

Steen, J., & Markert, K. (2019). Abstractive timeline summarization. *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, 21–31. https://doi.org/10.18653/v1/D19-5403

Strötgen, J., & Gertz, M. (2010). HeidelTime: High quality rule-based extraction and normalization of temporal expressions. *Proceedings of the 5th International Workshop on Semantic Evaluation*, 321–324. https://aclanthology.org/S10-1071

Subramani, N., Matton, A., Greaves, M., & Lam, A. (2021). A survey of deep learning approaches for ocr and document understanding.

Sun, C., Gong, Y., Wu, Y., Gong, M., Jiang, D., Lan, M., Sun, S., & Duan, N. (2019, July). Joint type inference on entities and relations via graph convolutional networks. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1361–1370). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1131

Swampillai, K., & Stevenson, M. (2011). Extracting relations within and across sentences. *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, 25–32. https://aclanthology.org/R11-1004

Swan, R., & Allan, J. (2000). Automatic generation of overview timelines. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 49–56. https://doi.org/10.1145/345508.345546

Taghva, K., Condit, A., Borsack, J., Kilburg, J., Wu, C., & Gilbreth, J. (1998). Manicure document processing system. *Document Recognition V*, *3305*, 179–184.

Thenmalar, S., Balaji, J., & Geetha, T. (2015). Semi-supervised bootstrapping approach for named entity recognition. *arXiv preprint arXiv:1511.06833*.

Tibbo, H. (2007). Primarily history in america: How u.s. historians search for primary materials at the dawn of the digital age. *American Archivist*, *66*, 9–50.

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. https://aclanthology.org/W02-2024

Toledo, J. I., Carbonell, M., Fornés, A., & Lladós, J. (2019). Information extraction from historical handwritten document images with a context-aware neural model. *Pattern Recognition*, *86*, 27–36.

Tran, G., Alrifai, M., & Herder, E. (2015). Timeline summarization from relevant headlines. In A. Hanbury, G. Kazai, A. Rauber, & N. Fuhr (Eds.), *Advances in information retrieval* (pp. 245–256). Springer International Publishing.

Tran, G., Herder, E., & Markert, K. (2015). Joint graphical models for date selection in timeline summarization. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1598–1607. https://doi.org/10.3115/v1/P15-1154

Tran, G. B., Tran, T., Tran, N. K., Alrifai, M., & Kanhabua, N. (2013). Leveraging learning to rank in an optimization framework for timeline summarization.

Tran, N. K., Ceroni, A., Kanhabua, N., & Niederée, C. (2015). Back to the past: Supporting interpretations of forgotten stories by time-aware re-contextualization. In X. Cheng, H. Li, E. Gabrilovich, & J. Tang (Eds.), *Proceedings of the eighth ACM international conference on web search and data mining, WSDM 2015, shanghai, china, february 2-6, 2015* (pp. 339–348). ACM.

Vasilopoulos, N., & Kavallieratou, E. (2016). Complex layout analysis based on contour classification and morphological operations. *Proceedings of the 9th Hellenic Conference on Artificial Intelligence*, 1–4.

Vobl, T., Gotscharek, A., Reffle, U., Ringlstetter, C., & Schulz, K. U. (2014). Pocoto-an open source system for efficient interactive postcorrection of ocred historical texts. *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 57–61.

Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). Recaptcha: Human-based character recognition via web security measures. *Science*, *321*(5895), 1465–1468.

Wang, H., Qin, K., Zakari, R. Y., Lu, G., & Yin, J. (2022). Deep neural network-based relation extraction: An overview. *Neural Computing and Applications*, 1–21.

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., & Wang, G. (2023). Gpt-ner: Named entity recognition via large language models. https://arxiv.org/abs/2304.10428

Wang, X., Guan, Y., Zhang, Y., Li, Q., & Han, J. (2020). Pattern-enhanced named entity recognition with distant supervision. *2020 IEEE International Conference on Big Data (Big Data)*, 818–827.

Wang, X., Hu, V., Song, X., Garg, S., Xiao, J., & Han, J. (2021). Chemner: Fine-grained chemistry named entity recognition with ontology-guided distant supervision. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Watkins, J. (2015). Spatial imaginaries research in geography: Synergies, tensions, and new directions. *Geography Compass*, *9*(9), 508–522.

Wei, H., Baechler, M., Slimane, F., & Ingold, R. (2013). Evaluation of svm, mlp and gmm classifiers for layout analysis of historical documents. *2013 12th International Conference on Document Analysis and Recognition*, 1220–1224.

White, A. S., Reisinger, D., Sakaguchi, K., Vieira, T., Zhang, S., Rudinger, R., Rawlins, K., & Van Durme, B. (2016). Universal decompositional semantics on universal dependencies. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1713–1723.

Wick, C., & Puppe, F. (2018). Fully convolutional neural networks for page segmentation of historical document images. *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, 287–292.

Won, M., Murrieta-Flores, P., & Martins, B. (2018). Ensemble named entity recognition (ner): Evaluating ner tools in the identification of place names in historical corpora. *Frontiers in Digital Humanities*, *5*, 2.

Wu, F., & Weld, D. S. (2010). Open information extraction using Wikipedia. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 118–127. https://aclanthology.org/P10-1013

Yadav, V., & Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.

Yahya, M., Whang, S., Gupta, R., & Halevy, A. (2014). ReNoun: Fact extraction for nominal attributes. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 325–335. https://doi.org/10.3115/v1/D14-1038

Yalniz, I. Z., & Manmatha, R. (2011). A fast alignment scheme for automatic ocr evaluation of books. *2011 International Conference on Document Analysis and Recognition*, 754–758.

Yan, H., Deng, B., Li, X., & Qiu, X. (2019). Tener: Adapting transformer encoder for named entity recognition. *arXiv preprint arXiv:1911.04474*.

Yan, Y., Okazaki, N., Matsuo, Y., Yang, Z., & Ishizuka, M. (2009). Unsupervised relation extraction by mining Wikipedia texts using information from the web. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 1021–1029. https://aclanthology.org/P09-1115

Yao, L., Haghighi, A., Riedel, S., & McCallum, A. (2011). Structured relation discovery using generative models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1456–1466. https://aclanthology.org/D11-1135

Yao, L., Riedel, S., & McCallum, A. (2012). Unsupervised relation discovery with sense disambiguation. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 712–720. https://aclanthology.org/P12-1075

Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu, Z., Liu, Z., Huang, L., Zhou, J., & Sun, M. (2019). DocRED: A large-scale document-level relation extraction dataset. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 764–777. https://doi.org/10.18653/v1/P19-1074

Yates, A., Banko, M., Broadhead, M., Cafarella, M., Etzioni, O., & Soderland, S. (2007). TextRunner: Open information extraction on the web. *Proceedings of Human Language Tech-*

nologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 25–26. https://aclanthology.org/N07-4013

Ye, J., Xu, N., Wang, Y., Zhou, J., Zhang, Q., Gui, T., & Huang, X. (2024). Llm-da: Data augmentation via large language models for few-shot named entity recognition. https://arxiv.org/abs/2402.14568

Yu, B., Zhang, Z., Shu, X., Wang, Y., Liu, T., Wang, B., & Li, S. (2020). Joint extraction of entities and relations based on a novel decomposition strategy.

Yu, Y., Jatowt, A., Doucet, A., Sugiyama, K., & Yoshikawa, M. (2021). Multi-TimeLine summarization (MTLS): Improving timeline summarization by generating multiple summaries. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 377–387. https://doi.org/10.18653/v1/2021.acl-long.32

Zaporojets, K., Deleu, J., Develder, C., & Demeester, T. (2021). Dwie: An entity-centric dataset for multi-task document-level information extraction.

Zeng, D., Sun, C., Lin, L., & Liu, B. (2017). Lstm-crf for drug-named entity recognition. *Entropy*, *19*(6), 283.

Zeng, W., Lin, Y., Liu, Z., & Sun, M. (2017). Incorporating relation paths in neural relation extraction. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1768–1777. https://doi.org/10.18653/v1/D17-1186

Zhai, Z., Nguyen, D. Q., Akhondi, S. A., Thorne, C., Druckenbrodt, C., Cohn, T., Gregory, M., & Verspoor, K. (2019). Improving chemical named entity recognition in patents with contextualized word embeddings. *arXiv preprint arXiv:1907.02679*.

Zhang, B., & Soh, H. (2024). Extract, define, canonicalize: An llm-based framework for knowledge graph construction. https://arxiv.org/abs/2404.03868

Zhang, S., & Elhadad, N. (2013). Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of biomedical informatics*, *46*(6), 1088–1098.

Zhang, Y., Qi, P., & Manning, C. D. (2018). Graph convolution over pruned dependency trees improves relation extraction. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2205–2215. https://doi.org/10.18653/v1/D18-1244

Zhang, Y., Zhong, V., Chen, D., Angeli, G., & Manning, C. D. (2017). Position-aware attention and supervised data improve slot filling. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 35–45. https://doi.org/10.18653/v1/D17-1004

Zheng, S., Wang, F., Bao, H., Hao, Y., Zhou, P., & Xu, B. (2017). Joint extraction of entities and relations based on a novel tagging scheme.

Zhila, A., & Gelbukh, A. (2014). Open information extraction for spanish language based on syntactic constraints. *Proceedings of the ACL 2014 Student Research Workshop*, 78–85.

Zhong, X., Tang, J., & Jimeno-Yepes, A. (2019). Publaynet: Largest dataset ever for document layout analysis. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1015–1022.

Zhou, H., Li, M., Xiao, Y., Yang, H., & Zhang, R. (2024). Leap: Llm instruction-example adaptive prompting framework for biomedical relation extraction. *Journal of the American Medical Informatics Association*, ocae147.

Zhou, S., Yu, B., Sun, A., Long, C., Li, J., & Sun, J. (2022). A survey on neural open information extraction: Current status and future directions. *arXiv preprint arXiv:2205.11725*.

Zhou, W., Huang, K., Ma, T., & Huang, J. (2020). Document-level relation extraction with adaptive thresholding and localized context pooling.

Zosa, E., & Granroth-Wilding, M. (2019). Multilingual dynamic topic model. *RANLP 2019-Natural Language Processing a Deep Learning World*.

Zosa, E., Pivovarova, L., Boggia, M., & Ivanova, S. (2022). Multilingual topic labelling of news topics using ontological mapping. *European Conference on Information Retrieval*, 248–256.

Zulfiqar, A., Ul-Hasan, A., & Shafait, F. (2019). Logical layout analysis using deep learning. *2019 Digital Image Computing: Techniques and Applications (DICTA)*, 1–5.