

**Hugo Scheithauer**

*Licencié ès histoire de l'art*

*Diplômé de master histoire de l'art*

# **La reconnaissance d'entités nommées appliquées à des données issues de la transcription automatique de documents manuscrits patrimoniaux.**

**Expérimentations et préconisations à partir du projet LECTAUREP**

Mémoire pour le diplôme de master  
« Technologies numériques appliquées à l'histoire »

2021



# Résumé

Ce mémoire rend compte et problématise un stage de quatre mois effectué au sein de l'équipe ALMAaCH - Inria, pour le projet LECTAUREP (INRIA - Archives nationales), actuellement dans la fin de sa troisième phase. Le projet a pour objectif de transcrire automatiquement les répertoires des notaires conservés au Département du Minutier Central des Archives nationales. Grâce à l'utilisation de la plate-forme eScriptorium et dans la mesure où des modèles de transcription d'écriture manuscrites ont été entraînés, des données textuelles peuvent être produites en masse. La reconnaissance d'entités nommées (REN) est une exploitation envisageable pour enrichir les transcriptions qui en sont faites, ainsi que pour fournir de nouvelles portes d'entrées aux documents sources par le biais des entités nommées qu'ils contiennent. L'existence de modèles génériques de REN facilite l'expérimentation, mais ceux-ci ne constituent pas forcément une solution clé en main. Le bruit des données générées par la REM et la perte de la structure logique suite à la REM perturbent en effet leurs performances.

Ce travail a pour but de donner des préconisations pour appliquer des outils de REN dans le cadre d'une campagne de REM, du pré-traitement des données brutes en sortie de transcription, au signalement des entités dans un fichier XML-TEI, tout en veillant à étudier les enjeux métiers de la REN dans un contexte patrimonial.

**Mots-clés :** Reconnaissance d'entités nommées ; Reconnaissance d'écriture manuscrite ; Transcription automatique ; Traitement automatique des langues ; TAL ; Natural Language Processing ; NLP ; Archives ; Archives nationales ; INRIA ; Minutier Central ; Répertoires des notaires ; LECTAUREP ; HTR ; REM ; NER ; REN ; XML ; TEI ; XSLT ; Python ; Intelligence artificielle ; machine learning ; apprentissage machine ; Humanités numériques.

**Informations bibliographiques :** Hugo Scheithauer, *La reconnaissance d'entités nommées appliquées à des données issues de la transcription automatique de documents manuscrits patrimoniaux. Expérimentations et préconisations à partir du projet LECTAUREP*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Alix Chagué et Thibault Clérice, École nationale des chartes, 2021.



# Remerciements

**J**E souhaitais tout d'abord remercier l'équipe pédagogique du Master 2 « Technologies numériques appliquées à l'histoire » de l'Ecole nationale des chartes pour les cours passionnantes et enrichissantes qui ont été dispensés à la promotion 2021 dont je fais partie, ainsi que pour leur disponibilité et soutien, et ce même à distance au cours d'une année universitaire presque intégralement effectuée en distanciel, marquée par deux confinements.

J'adresse mes plus sincères remerciements à ma tutrice de stage, Mme Alix Chagué, pour son accueil, son soutien, ses conseils, et toutes les opportunités qui m'ont été proposées durant cette expérience professionnelle. Je remercie également M. Laurent Romary pour son aide précieuse, ainsi que Mme Aurélia Rostaing, interlocutrice principale du côté des Archives nationales pour le projet LECTAUREP.

Ce stage ne se serait pas déroulé de la même manière si je n'avais pas eu le plaisir de rencontrer et échanger avec mes collègues de l'équipe ALMAñACH : je remercie M. Lucas Terriel et Mme Floriane Chiffolleau, ainsi que Mme Tanti Kristanti, M. Pedro Javier Ortiz Suárez et M. Benoît Sagot.

J'adresse mes chaleureux remerciements et exprime ma reconnaissance à Salma Chenguitti qui, grâce à son soutien indéfectible, m'épaule et m'aide à avancer. Mes remerciements vont également à mes camarades de promotion ainsi qu'à mes chers/chères ami-e-s Victorin Scache, Auriane Quoix, grâce à qui j'ai découvert le master TNAH, Calin Ganea, Alexandre Soyez, Solène Falk, Sacha Lampens et Thomas Bonnay. Enfin, merci à mes parents pour leur confiance dans mon parcours universitaire.



# Liste des sigles et abréviations

## Institutions et organismes

- AN : Archives Nationales
- ALMAaCH : *Automatic Language Modelling and Analysis & Computational Humanities*
- DMC : Département du Minutier Central des notaires de Paris
- DMOASI : Département de la Maîtrise d'Ouvrage du Système d'Information
- INRIA : Institut Nationale de Recherche en Informatique et Automatique

\*

## Domaines et disciplines

- DL : *Deep Learning*
- EI : Extraction d'information
- IA : Intelligence Artificielle
- ML : *Machine Learning*
- SHS : Sciences Humaines et Sociales
- TAL : Traitement Automatique des Langues - NLP : *Natural Language Processing*
- EI : Extraction d'information

\*

## Technologies

- ALTO : *Analysed Layout and Text Object*
- API : *Application Programming Interface*
- CSS : *Cascading Style Sheets*
- CSV : *Comma-separated values*
- EL : *Entity-Linking*

- HTML : *HyperText Markup Language*
- IIIF : *International Image Interoperability Framework*
- JSON : *JavaScript Object Notation*
- KWS : *Keyword Spotting*
- OCR : *Optical Character Recognition*
- ODD : *One Document Does it all*
- PAGE : *Page Analysis and Ground-truth Elements*
- REM : Reconnaissance d'Ecriture Manuscrite - HTR : *Handwritten Text Recognition*
- REN : Reconnaissance d'Entités Nommées - NER : *Named Entity Recognition*
- TEI : *Text Encoding Initiative*
- TSV : *Tab-separated values*
- XML : *eXtensible Markup Language*
- XSLT : *eXtensible Stylesheet Language Transformations*

\*

### **Concepts**

- EN : Entités Nommées

\*

### **Métriques**

- CER : *Character Error Rate* (Taux d'erreur de caractères)
- WER : *Word Error Rate* (Taux d'erreur de mots)

# **Introduction**



Le présent mémoire rend compte de mon expérience de stage au sein de l'équipe-projet ALMAaCH (*Automatic Language Modelling and Analysis & Computational Humanities*) de Inria, pour le projet LECTAUREP. Inria, autrefois Institut National de Recherche en Informatique et en Automatique, est un établissement public à caractère scientifique et technologique créé au début de l'année 1967 et placé sous la double tutelle du ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation ainsi que du ministère de l'Économie et des Finances.<sup>1</sup> Inria se définit comme un organisme moteur de la recherche et de l'innovation numérique s'appuyant sur les innovations apportées par les disciplines des mathématiques appliquées et de l'informatique. Depuis 2018, Inria fait partie intégrante de la stratégie nationale de recherche en intelligence artificielle (IA), menée par la ministre de l'Enseignement supérieur, de la Recherche et de l'Innovation Frédérique Vidal et Mounir Mahjoubi, alors secrétaire d'État au numérique. Inria s'inscrit dans l'un des six axes de cette stratégie en étant chargé de piloter le déploiement d'un programme national pour l'IA.<sup>2</sup> Inria est également un des acteurs français principaux du logiciel libre (*open source*) et du développement de l'accessibilité de ressources pédagogiques pour la société. Cela passe notamment par son soutien de la plate-forme FUN (France Université Numériques) et de son infrastructure technologique mettant à disposition des formations en lignes ouvertes à tous (MOOC, *Massive Open Online Course*).

Inria compte aujourd'hui 205 équipes-projets, dont ALMAaCH.<sup>3</sup> Celle-ci est spécialisée dans le traitement automatique des langues (TAL, ou NLP pour *Natural Language Processing*), et développe par exemple des modèles de langues, des solutions logicielles et s'occupe de la création de corpus, en s'appuyant sur les progrès réalisés dans le domaine de l'apprentissage profond (*deep learning*).<sup>4</sup> ALMAaCH porte et assure l'accompagnement de projets en humanités numériques, en partenariat avec des institutions patrimoniales. Benoît Sagot en est le responsable scientifique, et partage la coordination de l'équipe avec Laurent Romary, Djamé Seddah, Éric de la Clergerie et Rachel Bawden, chercheur-es permanent-es. L'équipe

---

1. Voir le décret n°79-1158 du 27 décembre 1979. <https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000519325> (consulté le 09/08/21) Voir également « Inria - Notre histoire. » <https://www.inria.fr/fr/notre-histoire> (consulté le 09/08/21)

2. Voir « Stratégie nationale de recherche en intelligence artificielle. » <https://www.enseignementsup-recherche.gouv.fr/cid136649/la-strategie-nationale-de-recherche-en-intelligence-artificielle.html> (consulté le 09/08/21) et le rapport du mathématicien et député Cédric Villani, *Donner un sens à l'intelligence artificielle.* [https://www.aiforhumanity.fr/pdfs/9782111457089\\_Rapport\\_Villani\\_accessible.pdf](https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf) (consulté le 09/08/21) Les 5 autres axes concernent le lancement d'un programme d'attractivité et de soutien aux talents ; la dynamisation de la recherche en IA à l'Agence nationale de Recherche (ANR) ; le renforcement des moyens de calcul ; le renforcement de la recherche partenariale ; et le renforcement des coopérations bilatérales, européennes et internationales.

3. Voir "ALMAaCH project-team Inria Paris' NLP research team" <http://almanach.inria.fr/index-en.html> (consulté le 09/08/21)

4. Voir [http://almanach.inria.fr/software\\_and\\_resources-en.html](http://almanach.inria.fr/software_and_resources-en.html) (consulté le 09/08/21) Nous pouvons notamment citer le modèle de langue française CamemBERT, <https://camembert-model.fr/> (consulté le 09/08/21) et le corpus multilingue OSCAR, <https://oscar-corpus.com/> (consulté le 09/08/21).

est en outre composée de chercheur-es postdoctoraux, d'ingénieur-es, de doctorant-es et de stagiaires.<sup>5</sup> Tout le dynamisme d'ALMANaCH réside dans la diversité de ses profils, où se mélangent linguistes, développeurs-euses, mathématiciens-nnes, et chercheur-es en humanités numériques.

Le projet LECTAUREP évolue dans ce contexte.<sup>6</sup> Il est mené depuis le premier semestre 2018 en partenariat avec les Archives nationales, et plus particulièrement avec le Département du Minutier Central (DMC). À terme, le projet vise à faciliter et fluidifier la consultation des répertoires d'actes des notaires conservés au DMC, grâce aux technologies de reconnaissance d'écriture manuscrite (REM, ou HTR pour *Handwritten text recognition*). Les actes des notaires cumulent 65% des communications journalières de la salle de lecture parisienne des Archives nationales, et représentent près de 26 kilomètres linéaires de documents pour les notaires de la capitale. L'automatisation de la transcription des répertoires de notaires permettrait de proposer une nouvelle voie d'accès à ces documents.<sup>7</sup> En effet, lorsqu'un-e chercheur-e souhaite retrouver un acte dans la salle des inventaires virtuelles (SIV) des Archives nationales, il est nécessaire que celui-ci ait été traité et inventorié par les équipes du DMC. Dans le cas contraire, il faut alors parcourir les répertoires des notaires, disponibles sous forme d'images dans la SIV grâce aux différentes campagnes de numérisations étalées depuis plusieurs dizaines d'années, et rangés par notaires ayant exercé.<sup>8</sup> La mise à disposition des textes issus de la transcription automatique permettrait de considérablement raccourcir la phase de dépouillement : la recherche des mentions des actes passés en minute et en brevet, inscrits dans les répertoires des notaires devient plus efficace qu'avec notre seul œil face à une image numérique. Une simple recherche plein texte pourrait par exemple renvoyer à une minute spécifique, ou un nom en particulier.

LECTAUREP est entré dans sa troisième phase au mois de novembre 2019. Le projet s'appuie sur le développement de l'application eScriptorium par l'équipe Scripta (PSL), une interface graphique pour la transcription et l'entraînement de modèles de REM grâce au moteur de REM et d'OCR Kraken, et de son déploiement sur un serveur d'Inria, « Traces6 ».<sup>9</sup>

---

5. Voir <http://almanach.inria.fr/people-en.html> (consulté le 09/08/21)

6. Voir le blog Hypotheses du projet : <https://lectaurep.hypotheses.org/> (consulté le 09/08/21)

7. INRIA, Archives nationales et Ministère de la culture, *Convention de recherche particulière relative au projet : LECTAUREP (phase 3) (LECTure Automatique de REPertoires)*, 2019

8. Aurélia Rostaing, *Méthodologie de recherche dans les archives notariales des Archives n...* Formation, URL : <https://fr.slideshare.net/AR2012/mthodologie-de-recherche-dans-les-archives-notariales-des-archives-nationales> (visité le 07/04/2021). Voir annexes A.1, A.2 et A.3.

9. Concernant la plate-forme eScriptorium, voir Benjamin Kiessling, Robin Tissot, Peter Stokes et Daniel Stökl Ben Ezra, « eScriptorium : An Open Source Platform for Historical Document Analysis », dans *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2019, t. 2, p. 19-19, DOI : 10.1109/ICDARW.2019.10032 et D. Stökl Ben Ezra, *L'infrastructure eScriptorium de reconnaissance automatique d'écriture manuscrite (HTR)*, Biblissima, 24 mars 2021, URL : <https://projet.biblissima.fr/fr/infrastructure-escritorium-reconnaissance-automatique-ecriture-manuscrite-htr> (visité le 27/04/2021). Concernant Kraken, voir le site internet consacré à sa documentation <http://kraken.re/> et B. Kiessling, *Kraken - an Universal Text Recognizer for the Humanities*, 2019, URL :

Le projet est aujourd’hui en passe de produire des données textuelles numériques en masse grâce à l’entraînement de modèles de REM robustes à partir de la vérité de terrain créée pour le projet.

La convention de recherche particulière de la phase 3 stipule que de nouvelles exploitations du corpus pourraient être rendues possibles grâce au passage à la REM. Avec une quantité de données suffisantes, de nouvelles lectures des répertoires des notaires pourraient être envisagées. Ils pourraient par exemple faire l’objet d’analyses quantitatives sur la nature des actes passés, le nombre de clients selon la chronologie, ou encore d’études sur la périodicité de l’activité professionnelle, sur l’histoire de la taxation et la fiscalité des différents types d’actes. On comprend par cet objectif que la REM est une technologie perçue comme un moyen ouvrant de nouveaux chemins grâce à la possibilité de traiter un volume de données conséquent, autrement difficile à atteindre par des moyens humains. La convention met également en avant la possibilité d’établir des liens avec des référentiels patronymiques, géographiques, etc., et avec des bases de données externes telles que Wikidata, par exemple.

x	<u>Théoule</u> (par Alfred Sylvain Gorgel) s à Paris, rue Brag. H, à la mère	22	Gratis
o	de Jeufville (par Jacob Guillaume) à Paris, rue Falguière, devant notaire apposé Mme de Boissieux	2	3,75
o	de Mbrangier (de Marie Gertrude Chon, épouse de Clodomir Gilbert) s à Paris, rue du Printemps, 22	4	3,75
e	<u>St<sup>e</sup> Blocq frères et Cie</u> (par la St <sup>e</sup> siège à Coul, à Gabriel Henri Coniat, rue du Printemps 7, de 7,875)	4	3,75
	<u>Henry</u> (par Clémentine Victoire Lecharpentier, épouse de Jean Baptiste) à Paris,	4	19,25

FIGURE 1 – Exemple d’un enregistrement au sein d’une page d’un répertoire de notaire, avec souligné en rouge, l’adresse de la personne concernée.

À partir de l’exemple de la figure 1, on pourrait imaginer l’établissement d’un lien entre « Paris, rue du Printemps 22 », souligné en rouge, avec sa notice Wikipedia.<sup>10</sup>

Cependant, l’enrichissement des données textuelles issues des transcriptions automatiques doit d’abord passer par une phase d’extraction d’information afin de rendre ce processus automatique. Afin de pouvoir cibler dans le document quels mots ou quelles expressions doivent être signalés, on utilise des outils de reconnaissance d’entités nommées (REN, ou NER pour *named entity recognition*), qui est une tâche propre à l’EI. Celle-ci, ainsi que le concept qu’elle traite, les entités nommées, méritent d’être définis plus longuement dans le développement de ce mémoire. De manière succincte, nous pouvons résumer la REN à la récupération

<https://dev.clariah.nl/files/dh2019/boa/0673.html> (visité le 19/07/2021). Enfin, pour avoir plus de renseignements sur le serveur Traces6 d’Inria, voir Alix Chagué, *Traces6 : notre serveur principal*, LECTAUREP, URL : <https://lectaurep.hypotheses.org/402> (visité le 09/08/2021).

10. Voir [https://fr.wikipedia.org/wiki/Rue\\_du\\_Printemps\\_\(Paris\)](https://fr.wikipedia.org/wiki/Rue_du_Printemps_(Paris)) (consulté le 10/08/21).

automatique de termes et expressions précises dans un texte. Une fois détectées et signalées, ces entités peuvent donc devenir les matériaux de l'enrichissement d'un texte.

La REN prend également son sens dans la notion de *distant reading* définie par Franco Moretti au début de l'année 2000.<sup>11</sup> Celui-ci proposait, dans le contexte des études littéraires, de se libérer du canon en privilégiant des accès aux textes « distants », différents d'une lecture linéaire de chaque ouvrage, permettant de traiter des corpus plus importants et d'inclure des œuvres n'appartenant pas au canon grâce à des méthodes d'analyse quantitative. La REN, selon ce principe, donnerait des clés d'accès distants aux textes en créant de nouvelles portes d'entrées avec les entités nommées.<sup>12</sup>

Mes missions de stage s'inscrivent dans ces thématiques. Après avoir découvert au début de l'année universitaire de Master 2 au sein du parcours « Technologies numériques appliquées à l'histoire » (TNAH) de l'École nationale des chartes les sujets de recherche d'ALMAnaCH, j'ai souhaité l'intégrer pour terminer ma formation dans son environnement technologique et scientifique, pour lequel je ressens un vif intérêt. J'ai ainsi été chargé, sous la supervision de ma tutrice de stage Mme Alix Chagué, de mener une réflexion prospective sur l'intégration d'outils de reconnaissance d'entités nommées dans la chaîne de traitement du projet LECTAUREP. Cette mission principale a été divisée en plusieurs sous-missions, à savoir :

- Identifier les informations que l'on souhaiterait pouvoir extraire automatiquement depuis les répertoires des notaires.
- Prendre connaissance et apprendre à manipuler les outils de TAL existant, et plus particulièrement les systèmes de reconnaissance d'entités nommées, ainsi que ce qui a déjà été développé par l'équipe ALMAnaCH dans ce domaine.
- Tester les performances de ces outils sur les données du projet LECTAUREP, sur des textes transcrits manuellement et des textes en sortie de REM, ainsi que sur des données textuelles de nature différente.
- Établir des préconisations sur les outils les mieux adaptés au projet, et comment les intégrer dans la chaîne de traitement actuellement définie. De plus, et en s'appuyant sur le projet LECTAUREP, essayer de trouver une solution généralisable à d'autres projets en humanités numériques.

Mon travail s'inscrit également dans la continuité de deux précédents stages réalisés au sein du projet LECTAUREP par deux anciens étudiants du master TNAH. Premièrement, Marie-Laurence Bonhomme a soutenu au mois de septembre 2018 un travail de recherche

11. Franco Moretti, « Conjectures on World Literature », *New Left Review*–1 (1<sup>er</sup> févr. 2000), p. 54–68.

12. Francesca Frontini, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos et Ranka Stanković, « Named Entity Recognition for Distant Reading in ELTeC », dans *CLARIN Annual Conference 2020*, Virtual Event, France, 2020, URL : <https://hal.archives-ouvertes.fr/hal-03160438> (visité le 10/08/2021). Voir également <https://www.distant-reading.net/> (consulté le 10/08/21).

portant sur la phase exploratoire du projet LECTAUREP.<sup>13</sup> A cette occasion, elle a pu traiter en profondeur de la nature des répertoires des notaires et de leur structure, dans l'objectif d'établir une méthodologie pour segmenter les documents en zones de texte et d'information. Son mémoire rend également compte de l'exploration de la reconnaissance d'écriture manuscrite avec la création de données d'entraînement et la manipulation du logiciel Transkribus.<sup>14</sup> Deuxièmement, Lucas Terriel s'est attaché en 2020 à concevoir un fichier pivot XML-TEI dans le but de structurer les données importées et exportées depuis eScriptorium, et à développer une application Python pour évaluer les modèles de transcription automatique, *Kraken-Benchmark*.<sup>15</sup> Mes missions de stage s'inscrivent donc au bout de la chaîne de traitement qui s'est mise en place au cours de ces trois dernières années, le matériau principal de cette exploration étant les textes produits par les modèles de transcription automatique.

La REN n'est pas une tâche étrangère aux organismes publics. Etalab, nom donné au département de la direction interministérielle du numérique (DINUM), a notamment conçu un guide pour pseudonymiser<sup>16</sup> des documents grâce à l'IA.<sup>17</sup> Ce guide intelligemment conçu présente comment utiliser la REN dans une application métier, dans le but de récupérer automatiquement les noms de personnes, les adresses, etc. pour les pseudonymiser. La REN est également employée par la Bibliothèque nationale de France (BnF) pour exploiter des contenus textuels avec le moteur sémantique Exalead qui est utilisé dans Gallica.<sup>18</sup> La difficulté majeure de mes missions de stage et la nouveauté qu'apporte l'exploration de la REN du

---

13. Marie-Laurence Bonhomme, *Défis et opportunités de la reconnaissance automatique d'écriture manuscrite pour les documents d'archives : l'exemple des répertoires des notaires de Paris*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Jean-Baptiste Camps, 2018

14. Voir <https://readcoop.eu/transkribus/> (consulté le 10/08/21)

15. Lucas Terriel, *Représenter et évaluer les données issues du traitement automatique d'un corpus de documents historiques. L'exemple de la reconnaissance des écritures manuscrites dans les répertoires de notaires du projet LectAuRep.*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Alix Chagué et Thibault Clérice, 2020. Voir également le répertoire Gitlab de *Kraken-Benchmark* : <https://gitlab.inria.fr/dh-projects/kraken-benchmark> (consulté le 10/08/21)

16. « La pseudonymisation est un traitement de données personnelles. Elle sert à ce qu'on ne puisse plus attribuer les données relatives à une personne physique sans avoir recours à des informations supplémentaires. En pratique la pseudonymisation consiste à remplacer les données directement identifiantes (nom, prénom, etc.) d'un jeu de données par des données indirectement identifiantes (alias, numéro dans un classement, etc.). [...] L'opération de pseudonymisation est réversible, contrairement à l'anonymisation. » Voir <https://guides.etalab.gouv.fr/pseudonymisation/pourquoi-comment/#qu'est-ce-que-la-pseudonymisation> (consulté le 10/08/21) L'anonymisation correspond à « l'identification et le neutralisation de références confidentielles dans un document ou un ensemble de documents. » Maud Ehrmann, *Les Entités Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*, Theses, Paris Diderot University, 2008, URL : <https://hal.archives-ouvertes.fr/tel-01639190> (visité le 25/08/2021), p. 27

17. Voir <https://www.etalab.gouv.fr/> (consulté le 10/09/21) et plus particulièrement *Etalab - Pseudonymiser des documents grâce à l'IA*, URL : <https://guides.etalab.gouv.fr/pseudonymisation/> (visité le 12/07/2021).

18. Emmanuelle Bermès et Eleonora Moiraghi, « Le patrimoine numérique national à l'heure de l'intelligence artificielle. Le programme de recherche Corpus comme espace d'expérimentation pour les humanités numériques », *Revue des Sciences et Technologies de l'Information - Série RIA : Revue d'Intelligence Artificielle* (, 2019), URL : <https://hal-bnf.archives-ouvertes.fr/hal-02122073> (visité le 10/08/2021), p. 7

projet LECTAUREP réside dans le matériau source : des données dites bruitées. Par bruit, on entend l'ensemble des erreurs de transcription ayant un impact sur l'intégrité des mots : des caractères absents ou en trop, des espaces mal placés, etc. Ce mémoire propose d'étudier une mise en application de la REN sur des données issues de la transcription automatique, de sorte à pouvoir établir des premières pistes pour mener de l'extraction d'information sur les corpus conséquents que la REM va produire. Les études portant sur la REN appliquées à des données bruitées se sont développées au cours de ces dernières années. On observe deux champs d'application majeurs, celles portant sur des données textuelles issues des réseaux sociaux (Twitter, par exemple) et d'Internet de façon plus générale<sup>19</sup>, et sur des données patrimoniale.<sup>20</sup> Mon travail de stage s'inscrit dans ce cadre scientifique.

Simon Hengchen et al., en 2015, se sont demandés si la REN était une opportunité pour le secteur culturel dans un contexte de développement de l'apprentissage machine, avec notamment comme problématique principale la baisse des budgets accordés aux institutions patrimoniales, entraînant par conséquent le besoin de recourir à un « catalogage informatisé semi-automatique ». <sup>21</sup> Pour y répondre, ils rassemblèrent un corpus de données pour évaluer les outils de REN alors disponibles au milieu de la décennie 2010, à savoir *Alchemy API*, *dataTXT*, *Wikimeta* et *Zemanta*.<sup>22</sup> Les résultats obtenus étaient moyens : sur un total de 744 entités présentes dans la vérité de terrain, seul *Wikimeta* a réussi à correctement identifier au moins la moitié du corpus, 378 pour en donner le nombre précis.<sup>23</sup> Précisons également que

19. Voir Diego Esteves, José Marcelino, Piyush Chawla, Asja Fischer et Jens Lehmann, « HORUS-NER : A Multimodal Named Entity Recognition Framework for Noisy Data », dans *IDA 2021 : Advances in Intelligent Data Analysis XIX*, 2021, pp. 89-100, URL : <https://openreview.net/forum?id=eoWnVtxS1su> (visité le 10/08/2021) ; Shubhangshu Mishra et Jana Diesner, « Semi-supervised Named Entity Recognition in noisy-text », dans *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, Osaka, Japan, 2016, p. 203-212, URL : <https://aclanthology.org/W16-3927> (visité le 10/08/2021) ; Vu Nguyen Hong, Hien Nguyen et Vaclav Snasel, « Text normalization for named entity recognition in Vietnamese tweets », *Computational Social Networks*, 3 (1<sup>er</sup> déc. 2016), DOI : 10.1186/s40649-016-0032-0

20. Ariane Pinche, Jean-Baptiste Camps et Thibault Clérice, « Stylometry for Noisy Medieval Data : Evaluating Paul Meyer's Hagiographic Hypothesis », dans *Digital Humanities Conference 2019 - DH2019*, Utrecht, Netherlands, 2019, URL : <https://hal.archives-ouvertes.fr/hal-02182737> (visité le 16/07/2021), Animesh Prasad, Hervé Déjean, Jean-Luc Meunier, Max Weidemann, Johannes Michael et Gundram Leifert, *Bench-Marking Information Extraction in Semi-Structured Historical Handwritten Records*, 2018. Ehrmann et al. disaient en ce sens : "it appears that historical texts pose new challenges to the application of NE processing, as they do for language technologies in general. First, inputs can be extremely noisy, with errors which do not resemble tweet misspellings or speech transcription hesitations [...]", voir Maud Ehrmann, et al. (éd.), « Extended Overview of CLEF HIPE 2020 : Named Entity Processing on Historical Newspapers », *CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings (oct. 2020), Meeting Name : 11th Conference and Labs of the Evaluation Forum (CLEF 2020) Num Pages : 38 Publisher : CEUR-WS Series Number : 2696, DOI : 10.5281/zenodo.4117566, p. 2

21. Simon Hengchen, Seth van Hooland, Ruben Verborgh et Max De Wilde, « L'extraction d'entités nommées : une opportunité pour le secteur culturel ? », *I2D - Information, donnees documents*, Volume 52–2 (7 juil. 2015), p. 70-79, URL : <https://www.cairn.info/revue-i2d-information-donnees-et-documents-2015-2-page-70.htm> (visité le 13/07/2021), p. 1.

22. *Ibid.*, p. 23

23. *Ibid.* Wikimeta est un outil d'annotation de texte compatible avec le web sémantique grâce à l'utilisation de DBpedia, ainsi qu'un système d'annotation des entités nommées et d'étiquetage grammatical.

les performances de ces quatre outils étaient mesurées en prenant en compte l'extraction de l'entité ainsi que l'établissement d'un lien entre l'entité et une base de connaissance externe. En effet, pour ce groupe de chercheurs, la REN pourrait prendre tout son sens dans le domaine patrimonial grâce au web sémantique, permettant de lier les données textuelles dont sont chargées les institutions patrimoniales et les projets de recherche, avec des bases de connaissance utilisables par le plus grand nombre. Leur conclusion, bien qu'étant en apparence positive<sup>24</sup>, semble donner une image mitigée et datée de la REN. Après avoir cité la chercheuse Johanna Drucker sur les risques qui existent à utiliser des technologies numériques qui n'ont pas été conçues pour les objets des humanités, et donc à potentiellement passer à côté d'anomalies et d'exceptions qui nourrissent la réflexion au profit d'une uniformisation des résultats, les chercheurs concluent que la REN « ne peut pas remplacer le travail du professionnel ». <sup>25</sup> La philosophie de LECTAUREP, dans l'utilisation et le développement de technologies permettant d'automatiser des tâches habituellement exécutées à la main, la transcription notamment, est toute autre : l'apport technologique ne vise pas à remplacer l'humain, mais à l'assister dans le volume de données qu'il peut traiter. Dans le cas de LECTAUREP, la transcription des répertoires des notaires n'aurait probablement pas pu être envisagée en dehors d'une campagne de transcription automatique, tant le volume de données est important.

Ce mémoire propose de ré-évaluer cette question et cette conclusion en s'appuyant sur les expériences que j'ai mené en me basant sur le projet LECTAUREP. Six ans après cet article, les technologies ont effet évolué, et méritent l'intérêt du secteur culturel, qui les utilise par ailleurs déjà, comme en témoigne le guide Etalab susmentionné. Nous nous demanderons comment la reconnaissance d'entités nommées appliquées aux données produites dans le cadre d'une campagne de reconnaissance d'écriture manuscrite peut-elle impacter les pratiques des métiers du patrimoine. Pour répondre à cette question, le contexte de la production des données du projet LECTAUREP, aussi bien sur le plan technologique que des humanités sera brièvement présenté. La REN étant une tâche d'extraction d'information, il convient d'exposer pour commencer quelle est la nature de l'information présente dans les répertoires des notaires, et comment est produit le matériau qui est soumis à la reconnaissance d'entités nommées, autrement dit, le résultat de la transcription automatique. Deuxièmement, nous

---

24. « En conclusion, nous pouvons affirmer que si les services obtiennent des mauvais scores sur des critères quantitatifs, leur apport reste non négligeable lors d'une analyse qualitative des résultats : l'étude quantitative se basant sur des critères métriques peu souples et difficilement applicables au réel. » *Ibid.*, p. 35.

25. *Ibid.*, p. 39. Les auteurs citent notamment Johanna Drucker, « Humanistic Theory and Digital Scholarship », dans *Debates in the Digital Humanities*, dir. Matthew K. Gold, NED - New edition, 2012, p. 85-95, URL : <https://www.jstor.org/stable/10.5749/j.ctttv8hq.9> (visité le 10/08/2021), p. 85 - 86. « We use tools from disciplines whose epistemological foundations are at odds with, or even hostile to, the humanities. Positivistic, quantitative and reductive, these techniques preclude humanistic methods because of the very assumptions on which they are designed : that objects of knowledge can be understood as ahistorical and autonomous. Probability is not the same as ambiguity or multivalent possibility within the field of humanistic inquiry. The task of calculating norms, medians, means and averages will never be the same as the task of engaging with anomalies and taking their details as the basis of an argument. »

rendrons compte des expériences que j'ai menées sur la reconnaissance d'entités nommées appliquées à ces données lors du stage. Cette partie présentera une chaîne de traitement idéal pour mettre en œuvre la REN, tout en présentant les difficultés existantes. Troisièmement, nous proposerons une discussion sur les impacts d'une campagne de REN dans le contexte des métiers du patrimoine. Cette partie visera à replacer ces expériences dans une application métier, en présentant ce qu'il est possible d'envisager pour l'après REN, notamment en termes de signalement des entités et de leurs exploitations possibles.

Sur le plan technologique, ce stage m'a permis de manipuler principalement des données textuelles grâce au langage de programmation Python. Le travail que j'ai produit a été stocké dans un répertoire Gitlab accessible avec le lien suivant : <https://gitlab.inria.fr/almanach/lectaurep/ner>.<sup>26</sup> Certains livrables ont été documentés dans des issues, ou n'ont pas pu être inclus dans ce répertoire, nous les mentionnerons le cas échéant au cours du développement et indiquerons un lien pour y accéder. Mon outil de travail principal a été les *notebooks* Jupyter.<sup>27</sup> Les *notebooks* sont définis par Fernando Pérez, leur créateur, ainsi :

Un environnement informatique lettré permet non seulement à ses utilisateurs d'exécuter des commandes, mais aussi d'enregistrer dans un document de format littéraire les résultats de ces commandes, des figures, du texte libre et même des expressions mathématiques. En pratique, cet environnement peut ressembler à un hybride entre une ligne de commande (comme celle de la coquille Unix) et un logiciel de traitement de texte, puisque les documents résultants peuvent être lus comme du texte et inclure des blocs de code exécutés par l'ordinateur sous-jacent.<sup>28</sup>

Ils sont visualisables directement en ligne, sur le Gitlab mentionné précédemment. Ils peuvent également être exécutés localement, en téléchargeant le répertoire Gitlab et en reproduisant les environnements de tests, soit un environnement virtuel basique dont les dépendances sont indiquées dans un fichier *requirements.txt*, soit dans un environnement Anaconda, dont les dépendances apparaissent dans un fichier *environment.yml*. Ils permettent d'accéder rapidement aux expériences réalisées, à leurs déroulés et à leurs résultats.

---

26. En cas de problème pour accéder à ce répertoire, celui-ci a été copié à l'identique dans le répertoire Github dans lequel est également stocké ce mémoire.

27. Voir <https://jupyter.org/> (consulté le 17/08/21).

28. <https://programminghistorian.org/fr/lecons/introduction-aux-carnets-jupyter-notebooks#fn:2> (consulté le 17/08/21).

N° DU REPERTOIRE	DATES DES ACTES	NATURE ET ESPÈCE DES ACTES :		NOMS, PRÉNOMS ET DOMICILES DES PARTIES INDICATIONS, SITUATIONS ET PRIX DES BIENS	RELATION DE L'Enregistrement.	
		EN BREVETS	EN MINUTES		DATES	DROITS
164	18	Donation	Dufusius (par César) à Paris, rue Eustache 11, à Catherine Béjal, veuve	"	"	"
165	18	2e	D° (par Jules) s.a. à son neveu	"	"	"
166	18	Contrat de mariage	Murc (de Paul Louis Georges) à Paris, 14 Rue Chabrolleau 68, et Marie Sophie Lemoine à Paris, veuve (déclaration de biens)	27	6,25	
167	18	Certificat de propriété	Duboscq (et fils de 380, de ville 3% au nom de) Delbisson, Justine Céline, V de Sylvie et autres, à Paris, rue de l'Oratoire 20	27	3,75	
168	20	Procuration	Fouj (de Charles) à Paris, rue de la Croix des Petits Champs 15, à Charles Véronique	21	7,50	
169	20 (du 16 Mars 1900)	Quittance	Chabrolleau (de Louis) à Paris, 14 Rue Chabrolleau, et Louise (de) Beloncourt, route de Louviers 128, à Saint-Léger-en-Yvelines	21	123,25	
170	20	Requisition d'actes	Gauthier (par Marie Alexandre) à Paris, rue de l'Oratoire 2, à son père et mère	1	Gauthier	
171	20	- 2e -	Cusson (par Georges Ernest) à Paris, rue du Marché 144, à son père	1	"	
172	20	Certificat de propriété	Leroy (de Lucien) à Paris, rue de l'Oratoire 20, à Mme de Leroy	1	"	
173	20	Historique rectificative	Lejeune (de Louis) à Paris, rue du Commerce 21, le 16 Septembre 1900	27	3,75	
174	20 (du 16 Janvier 1900)	Résiliation de bail	Lejeune (de Louis) à Paris, rue du Commerce 21, le 16 Septembre 1900	27	3,75	
175	21	Dépôt de pièces	Dupont (de Louis) à Paris, 14 Rue Chabrolleau, et Louise (de) Beloncourt, route de Louviers 128, à Saint-Léger-en-Yvelines	2	12,50	
176	21	notoriété rectificative	Elouet (de Manuel) à Paris, 28 Rue de l'Oratoire 20, le 25 Mai 1900	22	27,83	
177	21	Réquis (de Jeanne)	Thiéouf (de Alfred Eugène) à Paris, rue Brag, Hôtel nièce	22	3,75	
178	21	Procuration d'actes	de Nelly (de Jeanne) à Paris, rue Brag, Hôtel nièce	2	3,75	
179	21	Estat liquidation	de Morangis (de Félix Gobard) à Paris, épouse de Clémentine (de) à Paris, rue de l'Oratoire 20, à Paris, 14 Rue Chabrolleau 68	4	8,75	
180	21 (du 10 Juillet 1900)	Procès en réclame	Sté Blocq frères et Cie (de la) à Paris, 14 Rue Chabrolleau 68	4	8,75	
181	22	Quittance	Lejeune (de Louis) à Paris, 14 Rue Chabrolleau 68	4	49,25	
182	22	Consentement à mariage	Henry (de Auguste) à Paris, rue de l'Oratoire 20, à Paris, 14 Rue Chabrolleau 68	22	2,50	
183	22 (du 14 Février 1885)	Procuration d'actes	Morozzi Simon (de Gustave) à Paris, rue François Perrin, 12, à Paris, veuve (Baptiste), 14 Rue Chabrolleau 68	23	3,75	
184	22	Dépôt de pièces	Hazard (de Louis) à Paris, rue de l'Oratoire 20, à Paris, 14 Rue Chabrolleau 68	4	25,00	
185	22	Certificat de propriété	de Paul Eugène Boel et Marie Thérèse Julie à Paris, rue de l'Oratoire 20, le 10 Septembre 1900	4	25,00	
186	22	Dépôt de testament et codicilles	Pol (de) à Paris, 14 Rue Chabrolleau 68	2	Gauthier	
187	22	Procuration	Boudon (de Pauline) à Paris, 14 Rue Chabrolleau 68	23	2,50	
188	22 (du 2 Juin 1877)	Procuration d'actes	de Ste Goyen (de) à Paris, 14 Rue Chabrolleau 68, et Louise (de) à Paris, 14 Rue Chabrolleau 68	27	22,14	
189	22	Dépôt de pièces	de Goyen (de) à Paris, 14 Rue Chabrolleau 68, et Louise (de) à Paris, 14 Rue Chabrolleau 68	5	3,75	
190	22 (du 21 Janvier 1901)	Contrat de mariage	des Locomotives sans foyer (de la) à Paris, 14 Rue Chabrolleau 68	25	20,63	
191	22	Clarification de legs	Lebedel (de) à Paris, 14 Rue Chabrolleau 68, et Louise (de) à Paris, 14 Rue Chabrolleau 68	5	200,00	
192	22	Vente	Lebedel (de) à Paris, 14 Rue Chabrolleau 68, et Louise (de) à Paris, 14 Rue Chabrolleau 68	1	24,60	
193	23	Certificat de propriété	de Goyen (de) à Paris, 14 Rue Chabrolleau 68, et Louise (de) à Paris, 14 Rue Chabrolleau 68	5	1813,75	
194	23	Procuration	Guisse (de) à Paris, 14 Rue Chabrolleau 68	25	3,75	
195	23	- 2e -	Guisse (de) à Paris, 14 Rue Chabrolleau 68	27	Gauthier	
196	23	Discharge demandée	Dutry (de) à Paris, 14 Rue Chabrolleau 68	28	3,75	
197	23	Inventaire	Manoff (de Eugénie) à Paris, 14 Rue Chabrolleau 68	1	11,25	
198	23	Pré-conditionnel	Fournier (de) à Paris, 14 Rue Chabrolleau 68	2	8,75	

FIGURE 2 – Exemple d'une page d'un répertoire de notaire parisien, datée du mois de février 1901.



# **Première partie**

## **Exploiter les transcriptions automatiques avec la reconnaissance d'entités nommées : présentation des répertoires des notaires et des données textuelles produites avec la REM dans le cadre du projet LECTAUREP**



# **Chapitre 1**

## **La reconnaissance d'entités nommées, une sous-tâche de l'extraction d'information dans un texte : définitions**

### **1.1 la reconnaissance d'entités nommées, une tâche d'extraction d'information**

#### **1.1.1 Le traitement automatique des langues**

Le traitement automatique des langues est un domaine de recherche issu de la linguistique, l'informatique, les mathématiques et l'intelligence artificielle, et est « l'ensemble des méthodes permettant de traiter de manière automatique les données exprimées dans une langue. »<sup>1</sup> Le TAL possède quatre axes principaux : le traitement du signal, permettant par exemple d'effectuer de la synthèse vocale, de l'OCR, de la REM, « etc. » ; la syntaxe, notamment par le biais d'analyses syntaxiques ; l'extraction d'information, domaine pour lequel nous présenterons une définition plus détaillée ; et la sémantique, pour obtenir des résumés automatiques, des traductions automatiques, etc.<sup>2</sup>

Parmi les tâches les plus communes du TAL, on retrouve entre autres :

- La tokenisation, qui est le découpage d'un texte en unités plus petites, par exemple en phrases, en mots ou en caractères. Ces unités sont nommées tokens.

---

1. Yoann Dupont, *La structuration dans les entités nommées*, thèse de doct., Université Sorbonne Paris Cité, 2017, URL : <https://tel.archives-ouvertes.fr/tel-01772268> (visité le 13/07/2021), p. 16 et Catherine Fuchs et Benoit Habert, « le traitement automatique des langues : des modèles aux ressources », dans *Le Français Moderne - Revue de linguistique Française*, 2004, t. LXXII : 1, URL : <https://halshs.archives-ouvertes.fr/halshs-00067884> (visité le 08/08/2021)

2. Y. Dupont, *La structuration dans les entités nommées...*, p. 16

- L'étiquetage morpho-syntaxique (ou *Part-Of-Speech (POS) tagging*) qui consiste à attribuer à un mot la fonction grammaticale qui lui correspond dans une phrase donnée.
- Les plongements lexicaux (ou *word embeddings*) qui permettent de représenter les mots sous forme de vecteurs. Un même mot peut avoir des vecteurs différents selon le contexte grammatical et sémantique dans lequel il est situé.
- La REN, que nous allons présenter de façon plus détaillée dans ce mémoire.
- *Etc.*

Le TAL concerne également la création de modèles de langue, qui sont des modèles statistiques modélisant la distribution de séquences de mots dans une langue donnée. En 2018, un modèle de langue état de l'art basé sur des réseaux de neurones récurrents (RNN, *Recurrent Neural Network*) et sur la technologie *transformer* nommé BERT (*Bidirectional Encoder Representations from Transformers* a été publié par Jacob Devlin *et al.*.<sup>3</sup> BERT est a été entraîné sur un corpus de données conséquent (2500 millions de mots issus du Wikipedia anglais, ainsi que 800 millions de mots issus de livres en anglais) et peut servir par exemple à déterminer si une phrase en anglais exprime un sentiment négatif ou positif, à prédire, depuis une première phrase, la ou les phrases qui la suivent, *etc.*<sup>4</sup> Depuis, l'architecture de BERT a été utilisée sur des corpus de langues différentes. ELMo par exemple, qui est un modèle de langue pour le japonais, le portugais, l'allemand et le basque, et camemBERT, développé au sein de l'équipe ALMAnaCH, modèle de langue pour le français.<sup>5</sup>

### 1.1.2 L'extraction d'information

La REN est une des tâches réalisées par le TAL dans le cadre de l'extraction d'information (EI). L'EI est un domaine concerné par le TAL, elle est définie par la chercheuse Rosa Stern comme une « une opération de repérage et de structuration en classes sémantiques [...] »

---

3. Jacob Devlin, Ming-Wei Chang, Kenton Lee et Kristina Toutanova, « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding », (, 11 oct. 2018), URL : <https://arxiv.org/abs/1810.04805v2> (visité le 30/08/2021). Voir également le site internet "BERT for Humanists" qui met à disposition des définitions, articles et tutoriels pour utiliser ce modèle de langue dans le contexte des humanités numériques : <https://melaniewalsh.github.io/BERT-for-Humanists/> (consulté le 28/08/21). On y trouve la définition de *transformer* : "[...] transformers process all the inputs simultaneously (rather than sequentially, like in LSTMs) and are great at parallelization (breaking up our task into parallel pieces), allowing us to process more data more quickly." (<https://melaniewalsh.github.io/BERT-for-Humanists/glossary/#transformers> (consulté le 28/08/21).

4. Voir <https://melaniewalsh.github.io/BERT-for-Humanists/introductions/Introduction/> (consulté le 28/08/21).

5. Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Y. Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah et Benoît Sagot, « CamemBERT : a Tasty French Language Model », dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020, p. 7203-7219, DOI : 10.18653/v1/2020.acl-main.645, p. 7204

d'éléments informatifs spécifiques présents dans des données non structurées, notamment textuelles, menée dans le but de donner à l'information une forme adéquate pour des traitements automatiques. »<sup>6</sup> Aux côtés de la REN, l'EI concerne également l'extraction de relations, le fait d'extraire les relations sémantiques présentes dans un texte entre plusieurs entités, par exemple entre deux personnes, et de les classifier dans des catégories, et l'extraction d'événements.<sup>7</sup>

Comme l'explique le chercheur Yoann Dupont, « l'EI peut se décrire comme le remplissage automatique d'un formulaire aux champs prédéfinis dans le but d'alimenter une base de connaissances qui pourra par la suite être consultée par des êtres humains. » Par exemple, pour remplir un formulaire à partir d'un texte décrivant un acte criminel, on chercherait à obtenir des informations générales, comme la date et le lieu, ses acteurs, et les moyens ainsi que les conséquences.<sup>8</sup> Cet exemple illustre efficacement comment l'EI vient faciliter la lecture et l'apprehension d'un document, autre qu'avec une lecture linéaire. C'est à une plus grande échelle qu'il faut concevoir ce que R. Stern nomme des traitements automatiques des résultats de l'EI. Andrew McCallum explique que l'EI est une étape qui précède le « data mining », ou exploration de données, en autorisant la structuration l'information dans une base de données.<sup>9</sup> Pour ce faire, plusieurs étapes sont nécessaires selon lui : la segmentation, qui s'apparente à ce que R. Stern nomme repérage, qui consiste à cibler l'information ; la classification des mots qui ont été segmentés ; l'association qui permet de déterminer quelles informations appartiennent au même domaine, au même thème ; la normalisation, considérée dans ce cas comme l'action d'attribuer à une donnée un format standard (par exemple, l'heure 15h en 3 p.m) ; et la *deduplication*, qui élimine les doublons afin de ne pas créer de redondances dans une base de données.<sup>10</sup>

R. Stern explique également qu'un système d'EI est réalisée en deux phases. La première correspond l'entraînement, lorsque des données sont créées à partir de l'annotation de textes, par exemple dans le but d'entraîner un modèle à reconnaître des motifs et à les classifier dans des catégories. La seconde est le déploiement, qui correspond au moment où le système d'EI est appliqué à des données distinctes du corpus d'entraînement. Grâce à cela, les « éléments textuels pertinents sont ainsi extraits, organisés relativement aux classes définies pour la tâche

---

6. Rosa Stern, *Identification automatique d'entités pour l'enrichissement de contenus textuels*, Theses, Université Paris-Diderot - Paris VII, 2013, URL : <https://tel.archives-ouvertes.fr/tel-00939420> (visité le 12/07/2021), p. 59

7. Kiran Adnan et Rehan Akbar, « Limitations of information extraction methods and techniques for heterogeneous unstructured big data », *International Journal of Engineering Business Management*, 11 (1<sup>er</sup> janv. 2019), Publisher : SAGE Publications Ltd STM, p. 1847979019890771, DOI : 10.1177/1847979019890771, pp. 6 -7

8. Y. Dupont, *La structuration dans les entités nommées...*, p. 17

9. Andrew McCallum, « Information Extraction : Distilling structured data from unstructured text », *Queue*, 3-9 (1<sup>er</sup> nov. 2005), p. 48-57, DOI : 10.1145/1105664.1105679, pp. 49 - 50

10. *Ibid.*

puis retournés dans le format structuré correspondant. »<sup>11</sup> LECTAUREP est donc aujourd’hui dans la phase préliminaire de ce premier temps, où l’on cherche à définir la meilleure stratégie pour effectuer une campagne de REN sur les données textuelles produites par le projet.

L’EI est également un domaine appliqué à d’autres médias, notamment les fichiers audio, les fichiers vidéos et les images. En 2019, Kiran Adnan et Rehan Akbar estimaient que sur 95 articles sélectionnés traitant de l’EI, près de la moitié traitait de l’extraction d’information à partir de textes.<sup>12</sup> Avec l’objectif de déployer un système de REN, LECTAUREP s’inscrit donc dans un contexte scientifique dynamique.

### 1.1.3 La reconnaissance d’entités nommées et le concept d’entité nommée

La reconnaissance d’entités nommées est une tâche qui s’inscrit dans le domaine de l’extraction d’informations. Elle a pour but d’extraire et de classifier automatiquement les entités nommées contenues dans un texte grâce à un outil informatique.<sup>13</sup> Y. Dupont souligne l’importance de la REN ainsi :

« Il s’agit d’une tâche très importante du TAL qui sert généralement de point de départ à d’autres tâches telles que l’extraction de relations, la construction d’une base de connaissances, l’*entity-linking*, la résolution de chaînes de corréférence, le résumé automatique, les systèmes de questions-réponses, etc. Elle permet, plus largement, l’accès à l’information pertinente pour des tâches qui autrement seraient irréalisables. »<sup>14</sup>

Nous allons ainsi présenter une définition conceptuelle, et une définition technique du concept d’entité nommée. Conceptuellement, les linguistes définissent l’entité nommée comme une « unité linguistique de nature référentielle ».<sup>15</sup> Cette unité est la plupart du temps un mot, mais peut également être une expression à plusieurs mots. Les entités nommées les plus communes, par exemple, sont des noms propres faisant référence à des noms de personnes.

---

11. R. Stern, *Identification automatique d’entités pour l’enrichissement de contenus textuels...*, p. 59

12. K. Adnan et R. Akbar, « Limitations of information extraction methods and techniques for heterogeneous unstructured big data »..., p. 4 L’EI à partir de texte concernait 49% des 95 articles, à partir de fichiers audio 9%, à partir d’images 27%, à partir de vidéos 15%. Autant de médias qui induisent donc des objectifs et des tâches différentes.

13. Charles Riondet et Luca Foppiano, « History Fishing When engineering meets History », dans *Text as a Resource. Text Mining in Historical Science #dhiha7*, Paris, France, 2017, URL : <https://hal.inria.fr/hal-01830713> (visité le 07/04/2021), p. 3

14. Y. Dupont, *La structuration dans les entités nommées...*, pp. 17 - 18. Voir également M. Ehrmann, *Les Entités Nommées, de la linguistique au TAL...*, pp. 21 - 25 La résolution de corréférence permet, par exemple, de savoir que dans les phrases « Lucien a acheté un livre. Il est vert », le pronom personnel « il » fait référence au livre, et non à l’EN de type personne « Lucien ». La désambiguïsation lexicale est « l’opération consistant à déterminer le sens d’un mot en contexte. »

15. Y. Dupont, *La structuration dans les entités nommées...*, p. 18

On donne également le nom de « mention » à une instance d'une entité nommée dans un texte.<sup>16</sup> Nous adoptons la même définition que Yoann Dupont pour sa thèse de doctorat, citée de la thèse de Maud Ehrmann, en 2008 :

« Étant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus. »<sup>17</sup>

Cette définition exprime la nature changeante des entités nommées en fonction des corpus auxquelles elles appartiennent. Bien que la communauté scientifique s'accorde sur le fait que les entités nommées sont généralement des noms propres faisant référence à des personnes, des localisations ou des organisations, il se peut que celles-ci ne soient d'aucune utilité pour certains corpus, par exemple pour des corpus de textes de chimie.<sup>18</sup> En ce sens, cette définition rejoint celle de Daniel Jurafsky et James Martin, qui soutiennent que les entités nommées sont régulièrement étendues à d'autres types d'unités linguistiques qui ne sont pas des entités en soi ou des noms propres.<sup>19</sup>

Pedro Javier Ortiz Suárez *et al.* soulignent, en outre, que les entités nommées peuvent être analysées de deux façons : l'analyse intrinsèque, correspondant au fait que le mot « Maroc », par exemple, fera toujours référence à la localisation en question ; et l'analyse contextuelle, qui situe l'entité nommée dans son contexte pour la classifier. Par exemple, le mot « Orange », selon le contexte dans lequel il est placé, fera référence à la couleur, l'entreprise, la ville, etc. Il en va de même pour le mot « Maroc », qui peut faire référence à une organisation dans la phrase « Le Maroc a gagné une médaille d'or aux Jeux Olympiques de 2020. »<sup>20</sup>

Techniquement, les entités nommées se définissent comme une séquence de caractères ayant un index de début et un index de fin dans un texte. Par exemple, pour l'enregistrement suivant, tiré d'une page d'un répertoire de notaire : « Dupuis (par Pierre) à Paris, rue Turgot 4, à Catherine Bizial, sa femme » (figure 1.1), on peut attribuer un index à chaque caractère, débutant à 0. Le mot « Dupuis » est une entité nommée, plus précisément un nom de personne. Cette entité nommée possède un début, l'index 0, car débutant la chaîne de caractères, et une fin, l'index 5 (figure 1.2). Les index suivants, donc représentant tour à tour un espace, la lettre « à », etc., ne sont pas inclus dans cette entité nommée. À une échelle supérieure,

---

16. *Ibid.*, p. 19

17. *Ibid.*, p. 20, et M. Ehrmann, *Les Entités Nommées, de la linguistique au TAL...*

18. Y. Dupont, *La structuration dans les entités nommées...*, p. 20

19. "Named entities are words for proper nouns referring mainly to people, places, and organizations, but extended to many other types that aren't strictly entities or even proper nouns." Daniel Jurafsky et James Martin, *Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, t. 2, 2020, p. 169

20. P. J. Ortiz Suárez, Y. Dupont, B. Muller, L. Romary et B. Sagot, « Establishing a New State-of-the-Art for French Named Entity Recognition », dans *LREC 2020 - 12th Language Resources and Evaluation Conference*, Marseille, France, 2020, URL : <https://hal.inria.fr/hal-02617950> (visité le 13/04/2021), p. 1

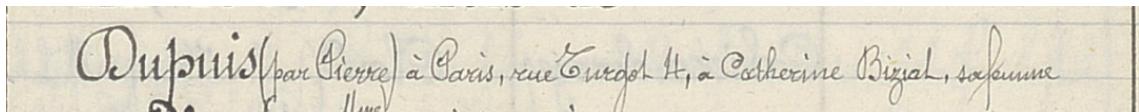


FIGURE 1.1 – Exemple d'un enregistrement situé dans la colonne 5 (noms, prénoms et domiciles des parties) d'une page d'un répertoire de notaire (1).

cette séquence de caractères est rassemblée en un token. Une entité nommée peut rassembler plusieurs tokens, comme par exemple l'entité « Parlement européen ».

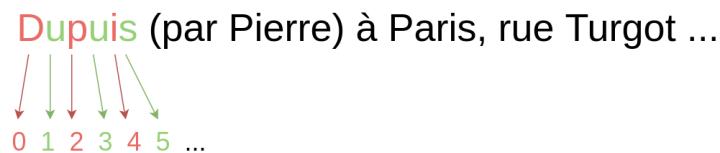


FIGURE 1.2 – Illustration des index de l'entité nommée « Dupuis », étant un nom de personne.

Lorsque cette séquence de caractères est classifiée, on lui attribue une étiquette, ou label, indiquant à quelle classe elle appartient. Pour le token « Dupuis », la bonne étiquette serait « PER », la rattachant à la catégorie commune des noms de personne. Il est possible d'extraire une entité nommée, mais de mal la classifier, la première étant condition de la deuxième. Sans extraction, il est en effet impossible de donner un label. Le nombre d'étiquettes possibles est large, et il ne serait pas pertinent d'en faire une liste exhaustive car comme nous l'avons vu, elles sont dépendantes du corpus d'application. Cependant, il existe des étiquettes communes aux modèles génériques de REN, qu'il convient de présenter :

- « PER », pour les noms de personne.
- « ORG », pour les noms d'organisation.
- « LOC », pour les noms de localisation.
- « MISC », pour *miscellaneous* (divers). Ce type d'entité indique de manière générale que la séquence de caractères a été détectée en une entité nommée, mais sans en déterminer la catégorie exacte : une entité nommée en devenir, en somme.

Le nombre d'étiquettes peut rapidement augmenter, comme en témoigne les 27 classes définies et reconnues par l'outil GROBID-NER. Parmi celles-ci, on retrouve l'étiquette « PERIOD », pour les dates, les ères historiques, etc. ; « TITLE », pour les titres honorifiques d'une personne ; ou encore « INSTITUTION », qui désigne une organisation de personnes et une localisation ou une structure qui partage le même nom (l'Université de Yale, par exemple). Cette dernière étiquette diffère de « ORG », présent dans les classes de GROBID-NER sous le nom d'« ORGANISATION », mais désignant un groupe organisé de personnes possédant une

dimension légale.<sup>21</sup> Les étiquettes varient donc d'un outil à l'autre.<sup>22</sup>

Dans l'exemple de la figure 1.1, on souhaiterait donc idéalement extraire :

- « Dupuis », en tant que « PER ».
- « Pierre », en tant que « PER ».
- « Paris », en tant que « LOC ».
- « rue Turgot 4 », en tant que « LOC ». Il s'agit donc ici d'une entités nommées constituées de plusieurs tokens.
- « Catherine Bizal », en tant que « PER ».

La reconnaissance d'entités nommées appliquées à des transcriptions automatiques est plus difficile à appréhender. Les transcriptions automatiques entraînent parfois le changement de la longueur de la séquence de caractères par rapport à la vérité de terrain. Par exemple, pour l'enregistrement reproduit dans la figure 1.3, on souhaite récupérer le nom de famille « de Neufville ». Mais selon les résultats de la transcription automatique, l'entité nommée peut grandement différer, comme par exemple dans ce résultat de REM : « Deeuit (par Jaéol Guillaume) dtà Paris, rue Falévy 6, deonbeusdtales appt à Mme de Boissieu », le nom de famille présent au début de l'enregistrement ne représente plus qu'un seul mot. Cependant, un système de REN devrait être en mesure d'identifier *Deeuit* comme un nom de personne.

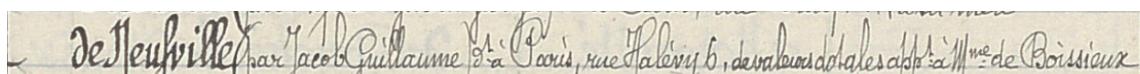


FIGURE 1.3 – Exemple d'un enregistrement situé dans la colonne 5 (noms, prénoms et domiciles des parties) d'une page d'un répertoire de notaire (2).

#### 1.1.4 La désambiguïsation des EN

La REN doit en outre faire face à une difficulté majeure : la désambiguïsation. En effet, un même mot, selon le contexte dans lequel il est inséré, peut être doté de sens différents. La polysémie est prise en compte dans les systèmes de REN. On rencontre parfois l'acronyme NERD, *Named Entity Recognition and Disambiguation* pour désigner ce système dans sa forme complète. Un système de REN devrait être capable, idéalement, de classifier les tokens suivants : « rue saint-Antoine » comme étant une localisation, et non deux tokens (« rue » et « saint ») et le token « Antoine » comme étant une personne.

21. Voir <https://grobid-ner.readthedocs.io/en/latest/class-and-senses/> (consulté le 24/08/21).

22. Voir également le guide d'annotation pour les entités nommées Quaero. Sophie Rosset, Cyril Grouin et Pierre Zweigenbaum, *Entités nommées structurées : guide d'annotation Quaero*, Google-Books-ID : iI9bMwEACAAJ, LIMSI, 2011

Un modèle statistique prend en compte le contexte dans lequel s'inscrit les différents types d'EN pour leur attribuer une étiquette.<sup>23</sup> Dans l'exemple ci-dessus, les tokens précédant le token « Antoine » peuvent faire figure d'indices contextuels pour un modèle statistique, l'amenant à segmenter l'EN correctement et à choisir l'étiquette « LOC », plutôt que « PER ».<sup>24</sup>

La désambiguïsation peut également s'appuyer sur de l'*entity linking*. L'EL consiste à désambiguïser une EN à l'aide d'une base de connaissances.<sup>25</sup> Pour l'exemple précédent, un référentiel des noms des rues de Paris pourrait être utilisé. Cette tâche s'effectue généralement en trois temps : la détection de mentions, la génération de candidats, et le classement de ceux-ci afin de déterminer lequel est le meilleur.<sup>26</sup>

Pour parvenir à extraire les EN, deux grandes approches existent : l'approche dite linguistique ou symbolique, et l'approche statistique grâce à l'apprentissage machine.<sup>27</sup> Nous les présenterons dans les deux sections suivantes.

## 1.2 Les systèmes de reconnaissance par règles

La première approche repose sur « l'intuition humaine, avec la construction manuelle des modèles d'analyse, sous la forme de règles contextuelles le plus souvent. »<sup>28</sup> Les règles agissent comme des « patrons d'extraction » pour récupérer les EN.<sup>29</sup> Par exemple, en s'appuyant sur un lexique de prénoms, si le système reconnaît un prénom et que le token suivant commence par une majuscule, alors on peut déduire qu'il correspond à un nom de personne.

D. Jurafsky et J. Martin expliquent qu'on retrouve fréquemment ces outils dans un contexte commercial. Ils sont basés sur des systèmes à règles et combinent fréquemment des listes d'EN à extraire et des règles, avec parfois l'usage de systèmes d'apprentissage machine supervisés. Ils donnent notamment l'exemple de l'*IBM System T architecture*, un outil qui permet à l'utilisateur-ice de déclarer des règles pour étiqueter des entités nommées en utilisant un langage de requête basé sur des expressions régulières, des dictionnaires, des contraintes

---

23. M. Ehrmann, *Les Entités Nommées, de la linguistique au TAL...*, p. 72

24. Y. Dupont, *La structuration dans les entités nommées...*, p. 21. Voir également M. Ehrmann, *Les Entités Nommées, de la linguistique au TAL...*, pp. 177 - 186. M. Ehrmann dédie un chapitre de sa thèse à la question de la polysémie et de son impact sur la REN.

25. Jan-Christoph Klie, Richard Eckart de Castilho et Iryna Gurevych, « From Zero to Hero : Human-In-The-Loop Entity Linking in Low Resource Domains », dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020, p. 6982-6993, DOI : 10.18653/v1/2020.acl-main.624, p. 6982 "Entity linking (EL) describes the task of disambiguating entity mentions in a text by linking them to a knowledge base (KB), e.g. the text span Earl of Orrery can be linked to the KB entry John Boyle, 5. Earl of Cork, thereby disambiguating it."

26. *Ibid.*, p. 6983

27. M. Ehrmann, *Les Entités Nommées, de la linguistique au TAL...*, p. 33

28. *Ibid.*, p. 32

29. *Ibid.*

sémantiques ("semantic constraints") et d'autres opérateurs.<sup>30</sup>

Y. Dupont présente également dans sa thèse quelques outils de REN semblables. Il mentionne notamment l'outil d'annotation en EN TM360, de la suite d'outils Luxid (Expert System France), utilisant un système de règles écrites en XML.<sup>31</sup> Ces outils ne sont cependant pas la norme dans le domaine de la recherche académique.<sup>32</sup>

### 1.3 La reconnaissance d'entités nommées et l'IA

La deuxième approche repose sur des systèmes automatiques reposant sur l'entraînement de modèles à partir de données annotées en EN. Ces méthodes statistiques correspondent à des arbres de décisions, des modèles probabilistes comme les CRF<sup>33</sup>, des chaînes de Markov cachées, ainsi que des architectures d'apprentissage profond, notamment les réseaux de neurones (récurrents, à convolution, les réseaux *Long Short-Term Memory* ou LSTM, et les systèmes hybrides).<sup>34</sup>

Il est possible d'entraîner un modèle en produisant des données d'entraînement -la vérité de terrain- prenant la forme de données textuelles annotées en EN, où on attribue manuellement à un mot une étiquette. Ces modèles sont des algorithmes auxquels sont donnés des exemples de tâches à reproduire qui leurs permettent d'inférer des « règles basées sur des statistiques afin de pouvoir adhérer au mieux aux exemples qui lui ont été donnés. »<sup>35</sup> Les modèles se construisent et s'améliorent grâce aux exemples créés et fournis par l'humain, on parle ainsi d'apprentissage automatique supervisé. Y. Dupont soutient en outre que les méthodes par apprentissage automatique sont plus adaptées pour les tâches de REN grâce à leur capacité d'adaptation et à leur performance souvent meilleure que les systèmes à règles.<sup>36</sup>

Il est possible d'affiner le modèle de langue BERT pour obtenir un modèle de classification en vue d'une tâche spécifique, par exemple classifier une phrase selon le sentiment

---

30. D. Jurafsky et J. Martin, *Speech and Language Processing...*, p. 168

31. Y. Dupont, *La structuration dans les entités nommées...*, p. 53. Y. Dupont mentionne aussi deux autres outils basés sur un système de règles, Essex, et CasEN.

32. D. Jurafsky et J. Martin, *Speech and Language Processing...*, p. 168

33. Un CRF modélise « une distribution conditionnelle d'un ensemble structuré d'étiquettes par rapport à un ensemble d'objets en entrée. Y. Dupont, "Exploration de traits pour la reconnaissance d'entités nommées du Français par apprentissage automatique", dans *TALN 2017*, Orléans, France, 2017, URL : <https://hal.archives-ouvertes.fr/hal-02448614> (visité le 13/04/2021), p. 60 »

34. M. Ehrmann, *Les Entités Nommées, de la linguistique au TAL...*, p. 33 et Y. Dupont, *La structuration dans les entités nommées...*, pp. 57 - 84 « Les réseaux de neurones sont des modèles inspirés du neurone formel [...] ainsi que des théories connectionnistes [...] dont le principe est souvent résumé à 'les neurones qui s'activent en même temps, se lient entre eux', principe théorisant que, lorsqu'un cerveau reçoit un stimulus particulier ou effectue une tâche particulière, les neurones utilisés tendent à se regrouper entre eux. Dans le domaine des réseaux de neurones artificiels, cette théorie peut se reformuler en 'les neurones s'activant en même temps représentent une même fonction'. *Ibid.*, p. 67 »

35. *Ibid.*, p. 57

36. *Ibid.*, p. 160

exprimé, ou classifier des tokens dans un texte en leur attribuant une classe particulière selon leurs natures respectives et le contexte dans lequel ils se trouvent, dans le but d'effectuer de la reconnaissance d'entités nommées.

Les modèles de REN se basent donc sur ces deux approches, à partir de règles ou à partir d'apprentissage machine, mais il existe également des systèmes hybrides mêlant les deux.<sup>37</sup>

## 1.4 Évaluer un modèle de reconnaissance d'entités nommées : présentation des métriques

Les extractions d'EN obtenues avec les modèles de REN constituent les principales portes d'entrées pour les usagers des bibliothèques en ligne. Hamdi *et al.* indiquaient en 2019 que 80% des 500 requêtes faites par les utilisateurs d'une librairie en ligne contiennent des EN.<sup>38</sup> Il semble donc nécessaire d'être en mesure d'évaluer ces modèles afin d'utiliser ceux qui sont les plus performants. Pour la conférence CoNLL 2003, les chercheurs Erik F. Tjong Kim Sang et Fien de Meuder considéraient que les EN étaient "non-recursive" et "non-overlapping".<sup>39</sup> Par conséquent, la bonne identification d'une EN avec un modèle de classification repose sur la délimitation correcte d'une chaîne de caractères et la correcte attribution à l'EN de l'étiquette de la classe à laquelle elle appartient.<sup>40</sup>

Soit la phrase suivante : « Léon III l'Isaurien était un empereur byzantin, né à Germanicia, en Turquie actuelle. » La suite de caractères « Léon III l'Isaurien » correspond à une EN, de type « PER », que le modèle devrait pouvoir identifier en tant que telle.

Les modèles de classification sont majoritairement évalués à l'aide de trois métriques : la précision (*precision*), le rappel (*recall*), et le F-score.<sup>41</sup>

---

37. K. Adnan et R. Akbar notaient, entre autres, les systèmes « ChemSpot, a chemical hybrid system with CRF (Conditional Random Fields) and chemIDplus dictionary, SVM (Support Vector Machine) with CRF for biological entities with 91% accuracy, and a semantic and statistical model for medical entity recognition with semantic method MEta-Map, chunker-based noun phrase extraction, SVP, and supervised learning CRF. », etc. K. Adnan et R. Akbar, « Limitations of information extraction methods and techniques for heterogeneous unstructured big data »..., p. 5

38. Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidère, Mickaël Coustaty et Antoine Doucet, « An Analysis of the Performance of Named Entity Recognition over OCRed Documents », dans *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Issue : 1, Champaign, United States, 2019, t. 24, p. 333-334, DOI : 10.1109/JCDL.2019.00057, p. 333

39. Erik F. Tjong Kim Sang et Fien De Meulder, « Introduction to the CoNLL-2003 Shared Task : Language-Independent Named Entity Recognition », *Proceedings of CoNLL-2003* (, 12 juin 2003), arXiv : cs/0306050, URL : <http://arxiv.org/abs/cs/0306050> (visité le 19/07/2021), p. 2

40. "A named entity is correct only if it is an exact match of the corresponding entity in the data file." *Ibid.*, p. 3

41. Notons que ces trois métriques sont utilisées dans le cadre la conférence annuelle CoNLL. *Ibid.* Il existe d'autres façons d'évaluer un modèle de REN. Là où CoNLL évalue la REN au niveau du token, le chercheur David S. Batista, a présenté un système d'évaluation au niveau de l'EN entière. Voir David Batista, *Named-Entity evaluation metrics based on entity-level*, 9 mai 2018, URL : <http://www.davidsbatista.com>.

Celles-ci s'appuient sur quatre concepts : vrai positif, faux positif, vrai négatif et faux négatif. Reprenons la phrase suivante. Soit le résultat de REN suivant : « [Léon III l'Isaurien, PER] était un [empereur, PER] byzantin, né à [Germinicia, LOC], en Turquie actuelle. »

1. Un vrai positif (*true positive*) correspond à une séquence de caractères parfaitement délimitée et correctement classifiée par le modèle. Par exemple : [Léon III l'Isaurien, PER]
2. Un faux positif (*false positive*) correspond à une séquence de caractères mal délimitée mais correctement classifiée par le modèle. Par exemple : [Léon III, PER]. Les faux positifs représentent du bruit.<sup>42</sup>
3. Un vrai négatif (*true negative*) correspond à une séquence de caractères qui n'a pas été délimitée ni classifiée par le modèle car elle n'est pas une EN. Par exemple, le verbe « était ». Les vrais négatifs sont aussi définis comme du silence.<sup>43</sup>
4. Un faux négatif (*false negative*) correspond à une séquence de caractères qui n'a pas été délimitée ni classifiée par le modèle, alors qu'elle correspond à une EN. Par exemple : « Turquie ».

Dans le cas d'un modèle de classification binaire, dont la tâche consiste à déterminer si oui ou non un objet appartient à une classe, ces quatre concepts peuvent se représenter dans une matrice de confusion<sup>44</sup> illustrée dans la table 1.1.

	Prédiction positive	Prédiction négative
Classe positive	Vrai positif	Faux négatif
Classe négative	Faux positif	Vrai négatif

TABLE 1.1 – Exemple d'une matrice de confusion pour un modèle de classification binaire. Source : <https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/> (consulté le 10/08/21).

Erik F. Tjong Kim Sang et Fien de Meuder définissent la précision comme étant le pourcentage d'EN trouvées correctement étiquetées par le modèle.<sup>45</sup> Elle concerne seulement les vrais positifs. Elle se calcule de la façon suivante :

---

net/blog/2018/05/09/Named\_Entity\_Evaluation/ (visité le 23/04/2021)

42. Y. Dupont, *La structuration dans les entités nommées...*, p. 48

43. *Ibid.*

44. Une matrice de confusion est une matrice, un tableau d'éléments, « qui mesure la qualité d'un système de classification. » Voir [https://fr.wikipedia.org/wiki/Matrice\\_de\\_confusion](https://fr.wikipedia.org/wiki/Matrice_de_confusion) (consulté le 10/08/21).

45. "Precision : when model said 'positive' class, was it right?", voir <https://developers.google.com/machine-learning/crash-course/classification/video-lecture> (consulté le 10/08/21). Voir également D. Jurafsky et J. Martin, *Speech and Language Processing...*, p. 167 : "[...] precision is the ratio of the number of correctly labeled responses to the total labeled [...]"

$$\frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

Le rappel correspond au pourcentage d'EN présents dans le corpus, trouvés par le modèle.<sup>46</sup> Il prend donc en compte les vrais positifs et les faux négatifs, et permet de prendre la mesure du taux d'EN qui ont été manquées.<sup>47</sup> Il se calcule de la façon suivante :

$$\frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

Enfin, le F-score est le nom donné à la moyenne harmonique de la précision et du rappel. Un modèle parfait a un F-score de 1, le score maximal. Il se calcule ainsi<sup>48</sup> :

$$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

D. Jurafsky et J. Martin expliquent également qu'évaluer la REN peut poser problème à cause de la nature des EN, se définissant en partie comme nous l'avons vu par leur segmentation. Ils donnent l'exemple suivant : "For example, a system that labeled Jane but not Jane Villanueva as a person would cause two errors, a false positive for O and a false negative for I-PER."<sup>49</sup>

Dominique Stutzmann *et al.* soutiennent que la précision est un facteur dont l'importance doit être relativisée, un-e chercheur-e profiterait en effet de dizaines de milliers de pages traitées automatiquement et gagnerait du temps même si un travail de correction manuelle devait être réalisé. Le rappel est, à l'opposé, un facteur essentiel. Il permet d'atteindre l'exhaustivité et évite un silence généré accidentellement. Par exemple, une recherche pour un type d'acte n'apparaissant que quelques fois dans les actes des notaires du projet LECTAUREP soit faite, et qu'elle ne renvoie aucun résultat à cause d'une erreur d'extraction.<sup>50</sup>

---

46. E. F. T. K. Sang et F. De Meulder, « Introduction to the CoNLL-2003 Shared Task... », p. 3

47. "Recall : out of all the possible positives, how many did the model correctly identify?", voir <https://developers.google.com/machine-learning/crash-course/classification/video-lecture> (consulté le 10/08/21). Voir également D. Jurafsky et J. Martin, *Speech and Language Processing...*, p. 167 : "[...] recall is the ratio of the number of correctly labeled responses to the total that should have been labeled [...]"

48. Voir <https://deeppai.org/machine-learning-glossary-and-terms/f-score> (consulté le 10/08/21). Voir également *Ibid.* : "[...] F-measure is the harmonic mean of the two [precision and recall]."

49. *Ibid.*

50. Dominique Stutzmann, Jean-François Moufflet et Sébastien Hamel, « Full Text Search in Medieval Manuscripts : Issues and Perspectives of the HIMANIS Project for Electronic Publishing », *Medievales -Paris-*, 73–73 (déc. 2017), Publisher : Puv, p. 67-96, DOI : 10.4000/medievales.8198, p. 92 - 93

## 1.5 Un exemple de l'implémentation de la reconnaissance d'entités nommées dans des plate-formes de publications numériques

### 1.5.1 "A machine learning tool for fishing entities" : présentation d'*Entity-fishing*

*Entity-fishing* est un système de NERD *open-source* utilisant pour la tâche d'EL Wikipedia, déployée dans le cadre du projet européen DARIAH (*Digital Research Infrastructure for the Arts and Humanities*).<sup>51</sup> Cet outil permet de traiter des textes en italien, allemand, français et l'anglais grâce à une architecture d'apprentissage machine ("Random Forest and Gradient Tree Boosting"). Pour ces deux dernières langues, *Entity-fishing* repose, pour la REN, sur l'utilisation de GROBID-NER, lui-même basé sur une architecture de champs aléatoires conditionnels (CRF, pour *Conditional Random Fields*).<sup>52</sup> GROBID-NER est issu de la suite d'outils de *text mining* GROBID.<sup>53</sup> Il est possible d'essayer cette application sur une instance déployée par la TGIR Huma-Num.<sup>54</sup>

Cet outil procède selon trois étapes : d'abord, la langue est identifiée afin de sélectionner les outils adéquats pour le traitement du texte (modèle de tokenisation, modèle de segmentation des phrases) et pour sélectionner la base de données Wikipedia dans la bonne langue. Les mentions sont ensuite reconnues et enfin l'identification des EN par confrontation avec Wikipedia (génération de candidats, classement, et sélection du candidat approprié).<sup>55</sup>

Dans le contexte de LECTAUREP, l'utilisation d'*Entity-fishing* pourrait être envisagée pour détecter et désambiguïser les mots matières et les noms de rue.<sup>56</sup> Pour reprendre l'exemple de la « rue saint-Antoine », Entity-Fishing serait théoriquement capable de repérer cette EN et de trouver sa notice Wikipedia pour la désambiguïser.<sup>57</sup> Pour les noms de per-

51. L. Foppiano et L. Romary, « Entity-fishing : a DARIAH entity recognition and disambiguation service », *Journal of the Japanese Association for Digital Humanities*, 5–1 (nov. 2020), Publisher : Japanese Association for Digital Humanities, p. 22-60, DOI : 10.17928/jjadh.5.1\_22. Voir également la documentation de l'outil : <https://nerd.readthedocs.io/en/latest/> (consulté le 29/08/21)

52. <https://github.com/kermitt2/grobid-ner> (consulté le 29/08/21).

53. Voir <https://github.com/kermitt2/grobid> (consulté le 29/08/21). Notons également que pour la conférence CLEF HIPE 2020, *Entity-Fishing* a été utilisé conjointement avec l'architecture DeLFT (*Deep Learning Framework for Text* pour la tâche de REN. Tanti Kristanti et L. Romary, « DeLFT and entity-fishing : Tools for CLEF HIPE 2020 Shared Task », dans, 2020, t. 2696, URL : <https://hal.inria.fr/hal-02974946> (visité le 30/08/2021)

54. Voir <http://nerd.huma-num.fr/nerd/> (consulté le 29/08/21).

55. L. Foppiano et L. Romary, « Entity-fishing... » et T. Kristanti et L. Romary, « DeLFT and entity-fishing... »

56. « Un mot-matière est un descripteur issu d'un vocabulaire contrôlé (thesaurus ou liste d'autorité), qui sert à caractériser le contenu d'un document. » Voir <https://www.enssib.fr/services-et-ressources/questions-reponses/quest-ce-quun-mot-matiere> (consulté le 29/08/21)

57. Voir [https://fr.wikipedia.org/wiki/Rue\\_Saint-Antoine\\_\(Paris\)](https://fr.wikipedia.org/wiki/Rue_Saint-Antoine_(Paris)) (consulté le 29/08/21)

sonnes, « Entity-Fishing » ne semble pas approprié, étant donné que les répertoires des notaires concernent la majorité du temps d'illustres inconnus. Les types d'acte pourraient également poser problème, tous n'ayant certainement pas une notice Wikipedia.<sup>58</sup> C'est notamment le problème qui a été rencontré par Charles Riondet et Luca Foppiano dans l'exploitation d'écrits personnels de soldats français durant la seconde guerre mondiale. Certaines mentions rencontrées étaient issues d'un vocabulaire spécifique utilisé par les locuteurs français de cette époque. Par exemple le terme « souris grise », qui désignait les auxiliaires de l'armée allemande qui étaient des femmes. D'autres mentions étaient en outre issues d'une terminologie propre aux personnes ayant écrit ces journaux, par exemple le terme « cocs », un diminutif péjoratif pour « coco » désignant le parti communiste français.<sup>59</sup> Pour désambiguïser ces mentions, les chercheurs ont créé des dictionnaires spécifiques, remplaçant ainsi Wikipedia lorsque ces mentions problématiques étaient rencontrées. Il serait cependant envisageable de seulement utiliser le moteur de REN utilisé par cette application, GROBID-NER, pour le projet LECTAUREP.

### 1.5.2 Utiliser *Entity-Fishing* pour améliorer le processus de publication et la recherche dans des documents publiés

Dans le cadre du projet HIRMEOS (*High Integration of Research Monographs in the European Open Science infrastructure*), Andrea Bertino *et al.* ont travaillé sur l'implémentation d'*Entity-fishing* pour améliorer l'intégration et l'accès aux monographies de sciences humaines et sociales dans cinq plates-formes de publication numérique, à savoir OpenEdition Books, OAPEN Library, EKT Open Book Press, The Universität Göttingen et Ubiquity Press.<sup>60</sup> Celles-ci ont toutes en commun de proposer des textes en accès ouvert. Les chercheurs expliquaient qu'il était critique pour les sciences humaines et sociales de développer de meilleurs accès en ligne à ces documents car ils contribuent activement au développement des savoirs en SHS.<sup>61</sup>

Dans ces cinq plates-formes, *Entity-Fishing* a été intégré pour ajouter un système de recherches à facettes basée sur les EN des types personne et localisation. Les facettes agissent

---

58. J'ai eu l'occasion de tester Entity-Fishing grâce à son client python créé dans le contexte du projet HIRMEOS : <https://github.com/Hirmeos/entity-fishing-client-python>. Le test a été effectué sur des enregistrements des répertoires des notaires et se trouve à l'adresse suivante : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner\\_python/ner\\_entityfishing/entity-fishing\\_test.ipynb](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_entityfishing/entity-fishing_test.ipynb) (consultés le 29/08/21). Entity-Fishing prend notamment en entrée une requête structurée en JSON et la retourne enrichie en sortie avec une liste des entités identifiées et désambiguies. Après observation des résultats, on observe que l'outil est efficace dans l'identification et l'EL des adresses.

59. C. Riondet et L. Foppiano, « History Fishing When engineering meets History »..., p. 6

60. Andrea Bertino, L. Foppiano, L. Romary et Pierre Mounier, « Leveraging Concepts in Open Access Publications », *Journal of Data Mining and Digital Humanities*, 2019 (15 juin 2020), URL : <https://hal.inria.fr/hal-01981922> (visité le 30/04/2021)

61. *Ibid.*, p. 2

comme des filtres qui permettent de trier les documents et de cibler une recherche.<sup>62</sup> Les chercheurs expliquaient par ailleurs qu'il était envisageable de créer pour chaque plate-forme un nuage de mots-clés résultant de la REN et l'utiliser comme facette. La REN a également été utilisée pour permettre la visualisation des concepts directement dans les textes. Dans EKT, cette fonctionnalité est utilisée dans les résumés et les titres. Il est possible de consulter une définition des informations les plus pertinentes à partir des notices Wikipedia qui ont été récupérées avec *Entity-fishing*.<sup>63</sup> Ils soulignent qu'avoir accès aux textes annotés, et donc enrichis avec les EN, permet de mieux et plus rapidement comprendre les textes, et encourage les utilisateur-ices à se confronter à des monographies issues de domaines scientifiques qui leurs sont étrangers grâce à un accès à l'information fluidifiée.<sup>64</sup>

Dans leur article, les chercheurs ont également proposé des idées pour améliorer les services déjà existants. Il serait par exemple possible de créer un moteur de recherche utilisant pleinement le service de désambiguïsation d'« Entity-fishing », permettant ainsi à un-e utilisateur-ice de choisir le sens d'un terme utilisé dans une recherche. Avec les EN, les textes d'une même plate-forme pourraient également être regroupés selon leur contenu, de sorte à créer automatiquement des collections à consulter pour les utilisateurs.<sup>65</sup>

---

62. *Ibid.*, pp. 12 - 15

63. *Ibid.*, p. 15

64. *Ibid.*, p. 17

65. *Ibid.*, pp. 17 - 19



# **Chapitre 2**

## **Des modèles de reconnaissance d'écriture manuscrite visant la perfection : étude du matériau source et de la production des données du projet LECTAUREP**

Afin d'établir une stratégie pour effectuer une campagne d'extraction d'information, en l'occurrence de la REN, il est important d'avoir connaissance de la nature des données afin de pouvoir les exploiter au mieux. Ce chapitre s'attachera à présenter la structure des répertoires des notaires, la nature linguistique des informations qui y sont renseignées, et comment sont produites les transcriptions automatiques qui en sont faites.

### **2.1 La réforme du notariat de 1803 : le début d'un enregistrement standardisé de l'information des activités des notaires**

Les répertoires des notaires sont des documents dans lesquels les clercs inscrivaient chaque jour les actes enregistrés de façon chronologique. En 1803, le notariat est réformé avec la loi du 25 ventôse an XI. L'article 29 stipule que « les notaires tiendront répertoire de tous les actes qu'ils recevront », et l'article 30 formalise les répertoires en indiquant que « la date, la nature et l'espèce de l'acte, les noms des parties, et la relation de l'enregistrement »

102	26	Procuration	Kinaxier (bar François Rouchard, époux de Chabane) rue la Ferraria 13, p. 102 toussomme de la Caisse d'Epargne de Paris	27	Gratias
-----	----	-------------	--	----	---------

FIGURE 2.1 – Exemple d'un enregistrement complet dans une page d'un répertoire de notaire.

devront obligatoirement être indiqués.<sup>1</sup> C'est également à ce moment qu'un formulaire pré-imprimé à sept colonnes avec en-tête est conçu à cet effet (voir la figure 2). De gauche à droite, les colonnes contiennent les informations suivantes :

- Le numéro d'ordre de l'acte dans le répertoire.
- La date à laquelle l'acte a été passé.
- La troisième colonne, indiquant la nature de l'acte, est subdivisée en deux. Une première pour les actes passés en brevet, et une seconde pour les actes passés en minute.<sup>2</sup>
- Les noms prénoms et domiciles des parties, ainsi que d'autres informations, telles que des indications, des situations et des prix de biens.
- La dernière colonne indiquant la relation de l'enregistrement est également subdivisée en deux. La sixième colonne indique la date de l'enregistrement de l'acte, et la septième les droits payés par les parties.

Les informations inscrites par les notaires sont donc réparties horizontalement dans un tableau, dans les colonnes qui les concernent. On appelle donc un enregistrement la totalité des informations contenues dans une rangée horizontale de la structure tabulaire des répertoires formant une unité et concernant un acte passé par le notaire (voir figure 2.1).

La structure tabulaire des répertoires des notaires se comprend donc en croisant l'en-tête, qui donne verticalement la nature de l'information, avec les rangées horizontales où est contenue l'information elle-même.

## 2.2 La nature des données textuelles des répertoires de notaires : un langage spécialisé ?

### 2.2.1 Les types d'actes : mots et mots complexes

Pour désigner les types d'actes enregistrés, les notaires employaient dans la troisième ou quatrième colonne des mots dits simples, ou des mots composés, aussi appelés mots

1. M.L. Bonhomme, *Défis et opportunités de la reconnaissance automatique d'écriture manuscrite pour les documents d'archives : l'exemple des répertoires des notaires de Paris*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Jean-Baptiste Camps..., p. 15. Voir également la loi 25 ventôse an XI sur le site de Legifrance : <https://www.legifrance.gouv.fr/loda/id/LEGISCTA000006118859> (consulté le 26/08/21).

2. « On parle d'actes en minute par opposition aux actes en brevet : un acte en brevet est un acte notarié dont l'original est dépourvu de formule exécutoire, et est remis aux parties. » *Ibid.*

complexes, ayant une valeur juridique, et une valeur informationnelle pour les usagers des AN lors de la consultation de ces documents. Les linguistes Martin Riegel *et al.* définissaient en 2004 le mot simple comme étant « l'unité de base du système grammatical et dénominatif que forme la langue ».<sup>3</sup> Dans les répertoires des notaires, l'information définissant le type d'acte peut être une suite de caractères graphiques, les lettres, formant une unité linguistique : le mot « procuration », par exemple. Les notaires utilisent également des « mots complexes », qui forment une unité non pas avec un seul mot, mais en étant « constitués de deux (ou de plusieurs) mots ou morphèmes [...] ».<sup>4</sup> Par exemple, « certificat de propriété », ou « cahier de charges ».

Ils expliquaient également que les mots complexes peuvent être formés de quatre façons différentes dans la langue française ; parmi elles, la composition qui concerne les types d'actes sous forme de mots complexes.<sup>5</sup> Ils définissent ce mécanisme en ces mots :

« La composition proprement dite regroupe tous les mots composés dont les éléments sont des mots français qui ont une existence autonome par ailleurs : *portefeuille*, *porte-monnaie*, *chaise longue*, etc. Les éléments réunis dans un mot composé forment une unité de sens nouvelle, dont la signification dépasse celle de ses éléments pris isolément : une *chaise longue* n'est pas littéralement une "chaise qui est longue" mais, globalement, un "fauteuil pliable destiné au repos en position allongée". Les rapports sémantiques entre les éléments d'un mot composé sont variés : il existe entre deux éléments nominaux, des rapports attributifs comme dans *député-maire* ("le député est maire") ou des rapports de détermination comme dans *pomme de terre*. [...] »<sup>6</sup>

Chaque mot présent dans le type d'acte *contrat de mariage*, par exemple, peut en effet être compris de manière indépendante, mais c'est leur union par rapport de détermination qui donne le sens au mot complexe, et qui permet de désigner un acte d'une nature particulière. Selon les exemples contenus dans la précédente citation, ils mettent en avant trois cas de

3. Martin Riegel, Jean-Christophe Pellat et René Rioul, *Grammaire méthodique du français*, ISSN : 0291-0489, Paris, France, 2004, p. 531

4. « Les mots complexes sont constitués de deux (ou de plusieurs) mots ou morphèmes : dans le premier cas, il s'agit ou bien de syntagmes lexicalisés qui figent une construction syntaxique (un *fil de fer barbelé* [...]) ou bien de mots formés par composition (*bébé-éprouvette*, *aide-mémoire*) [...]. » *Ibid.*, p. 540. On trouve également en français l'appellation « expression polylexicale », qui désigne le même phénomène et qui est souvent utilisée dans le domaine du TAL. Voir Caroline Pasquer, « Expressions polylexicales verbales : étude de la variabilité en corpus », dans *TALN-RECITAL 2017*, Orléans, France, 2017, URL : <https://hal.archives-ouvertes.fr/hal-01637355> (visité le 27/07/2021). Les mots complexes, ou mots composés, sont nommés en anglais *multi-word expressions*, voir Francesca Masini, *Multi-Word Expressions and Morphology*, Oxford Research Encyclopedia of Linguistics, ISBN : 9780199384655, 30 sept. 2019, DOI : 10.1093/acrefore/9780199384655.013.611.

5. Les trois autres mécanismes pour former les mots complexes sont : la dérivation suffixale, la dérivation préfixale et la conversion. M. Riegel, J.C. Pellat et R. Rioul, *Grammaire méthodique du français...*, p. 540

6. *Ibid.*, p. 547

composition orthographique des mots complexes : la soudure graphique (portefeuille), la liaison des éléments par un trait d'union (porte-monnaie), et la séparation des éléments par un blanc graphique, qui concerne directement la composition des types d'actes.<sup>7</sup> Dans le contexte de l'extraction d'information, il est important d'avoir connaissance de l'existence de mots complexes dans le corpus, afin d'extraire justement les unités linguistiques qui font sens. De plus, comme le souligne la chercheuse Caroline Pasquer, des outils de TAL permettent de les détecter, notamment pour les traduire automatiquement, sans faire l'erreur de traduire chaque élément du mot complexe comme une unité indépendante.<sup>8</sup>

## 2.2.2 Une syntaxe particulière induite par le notariat et la structure tabulaire des répertoires

Les colonnes un, deux, trois, quatre, six et sept des répertoires des notaires ne possèdent majoritairement qu'une seule unité d'information, un mot simple, un mot complexe, ou une information numérique. La cinquième colonne échappe à ce phénomène. L'information qui y est indiquée prend en effet la forme d'une phrase nominale, possédant une syntaxe propre au corps professionnel que représente le notariat.<sup>9</sup> Riegel *et al.* définissent ce type de phrase ainsi :

« On appelle *phrase nominale* une phrase sans verbe, par opposition à une phrase verbale. Cette phrase peut être déclarative (*Fin de l'épisode*), interrogrative (*Les toilettes ?*), ou impérative (*Vos papiers !*). [...] L'absence de verbe prive la phrase nominale du terme qui assure normalement la prédication et l'ancre situationnel. [...] En ce qui concerne l'ancre situationnel, E. Benveniste considère que la phrase nominale constitue une assertion "intemporelle, impersonnelle, non modale", ce qui la rend apte à exprimer une "vérité générale" [...]. Mais on peut observer qu'en l'absence de verbe, la phrase nominale est avant tout sensible aux variations de la situation d'énonciation particulière. [...] Pour la prédication, l'absence de verbe n'implique pas automatiquement l'absence de prédicat : le rôle de verbe est assuré par d'autres moyens. »<sup>10</sup>

Dans les phrases nominales de la cinquième colonne, il manque donc un prédicat, habituellement assuré par un verbe dans les phrases verbales. Le prédicat est un concept linguistique qui se définit comme une notion logique « souvent utilisée dans la représentation sémantique

7. *Ibid.*, p. 549

8. C. Pasquer, « Expressions polylexicales verbales... », pp. 1- 2

9. Une phrase est une « entité structurale abstraite que l'on peut caractériser par un ensemble de règles de bonne formation phonologique, morphologique et sémantique [...] » M. Riegel, J.C. Pellat et R. Rioul, *Grammaire méthodique du français...*, p. 26. La syntaxe « décrit la façon dont les mots se combinent pour former des mots et des phrases [...] » *Ibid.*, p. 22.

10. *Ibid.*, p. 457 - 458

des phrases pour symboliser la contrepartie relationnelle de leur verbe ou de leur attribut. »<sup>11</sup> En d'autres termes, c'est autour du prédicat que s'articule le sens d'une phrase et qu'on la comprend. L'information contenue dans un enregistrement est compréhensible autrement. Premièrement, c'est grâce à sa répartition dans la structure tabulaire des répertoires, dont l'en-tête indique pour chaque colonne sa nature, que l'enregistrement peut être appréhendé. Deuxièmement, le texte de la colonne 5 ne se comprend que par le prédicat présent dans la colonne 3 ou 4 où se situe donc cet « ancrage situationnel », qui est lui-même un mot ou un mot complexe, et qui induit une structure syntaxique qui lui est spécifique. C'est avec cette information que se comprend la cinquième colonne, comme on peut l'observer dans la variation de sa syntaxe en fonction du type d'acte (ou pour reprendre les mots de la citation précédente, dans la « variation de la situation d'énonciation ») (voir annexes B.1, B.2, B.3, B.4, B.5). Ces différentes syntaxes pourraient constituer des unités terminologiques d'une « langue de spécialité », ou langue spécialisée.<sup>12</sup> Comme le définit le linguiste Ross Charnock, « on parle de langue de spécialité lorsqu'il s'agit de se servir d'une langue naturelle (la langue de référence) pour rendre compte de connaissances particulières. »<sup>13</sup> Elle est un sous-ensemble de la langue de référence, ici le français, et est employée par des spécialistes la partageant, en l'occurrence, les notaires.

Grâce à la structure tabulaire, l'information est en ce sens normalisée, le langage n'est pas enclin à être modifié selon les différents notaires, et elle est inscrite beaucoup plus rapidement qu'avec une phrase verbale. Le langage spécialisé utilisé par les notaires convient ainsi à la tâche répétitive qui consistait pour les notaires à tenir quotidiennement ces documents à jour.

## 2.3 Les objectifs du projet LECTAUREP pour la reconnaissance d'entités nommées

Suite à un entretien avec Mme Aurélia Rostaing, archiviste paléographe, conservatrice du patrimoine et responsable du pôle instrument de recherche au DMC, une liste d'EN à extraire à partir des répertoires des notaires a été dressée :

- La date de l'acte et de son enregistrement (deuxième colonne)
- Les types d'actes (troisième et quatrième colonne)
- Dans la cinquième colonne :
  - Les noms des parties

---

11. *Ibid.*, p. 128

12. Ross Charnock, « Les langues de spécialité et le langage technique : considérations didactiques », *ASp. la revue du GERAIS-23* (1<sup>er</sup> déc. 1999), Number : 23-26 Publisher : Groupe d'étude et de recherche en anglais de spécialité, p. 281-302, DOI : 10.4000/asp.2566

13. *Ibid.*

- Les professions
- Les adresses et les lieux
- Les sommes d'argent
- Les mots matières pour les noms de société (par exemple, « cinéma », « assurances », etc.)

Cette liste permet d'établir les objectifs du projet en termes de REN, et de faire ressortir plusieurs difficultés et stratégies envisageables. Pour certaines EN, comme les types d'actes, il serait possible d'utiliser la structure tabulaire des répertoires afin de les cibler. Cependant, cette information sémantique doit être présente à l'issue de la REM, notamment grâce à la segmentation sémantique des régions de la page. Par ailleurs, la segmentation des colonnes des répertoires des notaires permettrait de reconstituer la structure logique du document après la REM, et ainsi de localiser l'information et de connaître sa nature sans avoir à la lire. Dans la cinquième colonne, là où il y a des phrases, un modèle de REN pourrait être appliqué. En effet, on peut distinguer statistiquement les noms de personnes des noms de société grâce au nombre de tokens, souvent plus nombreux pour ces derniers. De plus, les sociétés sont habituellement annoncées au début d'un enregistrement, par le mot « Société », ou l'abréviation qui lui correspond, donnant ainsi à un système de REN un élément structurant ou un indice contextuel sur lequel le modèle peut s'appuyer.<sup>14</sup> En termes de difficultés, reconstituer le nom complet (prénom(s) et nom de famille) des parties concernées par un enregistrement pourrait poser des problèmes. Généralement, lorsqu'un enregistrement concerne une personne, les deux premières EN de la cinquième colonne sont respectivement le nom de famille et le(s) prénom(s). Il suffirait ainsi de concaténer ces deux EN. Cependant, avant d'adopter cette solution, il faudrait s'assurer que cette observation est une règle, et qu'il n'y pas ou très peu d'exceptions. Dans le cas contraire, la REN risquerait de créer des personnes n'ayant jamais existé en associant un prénom à un nom de famille qui ne lui appartient pas.

## 2.4 Quelles solutions technologiques pour une campagne de reconnaissance d'écriture manuscrite réalisée dans le cadre d'un projet d'humanités numériques ?

### 2.4.1 La reconnaissance d'écriture manuscrite : définition

Les transcriptions automatiques produites dans le cadre du projet LECTAUREP sont donc obtenues grâce à la reconnaissance d'écriture manuscrite. Cette technologie permet

---

14. Y. Dupont, *La structuration dans les entités nommées...*, p. 21

d'obtenir automatiquement à partir d'une image une transcription du texte manuscrit qu'elle contient. En 2021, on compte notamment trois outils de REM état de l'art, à savoir Kraken, Tesseract 4 et Transkribus.<sup>15</sup> Ces systèmes fonctionnent en segmentant les lignes d'un texte (*layout analysis*) et en les transcrivant une par une grâce à l'utilisation d'un réseau de neurones récurrents ou hybrides.<sup>16</sup> En outre, ces réseaux sont préalablement entraînés sur des vérités de terrain correspondant au type d'écriture spécifique ou à une main particulière rencontré dans le texte sur lequel on souhaite utiliser la REM. Pour cela, la vérité de terrain prend la forme d'une transcription réalisée manuellement et alignée sur l'image qui lui correspond.

Günter Mühlberger *et al.* soutenaient en 2018 que la REM avait la capacité d'impacter durablement les missions d'archivage et la recherche en rendant la lecture, la transcription et l'exploitation des documents historiques plus simple et plus rapide. Ils estimaient également que cette technologie a le potentiel de se développer rapidement dans le milieu des humanités numériques grâce aux corpus de numérisations de documents patrimoniaux déjà disponibles et leur nombre grandissant.<sup>17</sup>

## 2.4.2 La démocratisation de la REM avec Transkribus

La REM est une technologie qui s'est démocratisée ces dernières années, comme en témoignent nombreux de projets l'ayant implanté dans leurs chaînes de traitement. Nous pouvons par exemple mentionner le projet ANR de la Bibliothèque Foucaldienne, clôturé en 2020, qui visait à rendre accessible aux communautés des SHS les carnets de note du philosophe Michel Foucault grâce à leur transcription avec le logiciel de REM Transkribus.<sup>18</sup> Chaque transcription a été exportée en TEI, un instrument de recherche en EAD a été produit pour faciliter leur consultation, et une interface permet aujourd'hui de les consulter. Le projet visait entre autres à mettre en avant dans les transcriptions les noms de personnes et noms d'oeuvres et à les relier à des informations biographiques et bibliographiques, afin de rattacher les travaux de Foucault à ses sources. Chaque EN a été pour cela annotée manuellement pour

15. Peter A. Stokes, B. Kiessling, D. Stökl Ben Ezra, R. Tissot et El Hassane Gargem, « The eScriptorium VRE for Manuscript Cultures, in Ancient Manuscripts and Virtual Research Environments », *Classic @ Journal*–18 (2021), URL : <https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/>

16. Cette technique est notamment différente des systèmes d'OCR, qui transcrivent lettre par lettre. *Ibid.*

17. Guenter Muehlberger, Louise Seaward, Melissa Terras, Oliveira Sofia Ares, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, H. Déjean, Markus Diem, Stefan Fiel, *et al.*, « Transforming scholarship in the archives through handwritten text recognition : Transkribus as a case study », *Journal of Documentation*, 75–5 (1<sup>er</sup> janv. 2019), Publisher : Emerald Publishing Limited, p. 954-976, DOI : 10.1108/JD-07-2018-0114, p. 955

18. Voir Marie-Laure Massot, Arianna Sforzini et Vincent Ventresque, « Transcribing Foucault's handwriting with Transkribus », *Journal of Data Mining and Digital Humanities*, Atelier Digit\_Hum (mars 2019), Publisher : Episciences.org, URL : <https://hal.archives-ouvertes.fr/hal-01913435> (visité le 07/04/2021) et <http://lbf-ehess.ens-lyon.fr/pages/infos.html> (consulté le 30/08/21).

une partie du corpus, et le reste a été annoté par un système de REN.<sup>19</sup>

Transkribus est une solution logicielle pour implémenter un système de REN mis à disposition du public depuis 2015, et financé par le programme-cadre pour la recherche et le développement technologique de la Commission européenne.<sup>20</sup> Le logiciel utilise des modèles basés sur des architectures d'apprentissage profond. Il propose une interface graphique pour produire des vérités de terrain à partir d'images, pour l'entraînement de modèles, et pour la production de transcriptions automatiques et leur exploitation. Transkribus propose une plate-forme conçue pour les communautés archivistiques, des SHS, des ingénieurs en informatique et pour le public afin de traiter les documents historiques. Cette solution pouvait être utilisée gratuitement par des projets et avait l'avantage de proposer un environnement pour travailler collaborativement. Outre la bibliothèque Foucaldienne, d'autres projets de REM ont vu le jour grâce à Transkribus, comme la transcription des carnets du juriste Eugène Wilhelm (1866 - 1951), menée par le chercheur Régis Schlagdenhauffen en 2020.<sup>21</sup> Notons également l'analyse stylométrique des textes attribués à l'écrivain Robert Musil (1880 - 1942), rendue possible grâce à l'utilisation de ce logiciel pour produire des transcriptions automatiques et menée par Simone Rebora en 2019.<sup>22</sup>

Cependant, Transkribus a aujourd'hui changé de modèle économique, rendant son service payant. LECTAUREP avait initialement envisagé d'utiliser Transkribus, mais cette décision a encouragé le projet à se tourner vers la solution de REM *open-source* nommée eScriptorium.

#### **2.4.3 Une interface open-source pour la transcription et l'entraînement de modèles de REM : eScriptorium**

eScriptorium est un VRE (*Virtual Research Environment*, ou environnement de recherche virtuel, c'est-à-dire une application en ligne qui permet à des chercheurs de collaborer, et qui propose un environnement ergonomique pour utiliser un outil de REM).<sup>23</sup> Elle est développée par l'équipe Scripta, à l'École Pratique des Hautes Études et à l'Université Paris Science et

---

19. *Ibid.*, pp. 7 - 8

20. G. Muehlberger, L. Seaward, M. Terras, *et al.*, « Transforming scholarship in the archives through handwritten text recognition... », p. 957. Voir également le site internet de Transkribus : <https://readcoop.eu/transkribus/> (consulté le 30/08/21).

21. Régis Schlagdenhauffen, « Optical Recognition Assisted Transcription with Transkribus : The Experiment concerning Eugène Wilhelm's Personal Diary (1885-1951) », *Journal of Data Mining and Digital Humanities*, Atelier Digit\_Hum (août 2020), Publisher : Episciences.org, URL : <https://hal.archives-ouvertes.fr/hal-02520508> (visité le 07/04/2021)

22. Simone Rebora, « A Digital Edition between Styliometry and OCR : The Klagenfurter Ausgabe of Robert Musil », *Textual Cultures*, 12-2 (2019), Publisher : [Society for Textual Scholarship, Indiana University Press], p. 71-90, URL : <https://www.jstor.org/stable/26821537> (visité le 08/04/2021)

23. P. A. Stokes, B. Kiessling, D. Stökl Ben Ezra, *et al.*, « The eScriptorium VRE for Manuscript Cultures, in Ancient Manuscripts and Virtual Research Environments »... Voir également B. Kiessling, R. Tissot, P. Stokes, *et al.*, « eScriptorium... »

Lettres (EPHE – PSL) et propose une interface graphique web et *open-source* au moteur de REM Kraken. Celui-ci est développé par Benjamin Kiessling, qui travaille également à l'EPHE-PSL.<sup>24</sup>

L'application permet de créer des collections, dans lesquelles peuvent être importées des images. Pour obtenir une transcription automatique, eScriptorium leur applique un modèle de segmentation générique ou affiné sur un type de données pour segmenter en priorité les lignes de texte (*baselines*), puis les régions, c'est-à-dire les zones de texte d'un document. Pour transcrire, Kraken n'a besoin que des *baselines*, mais les régions permettent notamment de conserver leur localisation dans l'image, sous forme de coordonnées  $x,y$ . Il est possible de contrôler manuellement les lignes de texte. Ensuite, appliquer un modèle de REM permet d'obtenir pour chaque image une transcription automatique, qu'il est également possible de corriger directement dans l'application. Enfin, eScriptorium compile les lignes de texte pour chaque image, et permet d'exporter le résultat.

Il est possible d'exporter les transcriptions depuis eScriptorium en plein texte, et dans deux formats XML standardisés : ALTO et PAGE. Ces deux derniers servent à stocker la mise en page d'un document et du texte obtenu avec de l'OCR ou de la REM qui en est issu.

Le format ALTO est un standard de la Bibliothèque du Congrès. Il possède trois balises parents : <Description>, indique les métadonnées du fichier XML, comme l'unité de mesure (habituellement en pixels), et le nom du fichier ; <tag>, qui indique les étiquettes des régions de la page et des *baselines*, si elles en ont ; et <layout>, qui contient autant de <TextBlock> qu'il y a de régions dans la segmentation. On trouve à l'intérieur de ces balises toutes les *baselines* qui leurs sont associées et qui sont représentées avec une balise <TextLine>, contenant entre autres les informations des coordonnées sur la page et la chaîne de caractères issue de la transcription (voir exemple de structure ALTO dans l'annexe C.1).<sup>25</sup>

Le format PAGE est différent de l'ALTO dans sa façon de présenter l'information. Il a été créé par l'université de Salford, à Manchester en 2010. Les métadonnées sont stockées dans une balise parent <Metadata>, et les transcriptions dans une balise <Page>. Il n'y a pas d'élément équivalent à la balise <tag> de l'ALTO. Dans la balise <Page>, il y a le même nombre de <TextRegion> qu'il y a de régions dans la segmentation. Un attribut *custom* vient indiquer son étiquette, si la région en a une. Les balises enfants <TextLine> représentent les *baselines*. Les transcriptions sont respectivement stockées sous forme de chaîne de caractères unicode dans une balise <TextEquiv> (voir exemple de structure PAGE dans l'annexe C.2).<sup>26</sup>

---

24. B. Kiessling, *Kraken - an Universal Text Recognizer for the Humanities...* Kraken a notamment été créé pour pouvoir traiter des alphabets non latins (comme l'arabe, le perse, le syriaque, le grec, l'hébreu, etc.), tout en ayant la capacité de traiter les alphabets latins.

25. Voir <https://www.loc.gov/standards/alto/description.html>. Voir également la structure d'un fichier ALTO à l'adresse suivante : <http://www.loc.gov/standards/alto/techcenter/structure.html> (consultés le 28/08/21).

26. Le format PAGE a de plus été conçu pour stocker les différentes métadonnées issues de la chaîne

Exporter une transcription au format plein texte implique la perte des informations de localisation du texte sur la page. Or, comme nous l'avons vu, cette donnée est essentielle pour comprendre la nature de l'information, notamment dans le cas d'une structure tabulaire. Ces deux formats XML servent donc de socle au post traitement des transcriptions automatiques et à leur exploitation. Ils permettent également d'établir une stratégie d'extraction d'information à partir de la segmentation sémantique et de l'annotation des *baselines* effectuée sur eScriptorium.

Ce VRE permet également d'entraîner des modèles de REM à partir de transcriptions manuelles produites depuis les images importées dans la plate-forme. L'alignement du texte par rapport à l'image se fait grâce aux *baselines*, auxquelles sont associées chaque transcription. Une fois un volume de données suffisant atteint, il est possible d'entraîner ou d'affiner un modèle de REM à partir de la vérité de terrain créée, et d'appliquer ensuite ce modèle sur d'autres images. Il est également possible d'entraîner ou d'affiner un modèle de segmentation depuis eScriptorium en créant des données d'entraînement, par exemple en annotant les régions de chaque image et en leur attribuant une classe sémantique, de sorte à pouvoir effectuer automatiquement ce type d'annotation sur d'autres images.

C'est donc sur eScriptorium que se déroulent les étapes de création de terrain, d'entraînement, de transcription automatique et d'export, tout en permettant à ALMAaCH et le DMC de travailler collaborativement.

## 2.5 Transcrire pour entraîner des modèles de reconnaissance d'écriture manuscrite

L'entraînement de modèle de REM se fait à partir de transcriptions manuelles alignées par rapport aux images auxquelles elles correspondent. Avec la démocratisation de la REM ces dernières années et l'augmentation du nombre de projets en humanités numériques employant cette technologie, nous voyons l'émergence de vérités de terrain créés à partir de domaines différents, ce qui permet d'accélérer la création de modèles de transcription. En théorie, un projet pourrait puiser dans des sets de données semblables au domaine des documents à transcrire, réduisant ainsi le nombre de transcriptions qu'il faudrait produire pour entraîner un nouveau modèle. Avec un modèle de REM du même domaine, un projet pourrait ensuite affiner celui-ci pour le spécialiser sur ses données. Un modèle affiné (*fine-tuned*) résulte du ré-entraînement d'un premier modèle à partir d'un autre jeu de données, avec comme objectif

---

de traitement pour l'analyse des images, comme la binarisation de l'image. Stefan Pletschacher et Apostolos Antonacopoulos, « The PAGE (Page Analysis and Ground-Truth Elements) Format Framework », dans *2010 20th International Conference on Pattern Recognition*, Istanbul, Turkey, 2010, p. 257-260, DOI : 10.1109/ICPR.2010.72. C'est avec ce format que j'ai travaillé durant le stage.

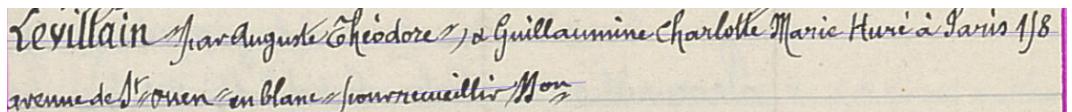


FIGURE 2.2 – Exemple d'un enregistrement avec l'abréviation « St// Ouen »

de le spécialiser dans une tâche spécifique et d'améliorer ses performances. Christian M. Dahl *et al.* ont par exemple publié au début de l'année 2021 un corpus nommé « HANA » (pour *HAndwritten NAme Database*), basé sur la transcription des noms contenus dans les formulaires pré-imprimés des registres de police danois datés de 1890 à 1923.<sup>27</sup> Il a été produit dans l'objectif de fournir des données d'entraînement basées sur des documents anciens, ainsi que pour servir à l'évaluation de modèles. Mentionnons également le corpus « OCR17 », créé à partir de documents imprimés du XVII<sup>e</sup> siècle par Simon Gabay *et al.*<sup>28</sup> Dans cette lignée, et dans le but de centraliser ces corpus, Mme Alix Chagué a créé une organisation Github pour constituer un catalogue de vérités de terrain et de modèles d'OCR et de REM, nommé « HTR-United ».<sup>29</sup> Le projet LECTAUREP participe également à cet effort commun dans la création de vérités de terrain à partir des répertoires des notaires.

Les données produites par un modèle de REM dépendent des règles d'annotation qui ont été adoptées pour transcrire les documents source.<sup>30</sup> Il est important d'en avoir connaissance afin d'appréhender au mieux l'exploitation de ces données textuelles. Il est utile de lister les choix de convention qui ont été adoptés pour la constitution de la vérité de terrain et qui peuvent avoir un effet direct dans l'exploitation des données produites automatiquement car se répercutant dans le texte produit :

- Dans le cas d'une écriture dense, donc avec des mots écrits de façon agglutiné, il a été décidé de transcrire sans coupure.
- Les abréviations n'ont pas toujours été transcrrites de la même façon : on trouve selon l'abréviation utilisée par le notaire « St// Ouen » ou « St Antoine », par exemple.

27. Christian M. Dahl, Torben Johansen, Emil N. Sørensen et Simon Wittrock, « HANA : A HAndwritten NAme Database for Offline Handwritten Text Recognition », *arXiv :2101.10862 [cs, econ]* (, 22 janv. 2021), arXiv : 2101.10862, URL : <http://arxiv.org/abs/2101.10862> (visité le 14/04/2021) HANA concentre 3 355 388 noms, distribués sur 1 106 020 images.

28. Simon Gabay, T. Clérice et Christian Reul, *OCR17 : Ground Truth and Models for 17th c. French Prints (and hopefully more)*, mai 2020, URL : <https://hal.archives-ouvertes.fr/hal-02577236> (visité le 19/07/2021) OCR17 propose une vérité de terrain composée de 30 000 lignes transcrrites à partir de 37 documents imprimés en français du XVII<sup>e</sup> siècle.

29. Voir <https://htr-united.github.io/> (consulté le 27/08/21).

30. A. Chagué, *Comment faire lire des gribouillis à mon ordinateur ?*, 2021, URL : <https://hal.archives-ouvertes.fr/hal-03170345> (visité le 06/04/2021) Les règles d'annotation servent par ailleurs à « expliciter les règles des cas 'normaux' ; définir les règles précises et univoques pour les cas particuliers ; définir des comportements pour les cas 'anormaux' imprévus ».

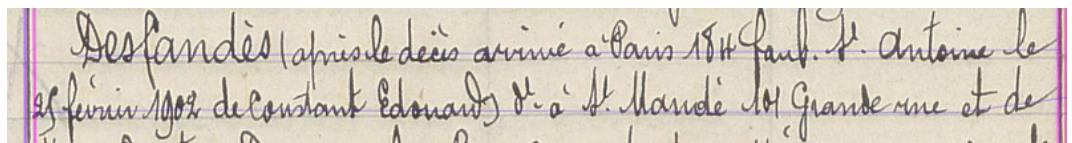


FIGURE 2.3 – Exemple d'un enregistrement avec l'abréviation « St Antoine »

## 2.6 Analyse des types d'erreurs générées par la reconnaissance d'écriture manuscrite

Enfin, avoir une bonne connaissance des erreurs typiques produites par la REM permet d'envisager des post traitements de correction pour préparer l'extraction d'information. D'après Thi-Tuyet-Hai Nguyen *et al.*, il existe plusieurs types communs d'erreurs d'OCR, qui peuvent être transposées à la REM.<sup>31</sup> En se basant sur la distance d'édition, c'est-à-dire le nombre d'opérations à effectuer pour corriger un token, on trouve les erreurs simples (*single-errors*), « noteire » au lieu de « notaire », et les erreurs multiples (*multi-errors*), par exemple « noteira ». Ces erreurs peuvent se retrouver à n'importe quelle place dans le token. Ensuite, on distingue les erreurs de mots non-réels (*non-word error*) et les erreurs de mots réels (*real-word error*). Dans le premier cas, l'erreur transforme le mot en un mot qui n'appartient à aucun lexique, dans le deuxième, le mot prend un autre sens à cause de l'erreur : « contrer », alors que dans la réalité de terrain on trouve « contrat », par exemple. Ce dernier type peut engendrer des contresens dans la transcription automatique. Enfin, on trouve également les erreurs de segmentation des mots, par exemple quand un espace a été ajouté dans un token, ou quand un espace a été supprimé entre deux tokens.

Pour corriger ces erreurs, quatre types d'opérations peuvent être réalisées au niveau du caractère : la suppression, l'insertion, la substitution d'un caractère par un autre et la transposition, correspondant à l'échange de deux caractères.

31. Thi-Tuyet-Hai Nguyen, Adam Jatowt, M. Coustaty, Nhu-Van Nguyen et A. Doucet, « Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing », dans *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Champaign, France, 2019, p. 29-38, DOI : 10.1109/jcdl.2019.00015, p. 30

## 2.7 Quel outil pour évaluer les données produites dans le cadre de LECTAUREP ?

### 2.7.1 Un outil développé dans l'équipe ALMAaCH : la librairie python KaMI

Afin de doter le projet LECTAUREP d'un outil de contrôle de la qualité des transcriptions produites par les modèles de REM, M. Lucas Terriel a développé une librairie python au cours de son stage au sein d'ALMAaCH en 2020, nommée « Kraken-benchmark », devenu par la suite « KaMI » et dont le développement continu avec Mme Alix Chagué.<sup>32</sup> La librairie permet de comparer deux chaînes de caractères et d'obtenir des métriques pour prendre la mesure des taux d'erreur d'un modèle de REM. KaMI est un outil que j'ai régulièrement utilisé au cours du stage pour constituer des données de test en fonction de la qualité de la transcription automatique.

### 2.7.2 Définition de trois métriques d'évaluation de la REM : la distance de Levenshtein, le CER, et le WER.

Pour réaliser cette évaluation, KaMI s'appuie entre autres sur trois métriques : la distance de Levenshtein, le taux d'erreur de caractères ou CER (pour *Character Error Rate*) et le taux d'erreur de mots ou WER (pour *Word Error Rate*).

La distance de Levenshtein, aussi nommée distance d'édition, est une distance mathématique qui compare deux chaînes de caractères de longueurs différentes. Elle permet d'obtenir « le coût minimal de transformation d'une chaîne de caractères  $R$  en une chaîne de caractère  $P$  » en fonction des opérations de substitution, d'insertion et de suppression, auxquelles est associé un coût de 1.<sup>33</sup> Pour reprendre l'exemple de L. Terriel utilisé dans son mémoire, la distance de Levenshtein entre le mot « magasin » et « megasinier » est de 4 « car il y a eu 3 insertions 'ier' et une substitution 'a' en 'e'. »<sup>34</sup>

Le CER et le WER se calcule à partir de 4 éléments.<sup>35</sup> Soit  $N$  « le nombre total de caractères ou de mots contenus dans la phrase de référence »,  $S$  le nombre de substitutions,  $D$  le nombre de suppressions, et  $I$  le nombre d'insertions, on obtient pour le CER :

---

32. Voir <https://gitlab.inria.fr/dh-projects/kraken-benchmark/-/tree/master/> et <https://gitlab.inria.fr/dh-projects/kami/kami-lib> (consultés le 28/08/21).

33. L. Terriel, *Représenter et évaluer les données issues du traitement automatique d'un corpus de documents historiques. L'exemple de la reconnaissance des écritures manuscrites dans les répertoires de notaires du projet LectAuRep., mémoire de master « Technologies numériques appliquées à l'histoire », dir. Alix Chagué et Thibault Clérice..., p. 117*

34. *Ibid.*

35. Les équations et l'explication de ces métriques sont citées à partir du mémoire de L. Terriel. Voir *Ibid.*, p. 118

$$\frac{S + D + I}{\text{Nombre total de caractères}}$$

Et pour le WER :

$$\frac{S + D + I}{\text{Nombre total de mots}}$$

La distance de Levenshtein peut être utilisée pour calculer plus rapidement ces taux. Soit  $D$  la distance de Levenshtein,  $R$  la phrase de référence et  $H$  la phrase de prédiction, on obtient pour le CER :

$$\frac{D(R, H)}{\text{Nombre total de caractères}}$$

Et pour le WER :

$$\frac{D(R, H)}{\text{Nombre total de mots}}$$

Avec ces deux taux, on arrive à prendre la mesure de la performance d'un modèle de REM. Plus il s'approche de 1, plus le modèle de REM a fait d'erreurs dans la transcription, plus il s'approche de 0, plus la transcription est satisfaisante. M.-L. Bonhomme indiquait, dans son mémoire, que les modèles de REM sont considérés comme performants lorsque le CER est inférieur à 10%, et comme excellents lorsqu'il est situé aux alentours de 5%.<sup>36</sup> Aujourd'hui, les deux modèles mixtes, donc destinés à être robustes à plusieurs mains<sup>37</sup>, entraînés pour le projet LECTAUREP atteignent respectivement environs 8 à 9% de CER.<sup>38</sup>

---

36. M.L. Bonhomme, *Défis et opportunités de la reconnaissance automatique d'écriture manuscrite pour les documents d'archives : l'exemple des répertoires des notaires de Paris*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Jean-Baptiste Camps..., p. 7

37. A. Chagué, *Création de modèles de transcription pour le projet LECTAUREP #1*, LECTAUREP, 6 sept. 2021, URL : <https://lectaurep.hypotheses.org/475> (visité le 28/08/2021)

38. Voir <https://gitlab.inria.fr/dh-projects/kraken-models/-/issues/17> et <https://gitlab.inria.fr/dh-projects/kraken-models/-/issues/8> (consultés le 27/08/21). Ce sont ces deux modèles qui ont été employés pour constituer des données de test lors de mon stage pour la tâche de REN.

# Chapitre 3

## Autres méthodes de recherche dans un document patrimonial

Nous avons présenté la REN comme une méthode pour accéder au texte sans avoir à le lire dans son intégralité, grâce aux EN qui lui servent de portes d'entrée. Cependant, il existe d'autres méthodes pour naviguer dans des textes que nous présenterons ici.

### 3.1 Les moteurs de recherche : recherches plein texte et recherches floues

La REN implique plusieurs étapes, notamment la reconnaissance en elle-même, puis le stockage des EN et leur exploitation. Elle n'est cependant pas le seul moyen d'accès à des données textuelles. Les moteurs de recherche peuvent constituer une solution différente, mais présentent le désavantage de ne pas stocker les EN, rendant ainsi le *data mining* plus difficile qu'avec la REN.

#### 3.1.1 Une première exploration du texte par la recherche plein texte

Pour commencer, nous pouvons mentionner la recherche plein texte. Cette fonctionnalité permet de chercher une chaîne de caractères dans un texte, et de retrouver son ou ses occurrences dans celui-ci. La recherche plein texte fonctionne très bien pour rechercher rapidement et efficacement des termes plus ou moins complexes dans des textes propres, par exemple des publications scientifiques. Elle devient moins robuste quand on l'applique à des textes bruités car ne permet pas de trouver autre chose que la chaîne de caractère exacte recherchée.

### 3.1.2 Un système de recherche flexible : les recherches floues

Plusieurs moteurs de recherche ont adopté un système de recherche floue (ou *fuzzy search*), pour être plus robustes face au bruit, notamment sur internet. Ce type de recherche permet de retourner, dans le contexte d'une recherche au sein d'une collection par exemple, les documents contenant des termes similaires au(x) terme(s) de recherche en utilisant la distance de Levenshtein.<sup>1</sup> Dans le cas des transcriptions des répertoires des notaires, un-e utilisateur-e pourrait chercher le terme « donation », et le système serait capable de lui renvoyer les documents contenant ce terme exact, mais aussi contenant des termes différents avec une distance d'édition de 1, « domation », de 2, « domalion », etc., selon son paramétrage.

Notons que cette fonctionnalité existe sur Transkribus.<sup>2</sup> Néanmoins, face au changement des conditions d'utilisation de cette plate-forme et de sa transformation en un service payant, certains projets de REM ont décidé d'implémenter cette solution différemment. C'est le cas du projet de la Bibliothèque Foucaldienne qui utilise la plate-forme Omeka/eman pour avoir un moteur de recherche floue, en complément de l'annotation des EN à l'aide d'un système de REN.<sup>3</sup>

## 3.2 Le *keyword spotting* : une affaire de vision par ordinateur

Il existe également une technologie permettant d'exploiter des documents image contenant du texte sans en produire une transcription, nommée *keyword spotting* (KWS). Le KWS est d'abord issu de travaux de recherche sur des fichiers sonores. Dans ce cas, il désigne des outils servant à détecter des mots clés ou des phrases spécifiques en analysant le signal audio.<sup>4</sup> Cette technologie a également été transposée à des documents image. La REM et le KWS sont fondamentalement différents. E. Vidal *et al.* explique, dans un article daté d'avril 2021, que la REM cherche à obtenir la meilleure transcription, ou chaîne de caractères, d'une région de texte située sur une image. Une fois obtenue, celle-ci est définitive et constitue le produit de sortie d'un modèle de REM. Le KWS, en contraste, propose de ne pas arrêter de transcription et de montrer, suite à la requête d'un-e utilisateur-ice, chaque région d'une

1. Voir la définition des recherches floues du moteur de recherche Elasticsearch : <https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-fuzzy-query.html> (consulté le 29/08/21).

2. Voir <https://readcoop.eu/glossary/fuzzy-search/> (consulté le 28/08/21).

3. M.L. Massot, A. Sforzini et V. Ventresque, « Transcribing Foucault's handwriting with Transkribus »..., p. 11

4. Angelos P. Giotis, Giorgos Sfikas, Basilis Gatos et Christophoros Nikou, « A survey of document image word spotting techniques », *Pattern Recognition*, 68 (1<sup>er</sup> août 2017), p. 310-332, DOI : 10.1016/j.patcog.2017.02.023

image contenant du texte s'apparentant au mieux à celle-ci.<sup>5</sup> Le KWS intervient directement à partir de l'image, et n'effectue pas de recherche dans une transcription complète et terminée.<sup>6</sup> Ce système relève davantage du domaine de la vision par ordinateur que du TAL. Néanmoins, bien qu'appartenant à deux paradigmes différents, le KWS et la REM partagent les mêmes architectures d'apprentissage machine et méthodes d'entraînement.<sup>7</sup> Les requêtes qui peuvent être faites avec un système de KWS peuvent être de deux natures<sup>8</sup> :

1. *Query-by-Example* (QbE) : l'utilisateur-ice donne une image d'un ou de plusieurs mots à partir d'une capture d'écran faite sur un document, et le système se charge de récupérer les exemples similaires trouvées dans le document soumis à la requête.
2. *Query-by-String* (QbS) : l'utilisateur-ice exprime sa requête en donnant au système une chaîne de caractères. Celui-ci se charge de récupérer les occurrences similaires à la chaîne de caractères dans le document soumis à la requête.

Le projet HIMANIS, mené par l'IRHT, a réalisé grâce au KWS l'indexation des registres de la chancellerie royale française des années 1302 à 1483, conservés aux AN, à partir de leurs numérisations.<sup>9</sup> Les statistiques du projet parlent d'elles-mêmes : pour 199 manuscrits, totalisant 83 320 pages, dont 83 141 aujourd'hui indexées, le système de KWS utilisé a identifié 285 791 425 endroits (*spots*) où se trouvent des termes différents. Pour un même mot écrit dans une page, le système peut détecter plusieurs *spots* correspondant à différents termes indexés. Par page, il est estimé qu'il y a en moyenne 3 437 endroits pouvant potentiellement répondre à la requête d'un-e utilisateur-ice.<sup>10</sup> L'indexation des termes présents dans ces documents, qui résultent en millions d'entrées d'index, permet aux utilisateurs de naviguer à travers ces documents à l'aide d'un système de QbS. Par exemple, pour le mot latin « mensis », le système renvoie un certain nombre de pages selon un seuil de confiance choisi. Ce mode de recherche est basé sur ce que D. Stutzmann *et al.* nomme la « négociation de contenu ».<sup>11</sup>

---

5. E. Vidal, A. H. Toselli et J. Puigcerver, « A Probabilistic Framework for Lexicon-based Keyword Spotting in Handwritten Text Images », *arXiv :2104.04556 [cs]* (, 9 avr. 2021), arXiv : 2104.04556, URL : <http://arxiv.org/abs/2104.04556> (visité le 23/04/2021), p. 3

6. Voir figure 3.1. Voir également A. P. Giotis, G. Sfikas, B. Gatos, *et al.*, « A survey of document image word spotting techniques »..., p. 311. En outre, cet article est une étude approfondie des techniques de KWS existantes, réalisée en 2017.

7. E. Vidal, A. H. Toselli et J. Puigcerver, « A Probabilistic Framework for Lexicon-based Keyword Spotting in Handwritten Text Images »..., p. 4

8. E. Rusakov, L. Rothacker, H. Mo et G. A. Fink, « A Probabilistic Retrieval Model for Word Spotting Based on Direct Attribute Prediction », dans *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2018, p. 38-43, DOI : 10.1109/ICFHR-2018.2018.00016, p.38

9. D. Stutzmann, J.F. Moufflet et S. Hamel, « Full Text Search in Medieval Manuscripts... » et Théodore Bluche, S. Hamel, Christopher Kermorvant, Joan Puigcerver, D. Stutzmann, Alejandro J. Toselli et Enrique Vidal, « Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript Collection in the HIMANIS Project », dans *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Kyoto, France, 2017, DOI : 10.1109/ICDAR.2017.59

10. Voir <http://himanis.huma-num.fr/app//views/stats.html> (consulté le 30/08/21).

11. D. Stutzmann, J.F. Moufflet et S. Hamel, « Full Text Search in Medieval Manuscripts... »

Elle permet ainsi de brasser plus ou moins large, tout en reconnaissant que la machine peut se tromper dans son traitement automatique. Par exemple, dans la recherche susmentionnée, en indiquant un taux de confiance à 50, le système ne renverra que les documents où « *mensis* » est mentionné avec un degré de certitude supérieur ou égal à ce taux ainsi que les endroits dans les images pour localiser ces résultats (voir annexes D.1 et D.2). Dans le cas d'HIMANIS, on parle d'« indexation probabiliste » car le système prend en compte toutes les hypothèses de reconnaissance textuelle, et non seulement la meilleure.<sup>12</sup>

Cette méthodologie est pertinente, mais il est regrettable de ne trouver aucun outil *open-source* permettant de l'expérimenter, voire de l'implémenter de manière libre sans passer par une prestation.<sup>13</sup> Le KWS n'a donc pas pu être expérimenté durant mon stage. Le développement d'un système semblable et *open-source* pourrait grandement bénéficier aux communautés des humanités numériques, d'autant plus pour les projets de REM.

### 3.3 Déetecter les entités nommées sans utiliser de données textuelles grâce à la vision par ordinateur

Chandranath Adak *et al.* ont publié en 2016 un système utilisant également la vision par ordinateur pour effectuer de la REN. Le système conçu segmente notamment les images en mots, et analyse leurs caractéristiques structurelles et positionnelles. C'est-à-dire, respectivement la présence de majuscules ou non, et la position des mots dans une phrase, ainsi que de leurs indices visuels présents dans l'image au niveau des pixels pour déterminer s'il est une EN.<sup>14</sup>

---

12. Étienne Cavalié (dir.), *L'indexation matière en transition : de la réforme de Rameau à l'indexation automatique*, ISSN : 0184-0886, Paris, France, 2019, pp. 121 - 122. Voir E. Rusakov, L. Rothacker, H. Mo, *et al.*, « A Probabilistic Retrieval Model for Word Spotting Based on Direct Attribute Prediction »... pour l'explication technique de ce système.

13. Le logiciel Transkribus implémente par exemple cette fonctionnalité. Alejandro H. Toselli et Enrique Vidal, *Transkribus User Conference. Keyword Spotting in Large Scale Documents*, nov. 2017, URL : [https://readcoop.eu/wp-content/uploads/2017/07/Toselli\\_Keyword\\_Spotting.pdf](https://readcoop.eu/wp-content/uploads/2017/07/Toselli_Keyword_Spotting.pdf) (visité le 08/09/2021)

14. Chandranath Adak, Bidyut B. Chaudhuri et Michael Blumenstein, « Named Entity Recognition from Unstructured Handwritten Document Images », dans *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, 2016, p. 375-380, DOI : 10.1109/DAS.2016.15

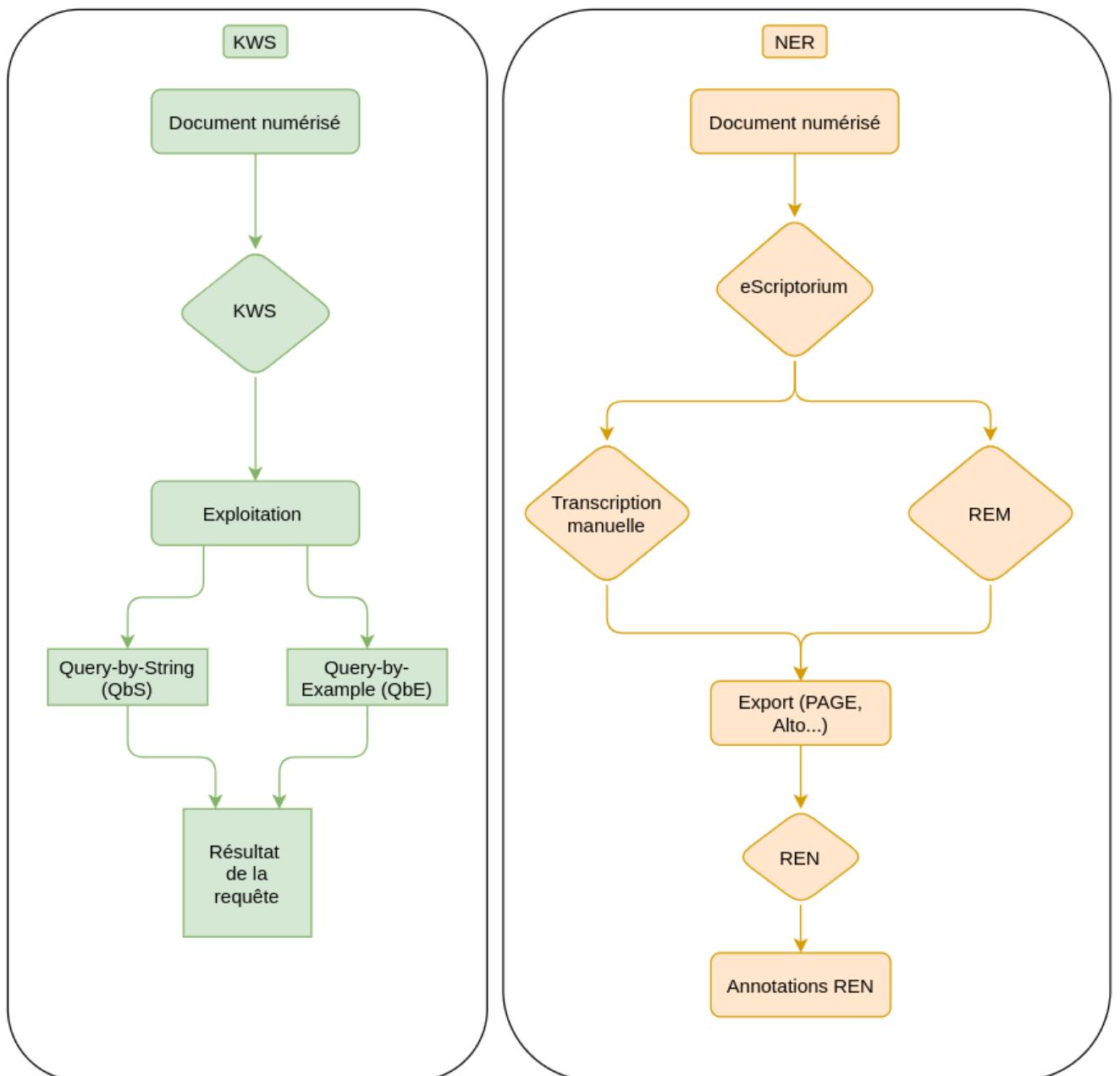


FIGURE 3.1 – Chaînes de traitement pour le KWS et la REM avec eScriptorium



## **Deuxième partie**

**La reconnaissance d'entités nommées appliquée à des données bruitées issues de la transcription automatique de documents patrimoniaux : expérimentations à partir des données du projet LECTAUREP**



# Chapitre 4

## Quels pré-traitements des données pour la reconnaissance d'entités nommées ?

Dans son article de 2003, le chercheur Alexander Clark définit le pré-traitement de données bruitées comme une séquence d'opérations commençant à partir d'un document numérique et ayant pour but d'aboutir à une représentation détaillée et structurée de celui-ci.<sup>1</sup>

À cette fin, il identifie plusieurs étapes :

- Le *parsing* d'un fichier texte pour obtenir une séquence de caractères qui peut être traitée.
- La segmentation des tokens.
- La segmentation des phrases qui forment le texte.
- La correction des mots mal orthographiés.
- Déterminer si un mot possède sa majuscule naturelle, pour un nom propre par exemple.<sup>2</sup>

Ce faisant, le texte traité retrouve une structure qui peut être exploitée. Ce chapitre aura pour but de décrire les expériences réalisées au cours du stage pour tenter de redonner une structure aux données textuelles issues de la REM. En structurant les données textuelles et en les corrigeant, on se donne les moyens pour appliquer des opérations de TAL telles que la REN, ainsi que pour envisager la production de données d'entraînement d'un modèle de REN.

---

1. "We conceive of the processing of texts as a sequence of operations starting from some sort of computer file or stream, and ending with some more detailed and structured representation." Alexander Clark et Alexander Clark Issco, « Pre-Processing Very Noisy Text », dans *Proc. of Workshop on Shallow Processing of Large Corpora*, 2003, p. 13

2. "Whether words have had their natural capitalisation altered at all." *Ibid.*

## 4.1 Reconstruire un document déconstruit par la reconnaissance d'écriture manuscrite

### 4.1.1 Reconstituer la structure logique des répertoires des notaires

Sans segmentation des régions, la REM induit la disparition de la structure logique tabulaire des pages des répertoires des notaires. Par structure logique, nous entendons la disposition des informations textuelles sur une page, et qui possède un sens pour la compréhension de l'information. Dans le cas d'un tableau, la structure logique correspond aux différentes colonnes et aux rangées qui le composent. Par exemple, dans la cinquième colonne des pages des répertoires des notaires, nous savons qu'il y aura systématiquement des informations relatives aux parties concernant les enregistrements.

Sur eScriptorium, la segmentation automatique des *baselines* et leur numérotation se fait de gauche à droite et de bas en haut. Chaque *baseline* implique la création d'une nouvelle ligne dans un fichier texte, mettant ainsi totalement à plat le document original (voir un export texte d'une transcription automatique dans l'annexe E.1). Il est impossible à ce stade de reconstituer la structure logique, le volume d'information présent dans les enregistrements connaissant d'importantes variations selon le notaire. On ne pourrait donc pas, par exemple, écrire de script découplant le texte toutes les  $x$  lignes. Cela entraînerait des résultats trop aléatoires, où les différents enregistrements seraient mélangés, la fin de l'un pouvant être intégrée au début d'un autre.<sup>3</sup> De plus, un seul découpage horizontal des enregistrements nous ferait perdre la structure des colonnes du tableau. Celles-ci ne seraient plus que matérialisées par des espaces dans un fichier texte, transformant ainsi, pour une machine, une rangée en une longue phrase sans structure syntaxique.<sup>4</sup>

Une segmentation verticale, symbolisant les colonnes, croisée à une segmentation horizontale, les rangées du tableau, semble donc plus appropriée. Reconstituer la structure logique tabulaire des pages permettrait de cibler l'extraction d'information dans une colonne. Par exemple, pour récupérer la date d'un enregistrement, il suffirait de cibler la deuxième colonne et d'en extraire une valeur. Selon cette idée, un modèle permettant d'annoter l'en-tête des répertoires des notaires et leurs colonnes avec un label "column\_pair" ou "column\_odd" (colonnes paires et impaire) a été entraîné par Alix Chagué. Il présente l'avantage de reconstituer la structure tabulaire verticale des pages des répertoires des notaires, en assignant à chaque *baseline* sa colonne respective. Cependant, ce modèle s'arrête au niveau de granularité de la page entière. Il n'est pas possible de reconstituer la structure logique à partir de ce niveau car les enregistrements ne sont pas toujours contenus sur une seule ligne, ils s'étalent bien souvent

---

3. Pour une illustration de la segmentation et de la numérotation automatique des lignes sur une page d'un répertoire de notaire, sans segmentation des régions, voir annexe E.1

4. Voir annexe E.2

sur plusieurs, rendant une concaténation systématique de 7 lignes consécutives difficile.<sup>5</sup> Nous avons déterminé qu'il était envisageable d'atteindre un autre niveau de granularité : celui des enregistrements, en s'appuyant sur d'autres caractéristiques des répertoires des notaires.

Dans ce but, Alix Chagué m'a proposé de travailler conjointement à l'élaboration d'une chaîne de traitement destinée à traiter automatiquement les pages des répertoires des notaires pour reconstituer leur structure logique, dans le but de transformer un export PAGE XML résultant de la REM depuis eScriptorium, en fichier XML-TEI, en vue de leur publication. Dans cette partie, nous détaillerons la première étape ; la transformation en TEI sera détaillée dans la troisième partie de ce mémoire, car répondant à d'autres problématiques.<sup>6</sup>

Notre idée se fonde sur la possibilité d'annoter les *baselines* dans eScriptorium. En annotant les premières lignes de chaque enregistrement présent dans une page, on se donne un repère exploitable dans le PAGE XML pour pouvoir découper horizontalement les enregistrements, tout en gardant l'information des régions.

#### **4.1.2 Description de l'ontologie utilisée pour l'annotation des régions et des *baselines***

Pour réaliser cette expérience, nous avons annoté manuellement dans eScriptorium 10 pages aléatoires issues des répertoires des notaires.

L'annotation des données a été effectuée selon une ontologie préalablement définie. Pour les régions, les étiquettes suivantes ont été choisies :

- Pour les colonnes, de gauche à droite :
  - « Col\_1 » pour la première colonne.
  - « Col\_2 » pour la deuxième colonne.
  - « Col\_3 » pour la troisième colonne.
  - « Col\_4 » pour la quatrième colonne.
  - « Col\_5 » pour la cinquième colonne.
  - « Col\_6 » pour la sixième colonne.
  - « Col\_7 » pour la septième colonne.
- "Header" pour l'en-tête
- "Marginal" pour les écritures marginales
- "Stamp" pour les timbres

---

5. Voir la numérotation des *baselines* sur une page avec segmentation des régions dans l'annexe E.3.

6. L'expérience a été documentée dans l'*issue* Gitlab suivante : <https://gitlab.inria.fr/almanach/lectaurep/documentation/-/issues/17> (consulté le 11/10/08).

Pour les *baselines*, les étiquettes suivantes ont été choisies :

- "First\_line" pour les premières lignes des enregistrements.
- "Main\_date" pour les lignes indiquant la date de la page au début de la cinquième colonne.
- "Printed" pour les lignes imprimées.

#### 4.1.3 Méthodologie

The figure shows a simplified schema of a ledger page from 1901. The page is divided into columns for dates (26, 27, 28), descriptions (Mainlevé, Motoriste, Décharge à mardi), and amounts (24.50, 3.75). Red dashed lines indicate the boundaries of the first-line segments. The first segment covers the first three columns and the first part of the descriptions. The second segment starts at the beginning of the fifth column ('An 1901, mois de Mars') and continues through the rest of the page. Annotations in red ink, such as 'Gandois' and 'Préfondue', are also present on the original document.

FIGURE 4.1 – Schéma simplifié illustrant le découpage horizontal des enregistrements.

Le schéma 4.1 illustre la méthodologie selon laquelle nous avons procédé. La ligne rouge épaisse symbolise une première ligne d'enregistrement, et correspond à une *baseline*. L'annotation manuelle décrite précédemment ajoute un attribut `<custom>structure {type:first_line;}</custom>` au noeud `<TextLine>` dans le PAGE XML. Lors de la segmentation, eScriptorium crée à partir de ces éléments des masques, dont nous prenons la coordonnée *y* la plus haute. Cette coordonnée nous donne un point, à partir duquel on peut trancher la page horizontalement, symbolisé par le trait rose en pointillé. Ce premier élément nous donne la borne haute de la rangée. Ensuite, on parcourt la cinquième colonne jusqu'à trouver une nouvelle première ligne. On cherche à nouveau le point le plus haut du masque généré, et on tranche à nouveau horizontalement, nous donnant la deuxième borne de la rangée. Cela nous permet de la reconstituer en récupérant toutes les noeuds `<TextLine>` du PAGE XML dont les coordonnées *y* sont contenues dans l'intervalle des deux bornes. Cette information se trouve dans l'attribut *points* de la balise enfant `<Coords>`.<sup>7</sup>

La reconstitution est effectuée à l'aide de scripts Python : chaque enregistrement est stocké dans un objet nommé *Row*. Le processus est itératif, pour chaque enregistrement présent dans une page, un objet *Row* est créé. De plus, nous ne récupérons pas la chaîne de caractères présente dans la balise enfant `<Unicode>`, mais l'identifiant du noeud `<TextLine>`. Pour relier les noeuds `<Textline>` à la région à laquelle ils appartiennent, il suffit de contrôler l'attribut `<custom>` du noeud parent `<TextRegion>` où l'étiquette de la région est indiquée. À partir de cette information, on distribue les identifiants de chaque `<Textline>`. Une fois

7. Un enregistrement dans sa version PAGE XML a été joint dans les annexes, voir E.1.

stocké, l'identifiant nous permet de récupérer la chaîne de caractères qui lui est associée si besoin.<sup>8</sup>

Il est possible de visualiser un objet *Row* grâce à une de ses méthodes nommée *show\_row()*, elle permet d'obtenir l'affichage d'un dictionnaire python représentant un enregistrement :

```
{ 'bottom\_limit': None,
  'date\_of\_act': ['eSc\_line\_c7f2a525'],
  'entry\_id': ['eSc\_line\_8ae6e948'],
  'head\_line': 'eSc\_line\_b6e1e9a5',
  'main\_paragraph': ['eSc\_line\_b6e1e9a5', 'eSc\_line\_cd9774d2'],
  'misc': [],
  'registration\_relation': { 'date': ['eSc\_line\_904d4325'],
                             'droits': ['eSc\_line\_a0fabfae']},
  'top\_limit': 3932,
  'type\_of\_act': {'brevet': [], 'minute': ['eSc\_line\_63f716fc']}}}
```

Une autre méthode, *show\_text\_in\_row()*, permet d'obtenir un affichage semblable, mais avec les identifiants des balises <Textline> remplacés par la chaîne de caractères présentes dans la balise enfant <Unicode>.

```
3932 < None
num de répertoire : ['206']
date de l'acte : ['6']
types de l'acte (brevet) : []
types de l'acte (minute) : ['Compte de tutelle']
paragraphe central : Deschamps (par Jacques Charles) à Ris Orangis (S et O)
+ rue du Pont 6bis à Emilie Deschamps à Paris rue Marine 21
date d'enregistrement : ['14']
droits d'enregistrement : ['3.75']
misc : []
---fin---
```

Néanmoins, certaines informations échappent à cette reconstitution automatique, possiblement car le découpage horizontal a raté une *baseline*, ou car l'annotation des régions ne s'est pas entièrement bien déroulée. Nous avons donc décidé de rajouter pour chaque rangée une catégorie *misc* (autres) qui récupère toutes les informations non classées afin de ne perdre aucune donnée lors de ce processus.

---

8. Le schéma présenté est une version simplifiée d'un autre schéma créé par Alix Chagué et joint en annexe, voir E.4.

La chaîne de traitement a été mise en forme dans un Google Colab par Alix Chagué.<sup>9</sup> En reconstituant la structure logique, donc en segmentant pour chaque rangée l'information en colonne, il sera possible de traiter l'information de manière spécifique à chacun de ces compartiments sémantiques. Grâce à cela, on peut par exemple reconstituer les phrases présentes en colonne 5 dans le but d'appliquer des outils de REN.

## 4.2 Entrainer un modèle de segmentation sémantique pour reconstruire automatiquement la structure logique

### 4.2.1 Objectifs et motivations du modèle de segmentation affiné

Dans la suite de l'expérience précédente, il a été décidé de constituer un set de données annotées en vue d'affiner (*fine-tuning*) le modèle de segmentation originellement entraîné par Alix Chagué.<sup>10</sup>

Cette décision a également été motivée par la variété des contenus présents dans les pages des répertoires de notaires<sup>11</sup>, et par les accidents parfois présents dans les colonnes, par exemple des informations non habituelles ou des timbres.

L'entraînement de ce modèle affiné repose sur les caractéristiques visuelles récurrentes que présentent les pages des répertoires des notaires, à savoir :

- La structure tabulaire pré-imprimée dans laquelle le notaire vient écrire.
- La formule indiquant la date au début de la cinquième colonne est toujours la même, et mélange écriture imprimée et écriture manuscrite. Voir la figure 4.2, par exemple.
- De manière générale, chaque enregistrement commence dans la cinquième colonne par un alinéa et un nom de famille écrit de manière différente du reste du paragraphe. Celui-ci est fréquemment écrit dans une taille plus grande avec un trait plus épais. Il est également parfois écrit de manière plus soignée. Voir les figures 4.3, 4.4, 4.5.

---

9. Voir <https://colab.research.google.com/drive/1UjT5Gw70rsmrBIYFxfQBHXKR18NWqT38?usp=sharing#>, jusqu'à la troisième partie. Le reste du Colab concerne la transformation du PAGE XML en XML-TEI, étape que nous expliquerons plus tard dans le déroulé de ce mémoire. Google Colab est un service de Google mettant à disposition des environnements pour exécuter du code sur une machine distante. Ce service a l'avantage de permettre le travail collaboratif, notamment le *peer coding*.

10. Cette expérience a été documentée dans une *issue* sur le Gitlab *Kraken Models* : <https://gitlab.inria.fr/dh-projects/kraken-models> (consulté le 10/08/21). Voir <https://gitlab.inria.fr/dh-projects/kraken-models/-/issues/22> (consulté le 10/08/21).

11. La densité de certaines pages aurait pu troubler une segmentation horizontale des enregistrements des répertoires de notaires. Voir annexe F.1.

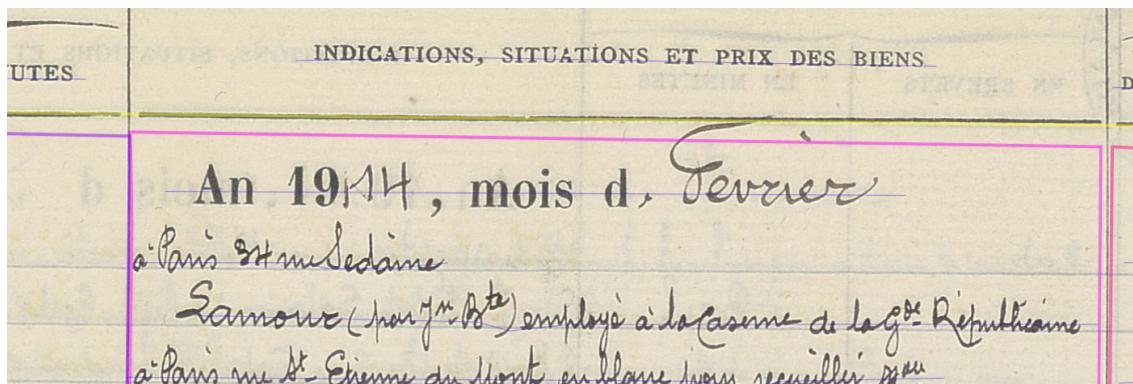


FIGURE 4.2 – Exemple d'une ligne indiquant la date d'une page d'un répertoires de notaire.

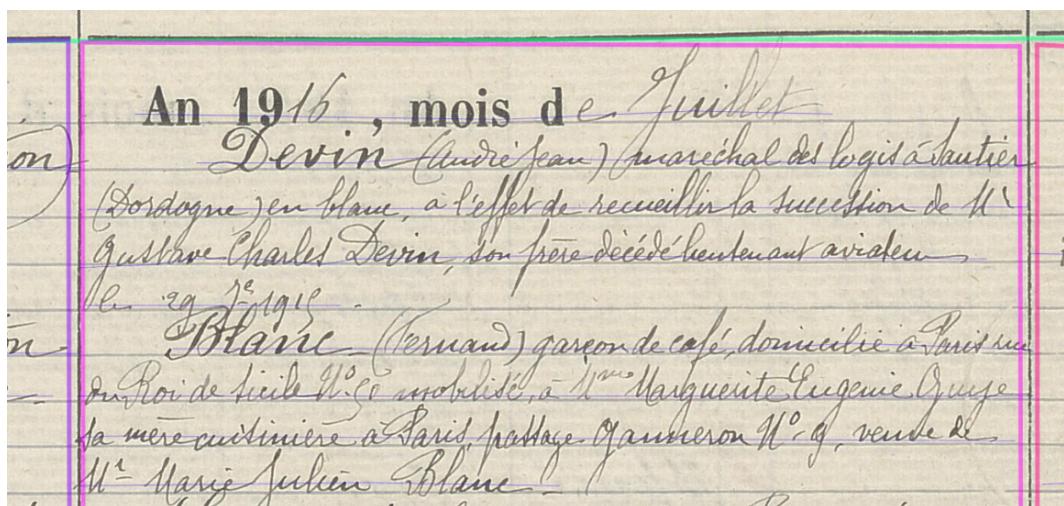


FIGURE 4.3 – Exemple générique des caractéristiques visuelles des premières lignes de chaque enregistrement. Ce cas de figure est rencontré fréquemment, avec parfois de faibles variations.

#### 4.2.2 Description du processus d'annotation du corpus d'entraînement

L'annotation des régions s'est effectuée dans eScriptorium, sur deux documents rassemblant respectivement 100 pages du *random set*, préalablement segmentées par le premier modèle de segmentation susmentionné.<sup>12</sup> J'ai utilisé pour cette tâche l'ontologie décrite dans la partie précédente. Les colonnes ont été ré-étiquetées, sauf pour les en-têtes, les timbres et les zones d'écriture marginales. Les *baselines* de chaque première ligne d'enregistrement et des lignes indiquant la date de la page ont été annotées, sauf pour les lignes imprimées pré-

12. Le *random set* est composé de mille doubles pages choisies aléatoirement depuis quatre campagnes de numérisation en couleur de répertoires datant des années 1880 à 1930. Il a été constitué dans le cadre du projet LECTAUREP pour entraîner et tester des modèles de transcription et de segmentation. L. Terriel, *Représenter et évaluer les données issues du traitement automatique d'un corpus de documents historiques. L'exemple de la reconnaissance des écritures manuscrites dans les répertoires de notaires du projet LectAuRep., mémoire de master « Technologies numériques appliquées à l'histoire », dir. Alix Chagué et Thibault Clérice... , p. 19*

An 1916, mois de Juillet (Suite)

Hugo Oberndoerffer (à la 15<sup>e</sup> de M) décédé  
rentré à Paris, avenue de Villiers 8 à Madrid Ida Wilmeidoerffer  
rentrée à Paris, avenue de Villiers 8)

Leroy (Marcel Théophile) après le décès de H.  
arrivé à Paris rue de Sèvres 13 le 29 février 1916 époux  
de M<sup>e</sup> Henriette Marie Brecher

Cail (Louis) avocat au conseil d'état et à la Cour  
de cassation à Paris, avenue Henri Martin 8<sup>e</sup> à M<sup>e</sup> Gabrielle  
Marie Devin son épouse à l'effet de recevoir la M<sup>e</sup> de H.  
Gustave Charles Devin frère de M<sup>e</sup> Devin

FIGURE 4.4 – Exemple d'une page où les premières lignes de chaque enregistrement sont très distinctes, du fait d'un alinéa prononcé et d'un nom de famille mis en avant.

annotées par l'ancien modèle.<sup>13</sup> Les 200 pages candidates n'ont pas pu toutes être annotées. Certaines présentaient des problèmes, tels que l'absence totale d'indices visuels permettant de distinguer les enregistrements ou le placement d'un alinéa sur une ligne autre que la première ligne de l'enregistrement. Toutes les pages annotées ont ensuite été rassemblées dans un nouveau document sur eScriptorium afin de faciliter l'accès aux données et de leur export.<sup>14</sup>

#### 4.2.3 Affinage du modèle de segmentation et évaluation

J'ai ensuite utilisé le module d'entraînement de Kraken en ligne de commande pour affiner le modèle que nous avons présenté au début de cette partie. Celui-ci génère un rapport d'entraînement affiché dans le terminal depuis lequel est lancée la commande qui a été joint dans les annexes.<sup>15</sup> À chaque epoch, c'est-à-dire à chaque fois que l'algorithme d'apprentissage a traité le corpus d'entraînement, est affichée l'évaluation du modèle grâce à des métriques obtenues avec un corpus de validation. Le corpus d'entraînement et le corpus de validation

13. Pour des illustrations de l'annotation, voir annexes F.2, F.3, F.4.

14. Le lien permettant de télécharger l'intégralité des pages annotées est disponible ici : [https://gitlab.inria.fr/dh-projects/kraken-models/-/issues/22#note\\_551920](https://gitlab.inria.fr/dh-projects/kraken-models/-/issues/22#note_551920) (consulté le 10/08/21). Il est possible, à partir de l'export, de les importer dans un nouveau document eScriptorium pour visualiser la segmentation manuelle.

15. Voir annexe F.1.

**An 19<sup>30</sup>, mois d** Juin  
 Agathe Alice N°6 Berdin à Cléchy 1 Bois Sainte  
 du Rêveau, licencié le 1er Septembre 1930 P.  
 Taxe. Génie et Angèle Marie Josephine Kraken  
 son épouse à Paris 27 av de la République d'inscriff  
 Taxe 30 Agathe 9 Juillet 1926 N°2988919 Adolphe  
 Kraken époux et son épouse  
 uation Sabadens Jean Marie Daffiste par Mad  
 pizere Chauvinier et de son fils Paris M. Victor  
 Kraken à son mariage pour vendre  
 10 immobilière du Sud-Est Parisien à  
 Paris 17 rue Cambon 10 achete et empruntee  
 10 immobilière de la Nation à Paris 10 meublée  
 10 immobilière du Sud-Est Parisien à Paris 10  
 une papette  
 Carré par Léonie Antoine à Paris 10 me Cambon

FIGURE 4.5 – Exemple de l'indice visuel minimum existant pour le début d'un enregistrement : l'alinéa.

sont générés automatiquement par Kraken en divisant le set de données initialement utilisé. Habituellement, le premier contient la grande majorité des données d'entraînement, et le second une petite quantité de données. Il nous a permis de générer des courbes rendant compte du processus d'entraînement, reproduites dans les figures 4.6, 4.7 et 4.8.

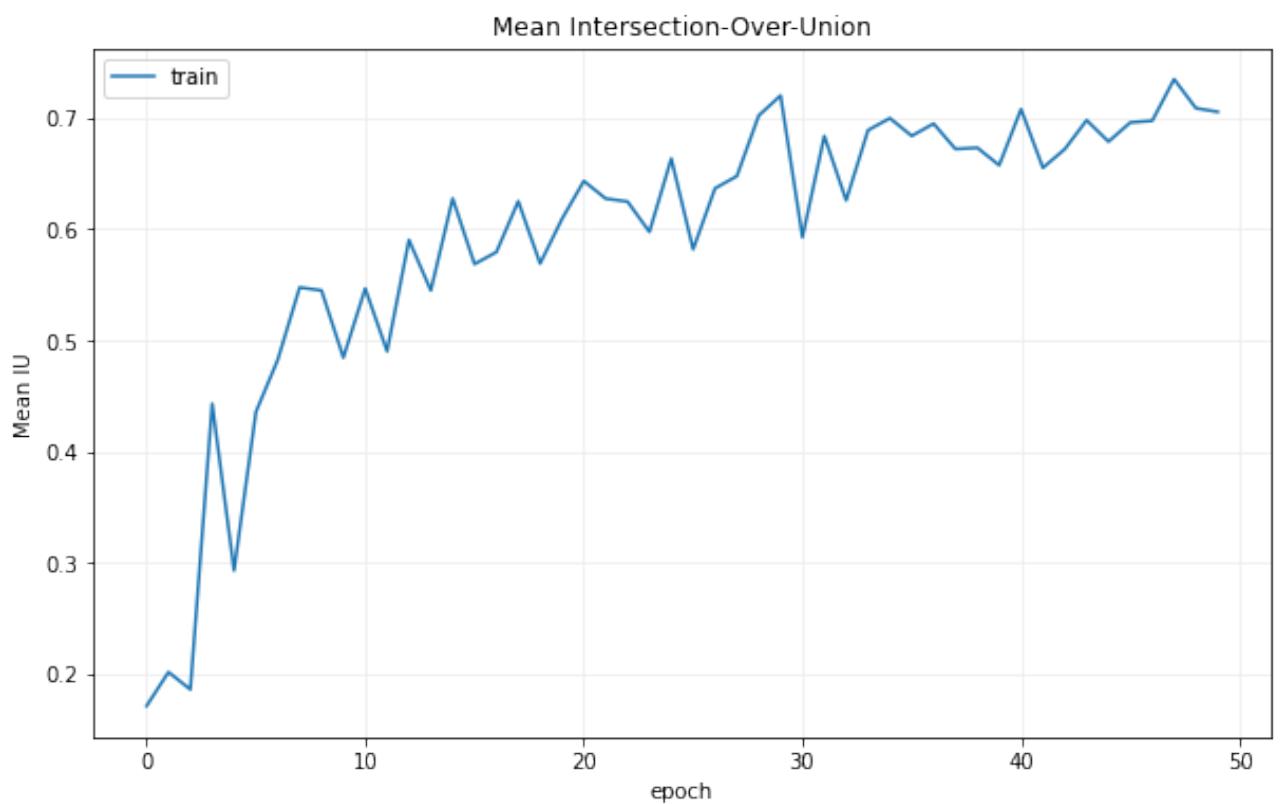


FIGURE 4.6 – Mean Intersection-Over-Union

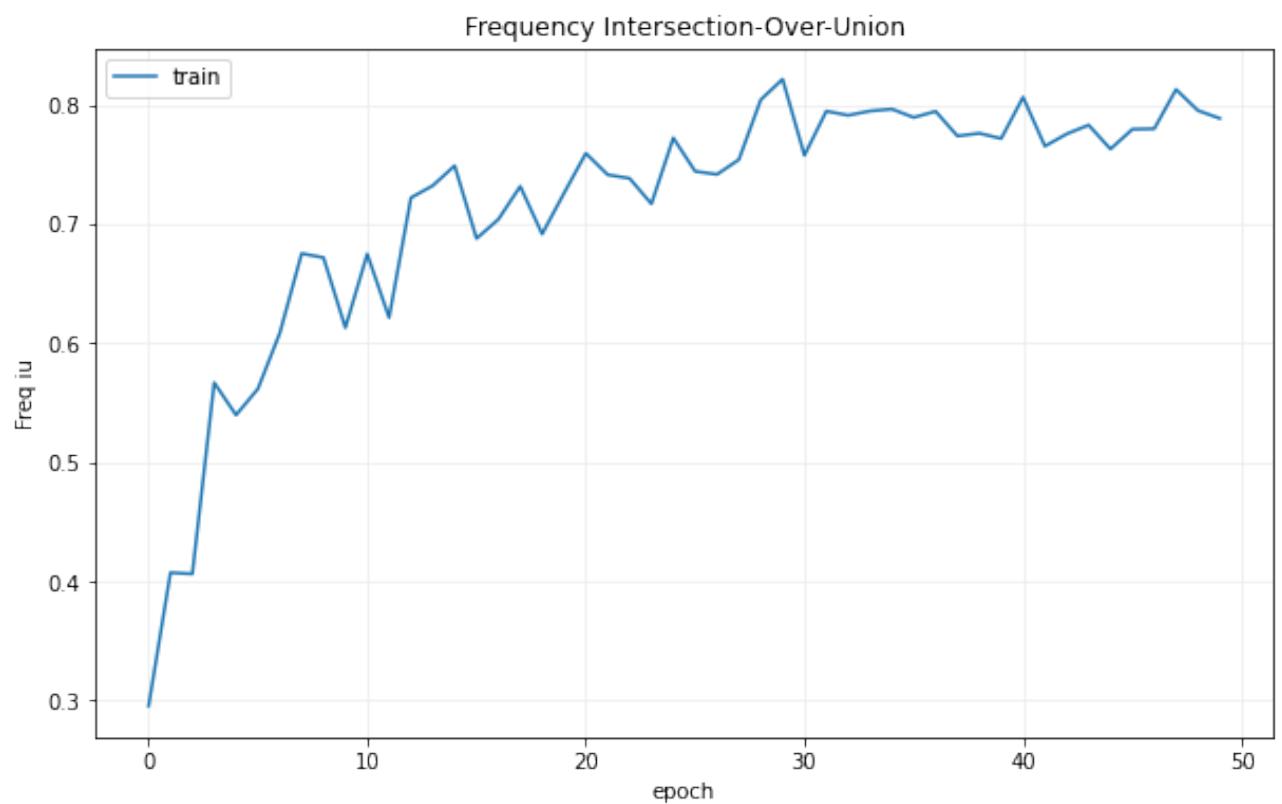


FIGURE 4.7 – Frequency Intersection-Over-Union

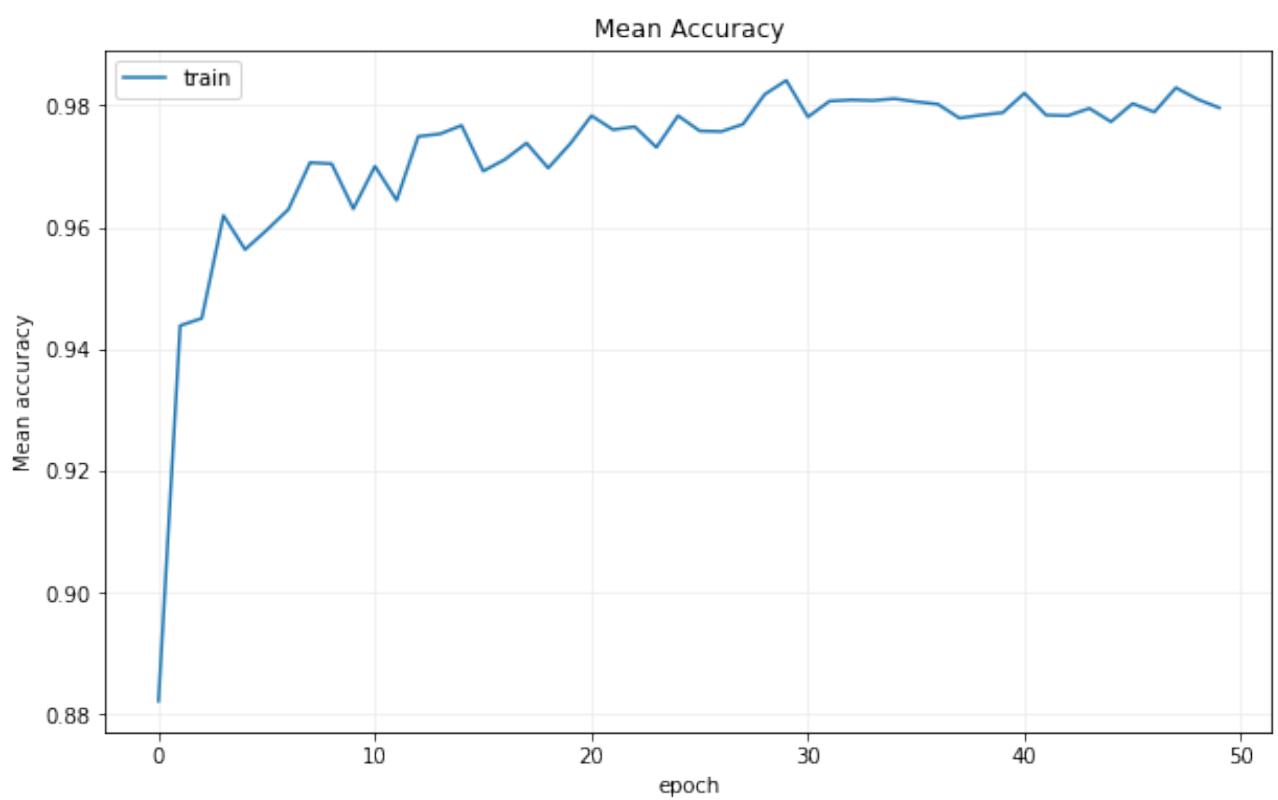


FIGURE 4.8 – Mean accuracy

#### 4.2.4 Explication des métriques d'évaluation du modèle de segmentation

Avec ces trois courbes d'entraînement, trois métriques différentes d'évaluation d'un modèle de segmentation sont illustrées : le *mean Intersection-Over-Union*, dont la traduction n'est pas aisée, mais est un dérivé de l'indice de Jaccard ; le *frequency Intersection-Over-Union*, un autre dérivé du même indice ; et le *mean accuracy*.

L'indice de Jaccard permet de calculer la similarité entre deux éléments. Il est le « rapport entre le cardinal de l'intersection des ensembles considérés et le cardinal de l'union des ensembles ».<sup>16</sup> Soit deux ensembles  $A$  et  $B$ , l'indice correspond à l'équation suivante :

$$J(A, B) = \frac{A \cap B}{A \cup B}$$

L'*Intersection-Over-Union* est un autre nom donné à l'indice de Jaccard, et sert dans l'évaluation à déterminer le taux d'alignement entre une vérité de terrain -le premier ensemble  $A$ -, une zone annotée manuellement par exemple, et une prédiction -le deuxième ensemble  $B$ -, le résultat d'un modèle de segmentation.<sup>17</sup> Plus le score se rapproche de 1, plus il est performant.

Le *mean Intersection-Over-Union* permet de prendre en compte toutes les classes d'un modèle de segmentation sémantique à la fois, dans notre cas toutes les colonnes, toutes les premières lignes, etc. On l'obtient en calculant l'*Intersection-Over-Union* de toutes les classes et en faisant la moyenne des résultats obtenus.<sup>18</sup>

Le *frequency Intersection-Over-Union* est une variation pondérée de la métrique précédente. Les poids sont assignés en fonction de la fréquence de chaque classe.<sup>19</sup>

Enfin, le *mean accuracy* est une variation de l'*accuracy*, métrique couramment employée pour mesurer les performances d'un modèle de *machine learning* pour une tâche donnée.<sup>20</sup> Cette métrique correspond au taux de prédictions correctement effectuées par le modèle, et se représente selon l'équation suivante :

---

16. Voir [https://fr.wikipedia.org/wiki/Indice\\_et\\_distance\\_de\\_Jaccard](https://fr.wikipedia.org/wiki/Indice_et_distance_de_Jaccard) (consulté le 10/08/21).

17. Pour une illustration de ce calcul, voir l'annexe F.8. Pour une illustration de la comparaison entre une vérité de terrain et une prédiction sur une photographie, voir l'annexe F.5. La même comparaison pourrait être effectuée avec les images du projet LECTAUREP, en comparant la vérité de terrain (annexe F.6) et la prédiction obtenue avec le modèle affiné (annexe F.7).

18. Voir <https://towardsdatascience.com/metrics-to-evaluate-your-semantic-segmentation-model-6bcb99639aa2> (consulté le 10/08/21), et [https://keras.io/api/metrics/segmentation\\_metrics/#meaniou-class](https://keras.io/api/metrics/segmentation_metrics/#meaniou-class) (consulté le 10/08/21).

19. Voir <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6472823/> (consulté le 10/08/21).

20. Voir <https://developers.google.com/machine-learning/crash-course/classification/accuracy> (consulté le 10/08/21).

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Le *mean accuracy* s'obtient en calculant l'*accuracy* du modèle pour toutes les classes et en faisant la moyenne des résultats obtenus.

#### 4.2.5 Analyse des scores et pistes d'améliorations pour le modèle

On constate dans le rapport d'entraînement que le meilleur modèle a été atteint à l'*epoch* 48, avec un *mean Intersection-Over-Union* s'élevant à 0.7349, un *frequency Intersection-Over-Union* à 0.8128, et une *accuracy* à 0.9829.<sup>21</sup> Nous pouvons donc considérer les scores du modèle comme bons, tendant vers l'excellent.<sup>22</sup> Cependant, nous observons que les courbes d'entraînement sont irrégulières. Nous pouvons supposer que le set de validation servant à évaluer le modèle durant l'entraînement contenait des images ne présentant pas assez de caractéristiques visuelles mentionnées précédemment, entraînant donc périodiquement une baisse des résultats, puis une hausse des résultats lorsque le modèle s'y accommodait.

Des captures d'écran de prédictions faites sur des pages choisies aléatoirement sur les-quelles le modèle à été appliqué ont été jointes en annexe.<sup>23</sup> Les résultats sont satisfaisants, on observe la régularité de l'annotation des *baselines* représentant la première ligne des enregistrements et de la segmentation des régions, sur des images en couleur et en nuances de gris. Cependant, les lignes indiquant la date sur la page sont pour le moment systématiquement mal segmentées. Cela est certainement dû au faible nombre de données annotées concernant cette classe, seulement 134, contre 2168 *First\_line*. On observe que les résultats d'annotation de la cinquième colonne sont réguliers, malgré des pages posant des difficultés et venant perturber la segmentation d'autres régions, notamment les colonnes alentours.<sup>24</sup>

La prochaine étape de cette expérience serait donc d'ajouter des données d'entraînement au corpus déjà existant afin de viser des scores plus hauts. Il serait aussi important de réaliser des tests sur plus de documents, dont des pages présentant des difficultés, notamment dans l'identification de la première ligne de chaque enregistrement, pour prendre mesure de la robustesse du modèle. Enfin, il serait également intéressant de se calquer sur des ontologies de segmentation déjà existantes, de sorte à pouvoir rejoindre un standard.<sup>25</sup>

---

21. Voir annexe F.1

22. Voir annexe F.9

23. Voir annexes F.10, F.11, F.12, F.13.

24. Voir annexe F.13.

25. Mentionnons ici le projet *SegmOnto*, qui s'attache à créer une ontologie de segmentation basé sur la norme TEI à partir de plusieurs types de documents historiques pour les projets d'HTR. Voir <https://github.com/SegmOnto> (consulté le 10/08/21). Voir également A. Chagué et Floriane Chiffolleau, *An accessible and transparent pipeline for publishing historical egodocuments*, mars 2021, URL : <https://hal>.

Pour conclure, ce modèle pourrait représenter une étape dans la reconstitution en masse de la structure logique des répertoires des notaires une fois la transcription automatique effectuée.

## 4.3 Vers un « débruitage » des données issues de la reconnaissance d'écriture manuscrite en vue de la tokenisation

### 4.3.1 La segmentation des mots, une étape essentielle pour l'exploitation automatique des données du projet LECTAUREP

Lors d'un entretien pour discuter de la REN appliquée aux données du projet LECTAUREP avec M. Benoît Sagot, il m'a été conseillé de créer un système pour *débruiter* au maximum les données issues de la REM. En attendant des solutions plus établies de correction post REM, il m'a indiqué que la priorité devait être donnée à retrouver une segmentation des mots correcte dans la cinquième colonne. Cela permettrait de se rapprocher des données sur lesquelles a été entraîné le modèle de langue française développé au sein de l'équipe ALMANACh, CamemBERT<sup>26</sup>, et donc d'envisager l'entraînement d'un modèle de classification. La préparation des données et sa normalisation en vue de la REN, est une étape courante dans l'extraction d'information. Nguyen *et al.* présentaient en 2016 une méthode pour normaliser des tweets en vietnamien pour cette tâche.<sup>27</sup>

La segmentation des mots ("word segmentation"<sup>28</sup>) consiste à séparer des mots agglutinés en incluant un caractère entre ceux-ci ("word boundary character"), un espace, etc..<sup>29</sup> Lire des mots agglutinés n'est pas un problème pour l'être humain qui peut facilement comprendre que la séquence de lettres « répertoiresdesnotaires » correspond aux trois mots « répertoires des notaires ». On arrive à séparer les mots contenus dans la séquence en analysant mentalement chaque suite de lettre formant potentiellement un mot.<sup>30</sup> Un ordinateur ne traitera

---

archives-ouvertes.fr/hal-03180669 (visité le 11/08/2021)

26. L. Martin, B. Muller, P. J. Ortiz Suárez, *et al.*, « CamemBERT... »

27. V. Nguyen Hong, H. Nguyen et V. Snasel, « Text normalization for named entity recognition in Vietnamese tweets »... Dans cet article, la normalisation des données résulte de la détection d'erreurs de saisie, d'orthographe ; de leur correction ; de la correction des majuscules manquantes -les majuscules servent en effet d'éléments pour détecter les EN, notamment les noms propres, mais elles peuvent être hasardeuses dans les tweets- ; de la segmentation des mots ; et du *POS tagging*.

28. Yerai Doval et Carlos Gómez-Rodríguez, « Comparing Neural- and N-Gram-Based Language Models for Word Segmentation », *Journal of the Association for Information Science and Technology*, 70-2 (févr. 2019), p. 187-197, DOI : 10.1002/asi.24082, arXiv : 1812.00815

29. *Ibid.*

30. j2kun, *Word Segmentation, or Makiingsenseofthis*, Math Programming, 15 janv. 2012, URL : <https://jeremykun.com/2012/01/15/word-segmentation/> (visité le 16/07/2021). Par exemple, après avoir

que la séquence de lettres, et non les mots qui la composent, à moins d'utiliser un modèle de langue entraîné sur des données du même domaine. La segmentation des mots est une tâche qui peut être exécutée par un ordinateur selon une approche basée sur un système de n-gramme ou une approche neuronale. Retrouver les frontières de chaque mot permettrait de tokeniser et de lemmatiser le texte, et donc de préparer les données à des traitements de TAL.<sup>31</sup>

Dans le cas de LECTAUREP, on retrouve fréquemment après la REM des mots agglutinés dans la colonne 5. Cela est probablement causé par la graphie du notaire et par la nature de la vérité de terrain (voir section 2.6) :

Oussonppar Geordes ErnertLéou dt à Paris, rue Lamarch 144, à son père

Dans le cas d'une segmentation automatique des mots, en prenant en compte les erreurs de transcription automatique, nous souhaiterions donc obtenir :

Oussonppar Geordes Ernert Léou dt à Paris, rue Lamarch 144, à son père

Similairement :

Dournié (a près décè deMarieMuller, Ve de Marcelin) décéie en sondonielles Paris, rue  
Beuret 20, le11 février 190



Dournié (a près décè de Marie Muller, Ve de Marcelin) décéie en son donielles Paris, rue  
Beuret 20, le 11 février 190

#### 4.3.2 Segmenter les mots avec des dictionnaires de fréquence de n-grammes

En 2018, les chercheurs Yerai Doval et Carlos Gómez-Rodríguez ont montré que les architectures neuronales et les systèmes de n-gramme ont des performances similaires, cette dernière solution étant même effectuée plus rapidement par la machine.<sup>32</sup> Il est donc pertinent de l'explorer.

---

déterminé que « répertoires » est le premier mot, on analyse la lettre « d », puis la séquence « de », et enfin la séquence *des* que nous reconnaissons comme un déterminant.

31. T. Clérice, « Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Texts in Old French and Latin », *Journal of Data Mining & Digital Humanities*, 2020 (Towards a Digital Ecosystem ...[2020]), p. 5581, DOI : 10.46298/jdmdh.5581, p. 1

32. Y. Doval et C. Gómez-Rodríguez, « Comparing Neural- and N-Gram-Based Language Models for Word Segmentation » ...

Un n-gramme, ou *n-gram*, « est une sous-séquence de *n* éléments [...] dans une séquence donnée. »<sup>33</sup> Par exemple, une sous-séquence de mots dans une phrase. Quand la sous-séquence comporte deux éléments, on parle de bigrammes, quand il y en a trois, on parle de trigrammes, etc. Par exemple, la phrase « Je vais à l'École des chartes. » peut être divisée en 4 bigrammes : « Je vais » ; « à l' » ; « École des » ; « chartes. ». Les éléments concernés par les n-grammes peuvent également être des lettres.<sup>34</sup> Il est possible de créer un modèle de langue en associant des n-grammes avec leurs fréquences d'apparition dans une langue -on parle alors de dictionnaire de fréquence-, en s'appuyant par exemple sur un corpus écrit. Les n-grammes servent alors de données pour générer des probabilités sur la construction d'une séquence de mots.<sup>35</sup>

L'avantage des bigrammes réside dans leur capacité à opérer une correction grâce au contexte dans lequel se place le mot, contrairement à un dictionnaire d'unigrammes. Un dictionnaire de fréquence de bigrammes prend généralement la forme suivante :

token 1	token2	fréquence
...	...	...

En utilisant un dictionnaire de fréquence de bigrammes, on peut construire un système probabiliste pour déterminer la probabilité qu'une suite de caractère soit ou non plusieurs mots agglutinés.

Il existe une librairie python nommée *Word Segment* qui utilise des dictionnaires de n-grammes pour effectuer la segmentation des mots. Cependant, celle-ci ne sert qu'à traiter des données textuelles en anglais, car les unigrammes et les bigrammes sont issus de la langue anglaise.<sup>36</sup> Selon l'équipe de développement, il est possible de facilement modifier la librairie afin de la faire fonctionner avec d'autres dictionnaires. Pour tester la segmentation des mots avec cette méthode, j'ai employé la librairie python *SymSpellpy*, qui est un portage de la librairie *SymSpell* écrite en C#, destinée à corriger et faire de la recherche floue dans un texte.<sup>37</sup> Les tests ont été réalisés dans un *notebook*, disponible sur Gitlab.<sup>38</sup>

---

33. Voir <https://blog.engineering.publicissapient.fr/2021/03/17/nlp-concepts-cles-et-etat-de-lart/> (consulté le 12/08/21).

34. Le mot « chartes » contient les bigrammes suivants : « ch » ; « ar » ; « te » ; « s ».

35. Voir <https://www.depends-on-the-definition.com/introduction-n-gram-language-models/> (consulté le 12/08/21), où Tobias Sterbak, mathématicien, crée un modèle de langue à partir d'un corpus de n-grammes.

36. Voir <http://www.grantjenks.com/docs/wordsegment/> (consulté le 12/08/21).

37. Pour SymSpellpy, voir <https://github.com/mammothb/symspellpy> et <https://symspellpy.readthedocs.io/en/latest/index.html#> (consultés le 13/08/21). Pour la librairie SymSpell, voir <https://github.com/wolfgarbe/SymSpell> (consulté le 13/08/21).

38. Voir [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/preprocessing/word\\_segmentation/test\\_symspellpy/Symspellpy\\_test\\_word\\_segmentation.ipynb](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/preprocessing/word_segmentation/test_symspellpy/Symspellpy_test_word_segmentation.ipynb) (consulté le 13/08/21).

Dans le *notebook*, la méthode *lookup\_compound* a été testée. Elle permet d'effectuer des corrections orthographiques et de segmenter les mots. Plus précisément, elle permet de supprimer les espaces insérés par erreur dans un mot entraînant la création de deux mots incorrects ; de traiter la séparation des mots agglutinés ; et de corriger les erreurs de substitution, de transposition, de suppression et d'insertion. La librairie autorise pour cela le chargement de deux ressources : un dictionnaire d'unigrammes servant à la correction, et un dictionnaire de fréquence de bigrammes, sur lequel s'appuie la segmentation des mots. On peut récupérer un dictionnaire des 100 000 unigrammes les plus courants de la langue française dans le répertoire Github de SymSpell<sup>39</sup>. Un répertoire Github nommé *frenchngrams* permet de télécharger un dictionnaire de fréquence de bigrammes contenant les 10 000 bigrammes les plus fréquents de la langue française.<sup>40</sup> Le dictionnaire a été composé à partir des 400 livres écrits en français les plus populaires sur le projet Gutenberg en 2016.<sup>41</sup> En utilisant ces ressources on peut donc traiter la phrase suivante, correspondant à une phrase dans la cinquième colonne :

Lavignolle (rectificative parherman) etrequistion derertificat depropriété

On souhaiterait, idéalement, obtenir : Lavignolle (rectificative par herman) et requisition de rertificat de propriété

On obtient, après utilisation de cette librairie :

La vignoble rectificative par hermann et réquisition d certificat approprie te

Comme nous pouvons l'observer, le résultat n'est pas entièrement satisfaisant, notamment à cause de la surcorrection du nom de famille au début de la chaîne de caractères, et de l'échec de la reconstitution de « derertificat de propriété » en « de certificat de propriété ». De plus, on observe la disparition des parenthèses. Il faudrait donc pouvoir être en mesure de cibler la correction et la segmentation des mots dans la phrase. Le token « etrequistion » a cependant bien été reconstitué en « et réquisition », et « parherman » en « par herman ».

Il serait intéressant de continuer cette expérience en utilisant des dictionnaires de fréquence plus conséquents, qu'il est possible de construire avec les sets de n-grammes de Google.<sup>42</sup> Ceux-ci sont constitués à partir des millions de livres disponibles avec Google Books. Nous pourrions également envisager la construction d'un dictionnaire de fréquence

---

39. Voir <https://github.com/wolfgarbe/SymSpell/tree/master/SymSpell.FrequencyDictionary> (consulté le 13/08/21).

40. Voir <https://github.com/orgtre/frenchngrams> (consulté le 13/08/21). Le bigramme le plus fréquent dans ce corpus est *de la*, avec un total de 132 940 occurrences, suivi en seconde place par *à la* avec 56 794 occurrences.

41. Le projet Gutenberg est une bibliothèque de livres électroniques issus du domaine public. Voir <https://www.gutenberg.org/> (consulté le 13/08/21).

42. Voir <http://storage.googleapis.com/books/ngrams/books/datasetsv2.html> (consulté le 13/08/21).

de bigrammes à partir des vérités de terrain du projet LECTAUREP. Cela présenterait l'avantage d'avoir un dictionnaire de bigrammes construit spécifiquement sur les données du projet. Hypothétiquement, il serait possible de coupler un système probabilistique basé sur ce dictionnaire avec une distance de Levenshtein pour pouvoir l'appliquer à des données issues de la transcription automatique. On chercherait ainsi à trouver, pour un bigramme dans la transcription automatique, quel bigramme dans le dictionnaire est le plus proche en terme de distance de Levenshtein.

Mais en utilisant des dictionnaires, on se heurte au fait qu'ils ne seront jamais exhaustifs. Il est possible de rencontrer un bigramme dans un texte non enregistré dans le dictionnaire. Pour contourner ce problème, il est possible d'utiliser la longueur des mots en se basant sur la loi de Zipf. Selon celle-ci, les mots les plus fréquemment utilisés ont tendance à être courts, par exemple les déterminants. En se basant sur cette information, le chercheur Peter Norvig a créé une équation pour calculer la probabilité qu'une séquence de caractères non connue dans un dictionnaire soit un mot ou non : "Not all unseen words are equally unlikely : a random sequence of 20 letters is less likely to be a word than a random sequence of 6 letters."<sup>43</sup>

### 4.3.3 Utiliser des expressions régulières pour segmenter les mots ? Le risque de la non-exhaustivité.

Une étude systématique des fautes caractéristiques générées par les modèles de REM entraînés avec les répertoires des notaires pourrait nous donner un catalogue d'erreurs récurrentes à traiter. Les deux exemples présentés au début de cette partie en contiennent : un déterminant agglutiné à un nom commun ou un nom propre, un déterminant agglutiné à une date, ou encore deux noms propres agglutinés.

Selon ces cas, pour segmenter les mots, il pourrait suffire de concevoir un système de règles basé sur des expressions régulières. Par exemple, pour deux noms propres agglutinés, on peut s'aider de la majuscule pour les découper avec l'expression régulière suivante :

1 (\w)([A–Z])

Le premier groupe capture n'importe quelle lettre, et le deuxième groupe capture n'importe quelle majuscule qui lui est collée. Il suffit alors d'inclure un espace entre ces deux groupes de capture.

---

43. Peter Norvig, « Natural Language Corpus Data », dans *Beautiful Data*, 2009, p. 219-242, p. 224. Voir également : Wolf Garbe, *Fast Word Segmentation of Noisy Text*, Medium, 23 sept. 2020, URL : <https://towardsdatascience.com/fast-word-segmentation-for-noisy-text-2c2c41f9e8da> (visité le 12/07/2021), et plus précisément les sections "Known words" et "Unknown words". Nous recommandons également le billet de blog "Word segmentation, or Makingsenseofthis", écrit sous le pseudonyme *j2kun*, qui est un très bon guide pas à pas pour développer soi-même un système probabiliste à partir de n-grammes : *j2kun, Word Segmentation, or Makingsenseofthis...*

Cette solution paraît néanmoins risquée à cause de sa non-exhaustivité. Il semble difficilement réalisable d'écrire un système de règles qui couvrirait tous les cas rencontrés lors de la transcription automatique de documents écrits par plusieurs mains différentes dans les réertoires des notaires, s'étalant sur plus d'un siècle et demi.<sup>44</sup>

#### 4.3.4 Segmenter les mots avec une approche neuronale

Le chercheur Thibault Clérice, dans un article de 2020, explique que la segmentation des mots est une opération essentielle pour exploiter un texte, et qu'elle peut être traitée comme une tâche de TAL.<sup>45</sup> Certains types d'écritures sont ininterrompues, comme par exemple un phénomène d'écriture occidental, disparu aux alentours du VIIIe siècle, nommé *Scripta Continua*. Dans cet article, le chercheur évalue trois architectures d'apprentissage profond possédant toutes un encodeur et un classifieur linéaire afin obtenir un modèle capable de segmenter les mots issus d'une transcription d'un texte en *Scripta Continua*. La plus performante d'entre elles utilise un encodeur convolutif (CNN) sans plongements lexicaux, et atteint dans cette tâche un score de 0.99 d'*accuracy*.<sup>46</sup>

Le modèle prend en entrée une chaîne de caractères, qui sera ensuite encodée en attribuant à chaque caractère un index. Le modèle crée en sortie un masque, qui une fois appliqué à cette chaîne de caractères peut être décodé. Dans le masque, « characters are classified either as word boundary or word content. »<sup>47</sup>

Le chercheur a développé une application python, nommée *Boudams* pour tokeniser les textes en latin et en français médiéval.<sup>48</sup> Il précise qu'il est possible de constituer un set de données d'entraînement prenant le format suivant pour annoter une phrase : « sa-mesentence<TAB>same sentence ».<sup>49</sup> D'abord la phrase avec les mots non segmentées, une tabulation, puis la même phrase mais avec la bonne segmentation des mots. En s'appuyant sur ce travail de recherche et de développement, il serait pertinent d'adapter cette méthodologie sur les données de LECTAUREP. Le projet pourrait ainsi se doter d'un outil automatique pour segmenter les mots issus de la REM.

---

44. L'expression régulière illustrée dans cette sous-section a été utilisée durant mon stage. Elle présentait peu de risques de surcorrection et permettait d'effectuer un pré-traitement minimal des données en sortie de la REM.

45. T. Clérice, « Evaluating Deep Learning Methods for Word Segmentation of *Scripta Continua* Texts in Old French and Latin »..., p. 1 et p. 9

46. La traduction française de l'architecture du modèle a été effectué grâce à l'article de Romain Benassi pour le blog Publicis Sapient. Voir <https://blog.engineering.publicissapient.fr/2021/03/17/nlp-concepts-cles-et-etat-de-lart/> (consulté le 13/08/21). En anglais, l'architecture se nomme : *Convolutional (CNN) encoder without position embeddings*. Un tableau récapitulant les scores des architectures testées est disponible dans l'article de T. Clérice à la page 5. *Ibid.*, p. 5.

47. *Ibid.*, p. 2

48. Voir <https://github.com/ponteineptique/boudams> (consulté le 10/08/21).

49. Voir <https://github.com/ponteineptique/boudams#how-to> (consulté le 10/08/21).

Dans un contexte scientifique où les systèmes d'apprentissage machine sont toujours plus performants et possèdent l'avantage s'adapter aux données, un système de segmentation des mots de ce type semble plus adapté.

#### **4.3.5 La tokenisation des données textuelles**

Une fois la structure interne des phrases reconstituée au maximum, il est nécessaire de tokeniser le texte pour préparer la REN. Cette tâche est importante pour le bon déroulé de la REN. Elle repose sur des modèles de tokenisation, et est donc dépendante du domaine sur lequel ils sont entraînés. Le modèle de langue française camemBERT, et les chaînes de traitement de la langue française des librairies de TAL pourraient certainement être des outils efficaces. La tokenisation varie en effet d'une langue à l'autre, ainsi que d'une période à l'autre. Le français utilisé dans les répertoires des notaires se rapprochant du français contemporain que nous connaissons n'entraînera pas de soucis importants de performance. Néanmoins, nous pouvons supposer que les abréviations utilisées dans les répertoires de notaires ne seront pas toujours bien tokenisées, car leurs étant spécifiques. Mal tokeniser un mot pourrait entraîner une identification erronée par un modèle de REN.

### **4.4 Une étape optionnelle : la normalisation des abréviations**

Le texte des répertoires des notaires est composé d'un grand nombre d'abréviations, qui ont été transcris dans la vérité de terrain utilisée pour entraîner les modèles de REM. Il est envisageable de normaliser ces abréviations en les transformant à l'aide d'un script dans leur forme non-abrégée.

Cela aurait pour effet immédiat de faciliter l'accès aux transcriptions des répertoires des notaires à tous les publics. Cette étape n'est pas obligatoire, elle relève davantage des questions de médiation. Cette décision doit être prise en amont de la campagne de REN, car si un corpus d'entraînement pour un modèle de REN est constitué en gardant les abréviations, puis que ce modèle est appliqué à des données normalisées, ses résultats risquent d'être perturbés et donc de baisser en performance.

Pour normaliser les abréviations, il est possible d'utiliser des référentiels. Il est cependant nécessaire de retenir comment les abréviations ont été transcris afin de pouvoir créer un système les identifiant et les associant à une forme normalisée (voir section 2.5).

## 4.5 La correction orthographique post transcription automatique

Corriger l'orthographe des transcriptions automatiques permettrait de les rapprocher au maximum d'un français propre, et par extension des modèles de langue et des données sur lesquels sont entraînés la plupart des modèles de REN. Cependant, ce processus présente le risque d'éloigner les données de la vérité de terrain en modifiant les lettres de façon erronée. Nous avons expliqué précédemment qu'un humain peut tout à fait reconstituer mentalement plusieurs mots collés. Une mauvaise correction rendra cette tâche plus difficile, le sens du mot corrigé pouvant être changé. Cette opération pourrait perturber la consultation des transcriptions automatiques, notamment si les noms de famille sont surcorrigés, comme nous l'avons observé précédemment. Il serait donc envisageable de ne corriger, pour commencer par exemple, que les mots vides, comme les déterminants. La REN pourrait aider dans la réalisation de cette correction ciblée. Une première étape de reconnaissance permettrait de faire ressortir les entités, et de les ignorer lors de la correction.

Il existe deux approches principales pour le traitement post REM :

1. Une approche basée sur les dictionnaires qui permet de corriger les termes isolés, en ne prenant pas compte du contexte dans lequel ceux-ci s'insèrent.
2. Une approche qui prend en compte le contexte grammatical et sémantique des erreurs, et cherche à les résoudre en utilisant des modèles de langues et des modèles d'apprentissage machine.<sup>50</sup>

Au cours du stage, j'ai manipulé la librairie python *pyspellchecker* pour expérimenter la correction post REM.<sup>51</sup> La correction est effectuée en se basant sur le calcul de la distance de Levenshtein entre le mot erroné et sa forme corrigée dans un dictionnaire de fréquence. Le test, disponible dans un *notebook* Jupyter a été découpé en deux étapes.<sup>52</sup> Dans la première, on tente de corriger les noms à l'aide d'un référentiel de prénoms en ciblant les entités *PER*, donc reconnues comme noms de personnes. Les résultats n'ont pas été concluants, de par la nature du référentiel utilisé où les noms propres composés sont attachés, on ne peut donc pas différencier un nom propre simple d'un nom propre composé, rendant leur correction difficile ; et car le modèle générique de REN utilisé a faussement identifié des mots comme

50. T.T.H. Nguyen, A. Jatowt, M. Coustaty, *et al.*, « Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing »..., p. 31. Voir également, au sujet de la correction post REM : T. Nguyen, A. Jatowt, M. Coustaty, N. Nguyen et A. Doucet, « Post-OCR Error Detection by Generating Plausible Candidates », dans *2019 International Conference on Document Analysis and Recognition (ICDAR)*, ISSN : 2379-2140, 2019, p. 876-881, DOI : 10.1109/ICDAR.2019.00145

51. Voir <https://github.com/barrust/pyspellchecker> (consulté le 13/08/21).

52. Le *notebook* est consultable à cette adresse : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/preprocessing/word\\_segmentation/test\\_spellchecker/test\\_spellchecker.ipynb](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/preprocessing/word_segmentation/test_spellchecker/test_spellchecker.ipynb) (consulté le 13/08/21).

des noms de personnes alors qu'ils ne l'étaient pas. La deuxième partie n'a également pas donné de résultats satisfaisants. J'ai utilisé le dictionnaire de fréquence français fourni par la librairie pour corriger les mots vides et les noms communs, mais cela a échoué. Cela s'explique peut-être par le fait que *pyspellchecker* ne corrige les mots erronés que lorsque la distance de Levenshtein est de deux avec un mot dans le dictionnaire de référence. Le bruit produit par la REM aboutit très certainement à une distance supérieure à 2. Il faudrait donc pouvoir modifier cette valeur en l'augmentant, ce qui n'est pas possible pour le moment avec cette librairie. On observe également la surcorrection des mots vides, où les « à » sont corrigés en *a*. Cela peut être problématique pour la REN, « *a* », correspondant au verbe « avoir » conjugué à la troisième personne du présent, et « à », préposition, interviennent en effet dans des contextes différents au sein d'une phrase et n'introduisent pas les mêmes informations.

L'expérimentation et la mise en place de la correction post REM est actuellement réalisée par Alix Chagué.<sup>53</sup> Parmi les outils expérimentés, le meilleur candidat est une librairie python, *OCRFixr*, qui corrige en utilisant un modèle de langue, BERT, et le contexte dans lequel se trouve le mot au sein d'une phrase.<sup>54</sup> Après discussion avec ma tutrice de stage, il serait envisageable de brancher le modèle de langue française camemBERT afin d'adapter l'outil à des données en français.<sup>55</sup>

---

53. Voir l'*issue* Gitlab suivante : <https://gitlab.inria.fr/almanach/lectaurep/documentation/-/issues/16> (consulté le 13/08/21).

54. Voir <https://github.com/ja-mcm/OCRFixr> (consulté le 13/08/21)

55. Notons également la possibilité de créer son propre système de correction orthographique grâce à la librairie python d'apprentissage machine TensorFlow, dont le processus est décrit dans David Currie, *Creating a Spell Checker with TensorFlow*, Medium, 18 mai 2017, URL : <https://towardsdatascience.com/creating-a-spell-checker-with-tensorflow-d35b23939f60> (visité le 12/07/2021).



# Chapitre 5

## Les modèles génériques de reconnaissance d'entités nommées : une solution clé en main ?

Aujourd'hui, de nombreuses librairies python de TAL proposent des modèles génériques de REN qui peuvent être rapidement appliqués à des données textuelles. Un modèle générique est un modèle que l'on pourrait hypothétiquement utiliser sur n'importe quel corpus, indépendamment du domaine auquel il appartient.<sup>1</sup> La praticité de ces modèles permet d'expérimenter rapidement et facilement, et même de les envisager comme une solution clé en main. Par cette expression, on entend un système qui peut être aisément mis en place, avec un minimum, voire aucun réglage nécessaire. Évidemment, ces librairies nécessitent de savoir manier python et ne passent pas par une interface graphique. Considérant que ces compétences sont acquises, ont-elles des performances suffisantes sur des données bruitées pour être déployées à grande échelle dans un projet tel que LECTAUREP ? Déjà en 2016, les chercheurs Shubhangshu Mishra et Jana Diesner publiaient un modèle de REN entraîné sur des données bruitées issues de tweets et de textes publiés sur internet, pour pallier aux modèles génériques entraînés sur des données propres, avec une syntaxe régulière, dont la performance était insuffisante sur des textes hors domaine.<sup>2</sup>

En outre, en 2019, un groupe de chercheurs des universités de Lorraine et du Luxembourg déplorait la disparité des méthodes d'évaluation des solutions de REN disponibles et de la difficulté que cela causait pour en choisir une pour les professionnels.<sup>3</sup> Ils ont proposé par

---

1. En anglais, "domain-independent." Ces modèles génériques s'opposent à des modèles entraînés à destination de domaines spécifiques, comme par exemple la santé.

2. S. Mishra et J. Diesner, « Semi-supervised Named Entity Recognition in noisy-text »...

3. Xavier Schmitt, Sylvain Kubler, Jérémie Robert, Mike Papadakis et Yves Le Traon, « A replicable comparison study of NER software : StanfordNLP, NLTK, OpenNLP, SpaCy, Gate », dans *Sixth International Conference on Social Networks Analysis, Management and Security, SNAMS 2019*, Grenada, Spain, 2019, DOI : 10.1109/SNAMS.2019.8931850

conséquent une méthodologie reproductible pour mener des analyses comparatives de ces outils. Deux ans après, avec le progrès des technologies et la mise à jour des outils, il est nécessaire de continuer à les comparer. Dans le cas de données issues de la REM à partir de textes en langue française, ce travail reste à faire. Ce chapitre propose de dresser une évaluation d'outils de REN dans ce contexte, réalisée au début de mon stage.

Après une exploration des données du projet LECTAUREP, il a été décidé de mener les expériences de REN dans la cinquième colonne, où sont donc renseignés les noms, les prénoms et domiciles des parties, contenant des informations rapportées sous forme de phrases, contrairement aux autres colonnes où l'information est généralement une valeur unique.

## 5.1 Tour d'horizon des modèles génériques de reconnaissance d'entités nommées disponibles avec les librairies de TAL python

Avant d'étudier les performances des modèles génériques de REN, il convient de présenter les librairies de TAL python disponibles en ligne. La grande majorité incluent un module générique de REN dans leurs chaînes de traitement. Les modèles génériques disponibles en ligne sont entraînés sur des corpus issus de différentes langues. On trouve donc des modèles de REN générique pour l'anglais, l'allemand, l'espagnol, le français, etc. Précisons que ces modèles n'ont pas été conçus, du moins à notre connaissance, en ayant considéré leur application sur des données historiques.

### 5.1.1 SpaCy

SpaCy est une librairie de TAL *open-source* écrite en Python et Cython (un langage de programmation incluant Python et un sous-ensemble du langage C/C++).<sup>4</sup> Elle peut être utilisée avec des chaînes de traitement différentes selon les langues disponibles.<sup>5</sup> SpaCy utilise des méthodes d'apprentissage profond tel que les *transformers* pour effectuer des tâches *tagging*, *parsing*, de la REN et de la classification de texte.

Il existe quatre chaînes de traitement pour le français. Pour tester la librairie, nous avons utilisé celle nommée « fr\_core\_news\_lg », pesant 545 MB, avec des modèles entraînés sur des textes issus de la presse et d'internet. Le corpus d'entraînement rassemble entre autres des articles de Wikipedia, le corpus WikiNER (un ensemble de textes de Wikipedia annotés

---

4. Voir <https://github.com/explosion/spaCy> et <https://spacy.io/> (consultés le 23/08/21).

5. SpaCy possède des chaînes de traitement pour plus de 60 langages. Voir <https://spacy.io/usage/models#languages> (consulté le 23/08/21).

en EN)<sup>6</sup>, et le corpus Sequoia.<sup>7</sup> Celui-ci est un corpus en français syntaxiquement annoté de textes issus du Parlement européen, du journal l'*Est Républicain*, du Wikipedia français, et de documents de l'Agence Européenne du Médicament, représentant un total de 3204 phrases et 69 246 tokens annotés.<sup>8</sup> Le modèle de REN permet d'extraire les EN de type « PER », « LOC », « ORG », et « MISC ». Quatre types d'entités, dont une difficilement exploitable en tant que telle (MISC), contre dix-huit pour la chaîne de traitement en langue anglaise. En plus des entités déjà présentes pour le modèle français, on retrouve les entités de date, de quantité, de temps, monétaires, etc.<sup>9</sup> Le seul pré-requis pour effectuer la REN avec SpaCy est la tokenisation.

### 5.1.2 Stanza

La deuxième librairie *open-source* écrite en Python que nous avons testé se nomme Stanza, elle a été créée par le Stanford NLP Group. À l'instar de SpaCy, cette librairie propose d'effectuer les tâches les plus courantes du TAL, et propose un modèle de REN générique pour le français.<sup>10</sup>

Le modèle de REN disponible pour le français a été entraîné avec WikiNER, et dispose des quatre mêmes étiquettes disponibles avec le modèle de REN de SpaCy. L'architecture du système de REN est composée d'un modèle de langue LSTM entraîné au niveau des caractères, qui fournit des représentations pour chacun d'eux, représentations qui sont ensuite concaténées de sorte à former des mots. Celles-ci sont ensuite traitées par un modèle mono-couche Bi-LSTM, puis par un décodeur CRF.<sup>11</sup> La REN avec Stanza ne nécessite également que la tokenisation.

### 5.1.3 Autres librairies de TAL n'ayant pas de modèle de REN pour le français

Nous mentionnons également ici deux autres librairies de TAL ne possédant pas de modèle adapté au français pour la REN directement implémenté dans la librairie, mais proposant la possibilité d'en entraîner un.

---

6. Voir <https://metatext.io/datasets/wikiner> (consulté le 23/08/21).

7. Voir [https://spacy.io/models/fr#fr\\_core\\_news\\_lg](https://spacy.io/models/fr#fr_core_news_lg).

8. Marie Candito et D. Seddah, « Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical », dans, 2012, URL : <https://hal.inria.fr/hal-00698938> (visité le 24/08/2021), p. 1

9. Voir [https://spacy.io/models/en#en\\_core\\_web\\_lg](https://spacy.io/models/en#en_core_web_lg) (consulté le 23/08/21).

10. Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton et Christopher D. Manning, « Stanza : A Python Natural Language Processing Toolkit for Many Human Languages », *arXiv :2003.07082 [cs]* (, 23 avr. 2020), arXiv : 2003.07082, URL : <http://arxiv.org/abs/2003.07082> (visité le 22/04/2021) et <https://stanfordnlp.github.io/stanza/index.html> (consulté le 23/08/21). Stanza permet de traiter 66 langues.

11. *Ibid.*, p. 3

La première se nomme FLAIR et est développée par l'université Humboldt de Berlin. Il est possible d'utiliser un modèle de REN pour le français possédant 4 étiquettes (PER, LOC, ORG, et MISC) en le téléchargeant sur le site de la librairie *Hugging Face*, mais il n'a pas été testé.<sup>12</sup>

La seconde est une suite de librairies réputées de TAL, proposant d'importantes ressources pour la langue anglaise, créée en 2001, nommée NLTK.<sup>13</sup> L'architecture de REN utilisée par NLTK commence par traiter une chaîne de caractères, qui est ensuite segmentée en phrases. Celles-ci sont tokénisées, puis les tokens passent par une phase de *part-of-speech tagging*. Après cela, les entités sont détectées.<sup>14</sup>

## 5.2 Une étude des performances des modèles génériques par seuil de taux d'erreur de caractères selon une performance idéale

Nous avons présenté les caractéristiques des corpus d'entraînement des chaînes de traitement pour la langue française de SpaCy et Stanza, dont la nature est donc différente des transcriptions automatiques des répertoires des notaires contenant du bruit généré par la REM. En évaluant les modèles génériques sur ces données, on met les d'emblée dans une position de difficulté. Il serait même nécessaire de revoir l'objectif de cette évaluation : nous n'évaluons ici pas seulement leurs performances, mais aussi leur capacité à s'adapter à des données hors domaines.

### 5.2.1 Évaluation de la chaîne de traitement de SpaCy

Une transcription issue de la REM avec un taux de CER de 21% a été choisie pour évaluer la chaîne de traitement de SpaCy.<sup>15</sup> L'évaluation a été réalisée dans un *notebook*<sup>16</sup>,

---

12. Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter et Roland Vollgraf, « FLAIR : An Easy-to-Use Framework for State-of-the-Art NLP », dans *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Minneapolis, Minnesota, 2019, p. 54-59, DOI : 10.18653/v1/N19-4010 et <https://github.com/flairNLP/flair>. Pour le modèle téléchargeable, voir <https://huggingface.co/flair/ner-french> (consultés le 23/08/21).

13. Voir le *NLTK Book*, <https://www.nltk.org/book/> (consulté le 23/08/21).

14. Voir chapitre 1.1 "Information Extraction Architecture", <https://www.nltk.org/book/ch07.html> (consulté le 23/08/21).

15. La transcription sur laquelle a été effectué le test se trouve à l'adresse suivante : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/corpus\\_test/lectaurep/doc\\_28\\_sample/HTR\\_cer\\_21/FRAN\\_0025\\_0029\\_L-0.xml](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/corpus_test/lectaurep/doc_28_sample/HTR_cer_21/FRAN_0025_0029_L-0.xml) (consulté le 18/08/21).

16. [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner\\_python/ner\\_spacy/ner\\_evaluation\\_HTR\\_cer\\_21.ipynb](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_spacy/ner_evaluation_HTR_cer_21.ipynb) (consulté le 18/08/21). Note : La visualisation des EN dans le texte testé est affichée à la fin des *notebooks* de test où SpaCy a été utilisée grâce à la librairie python DisplaCy. Voir <https://spacy.io/usage/visualizers/> (consulté le 18/08/21).

et suit le processus suivant :

1. Ouverture du fichier PAGE XML, récupération du texte de la cinquième colonne et segmentation minimale des mots à l'aide des majuscules.
2. Pré-annotation du texte en EN avec le modèle générique. Les prédictions ont ensuite été exportées et corrigées dans une plate-forme d'annotation, de sorte à constituer une vérité de terrain de REN.
3. Application du modèle de REN sur le texte
4. Comparaison de la prédition REN et de la vérité de terrain pour obtenir la précision, le rappel et le F-score.

La structure logique n'a pas été reconstituée pour ce test afin de mesurer les résultats d'un modèle générique sur des données brutes en sortie directe de REM.

Pour constituer la vérité de terrain, un premier résultat de REN obtenu avec SpaCy a été exporté. Les annotations ont été corrigées avec la plate-forme *open-source* d'annotation Doccano.<sup>17</sup> Pour l'utiliser, une version locale de l'application a été installée. Nous avons repris les étiquettes disponibles pour la chaîne de traitement testée, c'est-à-dire « PER », « LOC », « ORG », et « MISC » lors du paramétrage de l'application.<sup>18</sup> L'annotation s'effectue en sélectionnant une suite de caractères pour lui assigner une étiquette.<sup>19</sup> Au début du stage, l'installation simple et l'utilisation intuitive de Doccano a fait de celui-ci un outil de découverte de l'annotation d'EN efficace.

La création d'une vérité de terrain sur un échantillon en sortie de REM est une tâche particulière, car demande d'annoter des mots bruités. Cela s'apparente, à certains égards, à travailler à partir d'une autre langue proche du français. Pour corriger la pré-annotation, il a été décidé de se placer au niveau des tokens, et d'annoter ceux-ci dès lors que l'un d'eux aurait dû être détecté, par exemple, les noms de personnes au début d'une rangée dans la cinquième colonne. Les corrections de la pré-annotation concernent : les entités manquées, la segmentation des entités, la correction des entités faussement identifiées, la suppression des entités faussement détectées. Trois règles d'annotation ont été établies, les titres de civilité n'ont pas été relevés, les adresses ont été annotées dans leur forme complète, en incluant le mot « rue », et aucune EN de type *MISC* n'a été ajoutée à cause de sa définition trop large.

Doccano permet d'exporter les résultats de l'annotation au format JSONL, pour JSON Lines, où un objet JSON peut être contenu sur une seule ligne.<sup>20</sup>, qu'il est ensuite possible

---

17. Voir <https://github.com/doccano/doccano> (consulté le 18/08/21). Voir également les annexes H.1 et H.2.

18. Voir annexe H.3.

19. Voir annexe H.4.

20. Voir le fichier [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner\\_python/ner\\_spacy/spacy\\_ner\\_to\\_doccano/doc28\\_sample\\_cer\\_21.jsonl](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_spacy/spacy_ner_to_doccano/doc28_sample_cer_21.jsonl) (consulté le 18/08/21).

de convertir avec SpaCy dans le format JSON que la librairie utilise pour traiter les vérités de terrain. SpaCy propose un module pour évaluer des résultats de REN, que nous avons utilisé pour ce test.<sup>21</sup> Les scores obtenus après comparaison de la vérité de terrain et la prédiction ont été reportés dans le tableau suivant :

	Precision	Recall	F-score
Toutes entités	0.46	0.51	0.48
PER	<b>0.65</b>	0.48	<b>0.55</b>
LOC	0.40	<b>0.59</b>	0.48
ORG	0.16	0.12	0.14
MISC	0.12	0.5	0.2

TABLE 5.1 – Tableaux des scores du modèle générique de REN pour la chaîne de traitement *fr\_core\_news\_lg* de SpaCy sur une transcription automatique (CER de 21%) d'une page aléatoire d'un répertoire de notaire sans reconstitution de la structure logique.

On observe que les scores obtenus sont faibles. En prenant en compte toutes les entités, c'est environ 5 entités sur 10 qui sont correctement identifiées. Les EN de type « PER » ont le meilleur score de précision, indiquant que le modèle obtient de meilleures performances dans la bonne identification de cette étiquette. Le meilleur score de rappel est obtenu pour le type « LOC », le modèle rate donc moins d'EN de ce type. Pour les étiquettes « ORG » et « MISC », les très faibles scores s'expliquent par le peu d'EN de ce type contenues dans la vérité de terrain. Les scores obtenus, sur un taux de CER relativement élevés, ne sont donc pas suffisants pour envisager l'application du modèle générique à grande échelle sur les données de LECTAUREP, d'autant plus si le CER n'a pas été contrôlé. Ces résultats posent la question de savoir si les scores s'améliorent avec un CER plus bas.

### 5.2.2 Évaluation de la chaîne de traitement de Stanza

Une transcription ayant un taux de CER de 9% a été choisie pour réaliser une deuxième évaluation afin de répondre à la question précédente. Le texte a été pré-traité en vue de la REN, en réalisant une segmentation des mots minimale à l'aide d'expressions régulières et en normalisant les abréviations issues de la REM. Pour obtenir les scores d'évaluation, deux *notebooks* ont été utilisés.

Dans le premier, nous avons utilisé la chaîne de traitement de la langue française de la librairie Stanza sur le texte susmentionné pour obtenir une prédiction de REN à évaluer, selon plusieurs étapes<sup>22</sup> :

---

21. Voir l'objet *Example* de SpaCy, <https://spacy.io/api/example> (consulté le 18/08/21).

22. Voir le *notebook* suivant [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner\\_python/ner\\_stanza/create\\_bioes\\_annotation\\_text\\_file\\_with\\_stanza.ipynb](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_stanza/create_bioes_annotation_text_file_with_stanza.ipynb) (consulté le 18/08/21).

1. Ouverture du fichier texte contenant la transcription pré-traitée.<sup>23</sup>
2. Reconstitution de la cinquième colonne dans une liste python, où chaque index représente une séquence de caractères correspondant à un enregistrement.
3. Application du modèle générique de REN de Stanza à chaque index de la liste.
4. Création d'un fichier texte au format BIOES à partir de la REN.<sup>24</sup>

Un fichier texte BIOES est un format d'annotation pour les tokens utilisé pour la tâche de REN.<sup>25</sup> Il prend la forme suivante :

Token Tag

... ...

Chaque token est inscrit sur une ligne et suivi, après un espace, d'un *tag*. Ceux-ci sont les suivants :

- B : *beginning*, pour un token marquant le début d'une EN polylexicale.
- I : *inside*, pour un token contenu dans une EN polylexicale.
- O : *outside*, pour un token non concerné par une EN.
- E : *ending*, pour un token marquant la fin d'une EN polylexicale.
- S : *single element*, pour désigner une EN constituée d'un seul token.

À ces *tags* s'ajoutent l'étiquette d'EN concernée par le token. Pour la phrase suivante : « Léon III l'Isaurien était un empereur byzantin, né à Germanicia, en Turquie actuelle. », on aurait l'annotation BIOES suivante :

Listing 5.1 – Exemple d'annotations au format BIOES

- 1 Léon B–PER
- 2 III I–PER
- 3 l' I–PER
- 4 Isaurien E–PER
- 5 était O
- 6 un O
- 7 empereur O

---

23. Le pré-traitement du texte a été obtenu à l'aide d'un CLI créé pour réaliser cette tâche par lot que nous présenterons plus tard dans le déroulé de ce mémoire. Voir [https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/preprocessing/word\\_segmentation](https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/preprocessing/word_segmentation) (consulté le 18/08/21).

24. Le fichier résultant de la REN au format BIOES est disponible à l'adresse suivante : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/test\\_NERVAL/data/FRAN\\_0025\\_5038\\_L-0\\_htr\\_cer\\_9.bioes](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/test_NERVAL/data/FRAN_0025_5038_L-0_htr_cer_9.bioes) (consultée le 18/08/21).

25. Voir [https://en.wikipedia.org/wiki/Inside%20%93outside%20%93beginning\\_\(tagging\)](https://en.wikipedia.org/wiki/Inside%20%93outside%20%93beginning_(tagging)) (consulté le 18/08/21).

8 byzantin O  
9 , O  
10 né O  
11 à O  
12 Germanicia S—LOC  
13 , O  
14 en O  
15 Turquie S—LOC  
16 actuelle O  
17 . O

Pour évaluer cette prédiction, la librairie python NERVAL a été testée dans un second *notebook*.<sup>26</sup> NERVAL est développé par Teklia, une entreprise spécialisée dans l'intelligence artificielle, et plus particulièrement le TAL, l'OCR et la REM.<sup>27</sup> Elle a été conçue pour évaluer les résultats des modèles de NER appliqués à des données bruitées issues de la REM et de l'OCR.<sup>28</sup> NERVAL propose d'évaluer une prédiction REN réalisée sur une transcription automatique avec une vérité de terrain correspondant au texte original, donc transcrit manuellement, annotée en EN. Pour cela, NERVAL aligne la transcription automatique avec la vérité de terrain au niveau du caractère en minimisant la distance de Levenshtein ; chaque entité annotée dans la vérité de terrain est ensuite associée à une entité possédant la même étiquette et située au même endroit (index de début et index de fin) dans la transcription automatique alignée, ou avec un caractère vide si aucune association n'est trouvée ; une entité est enfin considérée comme reconnue dans la transcription si la distance d'édition avec l'entité de la vérité de terrain est, par défaut, de moins de 30%. La distance d'édition est paramétrable en déterminant un seuil lors de l'utilisation de NERVAL.

La vérité de terrain REN a donc été constituée en annotant la vérité de terrain de la transcription du texte. La prédiction a été obtenue en utilisant le modèle de Stanza sur la transcription automatique qui lui correspond. Le texte a été importé dans la plate-forme d'annotation *open-source* Inception, dont nous laissons la présentation pour une partie ultérieure de ce mémoire.<sup>29</sup> L'annotation du texte a été effectuée sur un texte non pré-annoté, et a suivi les mêmes directives que celles présentées précédemment. Inception permet d'exporter les annotations au format CONLL 2002, qui ont ensuite été converties au format BIOES.<sup>30</sup>

---

26. Voir [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/test\\_NERVAL/test\\_nerval.ipynb](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/test_NERVAL/test_nerval.ipynb) (consulté le 18/08/21).

27. Voir <https://teklia.com/company/> (consulté le 18/08/21).

28. Voir Blanche Miret, *Teklia - NERVAL : A NER evaluation python package for noisy text*, 2021, URL : <https://teklia.com/blog/202104-nerval/> (visité le 23/04/2021) et le répertoire Github de la librairie : <https://gitlab.com/teklia/nerval> (consulté le 18/08/21).

29. Voir le projet Inception, <https://inception-project.github.io/> (consulté le 18/08/21).

30. Voir [https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/test\\_NERVAL/](https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/test_NERVAL/)

Les résultats de l'évaluation obtenus avec NERVAL ont été reportés dans des tableaux reproduits ici.

	Precision	Recall	F-score	Predicted	Matched
Toutes entités	0.50	0.54	0.52	79	40
PER	<b>0.53</b>	<b>0.65</b>	<b>0.59</b>	<b>43</b>	<b>23</b>
LOC	<b>0.53</b>	0.48	0.50	32	17
ORG	0.0	0.0	0.0	2	0
MISC	0.0	None	None	2	0

TABLE 5.2 – Tableaux des scores du modèle générique de REN pour la chaîne de traitement de Stanza sur une transcription automatique (CER de 9%) d'une page aléatoire d'un répertoire de notaire avec reconstitution de la structure logique, obtenus avec la librairie NERVAL, avec un seuil de distance d'édition réglé à 0.30.

Ces premiers résultats (table 5.2) ont été obtenus avec les réglages par défaut, donc avec un seuil pour la distance d'édition de 30%. On observe qu'ils sont légèrement meilleurs que ceux obtenus avec SpaCy. Cela montre que les modèles génériques ont du mal à atteindre de bonnes performances sur des données hors domaines contenant du bruit généré par la REM. NERVAL donne, en complément de la précision, du rappel et du F-score, le nombre d'entités qui ont été prédites, et le nombre d'entités qui ont été correctement identifiées et associées (*matched*) avec une entité dans la vérité de terrain. En moyenne, le modèle arrive à correctement identifier et classifier la moitié des EN.

En descendant le seuil de distance d'édition à 20%, les performances baissent de façon logique (table 5.3). À 0%, NERVAL ne considère qu'une entité est correctement identifiée qu'au moment où elle est parfaitement transcrise, ce qui entraînent des scores très faibles (table 5.4). En augmentant ce seuil, donc en rendant l'évaluation plus flexible, les scores ne montent que de peu de points (table 5.5).

NERVAL est une librairie simple d'utilisation, ne nécessitant qu'une ligne de commande pour obtenir un résultat d'évaluation.<sup>31</sup> Néanmoins, pour être correctement utilisée, elle sous-entend que le modèle de REN utilisé ait été entraîné à partir de transcriptions manuelles annotées en EN, donc exemptes de bruit, pour ensuite mesurer son efficacité sur des textes de même nature obtenus grâce à de la REM, mais contenant du bruit. Deux chemins se dessinent donc : un premier consistant à entraîner un modèle de REN sur des transcriptions manuelles

---

data. L'export CONLL 2002 a été converti au format BIOES grâce à un script, trouvé à l'adresse suivante : <https://github.com/jiesutd/NCRFpp/blob/master/utils/tagSchemeConverter.py#L16>. L'export CONLL 2002 est présent à l'adresse suivante : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/test\\_NERVAL/data/FRAN\\_0025\\_5038\\_L-0\\_gt.conll](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/test_NERVAL/data/FRAN_0025_5038_L-0_gt.conll), et sa conversion au format BIOES ici : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/test\\_NERVAL/data/FRAN\\_0025\\_5038\\_L-0\\_gt.bioes](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/test_NERVAL/data/FRAN_0025_5038_L-0_gt.bioes) (consultés le 18/08/21).

31. Seulement deux arguments sont requis pour lancer une évaluation, un chemin de fichier vers la vérité de terrain REN et un chemin de fichier vers la prédiction REN.

	Precision	Recall	F-score	Predicted	Matched
Toutes entités	0.45	0.48	0.47	79	36
PER	0.46	<b>0.57</b>	<b>0.51</b>	<b>43</b>	<b>20</b>
LOC	<b>0.5</b>	0.45	0.47	32	16
ORG	0.0	0.0	0.0	2	0
MISC	0.0	None	None	2	0

TABLE 5.3 – Tableaux des scores du modèle générique de REN pour la chaîne de traitement de Stanza sur une transcription automatique (CER de 9%) d'une page aléatoire d'un répertoire de notaire avec reconstitution de la structure logique, obtenus avec la librairie NERVAL, avec un seuil de distance d'édition réglé à 0.20.

	Precision	Recall	F-score	Predicted	Matched
Toutes entités	0.29	0.31	0.30	79	23
PER	0.18	0.22	0.20	<b>43</b>	8
LOC	<b>0.46</b>	<b>0.42</b>	<b>0.44</b>	32	<b>15</b>
ORG	0.0	0.0	0.0	2	0
MISC	0.0	None	None	2	0

TABLE 5.4 – Tableaux des scores du modèle générique de REN pour la chaîne de traitement de Stanza sur une transcription automatique (CER de 9%) d'une page aléatoire d'un répertoire de notaire avec reconstitution de la structure logique, obtenus avec la librairie NERVAL, avec un seuil de distance d'édition réglé à 0.

	Precision	Recall	F-score	Predicted	Matched
Toutes entités	0.53	0.56	0.54	79	42
PER	<b>0.58</b>	<b>0.71</b>	<b>0.64</b>	<b>43</b>	<b>25</b>
LOC	0.53	0.48	0.50	32	17
ORG	0.0	0.0	0.0	2	0
MISC	0.0	None	None	2	0

TABLE 5.5 – Tableaux des scores du modèle générique de REN pour la chaîne de traitement de Stanza sur une transcription automatique (CER de 9%) d'une page aléatoire d'un répertoire de notaire avec reconstitution de la structure logique, obtenus avec la librairie NERVAL, avec un seuil de distance d'édition réglé à 0.40.

du projet LECTAUREP, et un deuxième portant l'entraînement d'un modèle de REN sur des données issues de la REM, débruitées au maximum mais contenant toujours un minimum de bruit.

### 5.2.3 Observation des performance des modèles génériques sur une vérité de terrain

Afin de terminer cette série d'expériences, il convient d'observer les résultats d'un modèle générique sur une transcription manuelle. Pour cela, nous utiliserons le modèle générique

de SpaCy, et suivrons le même processus expérimental décrit dans la sous-section 5.2.1. L'expérience sera réalisée dans un premier temps sur un texte issu de la cinquième colonne sans reconstitution des enregistrements, puis dans un deuxième temps sur ce même texte avec reconstitution des enregistrements en les segmentant. Ces deux expériences ont été réalisées dans deux *notebooks* différents.<sup>32</sup> Les résultats de la première évaluation ont été reportés dans la table 5.6, et ceux de la deuxième dans la table 5.7.

	Precision	Recall	F-score
Toutes entités	0.60	0.61	0.61
PER	<b>0.75</b>	<b>0.70</b>	<b>0.72</b>
LOC	0.57	0.54	0.56
ORG	0.42	0.42	0.42
MISC	None	None	None

TABLE 5.6 – Tableaux des scores du modèle générique de REN pour la chaîne de traitement *fr\_core\_news\_lg* de SpaCy sur une transcription manuelle (vérité de terrain) d'une page aléatoire d'un répertoire de notaire sans reconstitution de la structure logique.

	Precision	Recall	F-score
Toutes entités	0.60	0.62	0.61
PER	<b>0.80</b>	<b>0.70</b>	<b>0.75</b>
LOC	0.52	0.54	0.53
ORG	0.42	0.42	0.42
MISC	None	None	None

TABLE 5.7 – Tableaux des scores du modèle générique de REN pour la chaîne de traitement *fr\_core\_news\_lg* de SpaCy sur une transcription manuelle (vérité de terrain) d'une page aléatoire d'un répertoire de notaire avec reconstitution de la structure logique

On observe que, de manière générale, les deux expériences possèdent un score supérieur de 10 points pour la reconnaissance de toutes les entités confondues. Cette hausse n'est pas significative, mais nous indique que des données hors domaines « propres » mettent le modèle générique dans une difficulté moindre. Sans reconstitution des phrases des différentes rangées de la cinquième colonne, les meilleurs scores concernent les entités de type « PER ».

---

32. La première a été réalisée dans le *notebook* suivant : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner\\_python/ner\\_spacy/ner\\_evaluation\\_anno\\_manual.ipynb](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_spacy/ner_evaluation_anno_manual.ipynb). Les fichiers utilisés pour le test, le texte utilisé pour la prédiction, ainsi que la vérité de terrain REN obtenu avec Doccano sont disponibles ici : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/corpus\\_test/lectaurep/doc\\_28\\_sample/manual\\_transcription/no\\_sentence\\_segmentation](https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/corpus_test/lectaurep/doc_28_sample/manual_transcription/no_sentence_segmentation); et la deuxième : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner\\_python/ner\\_spacy/ner\\_evaluation\\_annotation\\_manuelle\\_sent\\_seg.ipynb](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_spacy/ner_evaluation_annotation_manuelle_sent_seg.ipynb). Les fichiers utilisés pour le test sont disponibles ici : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/corpus\\_test/lectaurep/doc\\_28\\_sample/manual\\_transcription/artificial\\_sentence\\_segmentation](https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/corpus_test/lectaurep/doc_28_sample/manual_transcription/artificial_sentence_segmentation) (consultés le 18/08/21).

La précision nous indique que le modèle identifie entre 7 et 8 EN sur 10 de ce type, et le rappel qu'il en rate 3 sur 10. Les scores atteints avec la segmentation des phrases sont à peine meilleurs, seulement de 0.01 point pour le rappel. Cependant, nous observons que la précision est meilleure pour la reconnaissance des « PER », entraînant un meilleur F-score. La performance globale pour les « LOC » quand à elle baisse légèrement. Il est difficile de prendre mesure de l'efficacité du modèle à reconnaître les entités de type « ORG » car elles ne sont que très peu représentées dans le texte. Il en va de même pour les « MISC ».

## 5.3 Des performances aléatoires pour les modèles génériques de reconnaissance d'entités nommées ?

### 5.3.1 Cas d'usage de l'annotation des noms de personnes et des noms d'organisation en début d'enregistrement

Cette expérience a été réalisée pour déterminer si oui ou non le modèle générique de la chaîne de traitement de la langue française de la librairie SpaCy présente des résultats aléatoires selon le taux de CER, et non des performances de plus en plus faibles plus le taux de CER augmente, comme ce à quoi on pourrait s'attendre logiquement.

Mon hypothèse de départ était que les modèles de REN, appliqués aux données hors domaines du projet LECTAUREP, présenteraient des résultats aléatoires selon des taux de CER différents, et que ces résultats ne seraient pas nécessairement meilleurs plus le CER est bas. Pour tester cela, nous pouvons imaginer une tâche, en apparence simple pour un humain, et évaluer la performance du modèle sur cette tâche particulière. En se projetant dans l'exploitation des résultats de la REN dans le cadre du projet LECTAUREP, nous pourrions supposer qu'il serait utile de récupérer de façon régulière les noms de personnes ou d'organisation concernés par les enregistrements, de sorte à envisager l'indexation de ces entités. Les enregistrements commencent systématiquement par ces entités. En prenant en compte cette caractéristique, un modèle performant propre aux réertoires des notaires devrait pouvoir les récupérer efficacement, d'autant plus que cette structure est fixe. Idéalement, pour une page avec 15 enregistrements, on souhaiterait récupérer 15 tokens en tout début de phrase reconnus comme des entités représentant des noms de personnes ou des noms d'organisation.

J'ai donc entrepris cette évaluation en calculant l'*accuracy* du modèle pour cette tâche. Cette expérience a été réalisée sur deux pages choisies aléatoirement, sur lesquelles ont été appliqués des modèles de REM dans le but d'obtenir des transcriptions avec des taux de CER différents. J'ai téléchargé plusieurs modèles disponibles sur le Gitlab *Kraken models*, qui sert à stocker et documenter les modèles de segmentation et de REM produits dans le cadre des

projets d'humanités numériques d'ALMAncH, et les ai appliqués sur ces pages.<sup>33</sup> L'export texte a ensuite été comparé à la vérité de terrain, une transcription manuelle, avec KaMI en comparant ces deux chaînes de caractères, afin d'obtenir un CER.<sup>34</sup> J'ai ainsi obtenu :

- Une première page possédant 33 enregistrements, avec des transcriptions ayant des taux de CER s'élevant à : 14%, 17%, 17%, 21%, et 31%. Les deux transcriptions à 17% ont été obtenues avec deux modèles différents, dans le but d'observer si la performance du modèle était strictement liée au CER, ou s'il s'agissait plutôt de la nature des erreurs qui varient donc d'un modèle à l'autre.
- Une deuxième page possédant 20 enregistrements avec des transcriptions ayant des taux de CER s'élevant à : 8%, 18%, 25%, 31% et 36%.

Bien que des taux de CER s'élevant à plus de 20% rendent les transcriptions inexploitables<sup>35</sup>, il était intéressant d'observer comment l'extraction des entités et l'attribution d'une étiquette variaient en fonction du CER. Les pages ont été testées avec deux *notebooks* différents.<sup>36</sup>

Le protocole expérimental se résume à quatre étapes :

1. Reconstitution de la structure logique de la cinquième colonne, et stockage de chaque enregistrement dans une liste python.
2. Segmentation des mots minimale dans chaque enregistrement à l'aide d'expressions régulières.
3. Récupération du nombre d'enregistrements présents sur la page et récupération du nombre du nombre d'entités *PER* ou *ORG* en première position dans tous les enregistrements.
4. Calcul de l'*accuracy*.

Soit  $y$  le nombre d'entités correctement étiquetées en tant que « PER » ou « ORG », et  $x$  le nombre d'enregistrements présents sur une page testée, pour obtenir l'*accuracy* en pourcentage on réalise l'opération suivante :

$$\frac{y}{x} \times 100$$

---

33. Voir <https://gitlab.inria.fr/dh-projects/kraken-models> (consulté le 16/08/21).

34. Pour la comparaison entre deux chaînes de caractères avec KaMI, voir <https://gitlab.inria.fr/dh-projects/kami/kami-lib/-/issues/20> (consulté le 16/08/21).

35. M.L. Bonhomme, *Défis et opportunités de la reconnaissance automatique d'écriture manuscrite pour les documents d'archives : l'exemple des répertoires des notaires de Paris*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Jean-Baptiste Camps..., p. 7

36. Le test de la première page choisie se trouve à l'adresse suivante : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner\\_python/ner\\_spacy/NER\\_first\\_token\\_experience\\_1.ipynb](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_spacy/NER_first_token_experience_1.ipynb), et le test de la deuxième page à l'adresse suivante : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner\\_python/ner\\_spacy/NER\\_first\\_token\\_experience\\_2.ipynb](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_spacy/NER_first_token_experience_2.ipynb) (consultés le 16/08/21).

Les résultats ont été ensuite représentés avec deux diagrammes en colonnes, que nous reproduisons ici (diagramme 5.1 et diagramme 5.2).

Accuracy pour l'extraction des premiers tokens PER et ORG de chaque enregistrement

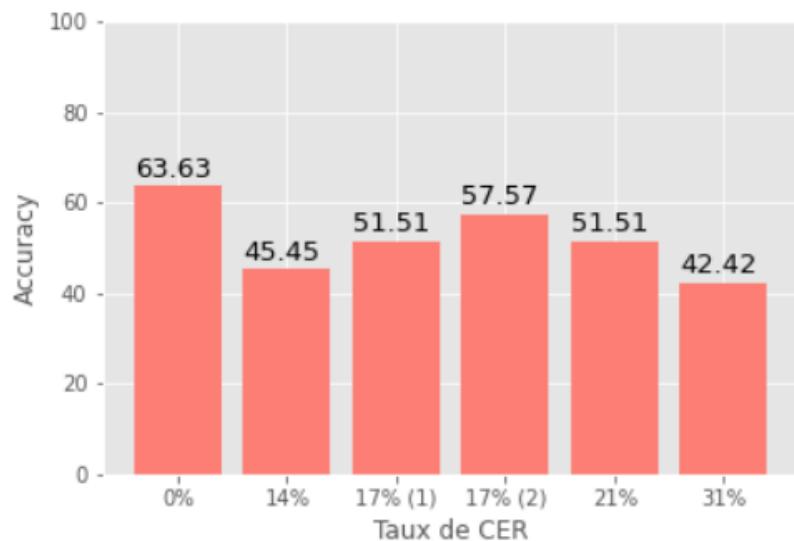


FIGURE 5.1 – Test d'accuracy pour l'extraction des premiers tokens PER et ORG de chaque enregistrement sur une page aléatoire d'un répertoire de notaire selon taux de CER (1)

Accuracy pour l'extraction des premiers tokens PER et ORG de chaque enregistrement

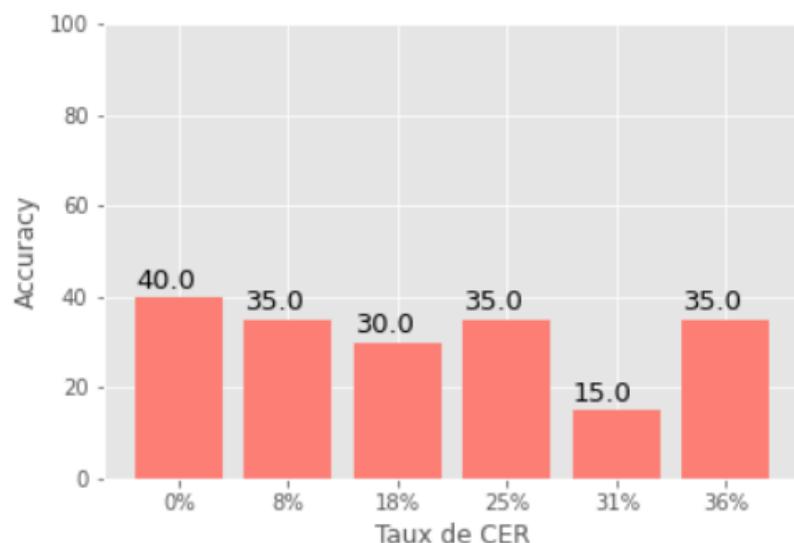


FIGURE 5.2 – Test d'accuracy pour l'extraction des premiers tokens PER et ORG de chaque enregistrement sur une page aléatoire d'un répertoire de notaire selon taux de CER (2)

Comme le montrent ces résultats, le taux de CER ne semble pas avoir d'impact sur la performance du modèle pour cette tâche. A titre de comparaison, elle a également été effectuée sur une transcription manuelle, donc avec un CER de 0%. Dans les deux cas, les résultats d'après une vérité de terrain restent bas, avec au maximum un score de 63%. De plus, comme le montre le premier diagramme, les résultats ne suivent pas une courbe régulières, le modèle obtient des performances moindres à 14% de CER qu'à 17%. Pour ce taux, l'*accuracy* est également la même que pour 31%. Les deux scores différents pour le CER à 17% montrent que la performance du modèle générique est impactée par la nature des erreurs générées par la REM. Dans le deuxième diagramme, le facteur aléatoire des résultats est bien illustré par les résultats pour les CER de 31% et 36%, passant de 15% d'*accuracy* à 35%. Cette expérience montre que les résultats sont trop aléatoires pour déployer un modèle générique sur les données de LECTAUREP, du moins pour cette tâche. Néanmoins, la structure régulière des répertoires des notaires encourage à envisager l'entraînement d'un modèle de REN pour le spécialiser sur cette structure.

Nous pouvons supposer que ce test aurait présenté des résultats similaires, voire plus aléatoires encore, si nous avions constitué une vérité de terrain annotée pour prendre en compte les noms de famille composés et les noms d'organisation, qui s'étendent souvent à plusieurs tokens. Par exemple, pour le nom de famille « De Goyenèche », récupérer seulement la particule « de » est loin d'être optimal : il faudrait pouvoir récupérer l'intégralité du nom de famille. La difficulté pour le modèle aurait alors été double : identifier l'entité concernée par l'enregistrement en entier, donc possiblement plusieurs tokens, et correctement l'identifier en tant que « PER » ou « ORG ». Enfin, cette évaluation est « souple » dans le sens où nous n'avons pas évalué le modèle sur sa capacité à discriminer les noms de personnes et les noms d'organisation.

### 5.3.2 Une variation dans le nombre d'entités extraites selon les taux de CER

Dans la continuité de la précédente expérience, on peut observer le nombre d'entités extraites par la chaîne de traitement de SpaCy selon le taux de CER pour savoir s'il y a une variation dans celui-ci, ce qui ne devrait pas normalement être le cas étant donné que les informations du texte ne changent pas. Mais le changement de sa forme causé par les erreurs de REM, risque de troubler le modèle. Les nombres d'entités ont été obtenus dans les mêmes *notebooks* où ont été réalisées les expériences précédentes.<sup>37</sup> Les nombre d'entités extraites

---

37. [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner\\_python/ner\\_spacy/NER\\_first\\_token\\_experience\\_1.ipynb](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_spacy/NER_first_token_experience_1.ipynb), et [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner\\_python/ner\\_spacy/NER\\_first\\_token\\_experience\\_2.ipynb](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_spacy/NER_first_token_experience_2.ipynb) (consultés le 16/08/21).

selon le taux de CER ont été rapportés dans deux diagrammes en colonnes, représentant respectivement les mêmes pages que celles sur lesquelles nous avons expérimenté précédemment (diagramme 5.3 et diagramme 5.4).

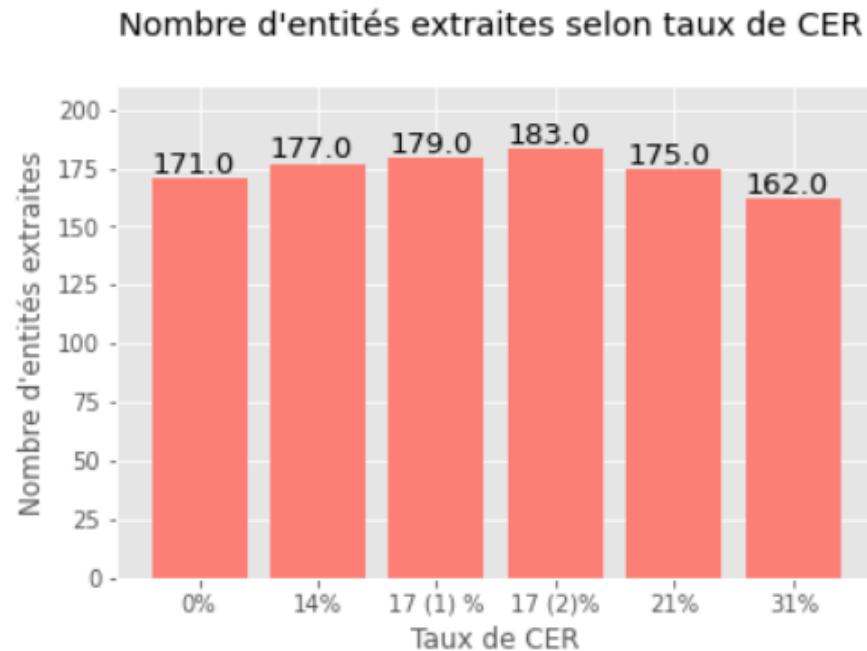


FIGURE 5.3 – Nombre d'entités extraites sur une page aléatoire d'un répertoire de notaire selon taux de CER (1)

Nous observons une variation dans le nombre d'entités extraites selon le taux de CER. Cela montre que les performances des modèles génériques sont aléatoires quand ils sont appliqués sur les données du projet LECTAUREP. Pour la première page (33 enregistrements), le nombre d'entités extraites commence à 171 sur une transcription manuelle, et augmente graduellement jusqu'à 17% de CER pour arriver à 183. Ce nombre baisse ensuite de façon conséquente à partir de 31%. Pour la deuxième page (20 enregistrements) il y a un écart de 11 entités entre le nombre d'entités extraites sur une transcription manuelle, 96 entités, pour 107 entités reconnues avec un taux de CER à 25%. On n'observe cependant pas le même phénomène à partir de 30% de taux de CER, où pour la deuxième page il redescend à 96 à 31% de CER, puis remonte à 103 à 36% de CER.

En conclusion, selon ces deux expériences, nous pouvons supposer que le bruit induit par la REM semble véritablement changer la nature du texte, du moins du point de vue du modèle générique, entraînant une variation dans le nombre d'entités extraites et des performances basses dans la tâche que nous avons étudié.

**Nombre d'entités extraites selon taux de CER**



FIGURE 5.4 – Nombre d'entités extraites sur une page aléatoire d'un répertoire de notaire selon taux de CER (2)

## 5.4 Quelles performances pour des textes propres présentant des caractéristiques linguistiques différentes des répertoires de notaires ?

Afin de mettre en comparaison les résultats que nous venons de présenter, il convient d'évaluer un modèle générique sur des données textuelles d'une nature différente des répertoires des notaires. Pour cela, ma collègue, Mme Floriane Chiffolleau, m'a autorisé à utiliser une lettre issue du corpus du projet DAHN, pour lequel elle travaille.<sup>38</sup> Ce document est issu de la correspondance entre Paul d'Estournelles de Constant (1852 - 1924), un homme politique français, et Nicholas Murray Butler (1862 - 1947), ancien président de l'université de Columbia, New York.<sup>39</sup> La lettre a été écrite par ce premier au mois d'octobre 1919. Dans le corpus constitué pour DAHN, la lettre porte le numéro 569, elle contient un nombre important de localisations et de noms de personnes. Les noms d'organisations sont également présents, mais en quantité moindre. La structure de la lettre se rapproche de la nature des textes sur lesquels sont entraînés les modèles de langues, c'est-à-dire des phrases verbales commençant

38. Voir <https://digitalintellectuals.hypotheses.org/category/dahn> (consulté le 18/08/21).

39. Pour plus d'informations biographiques sur Paul d'Estournelles de Constant et Nicholas Murray Butler, voir F. Chiffolleau, *Starting a new project – Discovering its source material*, Digital Intellectuals, mars 2021, URL : <https://digitalintellectuals.hypotheses.org/3398> (visité le 19/08/2021)

par une majuscule et se terminant par un point. Seule la fin de la lettre diffère du corps du texte, prenant la forme d'un tableau dans lequel Paul d'Estournelles de Constant a inscrit un itinéraire de voyage. Le texte est issu d'une transcription automatique corrigée lors d'une phase de post traitement, puis encodé en XML-TEI.<sup>40</sup>

Pour cette expérience, nous avons utilisé Les modèles génériques de REN de Stanza et SpaCy dans deux *notebooks*.<sup>41</sup> Une évaluation a été obtenue avec SpaCy, tandis que le test avec Stanza ne sert qu'à visualiser le résultat de la REN. Le protocole d'évaluation avec SpaCy est le même que celui décrit dans la sous-section 5.2.1. La segmentation automatique de SpaCy a été utilisée, le texte possédant des points pour les délimiter et ne posant donc pas de problème particulier. Les résultats de l'évaluation ont été reportés dans la table 5.8.

	Precision	Recall	F-score
Toutes entités	0.63	0.80	0.70
PER	0.33	0.71	0.45
LOC	<b>0.78</b>	0.79	<b>0.78</b>
ORG	0.69	0.9	<b>0.78</b>
MISC	0.04	<b>1</b>	0.08

TABLE 5.8 – Tableaux des scores du modèle générique de REN pour la chaîne de traitement *fr\_core\_news\_lg* de SpaCy sur une transcription automatique corrigée d'un document du corpus du projet DAHN, une lettre de Paul d'Estournelles de Constant.

On observe que les résultats obtenus sont meilleurs que ceux atteints avec le texte de la cinquième colonne des répertoires des notaires. Toutes EN confondues, le modèle générique possède une précision environ égale à ce que l'on a obtenu lors des expériences précédentes. Cependant, le rappel montre que le modèle ne rate qu'une EN sur deux. Les meilleurs scores sont atteints pour les entités de type « LOC ». Les résultats sont satisfaisants tout en laissant une marge d'amélioration en affinant le modèle sur ce domaine. Des erreurs caractéristiques, relevées lors de la correction de la pré-annotation pour constituer la vérité de terrain, pourraient certainement être corrigées ce faisant. Parmi ces erreurs, on retrouve :

- La mauvaise annotation des noms de villes composées. Le modèle générique annote souvent le nom d'une ville composée comme deux villes différentes, par exemple, pour « Sceaux-s/Huisnes » : [Sceaux LOC -s/ Huisnes PER]

40. L'encodage de la lettre 569 se trouve au lien suivant : [https://github.com/FloChiff/DAHNProject/blob/master/Correspondence/Paul\\_d\\_Estournelles\\_de\\_Constant/Corpus/Lettre569\\_3octobre1919.xml](https://github.com/FloChiff/DAHNProject/blob/master/Correspondence/Paul_d_Estournelles_de_Constant/Corpus/Lettre569_3octobre1919.xml) (consulté le 18/08/21).

41. Pour le test réalisé avec Stanza, voir [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner\\_python/ner\\_stanza/test\\_ner\\_dahn.ipynb](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_stanza/test_ner_dahn.ipynb), pour le test réalisé avec SpaCy, voir [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner\\_python/ner\\_spacy/test\\_dahn\\_spacy.ipynb](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_spacy/test_dahn_spacy.ipynb). Les fichiers texte utilisés pour les tests se trouvent à l'adresse suivante : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/corpus\\_test/dahn](https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/corpus_test/dahn) (encodage XML-TEI, extraction de toutes les balises <>p>> dans un fichier texte, et vérité de terrain REN obtenu avec Doccano. (consultés le 18/08/21).

- Les nationalités sont fréquemment étiquetées en tant que « LOC ».
- Les hyphénations perturbent la bonne identification des EN.
- Lorsque le texte change de structure et devient un tableau, l'extraction est moins régulière pour les entités de type « LOC », là où le modèle générique était pourtant efficace dans le corps du texte.

Nous remarquons également que le modèle générique a tendance à sur-annoter, avec 369 entités reconnues contre 293 dans la vérité de terrain. Des exemples de visualisation de la REN avec SpaCy ont été inclus en annexes.<sup>42</sup>

Les scores obtenus pour les entités « MISC » ne sont pas pertinents, leur nombre étant très faible.

Les résultats de la REN obtenus avec Stanza semblent similaires. Cependant, nous ne tirerons pas de conclusion ici car aucune information chiffrée n'a été obtenue.<sup>43</sup>

---

42. Voir annexes G.1, G.2, G.3. Dans ce dernier exemple, on observe l'extraction moins régulière des noms de lieux.

43. Un exemple de la visualisation de la REN obtenue avec Stanza a été inclu dans l'annexe G.1



# **Chapitre 6**

## **Une solution spécifique à un corpus : entraîner un modèle de reconnaissance d'entités nommées**

Suite aux résultats obtenus avec les modèles génériques, la possibilité d'entraîner un modèle de REN, ou d'en affiner un déjà existant pour des raisons de temps, a été discutée lors de mon stage. La solution d'affiner un modèle de REN obtenu préalablement avec le modèle de langue camemBERT a été une piste envisagée dans l'idée d'utiliser un outil développé par l'équipe ALMAnaCH.

### **6.1 Entrainer un modèle de classification avec le modèle de langue camemBERT**

#### **6.1.1 Présentation du modèle de langue camemBERT**

CamemBERT est un modèle de langue développé en 2020 au sein de l'équipe ALMAnaCH.<sup>1</sup> Il est basé sur le modèle RoBERTa, lui-même basé sur BERT, qui possède une meilleure performance grâce à une architecture modifiée.<sup>2</sup> CamemBERT est un modèle monolingue entraîné sur des données en langue française, deux versions existent à ce jour :

- CamemBERT base, entraîné avec « 12 couches, 768 dimensions cachées et 12 têtes d'attention, soit 110M de paramètres ».<sup>3</sup>

---

1. Voir <https://camembert-model.fr/> (consulté le 18/08/21).

2. L. Martin, B. Muller, P. J. Ortiz Suárez, *et al.*, « CamemBERT... », p. 2. "RoBERTa improves the original implementation of BERT by identifying key design choices for better performances, using dynamic masking, removing the next sentence prediction task, training with larger batches, on more data, and for longer."

3. Id., « Les modèles de langue contextuels Camembert pour le français : impact de la taille et de

- CamemBERT large, entraîné avec « 24 couches, 1024 dimensions cachées et 16 têtes d'attention, soit 340M paramètres. »<sup>4</sup>

Ce modèle a été entraîné à partir de données « à haute variabilité, éventuellement bruitées, plutôt que des données proprement éditées et stylistiquement homogènes. »<sup>5</sup> Ces données contiennent du bruit caractéristique de ce qu'il est possible de trouver sur n'importe quel site internet. Notons ici que le bruit des données textuelles rencontrées sur internet est différent du bruit généré par la REM. Dans le premier, on trouve des mots mal orthographiés, des fautes de grammaire, des expressions propres au langage français que l'on trouve sur internet, et des abréviations, tandis que dans le second, comme nous l'avons présenté, les fautes concernent des suppressions de lettres, des insertions, des substitutions et des transpositions.

CamemBERT a été entraîné sur trois corpus de données différents :

- OSCAR, un ensemble de sous-corpus constitué à partir de *Common Crawl*, un corpus multilingue de 20 TB constitué à partir de données en ligne. Pour entraîner CamemBERT, le sous-corpus français d'OSCAR a été utilisé.<sup>6</sup>
- CCNet, un autre corpus constitué à partir de *Common Crawl*, mais avec des documents en moyenne plus longs qu'OSCAR. Martin *et al.* estiment que ce corpus se positionne entre « OSCAR, peu filtré voire bruité, et Wikipedia, totalement édité. »<sup>7</sup>
- Wikipedia, qui constitue un corpus de données textuelles propres. Le *dump* français de Wikipedia, datant du mois d'avril 2019, a été utilisé pour l'entraînement de camemBERT.

Chaque modèle a ensuite été évalué sur plusieurs tâches de TAL, dont la REN.<sup>8</sup> Le corpus utilisé pour évaluer la REN avec CamemBERT est la *French TreeBank* annotée en EN.<sup>9</sup> Ces

---

l'hétérogénéité des données d'entraînement », dans *JEP-TALN-RECITAL 2020 - 33ème Journées d'Études sur la Parole, 27ème Conférence sur le Traitement Automatique des Langues Naturelles, 22ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, dir. Christophe Benitzoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni, Sylvain Pogodalla et Stéphane Schneider, Nancy / Virtuel, France, 2020, p. 54-65, URL : <https://hal.archives-ouvertes.fr/hal-02784755> (visité le 16/04/2021), p. 56

4. *Ibid.*

5. *Ibid.*, p. 55

6. Pedro Javier Ortiz Suárez, B. Sagot et L. Romary, « Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures », dans, 2019, DOI : 10.14618/IDS-PUB-9021 Voir également <https://oscar-corpus.com/> (consulté le 18/08/21).

7. L. Martin, B. Muller, P. J. Ortiz Suárez, *et al.*, « Les modèles de langue contextuels Camembert pour le français... », p. 56

8. CamemBERT a également été évalué sur les tâches d'étiquetage morpho-syntaxique, analyse syntaxique, et reconnaissance d'implication textuelle. Cette dernière « consiste à prédire la relation entre une phrase hypothèse et phrase prémissé (implication, contradiction, neutralité). » Voir *Ibid.*, p. 57

9. P. J. Ortiz Suárez, Y. Dupont, B. Muller, *et al.*, « Establishing a New State-of-the-Art for French Named Entity Recognition »... La *French TreeBank* a été développée par l'IUF, le CNRS et le CNRTL et est la principale ressource pour évaluer les modèles de TAL sur les tâches d'étiquetage morpho-syntaxiques et d'analyse syntaxique en langue française. Elle contient environ 21 550 phrases issues de numéros du journal

modèles ont été testés de deux façons différentes : premièrement, le modèle est affiné sur une tâche spécifique. Pour cela, une couche prédictive est ajoutée au modèle.<sup>11</sup> Deuxièmement les plongements lexicaux contextuels figés sont extraits du modèle pour classifier les tokens.<sup>12</sup>

Afin de donner la mesure des scores qu'il serait possible d'obtenir avec un modèle affiné de REN constitué avec camemBERT, nous reportons ici les scores d'évaluation pour cette tâche en fonction de la méthode d'évaluation :

- Avec un modèle affiné :
  - En utilisant Wikipedia : F-score de 89.96.
  - En utilisant CCNet : F-score de 90.46.
  - En utilisant la plus version d'OSCAR la plus conséquente pour le français (138 GB) : F-score de 91.55
- Avec l'extraction des plongements lexicaux :
  - En utilisant Wikipedia : F-score de 91.23
  - En utilisant CCNet : F-score de 92.30
  - En utilisant OSCAR (138 GB) : 91.83 ; OSCAR (4 GB) : 91.90.

Ces résultats sont donc beaucoup plus élevés que ce que nous avons obtenu avec les modèles génériques. Il serait donc tout à fait possible d'expérimenter avec ces différentes versions de camemBERT, disponibles sur son site de présentation, afin d'entraîner ou d'affiner un modèle de REN préexistant, entraîné avec camemBERT, sur des données du projet LECTAUREP.

### 6.1.2 Entrainement d'un modèle de classification avec camemBERT

Cependant, le bruit sur lequel a été entraîné camemBERT étant différent du bruit de la REM, il est nécessaire de débruiter au maximum ces données afin de ne pas entraîner une baisse des performances du modèle sur des données qui seraient trop hors domaine. Étant donné qu'il n'est pas possible d'assurer des scores satisfaisants avec un modèle de REN entraîné avec camemBERT uniquement sur des données bruitées issues de la REM, ce qui nécessiterait un volume de données annotées en EN conséquent, il serait plus adapté de commencer à expérimenter avec l'affinage d'un modèle de REN préalablement entraîné sur des données du même domaine, ou *a minima* sur des données s'y rapprochant.

---

« Le Monde » datés entre 1990 et 1993.<sup>10</sup> Afin de la compléter, Ortiz Suárez *et al.* ont annoté en 2020 ce corpus en EN, afin de pouvoir évaluer les modèles sur la REN.

11. L. Martin, B. Muller, P. J. Ortiz Suárez, *et al.*, « Les modèles de langue contextuels Camembert pour le français... », p. 57

12. *Ibid.*, p. 58 « Afin d'obtenir une représentation pour un token donné, nous calculons d'abord la moyenne des représentations de chaque sous-mot dans les quatre dernières couches du Transformer, puis faisons la moyenne des vecteurs des sous-mots résultants. »

Suite à une discussion à la fin du mois de juin, mon collègue, M. Ortiz Suárez, m'a partagé un *notebook* qu'il avait conçu pour un atelier dans le cadre de la conférence TALN 2021.<sup>13</sup> Le *notebook* principal<sup>14</sup> propose de guider pas à pas l'utilisateur-ice dans l'affinage d'un modèle de classification de tokens à partir de camemBERT, en expliquant comment charger un corpus de données annotées en EN, comment les pré-traiter en vue de l'affinage du modèle, notamment pour tokeniser le texte avec la librairie *Hugging Face Transformers*<sup>15</sup>, de sorte à avoir des données exploitables, et en présentant la phase d'entraînement elle-même.

Le *notebook* a servi à entraîner deux modèles de classification, chacun sur un corpus différent : CLEF HIPE, et LEM17.<sup>16</sup> Le premier est un corpus créé à l'occasion de la *shared task* CLEF HIPE 2020, destinée à évaluer les architectures de REN et d'EL sur des données de journaux historiques océrisés en français, allemand et anglais, datés de 1798 à 2018, et annotées en EN et en EL.<sup>17</sup> Cette tâche et ce corpus ont été créés pour fournir des données appropriées à l'évaluation des architectures de REN sur des données historiques. Le corpus n'est pas en français contemporain, et contient donc des conventions de nommage différentes, des noms de localisations qui n'existent plus, etc.<sup>18</sup> Il rassemble 563 documents, représentant 444 596 tokens.<sup>19</sup> Le guide d'annotation s'appuie sur la campagne Quaero, qui a défini en 2011 des directives pour annoter les EN en français à partir de transcriptions de discours, et a déjà pu être utilisé pour des données historiques.<sup>20</sup> Le deuxième est un corpus diachronique en français moderne composé à partir de textes datés entre le XVI<sup>e</sup> siècle et le XVIII<sup>e</sup> siècle, et annoté en EN (entre autres), créé par les chercheurs Thibault Clérice, Matthias Gille-Levenson, Jean-Baptiste Camps et Jean-Baptiste Tanguy.<sup>21</sup> Le sous-corpus d'entraînement se nomme *Presto*, et est issu du projet éponyme. On retrouve parmi les textes annotés, des œuvres littéraires telles que les *Essais*, écrits par Michel de Montaigne (1580), *Gargantua*, François Rabelais (1534), *Hernani*, Victor Hugo (1841), ou encore la *Lettre à M. Rousseau*,

---

13. TALN (Traitement Automatique des Langues Naturelles) est une conférence annuelle portant sur le TAL. Pour consulter le site et le programme de l'édition 2021, voir <https://talnrecital2021.inria.fr/en/>. L'atelier en question se nomme « CANTAL - Formats et ChAiNes de traitement de TAL », organisé par Yoann Dupont, Gaël Lejeune, Pedro Javier Ortiz Suárez, et Tian Tian. Voir <https://talnrecital2021.inria.fr/programme-2/> (consulté le 18/08/21).

14. [https://github.com/YoannDupont/taln2021tutorial/blob/main/with\\_transformers/CANTAL\\_With\\_Transformers.ipynb](https://github.com/YoannDupont/taln2021tutorial/blob/main/with_transformers/CANTAL_With_Transformers.ipynb) (consulté le 18/08/21).

15. Voir <https://github.com/huggingface/transformers> (consulté le 18/08/21).

16. Les modèles sont disponibles avec les liens de téléchargement disponibles dans le *readme* du répertoire Github CANTAL : [https://github.com/YoannDupont/taln2021tutorial/tree/main/with\\_transformers#predictions](https://github.com/YoannDupont/taln2021tutorial/tree/main/with_transformers#predictions)

17. Voir <https://impresso.github.io/CLEF-HIPE-2020/> (consulté le 23/06/21).

18. « Extended Overview of CLEF HIPE 2020... », p. 2

19. *Ibid.*, p. 11

20. *Ibid.*, p. 7 et S. Rosset, C. Grouin et P. Zweigenbaum, *Entités nommées structurées...*

21. Simon Gabay, Thibault Clérice, Matthias Gille-Levenson, Jean-Baptiste Camps, Jean-Baptiste Tanguy, LEM17 : data and models for modern French (16-18th c.), Neuchâtel : Université de Neuchâtel, 2020, <https://github.com/e-ditiones/LEM17>

de d'Alembert (1759).<sup>22</sup>

En attendant des corpus plus proches du français et du type de langage présent dans les pages des répertoires des notaires, ces deux corpus constituent des ressources essentielles pour expérimenter l'entraînement de modèles de REN pour LECTAUREP. Les deux modèles entraînés sur ceux-ci et mis à disposition par M. Ortiz Suárez pourraient être affinés sur des données annotées en EN du projet LECTAUREP. Par ailleurs, notons que la plupart des librairies de TAL précédemment présentées proposent d'entraîner des modèles de REN.<sup>23</sup>

## 6.2 Quels objectifs peut-on donner à un modèle affiné ?

À ma connaissance, il n'existe pas de statistiques réalisées pour renseigner sur le nombre de tokens annotés nécessaires pour entraîner un modèle de REN sur des données bruitées.<sup>24</sup> Avant de rassembler un corpus annoté aussi conséquent que les deux mentionnés précédemment, le projet LECTAUREP pourrait affiner les deux modèles précédemment présentés en respectant les étiquettes attribuées aux corpus d'entraînement. Cela serait l'occasion de fournir une première évaluation du modèle sur les transcriptions automatiques des répertoires des notaires. Nous pouvons supposer que le bruit et la façon aléatoire dont il est généré nécessiteront un volume de données d'entraînement plus important afin de fournir suffisamment de contexte pour reconnaître les EN.

En utilisant le modèle entraîné avec le corpus CLEF HIPE, les étiquettes disponibles

---

22. Voir [http://presto.ens-lyon.fr/?page\\_id=48](http://presto.ens-lyon.fr/?page_id=48) (consulté le 23/08/21).

23. Pour SpaCy, voir <https://spacy.io/usage/training#training-data> et Nishanth N, *Train NER with Custom training data using spaCy*. Medium, 29 juil. 2020, URL : <https://towardsdatascience.com/train-ner-with-custom-training-data-using-spacy-525ce748fab7> (visité le 12/07/2021), Michaël Benesty, *NER algo benchmark : spaCy, Flair, m-BERT and camemBERT on anonymizing French commercial legal cases*, Medium, 10 déc. 2019, URL : <https://towardsdatascience.com/benchmark-ner-algorithm-d4ab01b2d4c3> (visité le 23/04/2021); pour Stanza, voir <https://stanfordnlp.github.io/stanza/ner.html#training-only-options>; pour NLTK, voir <https://www.nltk.org/book/ch07.html> et Charles Bochet, *How to Train your Own Model with NLTK and Stanford NER Tagger? (for English, French, German...)* Medium, 10 mai 2018, URL : <https://medium.com/sicara/train-ner-model-with-nltk-stanford-tagger-english-french-german-6d90573a9486> (visité le 12/07/2021); pour Flair, voir [https://github.com/flairNLP/flair/blob/master/resources/docs/TUTORIAL\\_7\\_TRAINING\\_A\\_MODEL.md](https://github.com/flairNLP/flair/blob/master/resources/docs/TUTORIAL_7_TRAINING_A_MODEL.md) (consultés le 23/08/21). Notons également l'existence d'HORUS-NER, un système de REN conçu pour être appliqué à des données bruitées habituellement rencontrées sur internet : <https://github.com/SmartDataAnalytics/HORUS-NER>, voir également D. Esteves, J. Marcelino, P. Chawla, et al., « HORUS-NER... » (consulté le 23/08/21).

24. Des informations chiffrées peuvent être trouvées dans des forums tels que Stack Overflow. Par exemple, dans une réponse de 2018, un-e utilisateur-ice suggère d'annoter 2000 phrases (<https://stackoverflow.com/questions/52120487/what-is-the-amount-of-training-data-needed-for-additional-named-entity-recogniti>). Dans un article pour Deepnote, Isaac Aderogba affine un modèle de classification avec SpaCy dans le but d'extraire les EN concernant la nourriture, et utilise pour cela un corpus de 1 142 610 tokens annotés, voir <https://deepnote.com/@isaac-adelogba/Spacy-Food-Entities-LMLRnM0sQyGIUwvPLvVlsw> et <https://fdc.nal.usda.gov/download-datasets.html> (*Branded Foods dataset*, avril 2021, fichier CSV de 213 MB) (consultés le 23/08/21). À partir de ces rares informations, il semble que le plus de données, le mieux.

sont les suivantes<sup>25</sup> :

- « pers », pour les noms de personne.
- « org », pour les noms d'organisations.
- « prod », pour les « production humaines », c'est-à-dire les marques d'objet, les œuvres artistiques, les productions médiatiques, les produits financiers, les logiciels, les prix divers, les lignes de transport, les doctrines, et les règles.<sup>26</sup>
- « date », pour les dates.
- « loc », pour les localisations.

Concernant *Presto*, les étiquettes sont plus nombreuses<sup>27</sup> :

- « event », pour les événements.<sup>28</sup>
- « func », pour les métiers, les fonctions, les rôles sociaux, etc.<sup>29</sup>
- « loc ».
- « org ».
- « pers ».
- « prod ».
- « time », pour représenter un point dans le temps.<sup>30</sup>.

Les étiquettes du corpus *Presto* semblent plus adaptées à la nature des informations de la cinquième colonne des répertoires des notaires (section 2.3). Cependant, les performances d'un modèle affiné originellement entraîné sur ce corpus pourront être impactées par le français du XIXe et XXe siècle utilisé dans les pages des répertoires des notaires, non représenté dans ce corpus.

## 6.3 Rassemblement et préparation de données d'entraînement

Nous avons précédemment décrit les étapes nécessaires pour pré-traiter les textes en vue de la REN. Celles-ci servent également pour préparer des données à annoter en vue de l'entraînement d'un modèle de REN. Cette section aura pour but de rendre compte d'outils

---

25. « Extended Overview of CLEF HIPE 2020... », p. 6

26. S. Rosset, C. Grouin et P. Zweigenbaum, *Entités nommées structurées...*, p. 39

27. Voir [https://github.com/YoannDupont/taln2021tutorial/blob/main/with\\_transformers/presto.py#L75](https://github.com/YoannDupont/taln2021tutorial/blob/main/with_transformers/presto.py#L75) (consulté le 23/08/21).

28. *Ibid.*, pp. 62 - 63

29. *Ibid.*, p. 25

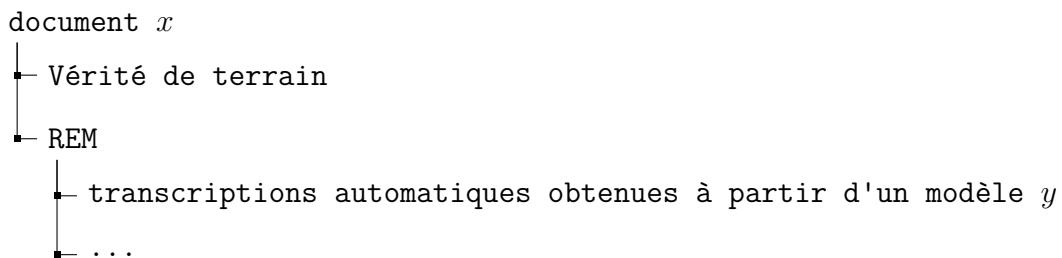
30. *Ibid.*, p. 56

expérimentaux que j'ai créé en vue de la constitution d'un corpus de données à annoter en EN pour le projet LECTAUREP.

### 6.3.1 Sélectionner des candidats pour l'annotation de données d'entraînement en évaluant par lot

Dans le but de sélectionner des candidats pour l'annotation, j'ai écrit lors de mon stage un outil de ligne de commande (*command line interface*, ou CLI) pour évaluer par lot les transcriptions issues de la REM grâce à la librairie KaMI.<sup>31</sup> Les transcriptions manuelles réalisées à partir du *random set* de LECTAUREP peuvent servir à évaluer les transcriptions automatiques de ce même set de données. À partir des deux documents eScriptorium issus de celui-ci et utilisés pour entraîner un modèle de segmentation, le projet LECTAUREP dispose potentiellement de 200 pages à annoter pour affiner un modèle de REN. Cependant, il est nécessaire de ne garder que les transcriptions ayant un score de CER exploitable. Afin d'obtenir plusieurs types de bruits différents générés par la REM, il serait possible d'annoter des données d'une qualité satisfaisante issues de deux modèles de REM, notamment avec les deux modèles mixtes créés pour le projet.

Ce CLI utilise KaMI pour comparer deux chaînes de caractères, la première étant la vérité de terrain, et la deuxième étant la transcription automatique. Pour fonctionner, il utilise une structure de répertoire spécifique pour traiter par lot les transcriptions :



Pour le document eScriptorium numéro 145 du serveur Traces6, par exemple, on aura donc un premier sous-répertoire contenant les fichiers PAGE XML de la transcription manuelle (vérité de terrain). Un deuxième sous-répertoire stocke les transcriptions automatiques obtenues à partir d'un modèle de REM, dont les fichiers PAGE XML sont stockés dans un dernier niveau de sous-répertoires. En plus des PAGE XML, on inclue un fichier texte nommé *model\_log.txt*, contenant le nom du modèle avec lequel la transcription automatique a été obtenue.<sup>32</sup>

31. Le CLI et son *readme* sont disponibles dans le répertoire Gitlab ici : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/preprocessing/HTR\\_batch\\_evaluation](https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/preprocessing/HTR_batch_evaluation), voir également l'*issue* documentant la création du CLI : <https://gitlab.inria.fr/almanach/lectaurep/ner/-/issues/3> (consultés le 23/08/21).

32. Voir <https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/preprocessing/>

Une fois les données organisées, le CLI évaluera chaque sous-répertoire de fichiers de transcription automatique face à la vérité de terrain selon plusieurs étapes :

1. La structure logique de la cinquième colonne est reconstituée grâce à l'annotation sémantique pour la vérité de terrain et la transcription automatique de chaque page présente dans le document eScriptorium, résultant en deux chaînes de caractères distinctes pour l'évaluation avec KaMI.
2. Les deux chaînes de caractères sont comparés avec KaMI.
3. Les résultats de chaque évaluation sont stockées dans un fichier JSON. Le fichier texte *model\_log.txt* permet de renseigner, dans le JSON, avec quel modèle les transcriptions ont été obtenues. La structure du fichier JSON a été reproduite dans l'annexe I.1.

Une fois le JSON obtenu, il est possible de l'exploiter grâce à un script pour obtenir tous les noms de fichiers dont les transcriptions possèdent un score de CER maximum par exemple. Ainsi, tous les PAGE XML dont la transcription est inférieure ou égale à 15% de CER ont été récupérés.<sup>33</sup>

### 6.3.2 Un CLI pour débruiter et normaliser les candidats sélectionnés

Un second CLI a été écrit dans l'idée de simplifier et de préparer rapidement des données à annoter en EN, et est basé sur les étapes de pré-traitement identifiées précédemment.<sup>34</sup> Celui-ci prépare des fichiers PAGE XML résultant de la REM présents dans un répertoire, et préférablement rassemblés grâce à la sélection de candidats selon le taux de CER avec le premier CLI, selon plusieurs étapes :

- La structure logique de la cinquième colonne est reconstituée, et chaque enregistrement est stocké dans une liste python.
- Les mots de chaque enregistrement sont segmentés au minimum à l'aide d'expressions régulières basées sur des erreurs de REM récurrentes identifiées après observation des transcriptions automatiques. Comme nous l'avons montré, le bruit généré par la REM

---

HTR\_batch\_evaluation/docs\_for\_HTR\_evaluation\_with\_kami/doc\_145 (consulté le 23/08/21).

33. Pour l'évaluation du document eScriptorium 145 réalisée à partir de transcriptions automatiques obtenues avec les deux modèles mixtes, on obtient le fichier JSON suivant : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/preprocessing/HTR\\_batch\\_evaluation/HTR\\_scores\\_with\\_kami/doc\\_145\\_HTR\\_scores.json](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/preprocessing/HTR_batch_evaluation/HTR_scores_with_kami/doc_145_HTR_scores.json). Pour l'exploiter, le script suivant a été écrit : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/preprocessing/HTR\\_batch\\_evaluation/get\\_pages\\_based\\_on\\_HTR\\_scores.py](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/preprocessing/HTR_batch_evaluation/get_pages_based_on_HTR_scores.py), résultant dans un répertoire avec des transcriptions ayant un CER inférieur ou égal à 15% : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/preprocessing/HTR\\_batch\\_evaluation/ner\\_annotation\\_candidates\\_for\\_preprocessing/CER\\_13%25/doc\\_145](https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/preprocessing/HTR_batch_evaluation/ner_annotation_candidates_for_preprocessing/CER_13%25/doc_145). Toutes les étapes décrites pour le document 145 ont également été réalisées pour le document 156 (consulté le 23/08/21).

34. Voir [https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/preprocessing/word\\_segmentation](https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/preprocessing/word_segmentation) (consulté le 23/08/21).

n'étant pas régulier, ni forcément prévisible, la solution des expressions régulières n'est pas la plus adaptée.

- Les enregistrements sont tokenisés afin de normaliser les abréviations. Chaque token est comparé à un référentiel d'abréviation au format JSON, constitué à partir de la lecture de plusieurs transcriptions manuelles des répertoires des notaires, pour expérimenter.<sup>35</sup>
- Dans un répertoire donné, le CLI génère un fichier texte pour chaque fichier PAGE XML pré-traité, où chaque ligne correspond à un enregistrement et se termine par le symbole unicode « « », afin de marquer la fin d'un enregistrement. Ce symbole peut facilement être supprimé, et sert de repère pour reconstituer la structure logique, notamment durant la phase d'annotation.<sup>36</sup>

Une fois le CLI utilisé pour pré-traiter les transcriptions issues de la sélection en fonction du taux de CER, il est possible de commencer leur annotation en EN en vue de l'affinage d'un modèle de classification propre à LECTAUREP.

### 6.3.3 Retour d'expérience sur l'utilisation de la plate-forme d'annotation *open-source* Inception

J'ai pu durant le stage, en plus de Doccano, essayer la plate-forme d'annotation *open-source* Inception.<sup>37</sup> J'ai reçu un accès grâce à mon collègue, M. Lucas Terriel, qui m'a ouvert un compte sur une instance de cette application déployée sur un serveur et utilisée dans le cadre du projet *NER4Archives* pour lequel il travaille.<sup>38</sup> Inception est une plate-forme spécialisé dans l'annotation d'EN qui propose un système permettant d'annoter plus rapidement des corpus issus de domaines où il existe peu ou pas de bases de connaissance.<sup>39</sup> Celles-ci peuvent aider à pré-annoter des textes, rendant la tâche considérablement moins longue que s'il fallait partir d'un document vierge en annotation. Les données textuelles issues de documents historiques, et plus particulièrement celles issus de l'OCR et de la REM, peuvent souffrir du manque de bases de connaissances, de corpus du même domaine déjà annotés, ou encore de modèles de REN pour annoter rapidement des corpus en EN. Inception propose pour cela un

---

35. Il n'existe à ce jour pas de guide pour la transcription des abréviations du projet LECTAUREP. Le référentiel a le format suivant, pour chaque ligne : "abréviation" : "forme normale". Voir [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/preprocessing/word\\_segmentation/referentiels/referentiel\\_abreviations.json](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/preprocessing/word_segmentation/referentiels/referentiel_abreviations.json) (consulté le 23/08/21).

36. Les fichiers pré-traités résultant de la sélection de transcriptions selon le taux de CER depuis les documents eScriptorium 145 et 156 se trouvent ici : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/preprocessing/word\\_segmentation/ner\\_annotation\\_candidates\\_preprocessed](https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/preprocessing/word_segmentation/ner_annotation_candidates_preprocessed) (consulté le 23/08/21).

37. Voir <https://inception-project.github.io/> (consulté le 23/08/21).

38. Voir <https://gitlab.inria.fr/almanach/ner4archives> (consulté le 23/08/21).

39. Inception permet aussi d'annoter d'autres types de données, tels que les lemmes, ou les corrections orthographiques dans le but d'entraîner un modèle. Voir annexe H.6.

système d'annotation sans base de connaissance ("domain-agnostic"), basé sur de l'apprentissage machine dont l'efficacité est ajustée à l'aide des utilisateurs ("human-in-the-loop").<sup>40</sup> Pour cela, Inception utilise des algorithmes, appelés « recommandeurs » ("recommenders"), qui suggèrent à l'annotateur-rice des EN.<sup>41</sup> Par exemple, lors de l'annotation d'un échantillon issu du projet LECTAUREP, le recommandeur proposait automatique d'annoter toutes les mentions de « Paris ». Ce type de recommandeur est un *string-matcher*, c'est-à-dire qu'après avoir annoté une fois le nom de la capitale française avec l'étiquette « LOC », un recommandeur était capable de retrouver dans le texte toutes les occurrences de cette même séquence de caractères et de suggérer l'attribution de l'étiquette « LOC ».<sup>42</sup> Il existe également un *string-matcher* utilisant la distance de Levenshtein, ce qui pourrait être utile pour l'annotation des transcriptions automatiques des répertoires des notaires.<sup>43</sup> Le principe est que plus les documents sont annotés, plus les recommandeurs apprendront et proposeront des candidats. Ensuite, c'est à l'annotateur-rice de valider ces résultats, permettant par conséquent de les ajuster et d'améliorer leurs performances.<sup>44</sup>

L'expertise de mon collègue sur cette plate-forme fait d'Inception, dans le cadre du projet LECTAUREP et d'ALMAnaCH, un atout dans la constitution d'un corpus en EN. De plus, l'avancée du projet NER4Archives, dont l'un des objectifs consiste à mener un campagne de REN sur les instruments de recherche XML-EAD des AN, et de lier les EN à des notices d'autorité, pourrait profiter à LECTAUREP si une couche d'*entity-linking* est ajoutée après l'obtention d'un modèle de REN robuste.

### 6.3.4 Un format d'annotation pour des données d'entraînement : CoNLL 2002

Avec Inception, il est possible d'exporter les annotations dans plusieurs formats différents.<sup>45</sup> L'un deux est le format texte ConLL 2002, dont la simplicité rend son exploitation et sa compréhension aisées. Celui-ci est très proche du format BIOES, présenté précédemment. Chaque ligne prend la forme de deux colonnes, séparées par un espace, et représente un mot d'une phrase, indiqué dans la première colonne. La deuxième colonne, si le mot est une EN, indique si le mot est le début d'une entité (B, pour *beginning*), ou s'il correspond à un mot

---

40. J.C. Klie, R. Eckart de Castilho et I. Gurevych, « From Zero to Hero... »

41. *Ibid.*, p. 6990

42. Voir annexe H.7 et H.5.

43. *Ibid.*, p. 6984

44. "By selecting an entity label from the candidate list, users express that the selected one was preferred over all other candidates. These preferences are used to train state-of-the-art pairwise learning-to-rank models from the literature : the gradient boosted trees variant [...]." *Ibid.*

45. 23 pour être exact. Parmi les formats d'export, on retrouve les formats CoNLL, texte, WebAnno TSV, etc. Voir [https://inception-project.github.io/releases/20.0/docs/user-guide.html#sect\\_formats](https://inception-project.github.io/releases/20.0/docs/user-guide.html#sect_formats) (consulté le 23/08/21). Voir également l'annexe H.9

composant l'entité et n'étant pas en première position (I, ou *inside*). Un mot n'étant pas une EN est indiqué par un O, pour *outside*. CoNLL 2002 est aussi connu sous le nom du format « BIO ».<sup>46</sup>

Un extrait d'une page d'un répertoire de notaire pré-traitée exportée au format CoNLL 2002 a été reproduite dans l'annexe H.<sup>47</sup>

## 6.4 Préconisations pour l'entraînement d'un modèle de reconnaissance d'entités nommées à partir des données de LECTAUREP

En conclusion, avant de démarrer une campagne d'annotation en EN, le projet LECTAUREP devra se doter d'un guide d'annotation exhaustif permettant de produire des données d'entraînement pour affiner un modèle, voire, si un volume conséquent de données annotées est atteint, entraîner un modèle. Le premier cas risque de poser des difficultés si les données du projet LECTAUREP sont trop hors domaine, le deuxième permettrait d'avoir un modèle spécialisé sur le domaine des répertoires des notaires. Le guide d'annotation servira de plus à donner des indications sur la façon dont doivent être annotées les EN, par exemple pour les adresses, si celles-ci doivent comporter le numéro, ou pour les noms, si ceux-ci doivent inclure les titres de civilité, etc. Le guide d'annotation Quaero pour les EN semble être une excellente base pour définir un ensemble d'étiquettes appropriés pour la REN dans le contexte de LECTAUREP.<sup>48</sup> Un guide d'annotation permettrait enfin de produire de la documentation pour donner des renseignements sur les données utilisées pour l'entraînement d'un modèle, et à cette tâche d'être attribuée à plusieurs annotateurs-rices afin de produire des annotations régulières.

Une fois un guide d'annotation conçu, il sera nécessaire de mettre au point une chaîne de traitement fixe afin d'obtenir des données structurées. Ce que nous avons décrit précédemment est une expérimentation de chaîne de traitement. À partir de celle-ci, nous pouvons suggérer que, pour pré-traiter les données textuelles issues de la REM, il serait nécessaire de :

1. Classer des candidats à l'annotation. Les modèles de REM permettent aujourd'hui d'obtenir des transcriptions avec un score de CER de moins de 10%. Avec le progrès des

---

46. Voir <https://simpletransformers.ai/docs/ner-data-formats/#text-file-in-conll-format> et <https://www.clips.uantwerpen.be/conll2002/ner/> (consultés le 23/08/21).

47. Le fichier intégral se trouve ici : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/exemples\\_training\\_data\\_ner](https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/exemples_training_data_ner). Un sous-répertoire a été créé dans le répertoire Gitlab utilisé lors de mon stage pour stocker ce fichier et le fichier texte non annoté : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/exemples\\_training\\_data\\_ner](https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/exemples_training_data_ner) (consultés le 23/08/21).

48. S. Rosset, C. Grouin et P. Zweigenbaum, *Entités nommées structurées...*

technologies, le travail fourni par ALMAAnCH, et les transcriptions manuelles fournies par les AN pour entraîner des modèles de REM, ces taux continuent de descendre.<sup>49</sup> En 2018, Animesh Prasad *et al.* montraient que des transcriptions automatiques avec un taux de CER de plus ou moins 5% permettait d'obtenir des résultats de REN aussi satisfaisants que sur la vérité de terrain, avec un F-score de 95.4%.<sup>50</sup> Hamdi *et al.* ont également démontré que les performances des modèles de REN dépendent fortement de la qualité des textes obtenus avec l'OCR, et donc par extension avec la REM. Les performances de la REN descendent de 90% à 60% d'*accuracy* lorsque le taux de CER et de WER monte de 1% à 7%, et de 8% et 20% respectivement.<sup>51</sup> Les données à annoter doivent donc tendre vers ces scores.

2. Reconstituer la structure logique afin de pouvoir annoter les mots dans leur contexte.
3. Segmenter les mots.
4. Éventuellement décider de normaliser les abréviations.
5. Annoter des données avec la plate-forme Inception.

Ces étapes ont été reproduites en annexes sous la forme d'un diagramme (annexe H.10).<sup>52</sup>

Entraîner un modèle spécifique au projet LECTAUREP permettra d'ajouter des étiquettes, et de le spécialiser sur le domaine des répertoires des notaires. Un modèle de REN, basé sur camemBERT, a été rendu disponible par un utilisateur du nom de Jean-Baptiste sur le site de la librairie *Hugging Face*, qui permet de se rendre compte de l'efficacité d'un modèle non affiné.<sup>53</sup> Pour cet enregistrement, choisi dans une page d'un répertoire de notaire transcrit automatique, avec un taux de CER de 9% :

Maisel (entre M. Joseph) aujet russe poaullier, bipoutier a Paris rue du fg M Antoine 71 et mademoiselle Gabrielle Sviadocht à Paris rue dufau-bourg saint Antoine 71 Apports futur  
500 future 500 constitution dot à la future. 41000 F—

Le résultat obtenu est satisfaisant (voir 6.1. Les noms ont correctement été extraits, ainsi que les adresses. Seul le titre de civilité n'a pas été reconnu. En affinant un modèle, on

49. Le premier modèle mixte de LECTAUREP atteint un score d'*accuracy* de 91%, et le deuxième un score d'*accuracy* de 90%. Voir <https://gitlab.inria.fr/dh-projects/kraken-models/-/issues/8> et <https://gitlab.inria.fr/dh-projects/kraken-models/-/issues/17> (consultés le 23/08/21).

50. A. Prasad, H. Déjean, J.L. Meunier, *et al.*, *Bench-Marking Information Extraction in Semi-Structured Historical Handwritten Records...*, p. 5

51. A. Hamdi, A. Jean-Caurant, N. Sidère, *et al.*, « An Analysis of the Performance of Named Entity Recognition over OCRed Documents »..., p. 334

52. Celles-ci ont également été indiquées dans une *issue* Gitlab : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/issues/3#note\\_552342](https://gitlab.inria.fr/almanach/lectaurep/ner/-/issues/3#note_552342) (consulté le 23/08/21).

53. Le modèle est disponible ici : <https://huggingface.co/Jean-Baptiste/camembert-ner>, il est également possible de le tester directement dans un navigateur internet (consulté le 23/08/21).

pourrait donc tendre vers ce résultat, tout en envisageant d'extraire d'autres EN, le token « bipoutier », « bijoutier » dans la vérité de terrain, en tant que *func* (fonction), par exemple.

⚡ Hosted inference API ⓘ

TokenName Classification

Maisel (entre M. Joseph) aujet russe poaullier, bipoutier a Paris rue du Compute

Computation time on cpu: 0.0911999999999999 s

Mais PER el PER (entre M PER . PER Joseph PER ) aujet russe poaullier, bipoutier a  
Paris LOC rue LOC du LOC f LOC g LOC M LOC Antoine LOC 71 LOC et  
mademoiselle Gabriel PER le PER S PER vi PER ado PER cht PER à Paris LOC  
rue LOC du LOC f LOC au LOC - LOC bourg LOC saint LOC Antoine LOC 71 LOC

Apports futur 500 future 500 constitution dot à la future. 41000 F-

JSON Output Maximize

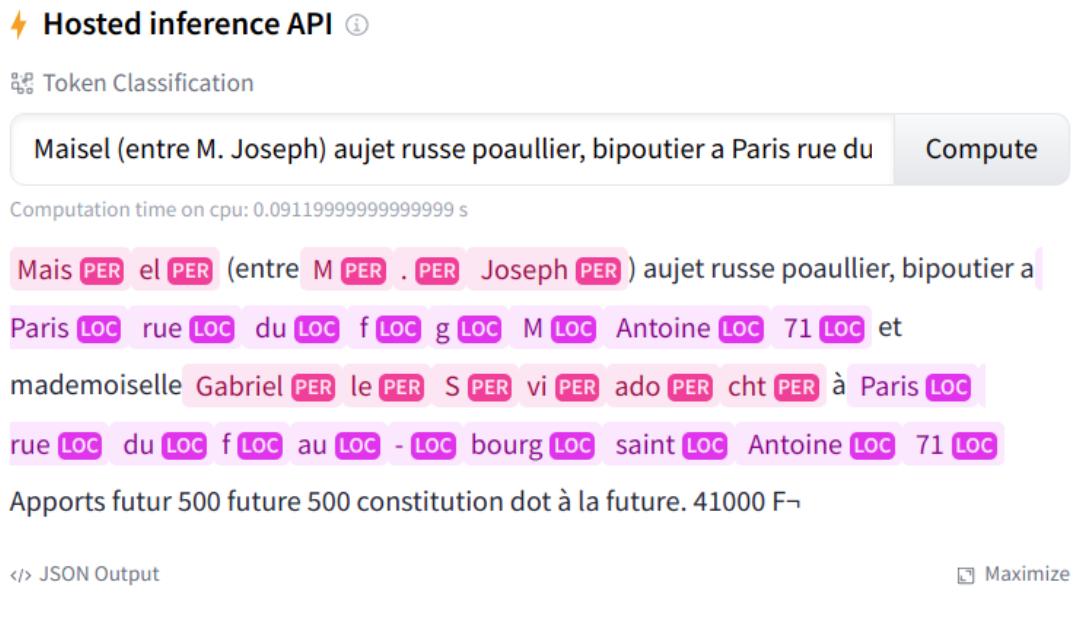


FIGURE 6.1 – Résultats de la classification des tokens d'un enregistrement transcrit automatique (9% de CER) avec un modèle de REN basé sur camemBERT. Source : <https://huggingface.co/Jean-Baptiste/camembert-ner> (consulté le 23/08/21).

LECTAUREP pourrait également envisager d'annoter des données issues de la vérité de terrain. Pour cela les mots devront être segmentés de sorte à se rapprocher des données sur lesquelles a été entraîné camemBERT.



# **Chapitre 7**

## **Une solution hétérodoxe : la reconnaissance d'entités nommées basée sur un système de règles**

La REN basée sur des règles est une solution que l'on peut qualifier d'« hétérodoxe » car peu utilisée dans une époque où les solutions faisant appel à l'apprentissage profond sont la norme. De plus, les systèmes à règles sont considérés comme moins adaptables que les systèmes d'apprentissage, et l'élaboration de règles complexes n'est pas une garantie de performance.<sup>1</sup> Pourtant, la nature des données du projet LECTAUREP pourrait se prêter à l'application d'un système à base de règles.

### **7.1 Une solution spécifique et non généralisable, profitant de la nature des données exploitées par le projet LECTAUREP**

M. Benoît Sagot, lors de l'entretien susmentionné, m'a conseillé d'envisager un système de REN basé sur des règles, adapté dans le cas de LECTAUREP car les données sont tabulaires et l'information est normalisée. Un référentiel de prénoms et de noms pourrait être utilisé pour récupérer les noms propres dans la cinquième colonne. De plus, les adresses sont toujours précédées de la préposition « à », qui annonce le nom de la ville, Paris, puis de la rue et de son numéro (voir figure 7.1).

On pourrait concevoir un système à base de règles reposant sur le système d'information présenté dans la section 2.2. Des règles pourraient être créées pour la cinquième colonne, en

---

1. Y. Dupont, *La structuration dans les entités nommées...*, pp. 56 - 57

Procuration	Létellier (par Jean Marie (par Victor Félix) à Paris, rue Blanche 12, à vendre immeuble)
-------------	--

FIGURE 7.1 – Exemple d'un enregistrement issu d'une page d'un répertoire de notaire, avec une indication d'adresse introduite par la préposition « à ».

Sté des Locomotives sans foyer (et là) par Léon Frangé, à Paris, 48, avenue Victor Hugo
Sté Blocq frères et Cie (par la Sté) siège à Coul, à Gabriel Henri Coniat, rue du Printemps 7, de 7.875

FIGURE 7.2 – Exemple de syntaxe pour les enregistrements concernant des organisations.

fonction du type d'acte présent en troisième ou quatrième colonne. Signalons néanmoins que sa mise en place risque d'être coûteuse en termes de temps, notamment pour couvrir tous les types d'actes de façon exhaustive, et pourrait ne pas être entièrement robuste s'il existe des variations linguistiques selon les notaires.<sup>2</sup>

Notons également la différence de syntaxe entre les enregistrements concernant des organisations et ceux concernant des personnes dans la parenthèse suivant directement ces EN respectives. Dans le premier cas, le notaire inscrit, par exemple « par la société ». Dans le second cas, on trouve habituellement le prénom de la première personne concernée par l'enregistrement (voir figures 7.2 et 7.3). Nous pourrions peut-être utiliser les tokens entre parenthèses situés juste après la première EN d'un enregistrement afin de distinguer si celui-ci débute par un nom de personne ou un nom d'organisation. Ceci dit, la parenthèse est un élément qui peut disparaître lors de la REM, ce qui pourrait rapidement rendre ce système non robuste.

La construction d'un système à base de règles signifie aussi que le projet LECTAUREP s'attacherait à produire un système non généralisable et spécifique à ses données.

2. On compte, selon un premier relevé réalisé dans les répertoires des notaires, 504 types d'actes différents. Cette liste doit très certainement être complétée. Voir [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/referentiels/referentiel\\_types\\_acte.json](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/referentiels/referentiel_types_acte.json) (consulté le 23/08/21).

*Henry par Augustine Victoire Lecharpentier, épouse de Jean Baptiste à Paris,  
rue d'Athènes, 24, au mariage de sa fille*

*Moreau Simon par Gustave Arthur Félix et Marie Françoise Pellerin ;  
ép., à Paris, rues des Batignolles 76, le vendredi 11 novembre*

FIGURE 7.3 – Exemple de syntaxe pour les enregistrements concernant des personnes.

NATURE ET ESPÈCE DES ACTES :	
EN BREVETS	EN MINUTES
<i>Mainlevée</i>	<i>(Subst. M. Moreau)</i>
<i>(Subst. M. Moreau)</i>	<i>Quittance</i>
<i>(— 9 —)</i>	<i>Dépol de décharge</i>
<i>(Subst. M. l. Baudrier)</i>	<i>Dépol de testament</i>
<i>Procuration</i>	

FIGURE 7.4 – Indications de substitutions dans la troisième et quatrième colonne.

## 7.2 De l'utilité des référentiels pour de la reconnaissance d'entités nommées avec un système de règles

Mis à part la construction de règles s'appuyant sur des indices linguistiques, certaines EN pourraient donc être extraites à l'aide de référentiels. Selon cet objectif, et pour expérimenter cette solution, un script a été mis au point pour tenter de récupérer les types d'actes des colonnes 3 et 4. Il n'est en effet pas possible de partir du principe que ces colonnes ne contiennent que ces informations, elles contiennent des occurrences d'informations différentes, lorsqu'un notaire indique une substitution, par exemple (voir figure 7.4).

Le script s'insère au bout de la chaîne de traitement décrite dans la section 4.1, à l'issue de la constitution d'un fichier XML-TEI à partir de la transcription au format PAGE XML, dont le processus sera expliqué dans la troisième partie.<sup>3</sup> Le script traite donc les données

3. Voir le Google Colab de la chaîne de traitement, section 3.3 : <https://colab.research.google.com/>.

textuelles issues des colonnes trois et quatre, puis les tokenise. Afin d'être flexible au bruit généré par la REM, la librairie Fuzzywuzzy a été utilisée. Elle permet de calculer un ratio de *fuzzyness*, ou de ressemblance, entre deux chaînes de caractères, basé sur la distance de Levenshtein.<sup>4</sup> Chaque token est ensuite comparé à un référentiel des types d'actes, qui a été créé à partir de la transcription manuelle de 160 pages des répertoires des notaires, où seules les troisième et quatrième colonnes ont été traitées.<sup>5</sup> Avec Fuzzywuzzy, on calcule le ratio entre chaque token et chaque type d'acte du référentiel. Lorsqu'un ratio supérieur ou égal à 91 est détecté, on estime qu'un token est identifié comme un type d'acte.

Cependant, ce système n'est pas encore complet : il ne traite que les tokens individuels, et ne prend pas en compte les mots complexes (voir section 2.2). Si ce genre de système est adopté, il faudrait pouvoir comparer un token et les n-grammes qui lui sont associés par rapport au référentiel. Si l'on se base sur un des types d'actes le plus long du référentiel, « acte de création d'obligations hypothécaires avec sociétés civile des obligations », il faudrait envisager de comparer jusqu'à 11-grammes. Cependant, la longueur des types d'actes est en moyenne entre 1 et 5 mots. On pourrait se contenter de cette longueur de n-grammes afin d'optimiser le système. L'utilisation d'un référentiel n'est pas parfaite car elle repose sur le degré d'exhaustivité dudit référentiel. Si un type d'acte de la colonne trois ou quatre n'est pas présent dans celui-ci, alors il ne pourrait pas être extrait. De plus, notons que le projet HIMANIS a également utilisé des référentiels pour extraire des EN comme des noms de lieux, de personnes, des concepts et des mots matières, en utilisant des référentiels créés par les AN, et des référentiels externes de données universelles, tels que GeoNames, Wikipedia et DBpedia, ainsi que des dictionnaires topographiques, comme par exemple Dico-topo.<sup>6</sup>

Enfin, il existe des référentiels de mots matières et de types de document, ainsi qu'un référentiel de notices producteurs, créés par les AN qu'il serait possible d'utiliser pour effectuer de la REN avec ce système.<sup>7</sup> Le deuxième pourrait venir compléter le référentiel des types d'actes, tandis que le troisième pourrait servir à identifier les noms de notaires substituants dans les colonnes 3 et 4.

---

com/drive/1utnnaUFQd3WMcSdbNt-qN-sjeT7rJvH2?usp=sharing#scrollTo=0zZl0pw00kN5 (consulté le 23/08/21).

4. Voir <https://github.com/seatgeek/fuzzywuzzy> (consulté le 23/08/21).

5. Un script a été écrit pour constituer le référentiel, il se trouve ici : <https://gitlab.inria.fr/almanach/lectaurep/ner/-/tree/master/referentiels/script>. Le référentiel en question se situe ici : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/referentiels/referentiel\\_types\\_acte.json](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/referentiels/referentiel_types_acte.json) (consultés le 23/08/21).

6. D. Stutzmann, J.F. Moufflet et S. Hamel, « Full Text Search in Medieval Manuscripts... »

7. Cyprien Henry et Florence Clavaud, *Vers un référentiel national des notaires ?*, fdocuments.fr, URL : <https://fdocuments.fr/document/relier-donnees-referentielsnotaireschenryfclavaud-final.html> (visité le 13/07/2021) et Marie-Françoise Limon-Bonnet et Gaetano Piraino, « Préparer l'innovation : l'informatisation des ressources du minutier central », *La Gazette des archives*, 2 (numéro 254 [2019]), p. 252-266, pp. 261 - 262

## **Troisième partie**

**Exploiter les entités nommées dans le  
contexte des métiers du patrimoine**



# **Chapitre 8**

## **Le signalement des entités nommées au sein d'un encodage en XML-TEI : l'après NER**

Une fois que le projet LECTAUREP aura une solution de REN pérenne, il sera nécessaire de réfléchir à une façon de stocker les EN extraites et classifiées en vue de les exploiter. Nous présenterons pour cela un scénario envisageable pour LECTAUREP, qui consiste à signaler les EN dans un encodage XML-TEI.

### **8.1 Modélisation de la structure logique des répertoires des notaires en TEI**

La chaîne de traitement présentée dans la section 4.1 permet, de transformer le fichier PAGE XML traité en fichier XML-TEI, après avoir reconstitué la structure logique d'une transcription automatique. Le choix de la TEI a été fait car ce format se prête à combiner les métadonnées d'un document et son texte, avec la possibilité de l'enrichir.<sup>1</sup>

#### **8.1.1 Transformer le PAGE XML en XML-TEI**

Mme Alix Chagué et moi-même avons commencé par utiliser une feuille XSLT écrite par notre collègue Mme Manon Ovide, stagiaire dans le cadre du projet DAHN sous la supervision de Mme Floriane Chiffoleau, qui permet de transformer un encodage PAGE XML en un

---

1. *Text Encoding Initiative*, voir <https://tei-c.org/> (consulté le 24/08/21). Voir également D. Stutzmann, J.F. Moufflet et S. Hamel, « Full Text Search in Medieval Manuscripts... », p. 77.

encodage XML-TEI.<sup>2</sup> Cette transformation permettait de générer un <teiHeader> basique suivi d'une arborescence ouverte par la balise <sourceDoc>. Celle-ci peut contenir une transcription, ou une représentation d'une source unique.<sup>3</sup> Dans le <sourceDoc> venait ensuite s'ouvrir une balise <surfaceGrp>, qui sert à indiquer un ensemble de zones contenant de l'écriture dans un document, par exemple le recto d'une feuille, contenant elle-même une balise <surface>, destinée à définir une zone écrite avec des coordonnées en deux dimensions.<sup>4</sup> Dans celle-ci, une balise <graphic> était insérée, indiquant avec un attribut *url* le nom du fichier image source, et avec les attributs *width* et *height* la hauteur et la largeur de l'image. Des balises <zone> viennent enfin symboliser chaque *baseline* du PAGE XML, en indiquant avec un attribut *xml :id* l'identifiant de ladite *baseline*, et ses coordonnées *x* et *y* avec un attribut *points*. Chaque chaîne de caractères issue de la transcription est contenue dans la balise enfant <line>. L'encodage TEI du <sourceDoc> obtenu avec cette transformation prenait donc la forme suivante :

```

<sourceDoc>
    <surfaceGrp xml:id="eSc_textblock_a5c6ac11 eSc_textblock_38d68d65"
        ↳ eSc_textblock_3f255e6a eSc_textblock_bf739b41
        ↳ eSc_textblock_653e37f9 eSc_textblock_65c23f81
        ↳ eSc_textblock_2a3815f7 eSc_textblock_d0d6cce a eSc_dummyblock_>
            <surface>
                <graphic url="DAFANCH96_023MIC07633_L-0.jpg" width="2048px"
                    ↳ height="2944px"/>
                <zone xml:id="eSc_line_746878ea"
                    points="289,586 317,556 348,555 356,559 388,554 388,603
                    ↳ 360,603 349,595 338,603 331,597 291,594">
                    <line>852</line>
                </zone>
                <zone xml:id="eSc_line_38ccd069"
                    points="294,694 318,674 336,670 341,676 372,674 385,688
                    ↳ 386,702 383,715 358,716 347,724 337,715 313,716
                    ↳ 296,695">
                    <line>853</line>
    
```

2. La feuille XSLT de Manon Ovide se trouve ici : [https://github.com/inoblivionem/xslt-playground/blob/main/xmlpage\\_to\\_tei/xmlpage\\_to\\_tei.xsl](https://github.com/inoblivionem/xslt-playground/blob/main/xmlpage_to_tei/xmlpage_to_tei.xsl) (consulté le 24/08/21).

3. "<<sourceDoc>> contains a transcription or other representation of a single source document potentially forming part of a dossier génétique or collection of sources." Voir <https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-sourceDoc.html> (consulté le 24/08/21).

4. Voir <https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-surfaceGrp.html> et <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-surface.html> (consulté le 24/08/21).

```

</zone>
...
</surface>
</surfaceGrp>
</sourceDoc>

```

Nous avons fait appel à M. Laurent Romary pour lui demander conseil afin d'adapter cette feuille de transformation aux besoins du projet LECTAUREP. Nous avons décidé qu'il était préférable de retrouver les différentes régions (`<TextRegion>` dans le PAGE XML) dans l'encodage TEI, ainsi que les *baselines* qui leurs étaient respectivement associées pour garder trace de la segmentation sémantique. De plus, nous avons déterminé qu'il fallait obtenir une feuille XSLT qui permettrait de créer, depuis un fichier PAGE XML en entrée, un fichier XML-TEI en sortie, possédant le même nom que le premier. J'ai donc modifié cette feuille en fonction.<sup>5</sup> Avec ces modifications, on retrouve dans le `<sourceDoc>` autant de `<surfaceGrp>` qu'il y a de `<TextRegion>` dans le PAGE XML. En outre, la balise `<surfaceGrp>` a un nouvel attribut *type*, prenant comme valeur celle de l'attribut *custom* de la balise `<TextRegion>`.<sup>6</sup> Ainsi, l'encodage TEI du `<sourceDoc>` ressemble à ceci :

```

<sourceDoc>
  <surfaceGrp xml:id="eSc_textblock_a5c6ac11"
    ↳ type="structure_{type:col_1;}">
    <surface>
      <graphic url="DAFANCH96_023MIC07633_L-0.jpg" width="2048px"
        ↳ height="2944px"/>
      <zone xml:id="eSc_line_746878ea"
        points="289,586 317,556 348,555 356,559 388,554 388,603
        ↳ 360,603 349,595 338,603 331,597 291,594">
        <line>852</line>
        ...
      </surface>
    </surfaceGrp>
    <surfaceGrp xml:id="eSc_textblock_3f255e6a"
      ↳ type="structure_{type:col_7;}">

```

5. La feuille XSLT modifiée est accessible ici : [https://github.com/HugoSchtr/xslt-playground/blob/modif\\_xmlpagetotei\\_xsl/xmlpage\\_to\\_tei/xmlpage\\_to\\_tei.xsl](https://github.com/HugoSchtr/xslt-playground/blob/modif_xmlpagetotei_xsl/xmlpage_to_tei/xmlpage_to_tei.xsl) (consulté le 24/08/21).

6. Un exemple de fichier XML-TEI transformé est disponible dans le commentaire de l'*issue* présentant la chaîne de traitement : [https://gitlab.inria.fr/almanach/lectaurep/documentation/-/issues/17#note\\_551369](https://gitlab.inria.fr/almanach/lectaurep/documentation/-/issues/17#note_551369), le lien pour télécharger le fichier XML-TEI directement est le suivant : [https://gitlab.inria.fr/almanach/lectaurep/documentation/uploads/e5287f18c9eaa6f3fa262e3c82a576fb/FRAN\\_0025\\_3056\\_L-0-tei.xml](https://gitlab.inria.fr/almanach/lectaurep/documentation/uploads/e5287f18c9eaa6f3fa262e3c82a576fb/FRAN_0025_3056_L-0-tei.xml) (consultés le 25/08/21).

```

<surface>
  <graphic url="DAFANCH96_023MIC07633_L-0.jpg" width="2048px"
    ↳ height="2944px"/>
  <zone xml:id="eSc_line_1feb9822"
    points="1913,652 1914,638 1925,626 1939,626 1950,634
    ↳ 1999,582 2003,650 1994,649 1961,685 1945,684
    ↳ 1915,660">
    <line>3.75</line>
  </zone>
  ...
</surface>
</surfaceGrp>
...
</sourceDoc>

```

### 8.1.2 Modéliser en TEI la structure tabulaire des réertoires des notaires dans la balise <text>

Nous avons ensuite fait le choix de modéliser les réertoires des notaires en TEI selon trois blocs différents : le <teiHeader> où sont encodées les métadonnées<sup>7</sup> ; le <sourceDoc> qui permet de conserver le texte brut issu de la REM ; la reconstruction de la structure logique est encodée dans la balise <text>. C'est également à ce niveau de l'arborescence TEI que les post traitements de la REM pourront être appliqués : corrections post REM et enrichissement des données avec la REM (voir figure 8.1). On garde ainsi la transcription automatique brute, et sa version reconstituée, permettant à n'importe qui d'avoir accès à ces informations dans un même fichier XML.

Dans la balise <text>, la structure tabulaire des réertoires des notaires a été reproduite grâce aux balises que propose la TEI. Nous avons choisi d'utiliser la balise <table> et sa balise enfant <row> pour chaque rangée, qui contient elle-même des balises <cell> pour chaque cellule, et donc colonnes.<sup>8</sup> Cette dernière balise possède un attribut *fac*, indiquant l'identifiant de la *baseline* du PAGE XML. L'arbre XML détaillé de la modélisation des réertoires des notaires en TEI est disponible dans l'annexe J.2. Enfin, il a été décidé de respecter la mise en page du document original en indiquant les sauts de ligne à l'aide de la balise <lb>. Avec

---

7. La forme du <teiHeader> n'est pas encore définitive et est encore en travail. Voir l'*issue* suivante : <https://gitlab.inria.fr/almanach/lectaurep/documentation/-/issues/18> (consulté le 25/08/21).

8. Voir <https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-table.html>, <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-row.html> et <https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-cell.html> (consultés le 25/08/21).

ces balises, le texte est aligné avec l'image. Une rangée d'une page d'un répertoire de notaire prend donc la forme suivante en TEI :

```
<text>
  <body>
    <table rows="12" cols="7">
      <row>
        <cell n="col1" role="data"
          ↳ facs="#eSc_line_746878ea">852</cell>
        <cell n="col2" role="data" facs="#eSc_line_b0ea30ca"><date
          ↳ when-iso="0000-00-29">29</date></cell>
        <cell n="col3" role="data"
          ↳ facs="#eSc_line_b62c40c8">Procuration</cell>
        <cell n="col4"/>
        <cell n="col5" role="data" facs="#eSc_line_337b98ba
          ↳ #eSc_line_7a7a45ab \#eSc\_line\_f48a5d51"><lb n="1"
            />Bonnet (par Germain Martial) dt à Paris rue <lb
            ↳ n="2"/>de Chateaudun n°23 à de
            Honorine Marie Therèse <lb n="3"/>Hardel sa fe dt avec
            ↳ lui à l'effet de XXX</cell>
        <cell n="col6" role="data" facs="#eSc_line_0433c14f"><date
          ↳ when-iso="0000-00-30">30</date></cell>
        <cell n="col7" role="data"
          ↳ facs="#eSc_line_1feb9822"><measure unit="francs"
            ↳ quantity="3.75"
            >3.75</measure></cell>
      </row>
    </table>
    <div type="misc"/>
  </body>
</text>
```

## 8.2 Automatisation de la transformation du PAGE XML résultant de la REM en XML-TEI

La chaîne de traitement applique la feuille XSLT modifiée à chaque fichier PAGE XML qui lui est soumis. Pour chacun d'eux, un nouveau fichier XML-TEI est créé, dans lequel on

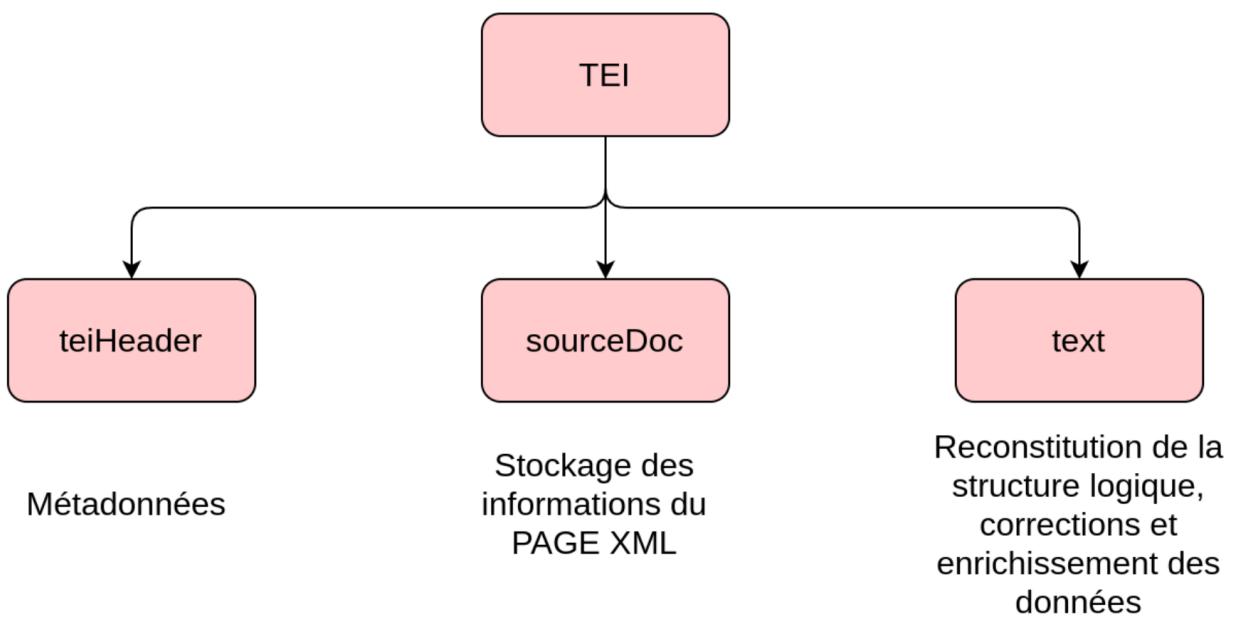


FIGURE 8.1 – Arbre XML-TEI simplifié de la modélisation des pages des répertoires des notaires.

ajoute, en plus du <sourceDoc>, un <teiHeader> et une balise <text> qui est remplie grâce à l'objet python *Row* que nous avons présenté dans la section 4.1. Pour chaque objet *Row* créé à partir d'un fichier PAGE XML, on crée une balise <row> dans l'arborescence <table> du fichier TEI, dont chaque balise <cell> est peuplée à partir de ce premier.<sup>9</sup> Tous les fichiers XML-TEI sont enfin rassemblés dans une archive, afin de les exporter facilement pour préparer leur exploitation et leur publication.

## 8.3 Signalement des entités nommées dans un encodage TEI

Carmen Brando et Gabriela Elgarrista ont présenté, au cours de la cinquième session du séminaire bimensuel des Sources aux Systèmes d'Information Géographique, le 6 avril 2021, une chaîne de traitement destinée à l'analyse des annuaires de propriétaires de Paris.<sup>10</sup> Elle permet de traiter des images et d'en proposer une transcription automatique à l'aide

9. Voir le schéma illustrant la chaîne de traitement dans l'annexe J.1.

10. Voir la Cinquième session du séminaire bimensuel des Sources aux Systèmes d'information Géographique, DYPAC UVSQ, [https://www.youtube.com/watch?v=4xkZHdy88DU&ab\\_channel=DYPACUVSQ](https://www.youtube.com/watch?v=4xkZHdy88DU&ab_channel=DYPACUVSQ) 11 :45 - 1 :15 :50.

d'eScriptorium.<sup>11</sup> La transcription est structurée en XML-TEI et permet ensuite de donner un cadre à l'extraction et la spatialisation des adresses grâce à la géolocalisation des entités récupérées.

Obtenir un encodage TEI des transcriptions des répertoires des notaires donne une opportunité similaire pour signaler les EN. En 2016, Solenn Le Pevedic et Denis Maurel ont écrit un article traitant du balisage des EN grâce à la TEI, dans lequel ils soutiennent que ce standard est adapté à cette tâche et préconisent par conséquent son utilisation.<sup>12</sup> Ils retiennent premièrement plusieurs balises générales, couramment utilisées avec la TEI. Pour commencer : la balise `<rs>` (*reference string*)<sup>13</sup>, qui permet de renvoyer un élément à un référent, notamment grâce à l'attribut *type* qui indique le type (une personne par exemple), et l'attribut *ref*, qui le lie à un référent dans un index. Ensuite, les balises `<persName>` et ses balises enfants `<forename>`, `<roleName>`, etc. qui peuvent être utilisées pour signaler les noms de personnes ; les balises `<placeName>` et `<geogName>` pour encoder respectivement les lieux administratifs et les lieux géographiques, et l'élément `<address>` et ses balises enfants pour encoder précisément les adresses ; pour les organisations, ils préconisent l'utilisation de la balise `<orgName>`, qui indique un regroupement de personnes considéré comme une unité.<sup>14</sup> Avec ce set de balises, il est ainsi possible de couvrir les EN basiques identifiées par les modèles génériques que nous avons présenté. Pour les dates et les mesures, ils proposent les balises `<date>` et `<num>`, cette dernière servant à signaler les valeurs numériques.<sup>15</sup>

Avec cet ensemble de balises, nous pouvons envisager un encodage des EN dans les répertoires des notaires. Nous pourrions encoder les types d'actes extraits des colonnes trois et quatre avec une balise `<rs>`, référant elle-même à un index TEI des types d'actes. De la sorte, chacune des balises `<rs>` pourrait pointer vers un type d'acte en particulier, et vers une forme normalisée de celui-ci. L'exploitation de ces balises permettrait une première analyse quantitative des actes enregistrés dans les répertoires des notaires. Les noms de la cinquième colonne, ainsi que les noms d'organisations, les dates et les adresses pourraient également être signalés avec les balises correspondantes. Enfin, les sommes d'argent inscrites dans les colonnes 6 et 7 pourraient être balisées avec `<num>`, néanmoins, la TEI propose également

---

11. Anciennement Transkribus.

12. Solenn Le Pevedic et Denis Maurel, « Retour sur les annotations des entités nommées dans les campagnes d'évaluation françaises et comparaison avec la TEI », *Corela. Cognition, représentation, langage*—14 (22 nov. 2016), Number : 14-2 Publisher : Université de Poitiers, DOI : 10.4000/corela.4644

13. Voir <https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-rs.html> (consulté le 25/08/21).

14. Pour la balise `<persName>`, voir <https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-persName.html>, pour `<placeName>`, voir <https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-placeName.html>, pour `<address>`, voir <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-address.html>, et pour `<orgName>`, voir <https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-orgName.html> (consultés le 25/08/21).

15. Voir <https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-date.html> et <https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-num.html> (consultés le 25/08/21).

<measure>, accompagnée d'un attribut *type* ayant pour valeur, par exemple, *currency*, pour indiquer les sommes d'argent. Nous l'utilisons dans la chaîne de traitement. Nous aurions ainsi, pour une cellule d'une rangée de la cinquième colonne, un encodage de base prenant la forme suivante :

```
<cell n="col5" role="data"
      facs="#eSc_line_337b98ba #eSc_line_7a7a45ab
      ↳ \#eSc\_line\_f48a5d51"><lb n="1"
      /><persName>Bonnet</persName> (par
      ↳ <persName>Germain Martial</persName>) dt à
      ↳ <address>
      <settlement>Paris</settlement>
      <street>rue <lb n="2"/>de Chateaudun n°23</street>
    </address> à de <persName>Honorine Marie Therèse <lb
    ↳ n="3"/>Hardel</persName> sa
    fe dt avec lui à l'effet de XXX
  </cell>
```

Dans cet article, S. Le Pevedic et D. Maurel ont également proposé des équivalences entre les types d'EN proposés par le guide d'annotation Quaero et la TEI. Ils estiment que le balisage en TEI peut être effectué sans problème de transposition à partir du guide Quaero. Leurs propositions pourraient donc profiter au projet LECTAUREP si un modèle de REN est entraîné avec davantage d'étiquettes que celles que proposent les modèles génériques de REN. Pour cela, ils suggèrent des conversions en TEI pour les personnes (« pers » dans le guide Quaero), pour les fonctions (« func »), pour les organisations (« org »), pour les localisation (« loc »), pour les productions humaines (« prod ») pour les quantités (« amount »<sup>16</sup>), pour les points dans le temps (« time »), et pour les événements (« event »). Cela aide les projets utilisant ce guide d'annotation pour signaler les EN extraites dans un fichier TEI. Ce guide de conversion serait donc une ressource importante pour le projet LECTAUREP, si les types d'EN du guide Quaero sont retenues.

Cette chaîne de traitement correspond aux premiers blocs de la construction d'un système d'extraction d'information, car elle crée une structure dans laquelle extraire les EN et les signaler. De plus, la TEI offre aux EN un environnement où les signaler et les stocker, et les ouvre potentiellement à des exploitations externes aux répertoires des notaires.

---

16. S. Rosset, C. Grouin et P. Zweigenbaum, *Entités nommées structurées...*, pp. 47 - 54

## **8.4 Une plate-forme de publication pour les transcriptions des répertoires des notaires : visualisation et recherches dans les pages des répertoires de notaires transcrits avec TEI Publisher.**

La chaîne de traitement présentée ci-dessus nous a servi de point de départ pour expérimenter la publication des transcriptions automatiques des répertoires des notaires en utilisant une application *open-source*, TEI Publisher que nous présenterons dans cette section.

### **8.4.1 Une interface configurable pour la publication des transcriptions des répertoires des notaires : l'application *open-source* TEI Publisher.**

Mme Floriane Chiffoleau et Mme Manon Ovide ont eu l'occasion de présenter la publication des fichiers TEI produits dans le cadre du projet DAHN au cours d'une réunion à la fin du mois de mai à laquelle Mme Alix Chagué et moi-même avons assisté. Nous avons alors pris connaissance de la plate-forme TEI Publisher et de son utilisation. Ayant à disposition des fichiers TEI avec la chaîne de traitement, nous avons décidé d'expérimenter avec cet outil.

TEI Publisher est une application *open-source* développée par une société internationale à but non lucratif nommée e-editiones.<sup>17</sup> Elle offre un environnement pour publier des fichiers XML-TEI, basé sur le standard TEI et le TEI *Processing Model*.<sup>18</sup> L'application fonctionne avec une base de données XML, exist-db, également *open-source*.<sup>19</sup> C'est avec ce système que sont stockés l'application et les encodages TEI.<sup>20</sup> Avec ceux-ci, TEI Publisher peuple des pages HTML à l'aide d'une ODD qui indique comment transformer les éléments TEI. L'application propose par ailleurs un éditeur d'ODD visuel, sans passer par l'écriture de celle-ci.<sup>21</sup>

TEI Publisher est aussi une interface permettant de naviguer à travers une collection de fichiers XML-TEI, en la parcourant avec un système de pagination, avec des recherches plein texte ou à filtres.<sup>22</sup> Il existe des versions de démonstration qui permettent de se rendre compte

---

17. Voir le site internet de TEI Publisher : <https://teipublisher.com/index.html>, ainsi que le répertoire Github de l'application : <https://github.com/eeditiones/tei-publisher-app> (consultés le 25/08/21).

18. Le TEI *Processing Model* déclare un modèle d'actions pour les éléments présents dans l'arbre XML. Voir <https://tei-c.org/release/doc/tei-p5-doc/fr/html/TD.html#TDPMPM> et F. Chiffoleau, *Publication of my digital edition – Working with TEI Publisher*, Digital Intellectuals, déc. 2020, URL : <https://digitalintellectuals.hypotheses.org/3912> (visité le 19/08/2021).

19. Voir <http://exist-db.org/exist/apps/homepage/index.html> (consulté le 25/08/21).

20. Voir annexe K.1.

21. Voir annexe K.2.

22. Voir annexe K.3.

de la puissance et de la diversité des affichages qu'il est possible d'obtenir. On trouve ainsi une édition de lettres écrites par Vincent Van Gogh<sup>23</sup>, une collection de pièces de Shakespeare<sup>24</sup>, ou encore une collection de documents commémorant le trentième anniversaire de la chute du mur de Berlin.<sup>25</sup>

L'application fonctionne comme un « système de lego », utilisant à la fois l'ODD et un système de *templates*, qui ne sont autres que des pages HTML reposant sur la technologies des *Web Components*, eux-mêmes faisant partie des spécifications HTML5 et qui fonctionnent avec de nombreux navigateurs internet.<sup>26</sup> Chaque *template* contient des éléments, qui une fois configurés, permettent d'afficher à la fois sur une même page, par exemple, un texte dans plusieurs langues, une image d'un fac-similé, une version d'un texte normalisé, etc.<sup>27</sup> TEI Publisher est de fait une application entièrement configurable et personnalisable selon les besoins spécifiques d'un projet. Des ODD et *templates* génériques sont fournies avec l'application, et permettent de visualiser des encodages TEI simplement.<sup>28</sup>

Pour expérimenter, j'ai installé une version locale de l'application et créé plusieurs affichages des encodages TEI résultant de la chaîne de traitement.<sup>29</sup> Le premier repose sur une très simple feuille CSS (voir annexe K.6) permettant d'afficher la transcription automatique dans une structure tabulaire, à l'instar des répertoires des notaires (voir annexe K.7). Cet affichage possède des défauts et n'a pas vocation à être déployé, les dimensions du tableau ne correspondent en effet pas à celles du tableau des répertoires, les colonnes 1, 2, 6 et 7 étant notamment plus larges. La deuxième visualisation résulte de la création d'un *template* permettant d'afficher sur une même page la transcription automatique à gauche, et à droite son fac-similé (Voir annexe K.8). Cette version semble plus adaptée au projet LECTAUREP, l'utilisateur-ice pouvant confronter la transcription automatique et la vérité de terrain en image. TEI Publisher propose d'afficher une image en passant par un serveur mettant à disposition le protocole IIIF avec le *Web component* nommé <pb-facsimile>.<sup>30</sup> J'ai ainsi créé une

---

23. Voir <https://teipublisher.com/exist/apps/vangogh/index.html> (consulté le 25/08/21).

24. Voir <https://teipublisher.com/exist/apps/shakespeare-pm/index.html> (consulté le 25/08/21).

25. Voir <https://teipublisher.com/exist/apps/dodis-facets/index.html> (consulté le 25/08/21).

26. Voir [https://developer.mozilla.org/en-US/docs/Web/Web\\_Components](https://developer.mozilla.org/en-US/docs/Web/Web_Components) (consulté le 25/08/21).

27. Voir "Page templates and pb-components" dans la documentation de TEI Publisher, <https://teipublisher.com/exist/apps/tei-publisher/doc/documentation.xml?odd=docbook.odd&id=webcomponents-intro> (consulté le 25/08/21).

28. Voir annexes K.4 et K.5.

29. Le test de l'application installée localement sur ma machine a été documenté dans un commentaire de l'*issue* dédiée à la chaîne de traitement : [https://gitlab.inria.fr/almanach/lectaurep/documentation/-/issues/17#note\\_552768](https://gitlab.inria.fr/almanach/lectaurep/documentation/-/issues/17#note_552768) (consulté le 25/08/21).

30. *International Image Interoperability Framework*. IIIF est un ensemble de spécifications techniques permettant la diffusion et l'échange d'images en haute résolution dans un soucis d'intéropérabilité. Voir <https://iiif.io/> (consulté le 25/08/21).

collection publique sur Nakala, un service développé et déployé par Huma-Num pour partager et publier des données scientifiques, dans laquelle ont été stockées les 10 numérisations des 10 transcriptions automatiques utilisées pour tester la chaîne de traitement.<sup>31</sup> Nakala met à disposition de ses utilisateurs une API IIIF qui a été utilisée pour obtenir cet affichage. Dans le fichier HTML, cet élément prend la forme suivante :

```
<pb-facsimile id="facsimile" base-uri="https://api.nakala.fr/iiif/"  
    ↳ default-zoom-level="0"  
    ↳ show-navigation-control="show-navigation-control"  
    ↳ show-navigator="show-navigator" subscribe="transcription"/>
```

L'attribut *base-uri* sert notamment à renseigner l'adresse du serveur IIIF depuis lequel les images sont récupérées. L'identifiant fourni par Nakala pour accéder à chacune des images a lui été inséré manuellement dans les encodages TEI en valeur d'un attribut *url*, au sein de la balise *<graphic>* dans le *<sourceDoc>*.

```
<graphic  
    ↳ url="10.34847/nkl.e7fbv097/682f5b9ab12c584f250186526a1d4b6198ba2856"/>
```

Le lien entre l'encodage TEI et le *template* est réalisé avec une ODD personnalisée pour cette tâche. Il manque cependant l'accès aux métadonnées de l'encodage, qu'il serait possible d'inclure dans le *template*. En outre, la création d'un lien entre l'encodage TEI et l'identifiant IIIF de la numérisation dont il est issu pourrait être automatisée. Il serait également possible de concevoir un affichage de l'encodage TEI où chaque EN, signalée avec sa balise appropriée, serait signalée, par exemple, par une couleur différente.<sup>32</sup>

L'exhaustivité de sa documentation fait de TEI Publisher une application facile à installer et à configurer.<sup>33</sup> L'expérience acquise au sein d'ALMAnaCH dans TEI Publisher à travers le projet DAHN, pourrait profiter au déploiement d'une application destinée aux fichiers XML-TEI de LECTAUREP. L'application permet également de générer une version propre à un projet de publication lorsqu'une ODD est fixée, et les *templates* définis, donc ne contenant pas les fichiers de tests génériques à l'application, ni les ODD et *templates* génériques.

31. Voir la collection en question à l'adresse suivante : <https://nakala.fr/10.34847/nkl.e7fbv097> (consulté le 25/08/21). Pour plus d'informations sur Nakala, voir le site internet <https://nakala.fr/> (consulté le 25/08/21).

32. Au sujet de la visualisation des EN, A. Bertino *et al.* insistaient sur le fait qu'un texte annoté avec des couleurs différentes peut convenir à certains utilisateurs, mais pas à d'autres, et qu'il est ainsi critique de pouvoir activer ou désactiver cette visualisation, voire même de proposer plusieurs degrés de visualisation, par exemple les plus pertinentes, puis le reste des EN extraites. TEI Publisher propose plusieurs visualisations d'un même document sur une même page, permettant donc de répondre à ce besoin utilisateur. A. Bertino, L. Foppiano, L. Romary, *et al.*, « Leveraging Concepts in Open Access Publications »..., p. 18 - 19

33. Voir la documentation de TEI Publisher : <https://teipublisher.com/exist/apps/tei-publisher/doc/documentation.xml?odd=docbook.odd> (consulté le 25/08/21).

## **8.4.2 TEI Publisher comme plate-forme de *crowdsourcing* pour la REN ?**

TEI Publisher est également une application dont le développement est soutenu et régulier, avec l'ajout de nouvelles fonctionnalités. Au cours du mois d'août 2021, la version 7.1.0 de TEI Publisher a été publiée, avec comme nouvelle fonctionnalité la possibilité d'annoter directement un texte dans l'application, la transformant donc en une plate-forme d'annotation potentielle.<sup>34</sup>

Avec cette fonctionnalité, LECTAUREP pourrait donc effectuer une première annotation des EN avec des systèmes de REN dans l'encodage TEI, les publier sur une instance personnalisée pour le projet, puis faire appel aux usagers qui seraient amenés à les consulter pour corriger les annotations, voire les compléter. L'annotation collaborative des EN pourrait par ailleurs être facilitée avec l'affichage du texte confronté à son image, de sorte à comparer la transcription automatique et la vérité de terrain.

## **8.4.3 Parcourir les transcriptions des répertoires des notaires avec TEI Publisher**

TEI Publisher permet donc de faire des recherches plein texte dans les documents stockés dans la base de données, il serait ainsi possible de rechercher un type d'acte, de retrouver toutes ses occurrences et de choisir dans quel document récupéré aller le consulter.<sup>35</sup> Il est aussi possible d'effectuer des recherches avec troncature pour donner plus de souplesse au système.<sup>36</sup>

Nous pouvons aussi paramétriser des recherches à facettes.<sup>37</sup> On pourrait ainsi trier les documents en s'appuyant sur le signalement des types d'actes dans l'encodage grâce à la REN. Un-e utilisateur-ice pourrait accéder à toutes les procurations en éliminant par exemple les autres types d'actes qui pourraient constituer du bruit dans une recherche.

Toutefois, il est important de noter que TEI Publisher ne possède à ce jour pas de système de négociation de contenu<sup>38</sup> ou de recherche floue. Cela pourrait éventuellement poser problème pour retrouver des mots bruités par la REM dans les encodages, cas de figure qui arrivera très probablement dans le contexte de la transcription automatique. Rappelons

---

34. Voir la page de la mise à jour sur le blog de TEI Publisher : <https://teipublisher.com/exist/apps/tei-publisher/doc/blog/tei-publisher-710.xml>. Une vidéo présentant cette fonctionnalité est inclue (consulté le 25/08/21).

35. Voir annexes K.9 et K.10

36. Voir annexe K.11.

37. Voir la documentation associée : <https://teipublisher.com/exist/apps/tei-publisher/doc/documentation.xml?odd=docbook.odd&id=facets> (consulté le 25/08/21).

38. D. Stutzmann, J.F. Moufflet et S. Hamel, « Full Text Search in Medieval Manuscripts... »

néanmoins que cette application est *open-source* et qu'il serait donc possible de développer cette fonctionnalité.

Enfin, LECTAUREP pourrait profiter de la publication des transcriptions des répertoires des notaires sur une instance de TEI Publisher pour établir un lien avec la SIV et les instruments de recherche EAD du DMC.

TEI Publisher se présente donc comme un espace de publication complet pour les transcriptions automatiques des répertoires des notaires, en proposant de les parcourir de façon linéaire en les lisant, ou en les consultant par d'autres portes d'entrées, selon les principes du *distant-reading*, en effectuant des recherches dans le corpus à partir ou non des EN. Cette application pourrait pourtant transformer l'expérience des usagers des AN dans la consultation de ces documents.



# Chapitre 9

## Indexer les réertoires des notaires grâce à la reconnaissance d'entités nommées

Au cours de ce mémoire, nous avons parlé sporadiquement d'indexation des EN. Cette dernière partie y sera consacrée.

### 9.1 Indexer des documents patrimoniaux : définitions

#### 9.1.1 L'indexation : définition

Un index, de manière générale, est défini dans l'*Abrégé d'archivistique* ainsi :

« INDEXATION : opération destinée à représenter par des éléments d'un langage documentaire ou naturel des données résultant de l'analyse du contenu d'un document ou du document lui-même. »

En ce sens, un index est une forme de déconstruction du texte, une mise à plat, qui permet de lui construire une nouvelle porte d'entrée. L'expert scientifique de Gallica, Jean-Philippe Moreux décrit l'indexation comme « l'organisation d'un ensemble de documents non structurés afin de faciliter la recherche d'information en son sein. »<sup>1</sup> C'est une tâche historiquement faite à la main, notamment lors de la création d'éditions critiques.<sup>2</sup>

1. É. Cavalié (dir.), *L'indexation matière en transition...*, p. 114. Voir également le chapitre écrit par Bruno Bachimont, pour une plus ample définition de l'indexation, dans le livre dirigé par É. Cavalié. *Ibid.*, pp. 23 - 31

2. Voir également la définition de D. Stutzmann *et al.* : « S'il désigne toujours un accès tabulaire à une information textuelle linéaire, l'index est à la fois le niveau zéro de l'édition et son degré ultime. C'est l'étape finale qui permet de créer un sens supplémentaire dans une édition et de révéler ce qui n'y est que sous-jacent : quand l'éditeur fournit un index des citations et réminiscences bibliques d'un texte, il ouvre une

Un index est organisé, habituellement, en référençant chaque document d'un ensemble à partir d'entrées construites à l'aide de métadonnées qui en sont issues.<sup>3</sup>

### **9.1.2 Automatiser l'indexation pour réaliser cette tâche à grande échelle**

L'indexation peut être une tâche réalisée automatiquement. Cela permet de traiter un volume de données beaucoup plus conséquent qu'un volume annoté à la main. Indexer automatiquement possède également l'avantage d'assurer la « cohérence de l'annotation », dès lors qu'un standard est créé et qu'on le fait appliquer à une machine.<sup>4</sup> Cette tâche se réalise notamment par l'extraction automatique du contenu des documents.

Mentionnons un exemple utilisé quotidiennement sur internet : les moteurs de recherche. Ceux-ci utilisent l'indexation plein texte. Dans un document textuel, il est possible de chercher n'importe quel terme à l'aide d'un moteur de recherche. Aussi nommée « recherche par mot-clé », l'indexation plein-texte souffre toutefois de la quantité de bruit non négligeable qu'elle provoque. Afin de dépasser cette faiblesse pouvant perdre les usagers, il existe d'autres méthodes d'indexation automatique que nous présenterons dans les sections suivantes.

## **9.2 Indexer automatiquement grâce à la reconnaissance d'entités nommées**

### **9.2.1 Retrouver la structure logique des documents issus de la REM pour améliorer la recherche d'information**

Pour J.-P. Moreux, la reconstitution de la structure logique des documents issus de l'OCR et de la REM permettent d'améliorer la recherche d'information. Il dit ainsi : « connaître la qualité d'un mot ou d'un passage au sein d'un texte, ou encore sa localisation dans le document, enrichit les modalités d'interrogation. »<sup>5</sup> Grâce à l'information de localisation d'un mot dans un texte, il est possible de construire un meilleur moteur de recherche à l'aide de trois paramètres. Premièrement, il est possible de pondérer la recherche. L'occurrence d'un mot à un endroit précis d'un document peut en effet posséder une valeur plus forte qu'à un autre endroit. Par exemple, dans le cas des répertoires des notaires, la date indiquée à chaque début de page est un élément important, permettant de situer les enregistrements

---

porte vers une compréhension renouvelée, non seulement de l'œuvre, mais aussi de la pensée de l'auteur. »D. Stutzmann, J.F. Moufflet et S. Hamel, « Full Text Search in Medieval Manuscripts... »

3. É. Cavalié (dir.), *L'indexation matière en transition...*, p. 114

4. *Ibid.*

5. *Ibid.*, p. 116

dans le temps. Pouvoir accéder à cette information en priorité lors d'une recherche par date permettrait de trier par année. Deuxièmement, elle permet de cibler la recherche. Par exemple, il serait possible de n'effectuer une recherche que dans la cinquième colonne des répertoires. Troisièmement, elle permet de filtrer. Nous renvoyons ici aux sections précédentes traitant de TEI Publisher.<sup>6</sup>

La chaîne de traitement exposés dans les sections 4.1 et 8.1 permet de produire une information structurée que l'indexation peut exploiter.

### 9.2.2 Les EN comme matériau d'indexation

J.-P. Moreux explique que les EN extraites par la REN peuvent constituer le matériau d'une indexation automatique. En ce sens, il soutient que la REN est un « enrichissement sémantique » qui favorise la recherche d'information.<sup>7</sup> Il ajoute que l'alignement des EN avec des référentiels d'autorités et/ou des bases de connaissances constitue un « intérêt majeur pour la recherche d'information et l'accessibilité à l'information, en particulier dans la dynamique actuelle du web de données et de ses promesses d'intéropérabilité. »<sup>8</sup> Grâce à ces ressources externes aux document, des corpus hétérogènes trouveraient « un espace de convergence et de correspondance. »<sup>9</sup> LECTAUREP pourrait participer à cet effort de mise en commun des données avec l'alignement des EN qui peuvent être trouvés dans d'autres corpus, à savoir les adresses, les types d'actes, les mots matières, et les noms de personne quand ceux-ci renvoient à un notaire ayant exercé.<sup>10</sup> En effet, les noms des enregistrements présents dans la cinquième colonne, à part pour quelques rares individus, n'ont pas de notices d'autorité. En ce qui concerne les adresses, il serait possible d'utiliser le référentiels des AN pour les noms de villes et pour les rues de Paris. Pour les types d'actes, il existe également un référentiel, aujourd'hui incomplet. Le référentiel créé et présenté lors de l'exploration de la REN (voir section 7.2) servirait d'une première base pour le compléter. Cela permettrait des lier les répertoires des notaires avec d'autres documents des AN, grâce à l'espace de convergence des référentiels.

6. *Ibid.*, pp. 115 - 116 et p. 118 Notamment, page 118, J.-P. Moreux traite de la reconstruction de la structure logique des documents tabulés en ces mots : « Cette catégorie recouvre les contenus représentés sous forme de tableaux. On pensera en premier lieu aux registres publics (naissance, décès, etc.), mais aussi aux livres de comptes, aux relevés scientifiques, aux cotations boursières, etc. Ici, la nature même de l'information portée sur les documents incline à la retranscrire sous la forme d'une base de données. Il faudra donc dans un premier temps détecter les tableaux présents dans le document, puis ses colonnes porteuses de types de données, et enfin scinder les contenus en lignes. Là aussi, différentes stratégies ont pu voir le jour, depuis la définition manuelle de grilles et modèles [...] appliqués ensuite automatiquement, jusqu'à l'apprentissage profond. »

7. *Ibid.*, p. 122

8. *Ibid.*, p. 123

9. *Ibid.*

10. M.F. Limon-Bonnet et G. Piraino, « Préparer l'innovation : l'informatisation des ressources du minutier central »..., pp. 261 - 262

### **9.2.3 Application de l'indexation automatique à partir des entités nommées pour les répertoires des notaires**

Dans le cas de LECTAUREP, après la campagne de REN, chaque EN pourrait être extraite des encodages TEI produits pour être regroupées par classe, fournissant ainsi un premier niveau d'indexation. Par exemple, regrouper toutes les EN d'une classe « mot matière », dans le cas où un modèle aurait été entraîné pour les extraire, permettrait d'accéder à chaque page contenant ce type d'EN. Nous pourrions également imaginer une indexation de toutes les EN extraites, donc potentiellement bruitées. Chaque EN serait associée dans l'index au fichier texte et à l'image dans lesquels elle se trouve, de sorte à être facilement consultable. Cela pourrait être mis en place à partir des métadonnées présentes dans l'encodage TEI.

Cependant, nous pourrions poser la question de la pertinence de l'indexation systématique de termes bruités. Peut-être serait-il souhaitable de créer un système de contrôle n'indexant que des EN qu'il est possible de lier à des bases de données, par exemple les noms d'adresses et les types d'acte, avec une distance de Levenshtein. Cela pose tout de même un problème que nous avons déjà soulevé : les EN ne pouvant être rattachées à ces référentiels externes seraient oubliées. Par ailleurs, l'idée d'un tel système repose sur l'idée que l'index serait exploitable par une machine, c'est-à-dire, pouvoir notamment faire de l'EL avec les EN indexées. Or, un index serait tout aussi adapté à la lecture humaine. Si le terme « procuraloon » est indexé, un-e utilisateur-ice pourrait comprendre que ce terme correspond très certainement à « procuration », d'autant plus que la distance de Levenshtein est ici de 2, donc faible. Ce terme bruité ne fait de plus pas partie du dictionnaire français, induisant donc qu'il y a une erreur de transcription. Cela indique donc qu'il faut « réparer » le mot. Des erreurs de mots réels pourraient poser plus de problèmes, mais à nouveau, il est possible de faire confiance à l'humain dans sa capacité à comprendre qu'en consultant une entrée « dotation », par exemple, cela puisse aboutir sur un document contenant « donation », d'autant plus que les utilisateurs sauraient qu'ils sont dans le contexte de la consultation d'un index des répertoires des notaires, donc sous-entendant un lexique spécifique. L'indexation de termes bruités entraînera toutefois la multiplication de formes d'un seul mot, générant ainsi du bruit dans l'index.

Il est peut-être préférable de générer du bruit, plutôt que de générer du silence. Dans le premier cas, le phénomène de multiplication des termes empêcherait de normaliser le vocabulaire d'indexation, tandis que dans le second, si les termes indexées sont liés à un référentiel, il serait envisageable de les normaliser avec celle-ci.<sup>11</sup> Cependant, dans le cas de termes issus

---

11. Le guide d'indexation pour le web du service interministériel des Archives de France indique à ce sujet que la normalisation du vocabulaire permet d'effectuer une indexation contrôlée, propre à l'intéropérabilité des données. Pierre-Frédéric Brau, Sylvie Boudaud, Alix Charpentier, Pauline Charbonnier, F. Clavaud, Louis Vignaud, Gaël Chenard, Céline Cros, Sylviane Follet-Clavreul, Alex-Adriana Grimont, et al., *Guide d'indexation pour le web*, 2021, URL : <https://francearchives.fr/file/6686af73e52bd3dd7d56cbad92228977cb>

de la REM, la normalisation des données est une tâche compliquée, et l'intéropérabilité des données produites par la REM est encore une difficulté à résoudre. L'imperfection des index automatiquement créés pourrait être considérée comme un frein. Pour J.-P. Moreux, cela n'en est pas tant que les acteurs qui mettront en place ces systèmes seront pragmatiques dans les choix des outils qui doivent être choisis en fonction des corpus, et transparents sur les performances des modèles utilisés.<sup>12</sup> De plus, il soutient que l'indexation automatique permet de traiter des corpus d'une ampleur bien plus importante que ceux qui peuvent être indexés à la main, ce qui est le cas pour le corpus de LECTAUREP.

Nous supposons par conséquent que l'indexation reposera sur les performances qu'atteindront les modèles de REM et la qualité des données qu'il sera possible de retrouver lorsqu'une chaîne de traitement post REM aura été mise en place. Enfin, un index des EN pourrait être un deuxième moyen d'accès aux transcriptions automatiques, en plus de leur consultation avec TEI Publisher, dont le moteur de recherche, notamment à facettes, constitue une première couche d'indexation.

Enfin, l'indexation automatique des répertoires des notaires permettrait de compléter dans la SIV les instruments de recherche de chacun de ces documents. Par exemple, pour le notaire Ernest Legay, ayant exercé entre le 24 février 1875 et le 14 mai 1902, nous trouvons aujourd'hui dans la SIV, dans l'instrument de recherche EAD FRAN\_IR\_041698 seulement un niveau de description archivistique pour le répertoire daté du 25 juillet 1893 au 23 août 1899, correspondant à « consentement à mariage ».<sup>13</sup> À l'aide des EN extraites depuis ce répertoire, il serait possible de peupler les niveaux de description archivistique concernant les minutes et les brevets enregistrés dans ce document, tout en établissant un lien entre les éléments créés et des référentiels.<sup>14</sup>

---

e576f5/GuideIndexation\_Web\_v202108.pdf, p. 23

12. É. Cavalié (dir.), *L'indexation matière en transition...*, p. 130

13. Voir [https://www.siv.archives-nationales.culture.gouv.fr/siv/rechercheconsultation/consultation/ir/consultationIR.action?irId=FRAN\\_IR\\_041698&udId=c1p6wqgn6wj7--1wafp63a894hv&details=true&gotoArchivesNums=false&auSeinIR=true](https://www.siv.archives-nationales.culture.gouv.fr/siv/rechercheconsultation/consultation/ir/consultationIR.action?irId=FRAN_IR_041698&udId=c1p6wqgn6wj7--1wafp63a894hv&details=true&gotoArchivesNums=false&auSeinIR=true) (consulté le 31/08/21).

14. Par exemple, pour la description archivistique d'une procuration, enregistrée dans un répertoire du notaire Louis Bronod daté de 1751 à 1760, on trouve dans l'instrument de recherche le type d'acte, ainsi que le nom des parties concernées. Voir [https://www.siv.archives-nationales.culture.gouv.fr/siv/rechercheconsultation/consultation/ir/consultationIR.action?irId=FRAN\\_IR\\_042895&udId=c1p72qqzs5ej-djio68oc4i3k&details=true&gotoArchivesNums=false&auSeinIR=true](https://www.siv.archives-nationales.culture.gouv.fr/siv/rechercheconsultation/consultation/ir/consultationIR.action?irId=FRAN_IR_042895&udId=c1p72qqzs5ej-djio68oc4i3k&details=true&gotoArchivesNums=false&auSeinIR=true) (consulté le 31/08/21). Les minutes et brevets non décrits dans les instruments de recherche pourraient l'être de la même façon que ceux présentés grâce à l'indexation automatique réalisé avec les EN.

## 9.2.4 Présentation d'autres méthodes d'indexation automatique applicable aux transcriptions des répertoires des notaires

J.-P. Moreux a discuté des autres méthodes d'indexation automatique dont peuvent se doter les établissements culturels, à savoir :

- L'indexation probabiliste, que nous avons présenté à travers la technologie du KWS dans la première partie de ce mémoire (voir section 3.2).<sup>15</sup>
- La pré-indexation lexicale, notamment mise en application avec la lemmatisation qui permet d'associer aux mots leurs forme canonique, ou lemme, dans « un texte sujets à flexion (substantif, adjectif, verbe) ». Elle permet la recherche par forme lexicale, et augmente le nombre de résultats d'une recherche.<sup>16</sup>
- L'extraction de terminologie, concepts et sujets, qui deviennent le matériau d'indexation. Ce type d'indexation est plus difficile à imaginer pour les répertoires des notaires, étant donné qu'ils n'ont pas de « sujet » à proprement parler, ni de concepts : ce ne sont ni des textes scientifiques, ni des textes littéraires.<sup>17</sup>
- Le classement automatique qui permet d'« attribuer une catégorie (une classe) à une collection d'objets à catégoriser ».<sup>18</sup> Par exemple, pour une collection de texte, il est possible d'envisager un classement par langue, par genre ou par mouvement littéraire. Dans le cas des répertoires des notaires, l'hétérogénéité rend difficile le classement automatique par catégorie. Les notaires ne rassemblaient en effet pas les enregistrements par types d'actes en listant toutes les procurations, puis tous les contrats de mariage, « etc. », passés dans une journée. Ce n'est par ailleurs pas le but de ces documents. De plus, il est pertinent de concevoir les répertoires des notaires comme une catégorie à part entière. Le classement automatique concernerait donc un niveau d'indexation plus vaste.

Toutes ces méthodes produisent un matériau permettant d'indexer les textes sur lesquels elles ont été appliquées.

---

15. *Ibid.*, pp. 121 - 122

16. *Ibid.*, p. 122 Voir également D. Stutzmann, J.F. Moufflet et S. Hamel, « Full Text Search in Medieval Manuscripts... ». Stutzmann *et al.* expliquent avoir utilisé ce mode d'indexation dans le cadre du projet HIMANIS.

17. É. Cavalié (dir.), *L'indexation matière en transition...*, p. 123

18. *Ibid.*, p. 124

## 9.3 L'indexation automatique dans le secteur culturel

### 9.3.1 Quels impacts métiers pour l'indexation automatique ?

Indexer grâce aux EN s'inscrit dans la continuité des pratiques d'indexation existantes, notamment par les bibliothèques avec la numérisation de leurs documents et leurs transcriptions automatiques, permettant la recherche plein texte.<sup>19</sup> J.-P. Moreux voit dans ce phénomène l'« automatisation (partielle) d'une des fonctions du bibliothécaire (orienter, conseiller), »<sup>20</sup> qui a entraîné la « reconfiguration » du rôle d'intermédiation de ces établissements. Par conséquent, l'indexation automatique questionne aujourd'hui plusieurs tâches des métiers du secteur culturel. Parmi elles le catalogage, qui pourrait être automatisé avec la création de notices à partir des EN pour chaque œuvre entrant dans une bibliothèque de façon numérique, par exemple à partir d'une numérisation. J.-P. Moreux mentionne aussi la modification des systèmes d'information des bibliothèques par la mise à plat qu'entraîne l'automatisation. C'est-à-dire la ré-indexation de toutes les données déjà existantes dans le but d'uniformiser les métadonnées produites dans le cadre d'un versement dans des portails en ligne, selon les principes du web de données.<sup>21</sup>

Concernant le secteur archivistique, Jean-François Moufflet et Gaetano Piraino expliquent que le rôle de médiation des archivistes restera toujours aussi important à l'heure du numérique, en particulier pour « orienter correctement la recherche du lecteur dans des sources dont la logique de production est bien loin de la sienne. »<sup>22</sup> En outre, ils affirment que les tâches fondamentales propres à ce métier, comme la description, la qualification, le classement et la diffusion des objets des archives, ne sont pas remises en question par l'avancée des technologies, et qu'elles rappellent « au contraire le haut niveau d'exigence » qui leur sont inhérentes.<sup>23</sup> Le web de données et les recherches à facettes mettent en avant l'indexation tout en remettant en question les recherches en plein texte. Pour indexer, « les ressources doivent être qualifiées de manière concise, que ce soit pour pouvoir être recherchées efficacement dans une masse de données qui ne cesse de croître, comme pour être reliées à des ressources parentes » et ce sera aux archivistes que ce travail de description incombera, notamment dans le contrôle des index, et dans la création de notices d'autorités auxquels les index pourront se référer.<sup>24</sup>

Dans le cas de LECTAUREP, l'indexation automatique viendrait théoriquement remplacer une indexation manuelle coûteuse en termes de temps et de moyens.

---

19. *Ibid.*, p. 129

20. *Ibid.*

21. *Ibid.*, pp. 129 - 130

22. J.F. Moufflet et G. Piraino, « Au cœur du document d'archives : le projet Himanis », *La Gazette des archives*, 2 (numéro 254[ 2019]), p. 267-281, p. 280

23. *Ibid.*, p. 279

24. *Ibid.*, p. 280

### **9.3.2 Un guide du service interministériel des Archives de France pour indexer**

Un guide d'indexation pour le web a été publié par le service interministériel des Archives de France au mois de juin 2021, dans le but de se détacher de l'hétérogénéité des métadonnées archivistiques d'indexation et pour présenter une méthodologie d'indexation dans le contexte du web de données et de l'intéropérabilité.<sup>25</sup> Ce guide vient confirmer que l'indexation plein texte est aujourd'hui insuffisante pour l'accès à l'information et que le bruit qu'elle génère peut être évité grâce à une indexation normalisée et uniformisée.<sup>26</sup>

Le guide s'attache également à décrire le processus de transition vers le nouveau standard de description des archives *Records in Contexts* (RiC), dont la portée dépasse le présent mémoire. Cependant, soutenons ici qu'il est critique que l'indexation automatique, notamment réalisée à partir de la REN, trouve toute sa place dans ce mouvement d'uniformisation des pratiques, afin de continuer à traiter des corpus de plus en plus conséquents et permettre à l'ensemble des usagers, professionnels et publics, de mieux naviguer, mieux comprendre, et mieux exploiter les documents traités par le secteur culturel.

---

25. P.F. Brau, S. Boudaud, A. Charpentier, *et al.*, *Guide d'indexation pour le web...*

26. *Ibid.*, p. 12

# **Conclusion**



En conclusion, nous avons montré que pour lancer une campagne de REN, le projet LECTAUREP devra préalablement arrêter une chaîne de traitement des transcriptions automatiques. Une fois celle-ci établie, des décisions scientifiques seront à prendre concernant les EN, notamment pour déterminer ce qui constitue une EN dans les répertoires des notaires et quelles sont les classes qui leurs seront attribuées.

L'arrivée d'outils d'apprentissage supervisé dans le secteur culturel implique la présence de professionnels le faisant vivre. Comme nous l'avons expliqué dans la première partie, les modèles de REM sont produits à partir de données d'entraînement, qui sont des transcriptions réalisées dans le cadre de LECTAUREP par les agents des AN. La transcription nécessite des savoirs paléographiques et historiques pour être menée à bien. Des transcriptions erronées pourraient induire des biais dans les modèles, par exemple mal transcrire certaines lettres, certains signes, etc. Nous soutenons qu'il en va de même pour les systèmes de REN appliqués aux documents patrimoniaux. La première partie s'est attachée à décrire la nature des informations contenues dans les répertoires des notaires selon l'idée que sans en avoir connaissance, il n'est pas possible d'établir des objectifs pour la REN. Cela aurait notamment pour risque de cantonner les campagnes de REN aux classes d'EN génériques, « PER », « LOC », « ORG » et « MISC ». Celles-ci ont certes du sens et représentent des éléments de compréhension des textes, mais les technologies actuelles permettent bien plus, notamment dans la création de nouvelles classes.<sup>27</sup> De plus, nous avons également présenté les données numériques créées à partir de ces documents sources. La REM transforme les données textuelles présentes dans les documents avec le bruit qu'elle produit, qu'il est possible de quantifier grâce à des outils de contrôle. Bien que nous puissions choisir les meilleurs modèles de transcription, il est critique de connaître la nature des données produites par les modèles afin de préparer leur exploitation. C'est en connaissant la nature du document original et en la comprenant qu'il est possible d'envisager des opérations de traitement pour le préparer à une exploitation automatique. Dans le cas de LECTAUREP, par exemple, cela concerne la segmentation des mots, mais aussi la normalisation des abréviations et la correction post REM. Autant d'étapes qui doivent être maîtrisées pour l'exploitation des données produites. Plus que jamais les acteurs et les experts des secteurs patrimoniaux ont leur place dans la conception et l'entraînement de modèles automatiques, notamment en apportant leur connaissances sur les documents et les objectifs d'extraction.

En outre, la réussite de l'extraction des EN est une étape et non une fin. Il est en effet nécessaire de penser à des systèmes pour pouvoir les exploiter. C'est ce que nous avons essayé de montrer en présentant TEI Publisher et la possibilité d'indexer les EN dans la troisième partie. Dans ce sens, les EN n'auront d'utilité que si un système pérenne est disponible pour les stocker, les signaler, et puisse servir de socle à leur exploitation. Le standard TEI s'y prête

---

27. Mentionnons à nouveau les 27 classes de GROBID-NER et celles du corpus « Presto ».

très bien, permettant de stocker les EN dans un fichier XML, de les signaler grâce au balisage, et de les récupérer efficacement pour les utiliser.

Enfin, du point de vue budgétaire et de l'accessibilité des données, le secteur culturel a tout intérêt à miser sur la REN. En effet, contrairement aux méthodes de KWS, la REN est bien documentée et fonctionne le plus souvent avec des méthodes *open-source*. La REN appliquée à des données bruitées nécessite cependant d'accepter l'imperfection, qui est le prix à payer pour traiter en masse. Nous pouvons arguer que Les exceptions et les anomalies ne sont pas un problème, et peuvent faire l'objet de traitements particuliers. À ce sujet, nous avons suggéré dans la troisième partie que celles causées par la REM et la REN pourraient être corrigées manuellement et de façon collaborative, notamment grâce à TEI Publisher. G. Mühlberger expliquait à ce sujet que les traitements automatiques devaient être vus comme une opportunité pour proposer un meilleur accès aux documents patrimoniaux. Il ajoutait également que la REM n'est pas une fin en soi, et qu'il est fort probable que des documents qui ont déjà été automatiquement transcrits le soient à nouveau lorsque de nouvelles architectures plus performantes verront le jour. Cette remarque peut tout à fait être appliquée aux modèles de REN.

"It will be a learning process for archivists to accept that even not fully accurate text will make access to their resources easier and that one should not focus upon the errors, but see every correct word as an opportunity to provide better access. Also it should become clear that the process of HTR is not done once and forever, but that it is repeated several times as soon as there is a reasonable chance of improving the results by applying new methods or by machine learning effects."<sup>28</sup>

Trouver une solution généralisable pour entreprendre une campagne de REN sur des documents patrimoniaux est une tâche encore difficile à entreprendre. K. Adnan et R. Akbar notait en 2019 que la REN devait encore faire face à de nombreux facteurs influençant la façon dont les modèles étaient produits, et ce à plusieurs niveaux. Par exemple, concernant les données, le taux de bruit impacte la qualité des performances des modèles de REN. Au niveau des entités, la désambiguïsation peut être difficile à effectuer, et certaines entités peuvent également être spécifiques à tel ou tel domaine, limitant donc la création de modèles applicables à des données de domaines différents.<sup>29</sup>

28. Günter Mühlberger, « Preprint : Handwritten Text Recognition (HTR) of Historical Documents as a Shared Task for Archivists, Computer Scientists and Humanities Scholars. The Model of a Transcription & Recognition Platform (TRP) » (, 2019), URL : [https://www.academia.edu/8601748/Preprint\\_Handwritten\\_Text\\_Recognition\\_HTR\\_of\\_Historical\\_Documents\\_as\\_a\\_Shared\\_Task\\_for\\_Archivists\\_Computer\\_Scientists\\_and\\_Humanities\\_Scholars\\_The\\_Model\\_of\\_a\\_Transcription\\_and\\_Recognition\\_Platform\\_TRP\\_](https://www.academia.edu/8601748/Preprint_Handwritten_Text_Recognition_HTR_of_Historical_Documents_as_a_Shared_Task_for_Archivists_Computer_Scientists_and_Humanities_Scholars_The_Model_of_a_Transcription_and_Recognition_Platform_TRP_) (visité le 21/07/2021), p. 5

29. K. Adnan et R. Akbar, « Limitations of information extraction methods and techniques for heterogeneous unstructured big data »..., p. 6

Comme nous l'avons vu, les modèles génératifs sont dépendants des domaines sur lesquels ils ont été entraînés. Ils peuvent servir au mieux de preuve de concept ou pour observer si leurs performances laisseraient envisager de les affiner sur les données testées. Ce dernier cas de figure peut arriver, comme nous l'avons vu dans la seconde partie en appliquant la chaîne de traitement de SpaCy à un texte structuré de façon plus « traditionnelle », ou en tout cas se rapprochant d'une structure fréquemment rencontrée en littérature : des phrases verbales commençant par une majuscule et se terminant par un point. Néanmoins, il semble qu'il restera toujours une marge d'amélioration des performances qu'il est possible d'obtenir en affinant un modèle.

Une difficulté supplémentaire est ajoutée dès lors que l'on cherche à appliquer des modèles des REN sur des données issues de la REM. La création de corpus annotés en EN à partir de données historiques est cependant un pas vers la création de plus en plus rapide de modèles de classification, et notamment de modèles affinés sur des données de domaines proches.<sup>30</sup> L'affinage d'un modèle de REN sur un corpus annoté en EN constitué à partir des données du projet LECTAUREP pourrait grandement profiter d'un premier modèle préalablement entraîné sur des données d'un domaine similaire, voire idéalement du même domaine. Le dernier obstacle serait le bruit généré par la REM. Obstacle qui, comme nous avons essayé de le montrer, n'est pas une fatalité.

Pour répondre à la question de la généralisation, nous soutenons qu'il est possible d'avoir des méthodologies communes d'analyse de l'information présente dans les documents sources, pour établir une stratégie de REN, et de trouver des solutions de normalisation généralisables pour préparer les données en vue de la REN. De plus, comme le disait M. Ehrmann *et al.*, les campagnes de REN vont très probablement être amenées à se multiplier durant les prochaines années grâce à la numérisation massive de documents patrimoniaux, et notamment de leur transcription, automatique ou non.<sup>31</sup> Les communautés du secteur culturel profiteront ainsi de l'expérience qui s'apprête à être acquise. De plus, nous insistons sur l'importance de la reconstitution de la structure logique des documents après une transcription automatique, qui permet de servir de socle à toute opération d'extraction d'information.<sup>32</sup>

Ce mémoire n'a pas été exhaustif dans son traitement des architectures de REN. Cela n'était en effet pas le but, nous avons préféré illustrer les outils qui ont pu être découverts et expérimentés durant le stage. Il existe de nombreuses autres architectures de REN que celles qui ont été présentées. Citons par exemple un article de Lilia Simeonova *et al.* documentant un système de REN qui s'appuie sur les informations morpho-syntactiques obtenues

---

30. Mentionnons ici le corpus CLEF HIPE, ainsi que « Presto ».

31. "NE processing has been called upon to contribute to the field of Digital Humanities (DH), where massive digitization of historical documents is producing huge amounts of texts [...]" « Extended Overview of CLEF HIPE 2020... », p. 2

32. É. Cavalié (dir.), *L'indexation matière en transition...*, pp. 115 - 116 et p. 118

avec les outils de TAL, plongements lexicaux, étiquetage morpho-syntaxiques et informations morphologiques, contenues dans le texte pour détecter et extraire les EN.<sup>33</sup> Nous pouvons également mentionner l'article de Manuel Carbonell *et al.* portant sur une architecture qui permet d'effectuer la REM et la REN conjointement, afin de minimiser l'impact des décisions de transcriptions prises lors cette première tâche sur la deuxième.<sup>34</sup> Les communautés scientifiques ont, dans ce sens, tout à gagner à mettre en relation les projets de REN et les expériences menées avec ces différents modèles afin de concevoir un ensemble de solutions optimales.

Concernant l'application de la REN elle-même, nous avons montré dans la seconde partie qu'elle ne peut pas être appliquée directement en sortie de REM. Il est nécessaire de concevoir une chaîne de post traitement des données de la REM, qui influencera la constitution d'un corpus d'entraînement annoté en EN. Il est important pour ne pas perdre le modèle que les données utilisées pour l'entraînement d'un modèle de REN ne soient pas différentes des données sur lesquels il sera appliqué. Une fois un modèle obtenu, alors LECTAUREP pourra lancer une campagne de REN. Les expériences réalisées durant le stage ont permis de dresser une première feuille de route, reproduite dans la figure 9.1. Une première étape rassemblera ainsi le processus de REM, l'export du résultat et de la reconstitution de la structure logique des répertoires des notaires. Deux chemins se présentent ensuite. Le premier consiste à profiter de la REN pour cibler la correction post REM, en évitant par exemple de corriger les noms de personnes, ce qui pourrait être une tâche compliquée à accomplir sans créer de fausses informations. Le second propose de corriger tout le texte issu de la REM. Ceux-ci ont en commun de profiter de la reconstitution de la structure logique et de la transformation du PAGE XML en TEI, en intervenant directement dans la balise <text> pour annoter les EN et corriger le texte, là où le <sourceDoc> permet de conserver la transcription automatique brute. Ils aboutissent également dans la publication des répertoires enrichis et corrigés sur TEI Publisher. Ainsi, la modélisation TEI présentée dans la troisième partie permet d'être transparent sur les modifications qui ont été faites sur les données.

LECTAUREP pourrait également envisager d'utiliser un système de REN hybride, mélangant de l'extraction par règles à l'aide de référentiels dans les colonnes 1, 2, 3, 4, 6 et 7. L'extraction d'information dans la colonne 5, étant donné que l'information est retranscrite sous forme de phrases, pourrait profiter d'un modèle de REN affiné sur un corpus annoté en EN constitué à partir des transcriptions automatiques des répertoires. Il reste cependant à ré-

33. Lilia Simeonova, Kiril Simov, Petya Osenova et Preslav Nakov, « A Morpho-Syntactically Informed LSTM-CRF Model for Named Entity Recognition », *arXiv :1908.10261 [cs]* (, 27 août 2019), arXiv : 1908.10261, URL : <http://arxiv.org/abs/1908.10261> (visité le 12/07/2021)

34. Manuel Carbonell, Mauricio Villegas, Alicia Fornés et Josep Lladós, « Joint Recognition of Handwritten Text and Named Entities with a Neural End-to-end Model » (, 16 avr. 2018)

soudre la question du volume de données d'entraînement nécessaire pour obtenir des résultats satisfaisants. En outre, LECTAUREP devra pour l'annotation de données se doter d'un guide d'annotation, similaire à ce qui a été produit pour la REM.<sup>35</sup> La REN pourrait également être l'occasion, grâce à l'EL, d'inclure les répertoires des notaires dans le web de données. Marie-Françoise Limon-Bonnet *et al.*, à ce sujet, remarquaient que le secteur archivistique, dans son développement numérique, doit envisager de « décloisonner les silos » pour « créer davantage de liens entre les ressources », aussi bien à l'intérieur des services d'archives qu'avec d'autres partenaires des secteurs patrimoniaux et universitaires.<sup>36</sup> Ce faisant, les documents traités par les archives n'en seraient que plus accessibles. Cela permettrait également de les publier selon les principes FAIR (*Findable, Accessible, Interoperable, Reusable*). Lier les EN à des référentiels communs contribue en effet, selon Lise Stork *et al.*, à l'annotation sémantique des textes, et constitue une étape du processus de publication FAIR.<sup>37</sup>

Le projet LECTAUREP s'apprête à créer un corpus conséquent de transcriptions de documents historiques propre au corps de métier du notariat. C'est une véritable opportunité pour la recherche en termes d'exploitation et d'études quantitatives. Pour les professionnels du patrimoine, notamment les AN, la consultation des répertoires des notaires serait considérablement améliorée.

Sur le plan professionnel, cette expérience au sein d'ALMAaCH a été très riche. J'ai terminé ma formation de Master en intégrant une équipe spécialisée dans le TAL et en me formant à l'extraction d'information grâce aux technologies de REN, sujet principal de ce mémoire. En outre, j'ai également pris connaissance de l'aspect gestion de projet grâce à plusieurs réunions réalisées avec Mme Alix Chagué et Mme Aurélia Rostaing, principale interlocutrice du projet LECTAUREP aux AN, ainsi qu'en participant à la réunion de bilan de la phase 3 au début du mois de juillet. Celle-ci a réuni les personnels des AN et d'ALMAaCH mobilisés sur le projet LECTAUREP, et j'y ai présenté la chaîne de traitement de reconstruction de la structure logique et la publication des répertoires des notaires sur TEI Publisher. J'ai également pu aider à administrer l'instance d'eScriptorium, déployé sur le serveur d'Inria Traces6, en créant des comptes et en assistant les utilisateurs en cas de problème. Mme Alix Chagué m'a également formé à l'utilisation de Kraken. Enfin, j'intégrerai ALMAaCH à compter du 1er octobre

---

35. A. Chagué et A. Rostaing, « Présentation du projet Lectaurep (Lecture automatique de répertoires) », dans *Atelier sur la transcription des écritures manuscrites - BnF DataLab*, Paris, France, 2021, URL : <https://hal.archives-ouvertes.fr/hal-03122019> (visité le 06/04/2021)

36. M.F. Limon-Bonnet, J.F. Moufflet et G. Piraino, « L'innovation numérique : un cercle vertueux pour l'archivistique », *La Gazette des archives*, 2 (numéro 254[ 2019]), p. 247-252, pp. 250 - 251

37. Lise Stork, Andreas Weber, Fons Verbeek et Katherine Wolstencroft, « From Historical Handwritten Manuscripts to Linked Data », *Digital Libraries for Open Knowledge - 22nd International Conference on Theory and Practice of Digital Libraries, TPDL 2018, Proceedings*, 11057 (2018), p. 5

2021 en tant qu'ingénieur recherche et développement dans le cadre d'un projet en partenariat avec l'Institut National d'Histoire de l'Art et la Bibliothèque Nationale de France. L'objectif de celui-ci est de travailler à des méthodes d'analyse automatique de contenus numérisés à partir des catalogues de vente de monnaies, pour pouvoir les structurer en travaillant avec la suite logicielle GROBID.

## 1.



## 2.

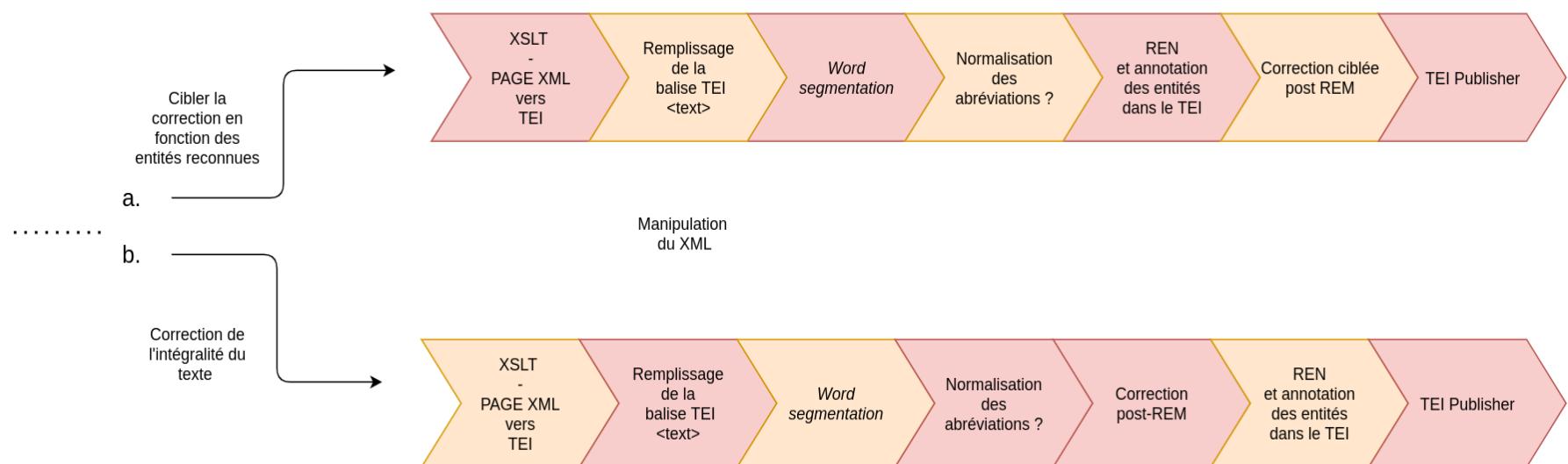


FIGURE 9.1 – Chaîne de traitement pour la REN dans le cadre du projet LECTAUREP



## **Annexes**



## **Annexe A**

### **Chercher un acte dans les réertoires de notaires dans la SIV**

The screenshot shows the SIV search interface. At the top, there are navigation links: "AIDE À LA RECHERCHE", "RECHERCHE AVANCÉE", "ARCHIVES NATIONALES" (with the subtitle "Salle des inventaires virtuelle"), "PARCOURIR LES FONDS", "PRODUCTEURS D'ARCHIVES", and a search icon.

The main search bar contains the text "Legay Ernest 18 -19". Below it, there are three tabs: "Toutes les archives", "Archives numérisées", and "Producteurs d'archives", with "Producteurs d'archives" being the active tab.

On the left, there is a sidebar titled "AFFINER AVEC PLUS DE CRITÈRES" which includes a date range selector from 1875 to 1935. The date range is set from 1875 to 1935, with the text "De 1875 à 1935" displayed. Below this is a button "Affiner sur cet intervalle".

The main search results area shows the query "Nom - prénom du notaire : Legay, Ernest (18..-19..)". There are three search options: "Tous les mots saisis" (radio button selected), "Au moins un des mots saisis", and "Expression exacte".

Below the query, there are fields for "Numéro d'étude" (with a placeholder box), "Intervalle de dates" (with two empty boxes separated by "à"), and "Date précise" (with an empty box). There is also a "Lieux" field with the link "Afficher les détails".

At the bottom of the search results area, there are buttons for "Effacer" and "Rechercher". Below these are filters "Trier par" (set to "Pertinence") and "Voir : 10, 20, 30, 50, 70", along with a "Affichage" dropdown menu.

The message "2 résultats." is displayed above the result list, which is currently empty.

FIGURE A.1 – Exemple de recherches des actes passés en minute et en brevet dans la SIV pour le notaire Ernest Legay.

The screenshot shows a web page from the 'Archives Nationales' website. At the top, there are navigation links: 'AIDE À LA RECHERCHE', 'RECHERCHE AVANCÉE', 'PARCOURIR LES FONDS', 'PRODUCTEURS D'ARCHIVES', and a search bar. The main title is 'PRODUCTEUR D'ARCHIVES' followed by the name 'Legay, Ernest Marie Joseph (18..-19..)'. On the left, there is a sidebar with a 'Retour à la page précédente' link and a decorative geometric pattern. On the right, there are links for 'Retour aux résultats de recherche', 'Haut de page', 'Télécharger la notice PDF', and 'Permalien'. The central content area displays three inventory entries for 'Ernest Marie Joseph Legay':

- INVENTAIRE**: Minutes et répertoires du notaire Ernest Marie Joseph LEGAY, 14 mai 1902 - 27 novembre 1935 (étude XXIII). Includes a 'Voir l'inventaire' button.
- INVENTAIRE**: Archives de l'office notarial XXIII (1767-1969). Includes a 'Voir l'inventaire' button.
- INVENTAIRE**: Images des répertoires du notaire Ernest Marie Joseph Legay pour l'étude XXIII. Includes a 'Voir l'inventaire' button and a red '★ Ajouter à mes favoris' button.

FIGURE A.2 – Liste des documents produits par le notaire Ernest Legay, inventoriés dans la SIV.

154

AIDE À LA RECHERCHE ▾ RECHERCHE AVANÇÉE

**ARCHIVES  
NATIONALES**  
Salle des inventaires virtuelle

PARCOURIR LES FONDS ▾ PRODUCTEURS D'ARCHIVES

INVENTAIRE ⓘ - Cotes : MC/RE/XXIII/37 - MC/RE/XXIII/48

Images des répertoires du notaire Ernest Marie Joseph Legay pour ...

Retour à la page précédente

Présentation générale Détail du contenu Archives numérisées

42 résultats. Voir : 15 30 45

Cote : 51 r<sup>o</sup>-62 v<sup>o</sup>

Liste chronologique des actes pour la période du 15 mai au 31 décembre 1902

15 mai - 31 décembre 1902

Producteur(s) :

Inventaire : **Images des répertoires du notaire Ernest Marie Joseph Legay**

Ajouter l'inventaire à mes favoris

Télécharger l'inventaire en PDF

Permalink de l'inventaire

Export XML de l'inventaire

Cote : 62 v<sup>o</sup>-79 v<sup>o</sup>

Liste chronologique des actes pour la période du 5 janvier au 31 décembre 1903

5 janvier - 31 décembre 1903

Producteur(s) :

FIGURE A.3 – Liste des numérisations des actes passés en minutes et en brevet dans les répertoires des notaires parisiens du notaire Ernest Legay

## **Annexe B**

### **Structuration de l'information dans la cinquième colonne : exemples**

*Inventaire | au 100, mois de juillet  
Conte (après décès de Joseph) à Paris, rue Rodier 70, du 14 Juin 1903*

{Nom} (après décès de {Prénom})

FIGURE B.1 – Structure syntaxique pour le type d'acte « inventaire ».

*(N° du 31 juillet 1900) Continuation à l'inventaire de Matrodetzki (par suite d'instance en divorce entre Abraham Boroch)  
et Alice Lehmann*

{Nom} (par suite d'instance  
en divorce entre {Prénom})

FIGURE B.2 – Structure syntaxique pour le type d'acte « continuation d'inventaire ».

Dépôt de testament | copie d'acte de vente, à la guayaquil, sur acte daté M<sup>e</sup> Legay, le 20 Juin 1890  
 L'adreï de Lacharrière (de Jules François René), décédé en son  
 domicile à Paris, quai Malaquais 3, le 14 Octobre 1903

{Nom} (de {Prénom})

FIGURE B.3 – Structure syntaxique pour le type d'acte « dépôt de testament ».

Obligation	Sterckeman (par Jules Charles Bertin Lucien) et Lucie Clémie Lebasble, ép <sup>e</sup> , à Paris, B <sup>e</sup> laumes 31 <sup>er</sup> à Odolphe Weil, rue Beaurepaire 27, de 28 000 <sup>f</sup> .
✓ - 2 <sup>e</sup> -	2 <sup>e</sup> - (par les mêmes) à Gabrielle Zoë Flora Henriette Brunelle, V <sup>e</sup> de l'éon Marcellin Albert Petit, à Paris, Faubourg St Denis, 210, de 30 000 <sup>f</sup> .

Idem (par les mêmes)

FIGURE B.4 – Structure syntaxique pour l'indication d'un « idem ».

Procuration - v" - Setellier (par M. J. J. Soumaré - v")  
111 : 0' (par Victor Félix) à Paris, rue Blanche 12, pr<sup>e</sup> vendre immuable

{Nom} (par {Prénom})

Valable également pour :

- Décharge
  - Donation
  - Mainlevée
  - Consentement à
  - Cession de bail
  - Reconnaissance de dette
  - Dépôt de pièces
  - Bail
  - Obligation
  - Réquisition d'acte respectueux
  - Vente

FIGURE B.5 – Structure syntaxique pour le type d'acte « procuration » et autres.

## **Annexe C**

### **Les formats d'export XML d'eScriptorium : ALTO et PAGE**

Listing C.1 – Exemple d'un export ALTO avec eScriptorium d'une transcription automatique d'une page d'un répertoire de notaire

```
<alto xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  ↳ xmlns="http://www.loc.gov/standards/alto/ns-v4#"
  ↳ xsi:schemaLocation="http://www.loc.gov/standards/alto/ns-v4#
  ↳ http://www.loc.gov/standards/alto/v4/alto-4-2.xsd">
<Description>
...
</Description>
<Tags>
<OtherTag ID="BT968" LABEL="Col_1" DESCRIPTION="block type Col_1"/>
<OtherTag ID="BT969" LABEL="Col_2" DESCRIPTION="block type Col_2"/>
<OtherTag ID="BT970" LABEL="Col_3" DESCRIPTION="block type Col_3"/>
<OtherTag ID="BT971" LABEL="Col_4" DESCRIPTION="block type Col_4"/>
<OtherTag ID="BT972" LABEL="Col_5" DESCRIPTION="block type Col_5"/>
<OtherTag ID="BT973" LABEL="Col_6" DESCRIPTION="block type Col_6"/>
<OtherTag ID="BT974" LABEL="Col_7" DESCRIPTION="block type Col_7"/>
<OtherTag ID="BT975" LABEL="Stamp" DESCRIPTION="block type Stamp"/>
<OtherTag ID="BT977" LABEL="header" DESCRIPTION="block type header"/>
<OtherTag ID="BT978" LABEL="marginal" DESCRIPTION="block type marginal"/>
<OtherTag ID="LT353" LABEL="First_line" DESCRIPTION="line type
  ↳ First_line"/>
<OtherTag ID="LT354" LABEL="Main_date" DESCRIPTION="line type
  ↳ Main_date"/>
<OtherTag ID="LT355" LABEL="printed" DESCRIPTION="line type printed"/>
</Tags>
<Layout>
<Page WIDTH="2999" HEIGHT="4420" PHYSICAL_IMG_NR="37"
  ↳ ID="eSc_dummypage_">
<PrintSpace HPOS="0" VPOS="0" WIDTH="2999" HEIGHT="4420">
<TextBlock HPOS="216" VPOS="279" WIDTH="2697" HEIGHT="385"
  ↳ ID="eSc_textblock_2e39cb10" TAGREFS="BT977">
...
</TextBlock>
<TextBlock HPOS="226" VPOS="681" WIDTH="211" HEIGHT="3595"
  ↳ ID="eSc_textblock_c3d77b3d" TAGREFS="BT968">
...

```

```

</TextBlock>
<TextBlock HPOS="430" VPOS="683" WIDTH="128" HEIGHT="3584"
↪ ID="eSc_textblock_08a57fef" TAGREFS="BT969">
...
</TextBlock>
<TextBlock HPOS="534" VPOS="682" WIDTH="359" HEIGHT="3585"
↪ ID="eSc_textblock_1845e379" TAGREFS="BT970">
...
</TextBlock>
<TextBlock HPOS="891" VPOS="680" WIDTH="324" HEIGHT="3589"
↪ ID="eSc_textblock_87f6c099" TAGREFS="BT971">
<Shape>
<Polygon POINTS="1215 4269 1207 680 891 684 897 4229 897 4261 1118
↪ 4267"/>
</Shape>
<TextLine ID="eSc_line_5cbee132" BASELINE="913 1366 1090 1366" HPOS="910"
↪ VPOS="1244" WIDTH="169" HEIGHT="174">
<Shape>
<Polygon POINTS="910 1362 917 1244 983 1244 1035 1296 1079 1296 1079 1418
↪ 917 1407"/>
</Shape>
<String CONTENT=" Prêt " HPOS="910" VPOS="1244" WIDTH="169"
↪ HEIGHT="174"/>
</TextLine>
...

```

Listing C.2 – Exemple d'un export PAGE avec eScriptorium d'une transcription automatique d'une page d'un répertoire de notaire

```
<PcGts
  ↳ xmlns="http://schema.primaresearch.org/PAGE/gts/pagecontent/2019-07-15"
  ↳ xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  ↳ xsi:schemaLocation="http://schema.primaresearch.org/PAGE/gts/pagecontent/2019-07-15
    http://schema.primaresearch.org/PAGE/gts/pagecontent/2019-07-15/pagecontent.xsd">
<Metadata>
  <Creator>escriptorium</Creator>
  <Created>2021-08-29T16:25:22.689956+00:00</Created>
  <LastChange>2021-08-29T16:25:22.689988+00:00</LastChange>
</Metadata>
<Page imageFilename="FRAN_0025_1671_L-0.jpg" imageWidth="2999"
  ↳ imageHeight="4420">
  <TextRegion id="eSc_textblock_2e39cb10" custom="structure
    ↳ {type:header;}">
  ...
  </TextRegion>
  <TextRegion id="eSc_textblock_c3d77b3d" custom="structure {type:Col_1;}">
  ...
  </TextRegion>
  <TextRegion id="eSc_textblock_08a57fef" custom="structure {type:Col_2;}">
  ...
  </TextRegion>
  <TextRegion id="eSc_textblock_1845e379" custom="structure {type:Col_3;}">
    <Coords points="560,686 534,4267 834,4267 893,4267 876,682"/>
    <TextLine id="eSc_line_bd55cc23">
      <Coords points="556,935 582,876 600,887 707,828 784,828 847,891 913,891
        ↳ 921,939 910,1020 884,1023 744,1023 700,979 560,979"/>
      <Baseline points="559,936 923,942"/>
      <TextEquiv>
        <Unicode> C^at de Proprieté </Unicode>
      </TextEquiv>
    </TextLine>
    <TextLine id="eSc_line_811a4ebf">
```

```
<Coords points="548,1152 556,1108 688,1108 740,1064 806,1086 851,1064
↪ 950,1082 950,1171 891,1193 847,1178 792,1193 747,1171 688,1186
↪ 630,1171 556,1193"/>
<Baseline points="552,1156 961,1156"/>
<TextEquiv>
<Unicode> C^at de Propriété </Unicode>
</TextEquiv>
</TextLine>
...

```



## **Annexe D**

**Le Keyword spotting : exemple de mise en oeuvre avec le projet HIMANIS**

# Himanis Chancery PrIx technology offered by *tranSkriptorium*

mensis

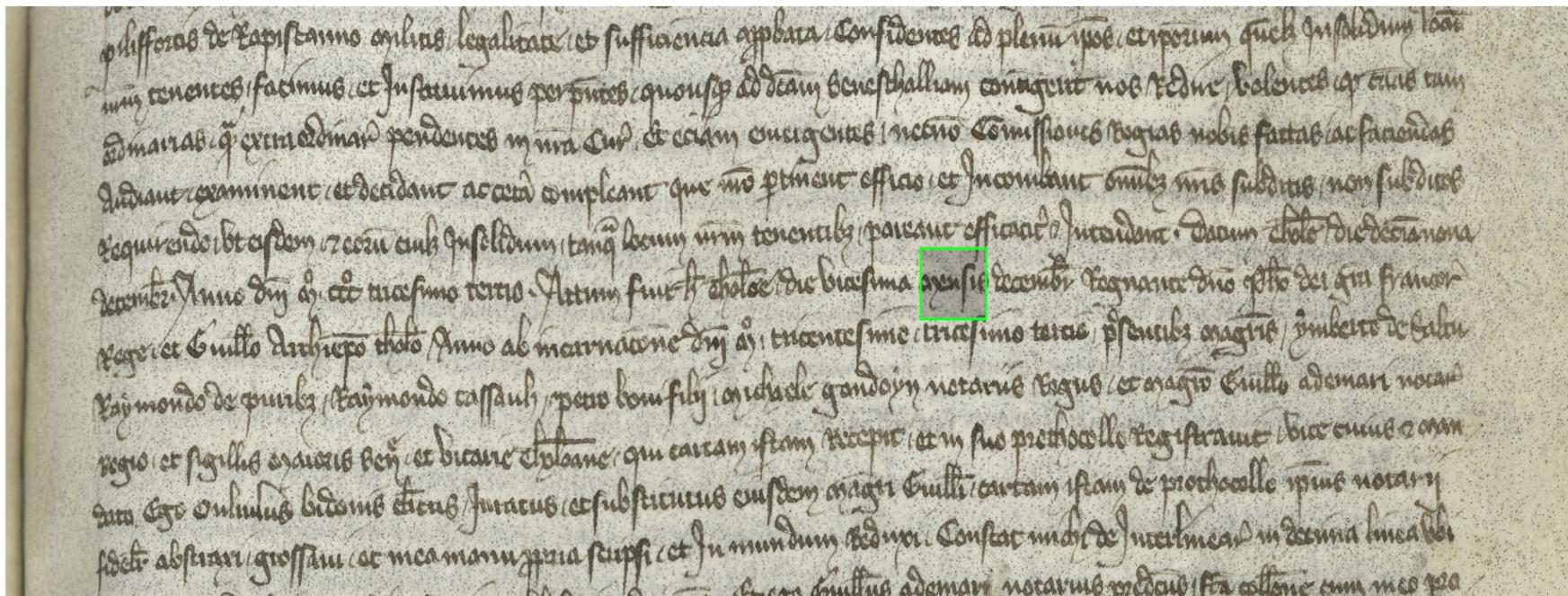
Confidence:   Max. results:

[Help & examples](#)  
[Indexing details](#)

You are here: [HOME](#) » [chancery](#) » [JJ066](#) » page 1127

2 matches found for "mensis" with a confidence of 67.2% !

[← PrevMatch](#) | [← Previous](#) | [Next →](#) | [NextMatch →](#)



© 2021 TS

HIMANIS project --- see also: [IRHT info](#), [additional help](#) and other PRHLT/*tranSkriptorium* PrIx demonstrators

FIGURE D.1 – Exemple d'un résultat de recherche donné par le système de KWS du projet HIMANIS pour le mot « mensis », avec un taux de confiance de 86%.

# Himanis Chancery PrIx technology offered by *tranSkriptorium*

mensis

[Help & examples](#)  
[Indexing details](#)

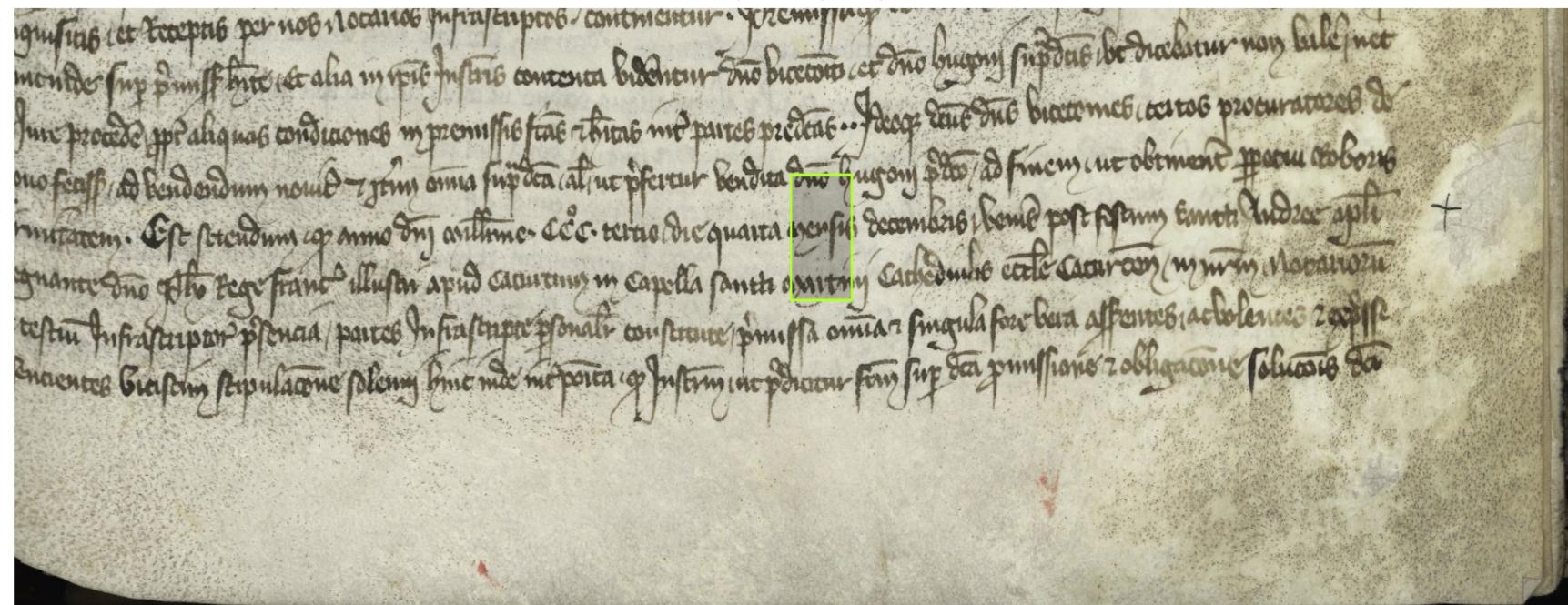
Confidence:

Max. results:

You are here: [HOME](#) » [chancery](#) » [JJ066](#) » page 1127

2 matches found for "mensis" with a confidence of 67.2% !

[- PrevMatch](#) | [- Previous](#) | [Next →](#) | [NextMatch →](#)



© 2021 TS  
HIMANIS project --- see also: [IRHT info](#), [additional help](#) and other [PRHLT/tranSkriptorium Prix demonstrators](#)

FIGURE D.2 – Exemple d'un résultat de recherche donné par le système de KWS du projet HIMANIS pour le mot « mensis », avec un taux de confiance de 63%.



## **Annexe E**

### **Reconstruire la structure logique des des répertoires des notaires après la REM**

An 1901, mois de Février

NATURE ET ESPÈCE		DROITS	
5	DATES	10	DES ACTES :
11	BREVETS	12	Enregistrement.
13	ACTES	14	RELATION
15	REPERTOIRE	16	INDICATIONS, SITUATIONS ET PRIX DES BIENS
17	BREVETS	18	MINUTES
19	DATES	20	DROITS
21	Ploud (de Vieux) à Malakoff (dans) devant des Frères Rougerie et autres, p <sup>r</sup> le 21 Janvier 1901	22	
23	Séjour à charge	24	
25		26	
27		28	
28		29	750
29		30	
31	133	32	
33		34	
34		35	
35		36	750
36		37	
37	131	38	
38		39	
39	Certificat d'appréciation	40	
41		42	
42		43	
43		44	
44	13	45	
45		46	
46	Liquidation	47	
48		49	
49		50	
50		51	
51	135	52	
52		53	
53	Procuration	54	
54		55	
55	13	56	
56		57	
57	138	58	
58	13	59	
59	2°	60	
60		61	
61		62	
62		63	
63		64	
64	13	65	
65		66	
66	Défis de testament	67	
67		68	
68		69	
69		70	
70		71	
71	13	72	
72		73	
73	Stéliquidat <sup>re</sup> (de la Cour de Paris, le 7 Janvier 1901)	74	
74		75	
75			

FIGURE E.1 – Illustration du numérotage des lignes sur eScriptorium sans segmentation des régions, et donc de l'ordre dans lequel les lignes transcrivent apparaissent dans un fichier texte.

Listing E.1 – Export d'un fichier texte résultant de la REM appliquée à une page d'un répertoire de notaire.

1  
2 RRITIO  
3 ATURE ETPOPECE  
4 des torrr  
5 NOMS 'PRETONS ET DOMCIES DES PARTIES  
6 L  
7 RArcoiOtremCNt.  
8 52  
9 (br  
10 5  
11 d  
12 e  
13 C Mree  
14 Cn00  
15 DROTT  
16 I mssois  
17 An 1913 mi de Pevrié  
18 nedu mariage de Charles Segmin et Louise Veturine Guilmart sa  
19 12  
20 St dt` a Paris 80 rue Didot  
21 3.75  
22 349  
23 20  
24 Substitution  
25 Corroy (par Chorles Jean) a Paris 36ape des Gobelins en blanc Paus  
26 21  
27 pouvens alui conferes par Veuve' Brea et Vve Hlerviy  
28 3.75  
29 188  
30 20  
31 Partage  
32 Wiquier (de la collvcation' eablie au profit de laComte d'entre  
33 d  
34 rvmt` a Henri Marcel aprican tissot dans 4vce 4. — Tissot  
35 20

36 65.65  
37 28  
38 281  
39 Cahier de Charges  
40 Boilleau (requete de Achille) et Janne Desombres safe` a Paris 73 rue  
41 Lafapolte pour vente de massons et terrains `a Tssonnes et` a Mnnecy  
42 1  
43 7.50  
44 282  
45 21  
46 Procuration  
47 BorinsetVie (par' Labte)` a Paris 45 Bd Hausmann` a Frederic Allest  
48 22  
49 Hoyan` a Buenos Anespour negir et poursenvie  
50 3.75  
51 283  
52 21  
53 Procuration  
54 Wormis et Cie (par la^meme Ste)` a Jean Andre Postin` a Duntergue  
55 12quai de la Voite 1  
56 22  
57 3.75  
58 284  
59 92  
60 Procuration  
61 Worns etlie (par ladilesS)` a Emile Giermneau dt a Pieppe  
62 1  
63 22  
64 3.75  
65 285  
66 21  
67 Procuration  
68 Chouillet (par Augustine MaiconnerVve de Jeane Paul)` a Paris 19  
69 M  
70 1  
71 rue du Val de Geace en blancepour recueillir 10u  
72 3.75

- 73 286  
74 94  
75 Procuration  
76 d par` Eugene Sabas` a Paris 28 rue des Prtante t l  
77 22  
78 3.75  
79 287  
80 1  
81 21  
82 3  
83 Procuration  
84 Anfdermann (par Raymond) et Leouise Marie Madeline Boulard  
85 22  
86 3.75  
87 sa ft.` a Paris 188 rue de la Convention en blanc pour recueillir 33au  
88 288  
89 24  
90 Inventaire  
91 Yeux (apres le`deces` avne en sondomte` a Paris 32 apte des Gobehus  
92 25  
93 le 20 Janvier 1913 deJeanne Marie Francoire)  
94 7.55  
95 289  
96 94  
97 Proces vecbal  
98 Hemberger (douverture de ligudre des seprises de Mlentine Denise  
99 Josephne de Parada` epouse de` Frederic Honri) dt` a Saint Maur des fosses  
100 25  
101 3.  
102 Bhaue pu  
103 298  
104 91  
105 (Ste du 20`fevier 193` Proces verbal  
106 Boilleau (au profit de divers de manons` etpieces de teure` a  
107 1  
108 559.43  
109 L adjudication Essonnes et Me t enemble 7800+

- 110 291  
111 21  
112 Procuration  
113 Mormis et Cu (par la Ste) sus nommee` a Henrie Doieu` a  
114 Biet  
115 22  
116 3.75  
117 292  
118 Docuration  
119 Boivrilhet (par Georges Marie Elierne)` a Paris 4 ue Berthollel  
120 222  
121 en blanc pour emprunter 6000+  
122 3.75  
123 293  
124 22  
125 Procuration  
126 Pourmer (par Jacques Paul)` a Paris 39 rue de d Echigner` a  
127 Man Pares 2lune de l'Arcade  
128 22  
129 3.75  
130 Woill (par Elie dit Emile)` a Pars 19 rue d'Amsterdam` a Henee  
131 294  
132 22  
133 Tonation  
134 Josephinee` Lattes sofe. dt avec lui, unvert en toute pppe  
135 25  
136 22  
137 Donation  
138 Weil (par lasite do oute` pte  
139 396  
140 22  
141 Mainlevee  
142 Pelsncheld (par Henri)` a Paris 18 pass des Petites Teuries de  
143 Mantrsst ep Maunce Ribonnet` a Montronge 10 grande vue  
144 297  
145 Procuration  
146 Leysens (par Victos)` a Paris 129 ape Paumentier et' Armee) Vve de

- 147 Dinee Piron`a Paris 89 Bd Voltaire en blanc pour vendre
- 148 24
- 149 898
- 150 22 1. duNofevrie 193) Rotification Rintzler`a Marie` Adele Wartman et 1a a Pierre) dt`a
- 151 Phopital de lasalpetriere du ese de Berthe Marie Rinteler`a Paris
- 152 994
- 153 Gratis
- 154 29 oine Chateau Laudon avec Arlhine Dubois
- 155 899
- 156 22
- 157 Vrstament
- 158 Bannen (de Fainy`Henier Vve de` Eugene Auguste)`decedeon
- 159 L dle`a Paris 4ruedt Maur le 13 Jaiu1913
- 160 1
- 161 9.38
- 162 308
- 163 24
- 164 Notorietè
- 165 Deslandes (apresle`deces`arvine`a Paris 184 faub. St Antoine le
- 166 25`fevrier 1988 deConstant Edouard) dt`a St`Mande M Grande rue et de
- 167 Maie Leontine Degny s° fe`a Paris 2 rue de snres`decedee lu son 30mt
- 168 le 3 n0vbre 1941
- 169 Bdlineau (par Piere Louis)`a Paris 5 be rue de B'asle lotincourt
- 170 Conguante`dempieme Feuillet
- 171 397
- 172 24
- 173 Procuration
- 174 3.
- 175 3.

501	15	Inventaire	Colombet (après décès de Albert Antonin Louis Jules / ép de) 18 Victorine Claire Ansberque de Paris 7 rue St Laurent y arrivé le 3 Février 1927	90	"
-----	----	------------	---	----	---

FIGURE E.2 – Exemple d'un enregistrement présent dans une page d'un répertoire de notaire.

Version texte, selon un découpage horizontal :

501 15 Inventaire Colombet (après décès de Albert Antonin Louis Jules / ép de Victorine Claire Ansberque de Paris 7 rue St Laurent y arrivé le 3 Février 1927

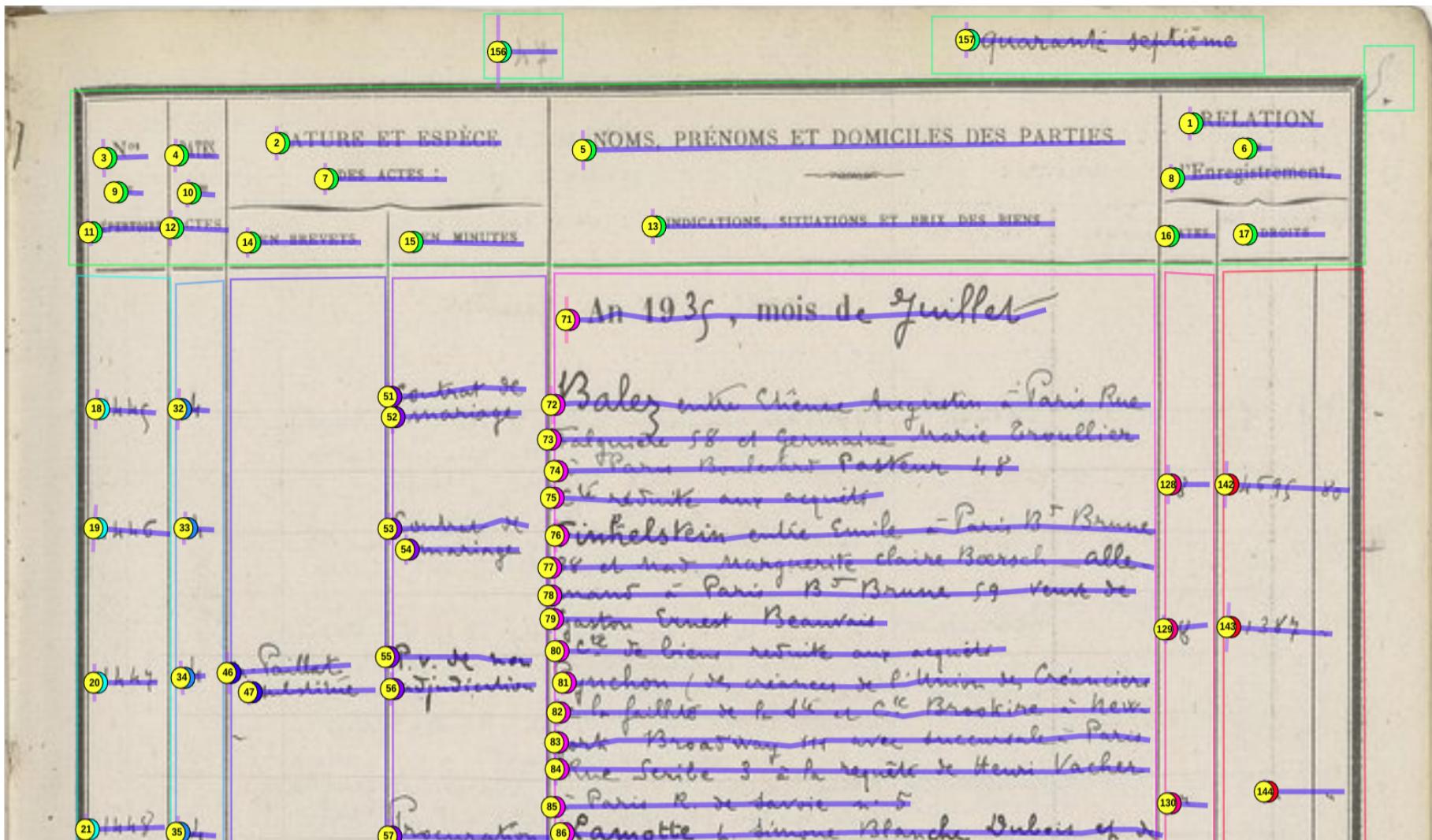
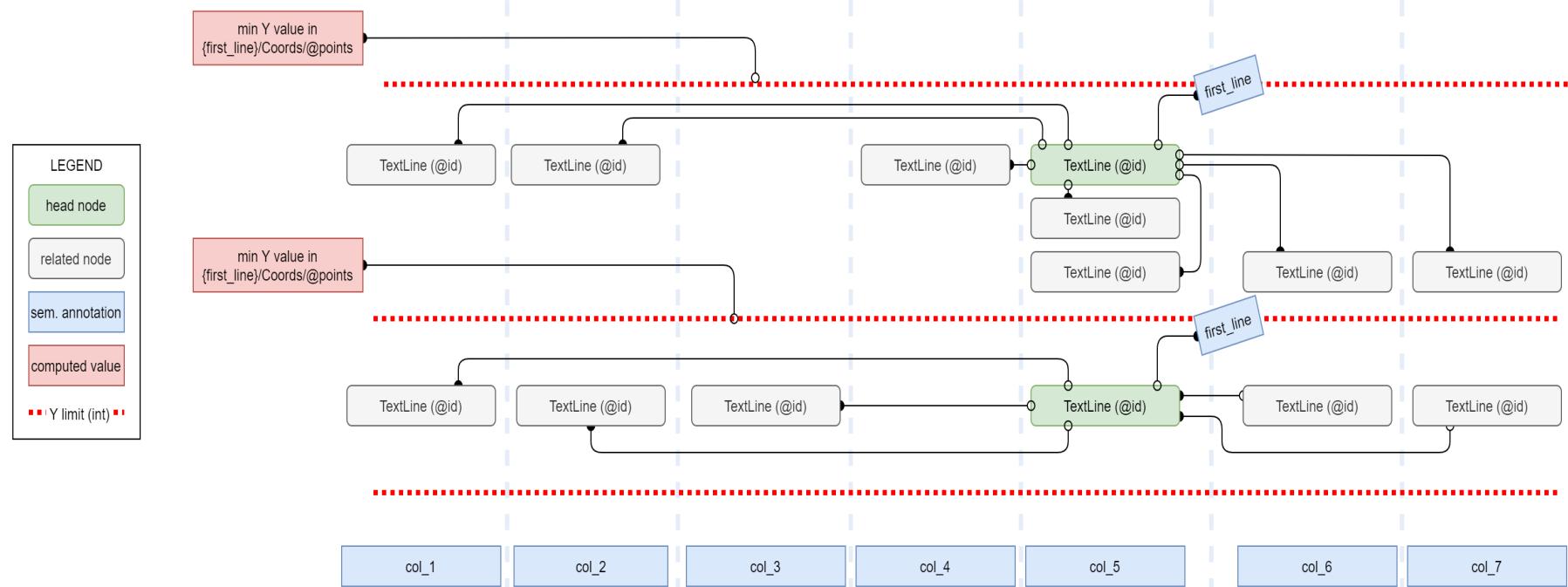


FIGURE E.3 – Illustration du numérotage des lignes sur eScriptorium avec segmentation des régions, et donc de l'ordre dans lequel les lignes transcrivent apparaissent dans un fichier texte.



#### COMPUTING PAGE XML FROM ESCRIPTORIUM TO REGROUP RELATED SEGMENTS OF TEXT (LECTAUREP)

FIGURE E.4 – Schéma illustrant le découpage horizontal des enregistrements présents sur une page d'un répertoire de notaire en s'appuyant sur l'annotation de leurs premières lignes. Schéma créé par Alix Chagué, voir [https://gitlab.inria.fr/almanach/lectaurep/documentation/-/issues/17#note\\_525626](https://gitlab.inria.fr/almanach/lectaurep/documentation/-/issues/17#note_525626) (consulté le 11/08/21).

Listing E.1 – Exemple d'un enregistrement présent dans une page de répertoire de notaire dans le PAGE XML résultant de la transcription, en l'occurrence manuelle.

```
1   <TextLine id="eSc_line_73523fc1" custom="structure"
2     ↳ {type:first_line;"}
3     <Coords points="1300,2741 1304,2712 1337,2687 1406,2712 1487,2698
4       ↳ 1652,2716 1677,2698 1703,2709 1751,2661 1780,2665 1831,2716
5       ↳ 1926,2719 1981,2665 2432,2661 2443,2756 2410,2782 2373,2767
6       ↳ 2245,2782 2128,2782 2077,2763 1882,2771 1816,2807 1740,2774
7       ↳ 1674,2789 1633,2763 1586,2782 1545,2760 1498,2778 1425,2763
8       ↳ 1304,2771"/>
9     <Baseline points="1304,2745 1461,2752 1498,2760 2446,2760"/>
10    <TextEquiv>
11      <Unicode>Boudier (par Abel Eugène) architecte &amp; Marie
12        ↳ Thérèse Louise</Unicode>
13    </TextEquiv>
14  </TextLine>
15
16
17  <TextLine id="eSc_line_d7e2b970" >
18    <Coords points="1249,2811 1256,2763 1322,2785 1428,2763 1483,2789
19      ↳ 1538,2767 1586,2789 1626,2767 1666,2789 1732,2771 1754,2789
20      ↳ 1842,2789 1882,2771 1948,2785 1978,2771 2113,2771 2135,2793
21      ↳ 2183,2771 2238,2793 2267,2771 2384,2771 2406,2793 2443,2793
22      ↳ 2450,2826 2439,2848 2003,2848 1945,2833 1523,2844 1414,2829
23      ↳ 1252,2840"/>
24    <Baseline points="1250,2815 1527,2815 1589,2822 1886,2826
25      ↳ 2454,2828"/>
26    <TextEquiv>
27      <Unicode>Jenny Belin safe- à Paris 15 rue Picot à Abel à Paris
28        ↳ 24 BdSt Denis</Unicode>
29    </TextEquiv>
30  </TextLine>
```



## **Annexe F**

**Affinage d'un modèle de segmentation  
en vue de la reconstitution de la  
structure logique des pages des  
répertoires de notaire**

An 1939 , mois de Mars									
434	3	Donation partage anticipé						8	5.953 80
435	3	Vente						3	30.941 46
436	4	Procuration						4	35
437	4	Recoup de Saint- Paul						6	35
438	4	Dépot notoriété X						6	35
439	4	Adoption notoriété						?	?
440	4	Paul						6	35
441	4	Vente						8	35
442	4	Reconnaisance de dette						8	00.825 00
443	4	Certificat de possession						8	60.472 50
444	4	Procuration						3	82 50
445	6	Procuration						8	35
446	6	Procuration						8	35
447	6	Flaix levée						8	170

FIGURE F.1 – Exemple d'une page d'un répertoire de notaire avec un texte dense, pouvant potentiellement troubler une segmentation horizontale des différents enregistrements. Les traits de couleurs correspondent à la segmentation des colonnes sur eScriptorium.

description images EDITION MODELS Element 1 - FRAN\_0025\_0073.l-0.jpg (2997x4293) ZIP IMPORT ? >

N° DU RÉPERTOIRE	DATES DES ACTES	NATURE ET ESPÈCE DES ACTES :		NOMS, PRÉNOMS ET DOMICILES DES PARTIES	INDICATIONS, SITUATIONS ET PRIX DES BIENS	RELATION DE l'Enregistrement.	
		EN BREVETS	EN MINUTES			DATES	DROITS
576	24		Inventaire		An 1903 , mois de Juillet		
577	24	Décharge		Jouffret (acte déclaré à Jouffret à Paris, rue Bodinier, 6, du 14 Juin 1903)		29	11.25
578	24	Procuration		Jouffret (par Pierre Joseph Marie à Paris, rue Duhautin, 27, à M. Carrere, de Mandat, et de 8.400...)		25	3.75
579	24		Donation	Dedourve (par Louis Jaurand à Paris, avenue Erdmann, 14, 1 <sup>re</sup> étage, à M. Bouveret (aux Claude Pierret à Paris, rue de la Sorbonne, 12, à Clugdine Imbert, son épouse (universelle en toute propriété))		25	3.75
580	24		- 2 <sup>e</sup> -	- 2 <sup>e</sup> (par Mme à son mari ( 2 <sup>e</sup> ))		"	"
581	25	Procuration		Letellier (par Victor Félix à Paris, rue Blanche, 2, 1 <sup>re</sup> étage immobile)		25	3.75
582	27	(du 31 Juillet 1900)	Mainlevée	Frizon (par Siegfried, rue de Rival, 22, 3 <sup>e</sup> étage à Louis Olmion, à Plaques et Matodetzki (par Siegfried, divorce entre Oberläufer Berck) et Alice Schmaïn)		3	2.55
583			Continuation 2 <sup>e</sup> inventaire	Jaslin (par Anne Marie Cadet, 1 <sup>re</sup> de Charles à Paris, rue de la Roquette, 16, à sa fille Clugdine Louise Gabrielle à Levallois Perret)		14	7.50
584	28	Consentement à mariage				30	Grecias

FIGURE F.2 – Attribution d'une nouvelle étiquette pour les colonnes pour les régions pré-annotée par le modèle de segmentation entraîné par Alix Chagué. Chaque couleur représente une étiquette et une région différente.

De gauche à droite :

- « Col\_1 » pour la première colonne
- « Col\_2 » pour la deuxième colonne
- « Col\_3 » pour la troisième colonne
- « Col\_4 » pour la quatrième colonne
- « Col\_5 » pour la cinquième colonne
- « Col\_6 » pour la sixième colonne
- « Col\_7 » pour la septième colonne

RE ET ESPÈCE DES ACTES :	NOMS, PRÉNOMS ET DOMICILES DES PARTIES	RELATI DE l'Enregistre
ETS	INDICATIONS, SITUATIONS ET PRIX DES BIENS	DATES DR
Inventaire	An 1903 , mois de Juillet Contel (après décès de Joseph) à Paris, rue Rodier 70, du 11 Juin 1903 Jouffrel (par Pierre Joseph Marie) à Paris, rue Duvauzin 27, à III. Carrère, de mandat, et de 8.100 <sup>e</sup>	29 H 25 H

FIGURE F.3 – Annotation des lignes indiquant la date sur une page d'un répertoire de notaire. L'attribution d'un label à cette *baseline* est symbolisé par le trait vertical rose au début de celle-ci.

<u>inventaire</u>	el Alice Lehmann	4
	Jaslin (par Anne Marie Cadet, 4 <sup>e</sup> de Charles) à Paris, rue des Ruisseaux, 46 à sa fille Augustine Louise Gabrielle à Sevallis Perrel	30
<u>Mainlevée</u>	du Bois (par Marie Éléonore Enery, 4 <sup>e</sup> de Edouard) d'inscription Gabriel Prosper Déchauny, à Paris, rue Corbeau 13, et Berthe Marie Misché, ép <sup>e</sup>	1
<u>cession de bail</u>	Sauvionier (par Henriette Caroline Chaupeulier, à Paris, rue Oberkampf, 1 <sup>e</sup> de Alphonse Gustave Adrien) à Ernest Félicien Ludovic Victorien Pommerec, et Marie Josephine Rufin, même ad.	11
<u>reconnaissance</u>	Cléolas (par Jean Obernai, Adolphe Adélaïde Lebel, à Paris, rue Daunou 63, à Alphonse Adolphe Sachau), même adresse, de 3.800	3
<u>dette</u>	" <u>T. O. Vallaud et Cie</u> " (de la cité en commandite simple)	3
<u>défaut de pieces de papier</u>	Finch (par Francis Trappes, à Paris, rue des Lazaristes, 71) contre M. M.	3
<u>Défaut de pieces</u>	Autres fidéi-commissaires de Marie Margaret Pollen, 4 <sup>e</sup> de John Hungerford Pollen, 2 <sup>e</sup> arrondissement, 11 Cambridge Crescent Bayntree	3
	217 ans 2' 1902	

FIGURE F.4 – Annotation des premières lignes de chaque enregistrement dans une page d'un répertoire de notaire. L'attribution d'un label à ces *baseline* est symbolisé par le trait vertical rose au début de celles-ci. Les *baselines* non annotées possèdent un trait vertical violet.

Listing F.1 – Rapport d’entraînement obtenu avec Kraken pour le modèle de segmentation sémantique visant à reconstituer la structure logique des pages des répertoires des notaires.

```
1 ketos segrain --o semantic_seg --load blla_ft_reg.mlmodel --f page --t allxmls --device
    cuda :0 --augment --threads 16 --resize both
2 Loading existing model from blla_ft_reg.mlmodel [1.5974] Region eSc_dummyblock_
    without coordinates
3 [1.6354] Region eSc_dummyblock_ without coordinates
4 [1.6652] Region eSc_dummyblock_ without coordinates
5 [1.7121] Region eSc_dummyblock_ without coordinates
6 [1.7388] Region eSc_dummyblock_ without coordinates
7 [1.7860] Region eSc_dummyblock_ without coordinates
8 [1.9808] Region eSc_dummyblock_ without coordinates
9 [2.0137] Region eSc_dummyblock_ without coordinates
10 [2.0196] Region eSc_dummyblock_ without coordinates
11 [2.3338] Region eSc_dummyblock_ without coordinates
12 Fitting network exactly to training set Adding 11 missing types and removing 3 to network
    output layer
13 [2.3567] Setting baseline location to baseline from unset model.
14 Training line types :
15 default 2 15393
16 printed 3 1864
17 First_line 6 2168
18 Main_date 7 134
19 Training region types :
20 header 4 113
21 marginal 5 73
22 Col_7 8 110
23 Col_3 9 110
24 Col_4 10 110
25 Stamp 11 4
26 Col_2 12 109
27 Col_1 13 110
28 text 14 18
29 Col_6 15 110
30 Col_5 16 110
31 stage 1/50 [#####
  110/110 Accuracy report (1) mean_iu : 0.1712 freq_iu : 0.2955 mean_acc : 0.8821
```

accuracy : 0.8821  
 32 stage 2/50 [#####]  
     110/110 Accuracy report (2) mean\_iu : 0.2016 freq\_iu : 0.4073 mean\_acc : 0.9438  
     accuracy : 0.9438  
 33 stage 3/50 [#####]  
     110/110 Accuracy report (3) mean\_iu : 0.1860 freq\_iu : 0.4063 mean\_acc : 0.9450  
     accuracy : 0.9450  
 34 stage 4/50 [#####]  
     110/110 Accuracy report (4) mean\_iu : 0.4430 freq\_iu : 0.5667 mean\_acc : 0.9619  
     accuracy : 0.9619  
 35 stage 5/50 [#####]  
     110/110 Accuracy report (5) mean\_iu : 0.2930 freq\_iu : 0.5396 mean\_acc : 0.9563  
     accuracy : 0.9563  
 36 stage 6/50 [#####]  
     110/110 Accuracy report (6) mean\_iu : 0.4356 freq\_iu : 0.5616 mean\_acc : 0.9595  
     accuracy : 0.9595  
 37 stage 7/50 [#####]  
     110/110 Accuracy report (7) mean\_iu : 0.4826 freq\_iu : 0.6083 mean\_acc : 0.9629  
     accuracy : 0.9629  
 38 stage 8/50 [#####]  
     110/110 Accuracy report (8) mean\_iu : 0.5476 freq\_iu : 0.6751 mean\_acc : 0.9706  
     accuracy : 0.9706  
 39 stage 9/50 [#####]  
     110/110 Accuracy report (9) mean\_iu : 0.5449 freq\_iu : 0.6717 mean\_acc : 0.9704  
     accuracy : 0.9704  
 40 stage 10/50 [#####]  
     110/110 Accuracy report (10) mean\_iu : 0.4845 freq\_iu : 0.6129 mean\_acc : 0.9630  
     accuracy : 0.9630  
 41 stage 11/50 [#####]  
     110/110 Accuracy report (11) mean\_iu : 0.5466 freq\_iu : 0.6747 mean\_acc : 0.9700  
     accuracy : 0.9700  
 42 stage 12/50 [#####]  
     110/110 Accuracy report (12) mean\_iu : 0.4902 freq\_iu : 0.6214 mean\_acc : 0.9644  
     accuracy : 0.9644  
 43 stage 13/50 [#####]  
     110/110 Accuracy report (13) mean\_iu : 0.5903 freq\_iu : 0.7219 mean\_acc : 0.9749  
     accuracy : 0.9749

44 stage 14/50 [#####] 110/110 Accuracy report (14) mean\_iu : 0.5449 freq\_iu : 0.7321 mean\_acc : 0.9753 accuracy : 0.9753

45 stage 15/50 [#####] 110/110 Accuracy report (15) mean\_iu : 0.6276 freq\_iu : 0.7488 mean\_acc : 0.9767 accuracy : 0.9767

46 stage 16/50 [#####] 110/110 Accuracy report (16) mean\_iu : 0.5686 freq\_iu : 0.6877 mean\_acc : 0.9692 accuracy : 0.9692

47 stage 17/50 [#####] 110/110 Accuracy report (17) mean\_iu : 0.5794 freq\_iu : 0.7038 mean\_acc : 0.9711 accuracy : 0.9711

48 stage 18/50 [#####] 110/110 Accuracy report (18) mean\_iu : 0.6251 freq\_iu : 0.7315 mean\_acc : 0.9738 accuracy : 0.9738

49 stage 19/50 [#####] 110/110 Accuracy report (19) mean\_iu : 0.5691 freq\_iu : 0.6915 mean\_acc : 0.9697 accuracy : 0.9697

50 stage 20/50 [#####] 110/110 Accuracy report (20) mean\_iu : 0.6092 freq\_iu : 0.7257 mean\_acc : 0.9736 accuracy : 0.9736

51 stage 21/50 [#####] 110/110 Accuracy report (21) mean\_iu : 0.6433 freq\_iu : 0.7592 mean\_acc : 0.9783 accuracy : 0.9783

52 stage 22/50 [#####] 110/110 Accuracy report (22) mean\_iu : 0.6275 freq\_iu : 0.7412 mean\_acc : 0.9760 accuracy : 0.9760

53 stage 23/50 [#####] 110/110 Accuracy report (23) mean\_iu : 0.6248 freq\_iu : 0.7382 mean\_acc : 0.9765 accuracy : 0.9765

54 stage 24/50 [#####] 110/110 Accuracy report (24) mean\_iu : 0.5976 freq\_iu : 0.7168 mean\_acc : 0.9731 accuracy : 0.9731

55 stage 25/50 [#####] 110/110 Accuracy report (25) mean\_iu : 0.6635 freq\_iu : 0.7721 mean\_acc : 0.9783 accuracy : 0.9783

56 stage 26/50 [#####]

110/110 Accuracy report (26) mean\_iu : 0.5819 freq\_iu : 0.7441 mean\_acc : 0.9758  
accuracy : 0.9758

57 stage 27/50 [#####]  
110/110 Accuracy report (27) mean\_iu : 0.6367 freq\_iu : 0.7415 mean\_acc : 0.9757  
accuracy : 0.9757

58 stage 28/50 [#####]  
110/110 Accuracy report (28) mean\_iu : 0.6477 freq\_iu : 0.7539 mean\_acc : 0.9769  
accuracy : 0.9769

59 stage 29/50 [#####]  
110/110 Accuracy report (29) mean\_iu : 0.7022 freq\_iu : 0.8041 mean\_acc : 0.9818  
accuracy : 0.9818

60 stage 30/50 [#####]  
110/110 Accuracy report (30) mean\_iu : 0.7202 freq\_iu : 0.8214 mean\_acc : 0.9841  
accuracy : 0.9841

61 stage 31/50 [#####]  
110/110 Accuracy report (31) mean\_iu : 0.5927 freq\_iu : 0.7575 mean\_acc : 0.9781  
accuracy : 0.9781

62 stage 32/50 [#####]  
110/110 Accuracy report (32) mean\_iu : 0.6838 freq\_iu : 0.7946 mean\_acc : 0.9807  
accuracy : 0.9807

63 stage 33/50 [#####]  
110/110 Accuracy report (33) mean\_iu : 0.6259 freq\_iu : 0.7911 mean\_acc : 0.9809  
accuracy : 0.9809

64 stage 34/50 [#####]  
110/110 Accuracy report (34) mean\_iu : 0.6890 freq\_iu : 0.7947 mean\_acc : 0.9808  
accuracy : 0.9808

65 stage 35/50 [#####]  
110/110 Accuracy report (35) mean\_iu : 0.6998 freq\_iu : 0.7963 mean\_acc : 0.9811  
accuracy : 0.9811

66 stage 36/50 [#####]  
110/110 Accuracy report (36) mean\_iu : 0.6839 freq\_iu : 0.7893 mean\_acc : 0.9806  
accuracy : 0.9806

67 stage 37/50 [#####]  
110/110 Accuracy report (37) mean\_iu : 0.6949 freq\_iu : 0.7944 mean\_acc : 0.9802  
accuracy : 0.9802

68 stage 38/50 [#####]  
110/110 Accuracy report (38) mean\_iu : 0.6721 freq\_iu : 0.7737 mean\_acc : 0.9779

accuracy : 0.9779  
 69 stage 39/50 [#####]  
     110/110 Accuracy report (39) mean\_iu : 0.6732 freq\_iu : 0.7760 mean\_acc : 0.9784  
     accuracy : 0.9784  
 70 stage 40/50 [#####]  
     110/110 Accuracy report (40) mean\_iu : 0.6575 freq\_iu : 0.7716 mean\_acc : 0.9788  
     accuracy : 0.9788  
 71 stage 41/50 [#####]  
     110/110 Accuracy report (41) mean\_iu : 0.7079 freq\_iu : 0.8063 mean\_acc : 0.9820  
     accuracy : 0.9820  
 72 stage 42/50 [#####]  
     110/110 Accuracy report (42) mean\_iu : 0.6552 freq\_iu : 0.7652 mean\_acc : 0.9784  
     accuracy : 0.9784  
 73 stage 43/50 [#####]  
     110/110 Accuracy report (43) mean\_iu : 0.6720 freq\_iu : 0.7755 mean\_acc : 0.9783  
     accuracy : 0.9783  
 74 stage 44/50 [#####]  
     110/110 Accuracy report (44) mean\_iu : 0.6980 freq\_iu : 0.7829 mean\_acc : 0.9795  
     accuracy : 0.9795  
 75 stage 45/50 [#####]  
     110/110 Accuracy report (45) mean\_iu : 0.6788 freq\_iu : 0.7628 mean\_acc : 0.9773  
     accuracy : 0.9773  
 76 stage 46/50 [#####]  
     110/110 Accuracy report (46) mean\_iu : 0.6960 freq\_iu : 0.7794 mean\_acc : 0.9803  
     accuracy : 0.9803  
 77 stage 47/50 [#####]  
     110/110 Accuracy report (47) mean\_iu : 0.6975 freq\_iu : 0.7798 mean\_acc : 0.9789  
     accuracy : 0.9789  
 78 stage 48/50 [#####]  
     110/110 Accuracy report (48) mean\_iu : 0.7349 freq\_iu : 0.8128 mean\_acc : 0.9829  
     accuracy : 0.9829  
 79 stage 49/50 [#####]  
     110/110 Accuracy report (49) mean\_iu : 0.7088 freq\_iu : 0.7951 mean\_acc : 0.9810  
     accuracy : 0.9810  
 80 stage 50/50 [#####]  
     110/110 Accuracy report (50) mean\_iu : 0.7055 freq\_iu : 0.7885 mean\_acc : 0.9796  
     accuracy : 0.9796

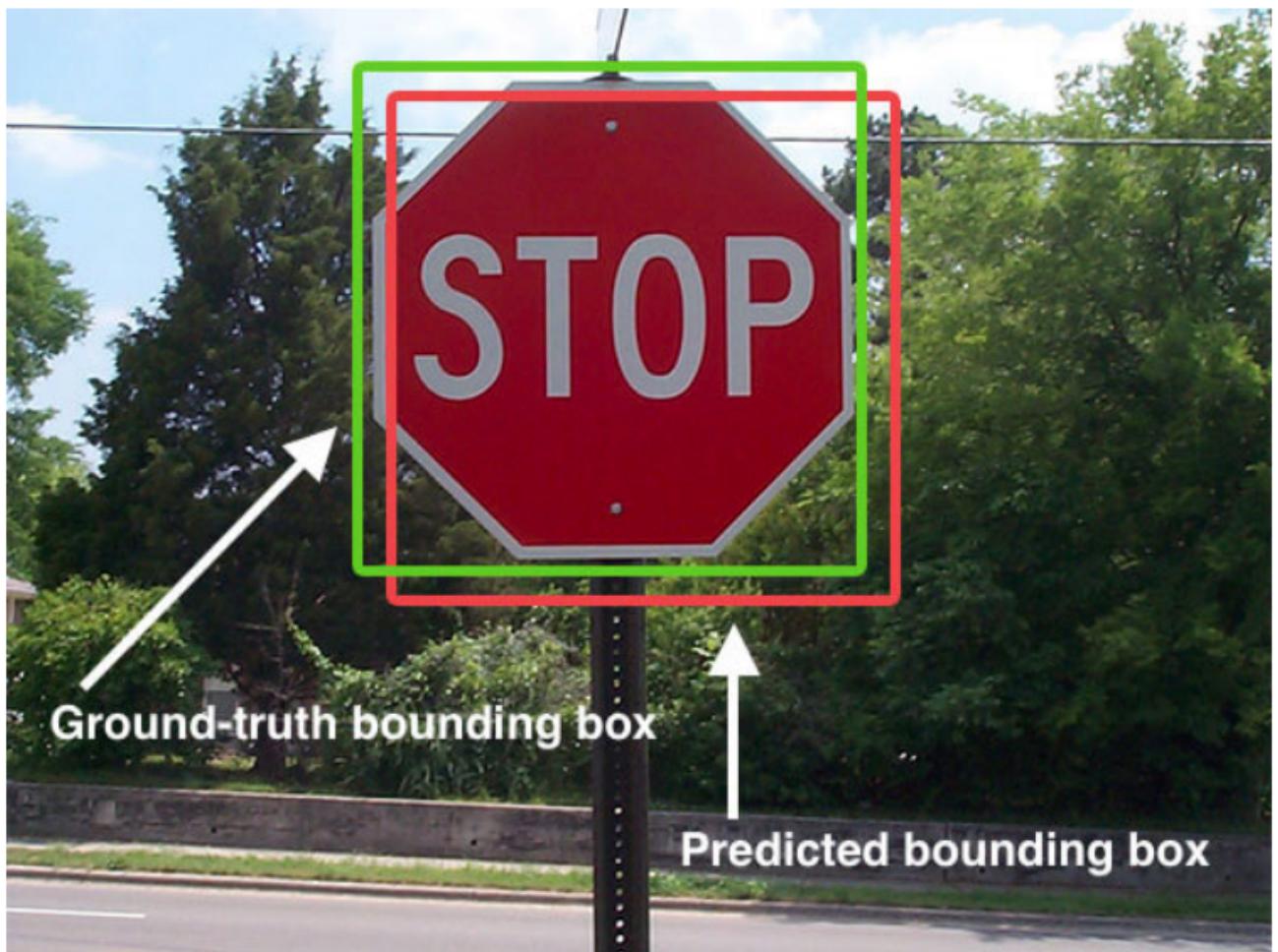


FIGURE F.5 – Exemple de visualisation de l'intersection de deux cadre de délimitation, le vert étant la vérité de terrain et le rouge étant la prédiction d'un modèle. Source : <https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/> (consulté le 10/08/21).

N° DU RÉPERTOIRE	DATES DES ACTES	NATURE ET ESPÈCE DES ACTES		NOMS, PRÉNOMS ET DOMICILES DES PARTIES		RELATION DU RE
		EN BREVETS	EN MINUTES	INDICATIONS, SITUATIONS ET PRIX DES BIENS		
1344	12 Juillet 1929	Dépôt de Pièces de Publicité		An 1929 , mois de Juillet Union Patriotique Fencière (rig. à Paris) 8 rue Boulot		
1345	12 Juillet 1929	Dépôt de Pièces de Publicité		St amayen Immobilier dont le siège est à Paris. Le tout bâti au 5 <sup>e</sup> étage du 50 rue Raffet		
1346	12 Juillet 1929	do. do.		St amayen Immobilier au 11 <sup>e</sup> étage du 13 boulevard Montmorency (1 <sup>e</sup> arr mme) dont le siège est à Paris 11 <sup>e</sup> arrondissement		
1347	12 Juillet 1929	do. do.		Union Métallurgique de la Marne Seine (1 <sup>e</sup> arr) Fabrice Léonel dit do le siège est à Paris 1 <sup>e</sup> arrondissement de la République		
1348	12 Juillet 1929	Certification de Signature		Messager par Fabrice Léonel à Paris 1 <sup>e</sup> arrondissement de la République do le siège à l'angle de l'avenue de l'Opéra et de l'avenue de l'Opéra au n <sup>o</sup> 11 avenue Auguste Renoir		
1349	10	do. do.		Messager rappel au rig. à Paris		
1350	10	do. do.		Messager do. do.		
1351	10	do. do.		Benton (Louis) a fait son décret à la date ci-dessous, l'ensemble Henri Schmit à Chelles, ville France Verbeau, Villa Gentil et autres.		

FIGURE F.6 – Exemple d'une annotation vérifié de terrain des régions et des baselines dans une page d'un répertoire de notaire.

N° DU RÉPERTOIRE	DATES DES ACTES	NATURE ET ESPÈCE DES ACTES		NOMS, PRÉNOMS ET DOMICILES DES PARTIES		RELATION DE l'Enregistrement	
		EN BREVETS	EN MINUTES	INDICATIONS, SITUATIONS ET PRIX DES BIENS		DATES	DROITS
1344	12. Juillet 1929	<del>Dépôt à Picard</del>		Union Parisienne Foncière (siège à Paris) 8 rue de la Publicité 2000.			
1345	12 Juillet 1929	<del>Dépôt à Picard</del>		St Améry Union Bâtie dont le siège est à Paris 16 rue de la Publicité n° 50 rue Raffet			
1346	12 Juillet 1929			St Jean bâtie au 11 et 13 boulevard de Montmorency 1 <sup>er</sup> arr. (siège) dont le siège est à Paris 1 <sup>er</sup> arrondissement			
1347	12 Juillet 1929			Union Métallurgique de la Haute Seine (siège) bâtie bâtie dont le siège est à Chelles place de la République			
1348	12. Octobre 1929	<del>Classification et signature</del>		Messager par trois bâties Chelles à Paris 10 <sup>e</sup> arrondissement (siège) bâtie de l'agence Jean Charles Gayot à Paris 1 <sup>er</sup> arrondissement Auguste 2 <sup>e</sup> arrondissement			
1349	12	do	do	Messager appelle au registre			
1350	12	do	do	Messager	do		
1351	12	do	do	Besson (faire) à Paris 1 <sup>er</sup> arrondissement (bâtie)			
				Henri Besson à Chelles, affilié à Paris, Villa Jeante et autres			

FIGURE F.7 – Exemple d'une prédition de l'annotation des régions et des baselines dans une page d'un répertoire de notaire.

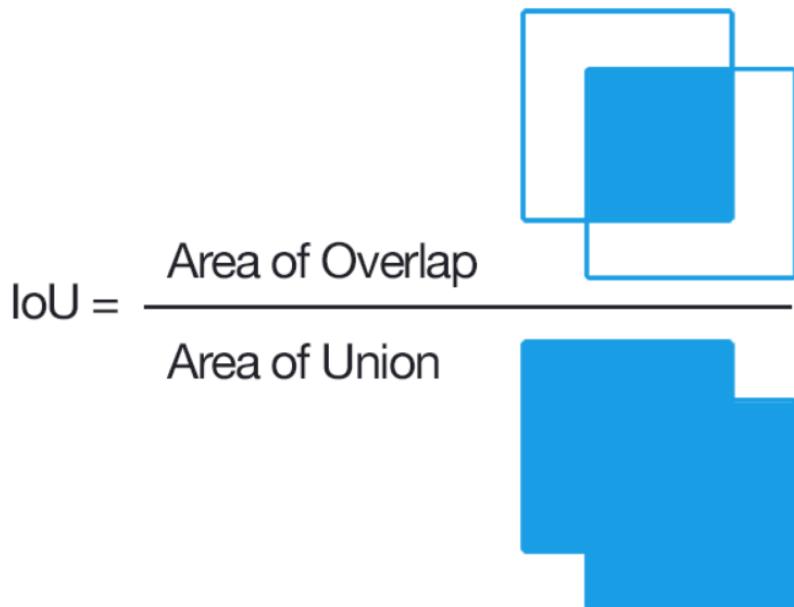


FIGURE F.8 – Calcul de la métrique *Intersection over Union*, ou indice de Jaccard.  
Source : <https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/> (consulté le 10/08/21).



FIGURE F.9 – Visualisation des scores de la métrique *Intersection over Union*, ou indice de Jaccard. Source : <https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/> (consulté le 10/08/21).

				INDICATIONS, SITUATIONS ET PRIX DES BIENS.	
		EX-BREVES.	EX-MINUTES.		
117	+		Vente	An 1875 mois d'Avril. Beaujardin par Jean Thibaut, en deux lots toutes les deux dominances à Paris, parage bivali 16 à Beaujardin Athen. Beaujardin (une ferme grande de fermes), dominante à Paris, jusqu'à l'heure 16, à un pied de hauteur à Paris parage bivali 16, négociant 149.000	
260	o		Souhait	Institut agricole de Gén. pour la formation d'un nouvel agriculteur dans le monde, dont le budget annuel est établi à Gén. bras, pour développer cette école en deux cours de classe fermées.	
261	z		Terrière	Lavignolle, résidence pour homme, et auquel deux mètres peuvent être	100
262	z	l'abbé (pays) Dijon		Lebeaux (un des François Louis, dominante à Paris, en deux dominances à Paris, et l'offre de vente au prix 2 million dans lequel que l'abbé pour les marchés de la ville).	1.10
263	z	Dominante		Mayre, partie de François, nomme Bonne, issue de Poissons, dominante à Paris, en la Haute Marne, et l'offre de vente	1.11
264	z	Pavarini		Audin, dominante à Paris, issue de la Haute Marne et l'offre de vente au prix 100.000 francs.	1.11
265	z	Gaudin		Sanguinelle (par Homme à la ferme autrement nommée, en deux dominances à Paris (quatre fermes) à Paris, et l'offre de vente.	1.11
266	z	Omnium		Sanguinelle (partie de la ferme autrement nommée, à Paris, et l'offre de vente).	1.11

FIGURE F.10 – Résultats d'une prédiction faite par le modèle affiné de segmentation sémantique. Sur cet échantillon, toutes les premières ligne de chaque enregistrement ont été correctement identifiées, et la segmentation des régions est également bonne. La ligne indiquant la date de la page est cependant mal segmentée.

				An 1903 , mois de Juillet		
576	24	Décharge	Inventaire	Contel (abbé dédié à Joseph) à Paris, rue Radier 56, du 11 Juin 1903	29	11.25
577	24	Procuration		Jouffrel (pour Pierre Joseph Marie) à Paris, rue D'Assas 17, ill. Carrere de manuel, et de 8.100	25	3.75
578	24		Donation	Dedourze (par Louis Paul) à Paris, avenue Trudaine 11, 1 <sup>re</sup> étage	25	3.75
579	24			Bouveret (par Claude Pierre) à Paris, rue de la Chaussine 12, à Chiquine Embert, son épouse (universelle en toute propriété)	25	3.75
580	24	Procuration		29 (part 1 <sup>re</sup> ) à son mari ( " )	"	"
581	25			Séteillicier (par Félix Félix) à Paris, rue Blanche 11, 1 <sup>re</sup> étage immobile	25	3.75
582	27		Mainlevée	Fitzoulis (partie d'une partie de 82.000 francs) à Louis Onion, à Staines	3	2.55
583	27	(du 31 juillet 1903) Continuation 2 <sup>e</sup>	inventaire	Mercodetzki (partie d'instruction suivante entre Ibrahim Borod) à Miss Lehmann	4	7.50
584	28	Consentement à mariage		Jaslin (par Clémie Marie Cadet, 1 <sup>re</sup> de Charles) à Paris, rue de l'Assomption 16, à sa fille Chiqueline Louise Gabrielle) à Lavallois Perret	30	Grosjean
585	29		Mainlevée	du Bois (par Marie Eleonore Emery, 1 <sup>re</sup> de Edouard) à inscription C	1	12.45
586	29		Ression de bail	Gabriel Pothier Dechamp, à Paris, rue Corbeau 13, à Bernhardine M. Schiebel	1	16.55
587	30		Connaissance	Laumonier (par Félicité Caroline Chaperon, à Paris, rue Chêne 1 <sup>re</sup> de Claude, Giuliano, à Eustache Lucien Sébastien Cormier, et Marie Joseph Ralieu, née à Dr	7	47.50
			acte	Nicolas (par Jean Chaudron, le fils Charles de l'épicerie à Paris, rue Damrémont 63, à Châlon-sur-Saône La Châlon, même adresse, des 3.800 "	3	3.75
588	31	(du 10 Juin 1903) Défense de faire défaillir	Défense de pieces	O. Talland et Cie (de la 1 <sup>re</sup> au communiqué de simple)	3	239.50
589	31			Finch (par Francis Frappes, à Paris, rue de Lazarre 11) contre M. H. Collet, à Londres, 11 Cambridge Crescent Bayswater	3	
				La autre partie : commissaires de Marie Margarete Collen, 1 <sup>re</sup> de John Edward Collet, à Londres, 11 Cambridge Crescent Bayswater		
590	1	Procuration		Mois d'Août 1903		
				Charmeil (par Louis Jean Charles Chauvin Thibaut) à Paris, rue Monge 89, Châlon Charles Victor, Chuchot la Peuline de Bacchus, rue de la Cour 11, et Maurice Minet		

FIGURE F.11 – Résultats d'une prédiction faite par le modèle affiné de segmentation sémantique, similaire à l'exemple précédent.

N° DU RÉPERTOIRE	DATES ACTES	NATURE ET ESPÈCE DES ACTES :		NOMS, PRÉNOMS ET DOMICILES DES PARTIES		RELATION DE L'Enregistrement.
		EN BREVETS	EN MINUTES	INDICATIONS, SITUATIONS ET PRIX DES BIENS		
500	6	Etat liquidat		An 1911  , mois d e Mai Gauvin, au temps de la vente deux fermes au moin (barbe) au des chalos St éponne et paroisse de corps de M. Auguste Alexandre	9	150
501	6	Proc Verbal de lecture		Gauvin de l'état liquidatif no cience	9	31.70
502	6	18 8 1903	Realisation d'un Prat	Ericon (au temps connaît sur le Crédit foncier à Benjamin Joseph Maurice à Paris boulevard de Courcelles 40 d'un prêt de 9000 Roche Jean & Charlotte Guichard Edouard Baron inaugurante à Paris rue leon Coquet 11 n° 100 Willy de Cornuau en bl. pour intervenir à un acte succession et de bons fonds	11	119.00
503	8	Procuration		Daleine (marie) sa femme Nathalie Roche autre à Paris rue leon Coquet 11 n° 100 venue de M. le Baron Jean Baptiste Alfred en bl pour cesser et transcrire	9	371
504	8	Procuration				221

FIGURE F.12 – Résultats d'une prédiction faite par le modèle affiné de segmentation sémantique, similaire à l'exemple précédent.

Rép. <sup>n</sup>	Actes.	NATURE ET ESPÈCE DES ACTES		NOMS, PRÉNOMS ET DOMICILES DES PARTIES.		RELATION de l'enregistrement
		EN BREVETS.	EN MINUTES.	INDICATIONS, SITUATIONS ET PRIX DES BIENS.		
401	14.			An 1836 mois de Juillet Balabot (par M <sup>r</sup> Joseph Léon) Dem <sup>t</sup> à Paris rue Blanche n <sup>o</sup> 17 à M <sup>r</sup> Auguste François Balaille Dem <sup>t</sup> à Paris au ministère des finances.		Demandeur 15 2.20
402	14.	Procuration		Balabot ( par mand. f ) à M <sup>r</sup> Jean Laine Dem <sup>t</sup> à la Villette, près Paris.		15 2.20
403	15.			Vaillant ( par Daniel Narcisse ) Dem <sup>t</sup> à Paris rue Grange batelière n <sup>o</sup> 1, de procuration par M <sup>me</sup> Léophasie Brocard Dem <sup>t</sup> à Brux <sup>s</sup> à Mr Jean Frédéric Deinturier Dem <sup>t</sup> à Paris, rue Grange batelière n <sup>o</sup> 1		16 2.20
404	15.18.			Quittance Diney ( par Mr Jean Baptiste ) Dem <sup>t</sup> à Paris rue des Bourrelles n <sup>o</sup> 18 à Mr Charles Comte Féon Dem <sup>t</sup> à Paris rue Haubout n <sup>o</sup>		

FIGURE F.13 – Exemple de problèmes rencontrés dans une prédiction faites par le modèle affiné de segmentation sur une page de répertoire de notaire. Ici, la segmentation des régions a été fortement perturbée, mais l'annotation des premières lignes dans la colonne 5 reste régulière.

## **Annexe G**

### **Visualiser la reconnaissance d'entités nommées**

The screenshot shows a GitLab project interface for a 'NER' (Named Entity Recognition) application. The left sidebar includes links for 'Project overview', 'Repository' (selected), 'Files', 'Commits', 'Branches', 'Tags', 'Contributors', 'Graph', 'Compare', 'Issues' (2), 'Merge requests' (0), 'Security & Compliance', 'Operations', 'Packages & Registries', and 'Collapse sidebar'. The main content area displays a text document with several entities highlighted in orange boxes with 'LOC' suffixes. The text discusses a social gathering in Mamers, Muni, Clermont, Alençon, and various villages like Oisseau, Chérisay, Bourg-le-Roy, Louvigny, Roussé-Fontaine, Saint-Rémy-du-Plein, and Saint-Rémy-des-Ponts. It also mentions 'Pré en Pail', 'Pooté', 'Mieuxé', 'Héloup', 'Champleur', 'Saint-Rigormer-des-Bois', 'Marollette', 'Douillet-le-Joly', 'Montreuil-le-Chétif', 'Saint-Léonard-des-Bois', 'Saint-Onu-de-Mé', 'Saint-Paul-le-Gandelin', 'Saint-Pierre-des-Ormes', 'Saint-Vincent-des-Prés', 'Saint-Christophe-du-Jambet', 'Boucelles', 'Saint-Germain-de-la-Coudre', 'Bonnétable', 'Saint-Georges-du-Rocay', 'Rouperoux-le-Coquet', 'Chapelle des-Bois', 'Saint-Martin-des-Monts', 'Saint-Aubin', 'Loquenay', 'Saint-Ouen', 'Mimbré', 'Greez-sur-Roc', 'Saint-Jean-des-Echelles', 'Saint-Hilaire-le-Lierru', and ends with a note about the variety and history of the region.

FIGURE G.1 – Visualisation du résultat de la REN avec la librairie DisplaCy d'une lettre de Paul d'Estournelles de Constant, octobre 1919, [https://github.com/FloChiff/DAHNProject/blob/master/Correspondence/Paul\\_d\\_Estournelles\\_de\\_Constant/Corpus/Lettre569\\_3octobre1919.xml](https://github.com/FloChiff/DAHNProject/blob/master/Correspondence/Paul_d_Estournelles_de_Constant/Corpus/Lettre569_3octobre1919.xml), corpus du projet DAHN (1). Visualisation issue du notebook suivant : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner\\_python/ner\\_spacy/test\\_dahn\\_spacy.ipynb](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_spacy/test_dahn_spacy.ipynb).

201

l'heure du dîner. 5 Octobre MISC 1919. Le surlendemain je suis à Paris LOC ; je me rends compte que le traité de Paix MISC va enfin être approuvé par la Chambre ORG ; il sera voté la semaine prochaine par le Sénat ORG . J'ai juste le temps de retourner au Mans LOC où La Ligue des Droits de l'Homme ORG m'a fait promettre de lui faire, le Samedi 4 MISC , une conférence, et où je surveille les progrès de notre monument Wilbur Wright PER tout en préparant mon installation matérielle pour la période électorale imminente. Du Mans, je retourne à Créans LOC puis, le lendemain matin je préside mon conseil Municipal PER à Clermont LOC - Créans LOC et je mets la dernière main aux préparatifs de l'inauguration de notre monument aux morts de la guerre. Cela fait, j'arrive à La Flèche MISC où j'ai, malgré tout, réussi à réorganisé le Comice Agricole ORG du canton. C'est le Dimanche LOC , 5 octobre. M'y voici. Et c'est par là que je termine et que je boucle mon tour de la Sarthe. La Flèche LOC est, comme vous savez, la plus réactionnaire de toutes les villes du département, après ma commune de Clermont LOC - Créans LOC sa digne voisine. Mais comme La Flèche LOC est ma ville natale, je ne puis en désespérer. Et c'est justice. La Flèche LOC est une ville de vieux retraités, de rentiers, de séminaires, de couvents, et par conséquent de bigotes. L'école militaire du Prytanée LOC y entretient, en outre, le plus détestable esprit de caste. C'est charmant de voir ces enfants et ces jeunes gens en-27 - pantalon rouge, comme avant la guerre, comme sous le Premier Empire MISC c'est charmant et c'est désolant ; c'est toujours à recommencer ; c'est un nid de militarisme et de tout ce que je combats. Raison de plus pour m'obstiner dans ma tâche. J'en ai été, cette fois récompensé. Mon Comice MISC , dont on prédisait l'échec, a parfaitement réussi ; la journée, plus belle que celle du Dimanche précédent MISC , s'est passée admirablement. On n'avait jamais tant vu de monde dans les rues, sur les places. Les visages étaient souriants. La fête a duré très tard dans la nuit. Je rentre à Paris LOC maintenant, la conscience tranquille. J'ai rattrapé le temps que j'avais consacré à mes voyages aux Etats-Unis LOC et en Angleterre LOC . J'ai dû renoncer à me rendre à Genève LOC où m'appelait avec insistance l' Union Interparlementaire ORG et nous avons ajourné la réunion des Associations pour la société des Nations ORG qui devait se tenir à Bruxelles LOC . Maintenant, une fois réinstallé à Paris LOC , aussitôt le traité de paix ratifié, pour ce qu'il vaut, et la période

FIGURE G.2 – Visualisation du résultat de la REN avec la librairie DisplaCy d'une lettre de Paul d'Estournelles de Constant, octobre 1919, [https://github.com/FloChiff/DAHNProject/blob/master/Correspondence/Paul\\_d\\_Estournelles\\_de\\_Constant/Corpus/Lettre569\\_3octobre1919.xml](https://github.com/FloChiff/DAHNProject/blob/master/Correspondence/Paul_d_Estournelles_de_Constant/Corpus/Lettre569_3octobre1919.xml), corpus du projet DAHN (2). Visualisation issue du notebook suivant : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner\\_python/ner\\_spacy/test\\_dahn\\_spacy.ipynb](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_spacy/test_dahn_spacy.ipynb).

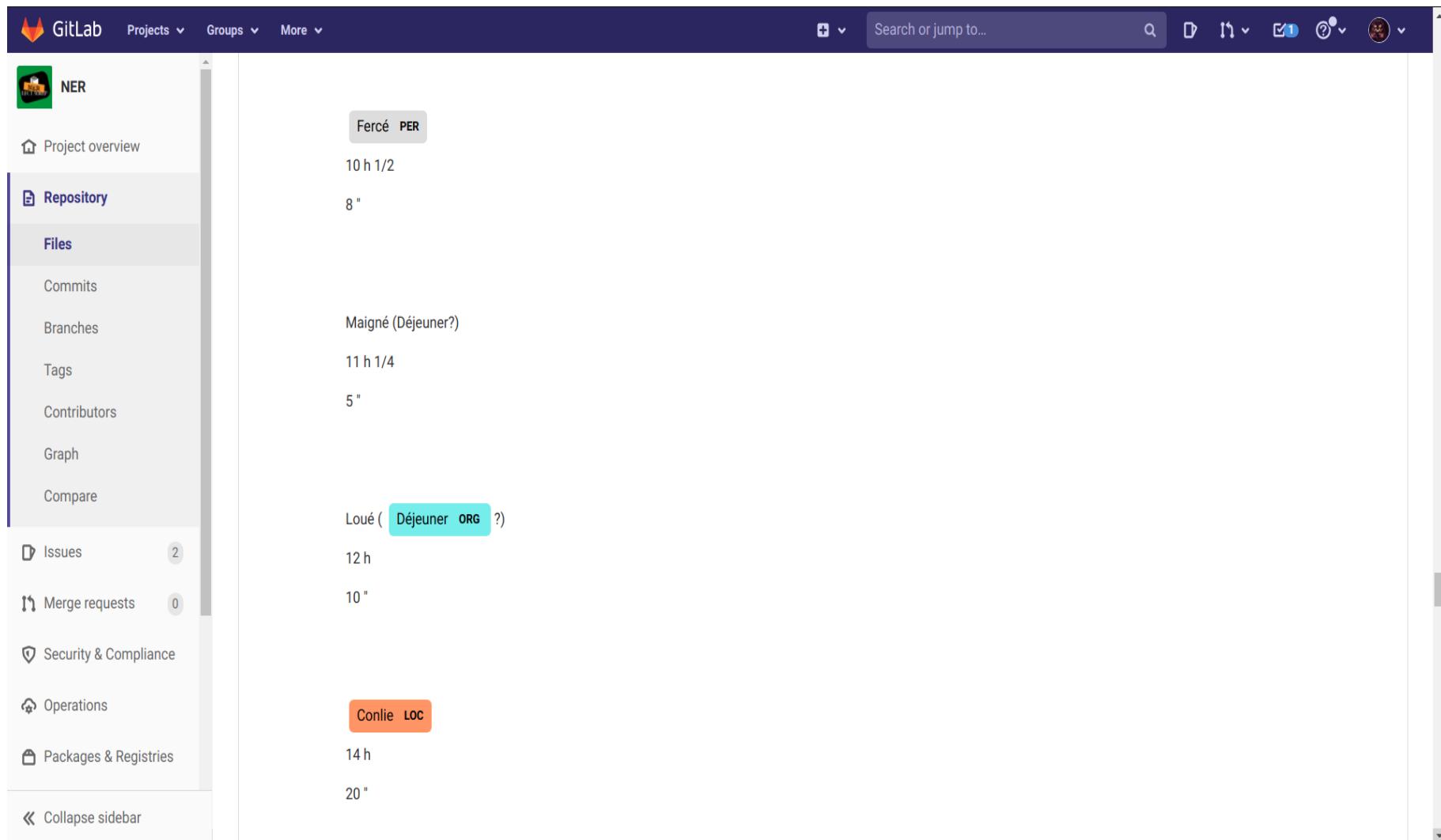


FIGURE G.3 – Visualisation avec la librairie DisplaCy du résultat de la REN avec la librairie DisplaCy d'une lettre de Paul d'Estournelles de Constant, octobre 1919, [https://github.com/FloChiff/DAHNProject/blob/master/Correspondence/Paul\\_d\\_Estournelles\\_de\\_Constant/Corpus/Lettre569\\_3octobre1919.xml](https://github.com/FloChiff/DAHNProject/blob/master/Correspondence/Paul_d_Estournelles_de_Constant/Corpus/Lettre569_3octobre1919.xml), corpus du projet DAHN (3). Visualisation issue du *notebook* suivant : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner\\_python/ner\\_spacy/test\\_dahn\\_spacy.ipynb](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_spacy/test_dahn_spacy.ipynb).

Listing G.1 – Visualisation avec la librairie DisplaCy du résultat de la REN avec la librairie DisplaCy d'une lettre de Paul d'Estournelles de Constant, corpus du projet DAHN (1). Visualisation issue du *notebook* suivant : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner\\_python/ner\\_stanza/test\\_ner\\_dahn.ipynb](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_stanza/test_ner_dahn.ipynb).

```

1 token : je ner : O
2 token : franchis ner : O
3 token : cette ner : O
4 token : nouvelle ner : O
5 token : étape ner : O
6 token : avec ner : O
7 token : une ner : O
8 token : régularité ner : O
9 token : chronométrique ner : O
10 token : , ner : O
11 token : laissant ner : O
12 token : la ner : O
13 token : grande ner : O
14 token : route ner : O
15 token : d' ner : O
16 token : Alençon ner : S-LOC
17 token : et ner : O
18 token : celle ner : O
19 token : de ner : O
20 token : la ner : O
21 token : belle ner : O
22 token : forêt ner : B-LOC
23 token : de ner : I-LOC
24 token : Perseigne ner : E-LOC
25 token : pour ner : O

```

Listing G.2 – Visualisation avec la librairie DisplaCy du résultat de la REN avec la librairie DisplaCy d'une lettre de Paul d'Estournelles de Constant, corpus du projet DAHN (2). Visualisation issue du *notebook* suivant : [https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner\\_python/ner\\_stanza/test\\_ner\\_dahn.ipynb](https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_stanza/test_ner_dahn.ipynb).

```

1 token : couper ner : O
2 token : ou ner : O
3 token : plus ner : O
4 token : court ner : O
5 token : , ner : O
6 token : par ner : O
7 token : ces ner : O
8 token : jolis ner : O
9 token : villages ner : O
10 token : aux ner : O
11 token : vieux ner : O
12 token : noms ner : O
13 token : parlants ner : O
14 token : : ner : O
15 token : le ner : O
16 token : petit ner : O
17 token : Oiseau ner : S-LOC
18 token : , ner : O
19 token : Chérisay ner : S-LOC
20 token : , ner : O
21 token : Bourg-le-Roy ner : S-LOC
22 token : , ner : O
23 token : Louvigny ner : S-LOC
24 token : , ner : O
25 token : ou ner : O

```



## **Annexe H**

**Annoter des entités nommées : les plate-formes d'annotation Doccano et Inception**

The screenshot shows a dark-themed user interface for managing projects. At the top right, there are buttons for language selection (EN), project navigation (Projects), and a more options menu. Below the header, there are two buttons: 'Create' (highlighted in blue) and 'Delete'. A search bar labeled 'Search' is present. The main area displays a table of projects with the following columns: Name, Description, Type, Updated, and Tags. The table contains six rows of data:

	Name	Description	Type	Updated ↓	Tags
<input type="checkbox"/>	Echantillon REM	Echantillon issu de la REM	SequenceLabeling	18/08/2021 09:56	lectaurep REM
<input type="checkbox"/>	Transcription manuelle	Transcription manuelle - Lectaurep	SequenceLabeling	18/08/2021 09:56	lectaurep GT
<input type="checkbox"/>	Transcription manuelle - démo	Démo Doccano	SequenceLabeling	18/08/2021 09:55	lectaurep démonstration GT
<input type="checkbox"/>	Segmentation phrases - GT	Transcription manuelle avec segmentation des phrases artificielle	SequenceLabeling	18/08/2021 09:55	lectaurep GT
<input type="checkbox"/>	DAHN - GT test	GT pour tester REN sur document épistolaire	SequenceLabeling	18/08/2021 09:53	correspondance rem corrigée

At the bottom right, there are buttons for 'Rows per Page' (set to 10), page navigation ('1-5 of 5'), and other controls.

FIGURE H.1 – Page des projets de la plate-forme Doccano.

The screenshot shows the Doccano web application interface. On the left is a dark sidebar with white icons and text, listing navigation options: Home, Dataset (selected), Labels, Members, Comments, Guideline, Statistics, and Settings. A prominent blue button labeled "Start Annotation" is at the top of this sidebar. To its right is the main content area. At the very top of the content area are three buttons: "Actions ▾", "Delete", and a red "Delete All" button. Below these are search and filter controls: a search bar with placeholder "Search", a checkbox for "Text", and columns for "Metadata", "Comments", and "Action". A single row of data is listed, containing the text "An 1901, mois de Étévrier Dupuis par Pierre) à Paris, rue Turot 4, à Catherine Bigol, safes a Oo (par Mve) s.n. à Muret (de Paul Louis Georges) dt à Paris, Bd Rochechouart 68, et Marie Joséphin...", a count of 0 in the Metadata column, a count of 0 in the Comments column, and a blue "Annotate" button. At the bottom of the main area are pagination controls: "Rows per Page" set to 10, "1-1 of 1", and navigation arrows. The top right corner of the interface includes language selection ("EN ▾"), project management ("Projects"), and a more options menu (three dots).

FIGURE H.2 – Page du corpus de la plate-forme Doccano.

The screenshot shows the Doccano web interface. On the left, a sidebar menu includes options like Home, Dataset, Labels (which is selected), Members, Comments, Guideline, Statistics, and Settings. A prominent blue button labeled "Start Annotation" is at the top of the sidebar. The main content area is titled "Echantillon REM". It features a table for managing labels, with columns for Name, Shortkey, Color, and Actions. The table contains four entries: PER (Shortkey: #2196F3, Color: blue), ORG (Shortkey: #B71C1C, Color: red), LOC (Shortkey: #FFCA28, Color: yellow), and MISC (Shortkey: #2E7D32, Color: green). The interface also includes a search bar, a "Rows per Page" dropdown set to 10, and navigation arrows for pagination.

	Name	Shortkey	Color	Actions
<input type="checkbox"/>	PER	#2196F3	<span style="background-color: #2196F3; width: 15px; height: 15px; display: inline-block;"></span>	<span style="color: #2196F3;">✎</span>
<input type="checkbox"/>	ORG	#B71C1C	<span style="background-color: #B71C1C; width: 15px; height: 15px; display: inline-block;"></span>	<span style="color: #B71C1C;">✎</span>
<input type="checkbox"/>	LOC	#FFCA28	<span style="background-color: #FFCA28; width: 15px; height: 15px; display: inline-block;"></span>	<span style="color: #FFCA28;">✎</span>
<input type="checkbox"/>	MISC	#2E7D32	<span style="background-color: #2E7D32; width: 15px; height: 15px; display: inline-block;"></span>	<span style="color: #2E7D32;">✎</span>

FIGURE H.3 – Page des étiquettes pour les EN de la plate-forme Doccano.

☰ Echantillon REM

EN Projects ⋮

**Start Annotation**

LOC PER LOC LOC  
**Oussonpar** **Geordes Ernert Léou** dt à **Paris**, **rue Lamarch 144**, à sonpère

PER PER LOC LOC  
**Lecrosnier** (etlivret de **Caisste d'Epargne de Paris**, de157.19, aunonde **Pierre Pugute**

PER ORG PER  
) décédé en sondomicilé à **P Paris**, **rue Descartes 21**, le 16 8bre 1900

PER PER LOC LOC  
**Lettéron** ctlespiénond de **Louis Cmille** ) décédé ensondouï à **Villneuse Heorges**

PER LOC  
, le 15 7bre 187

PER PER LOC PER  
**Oupont** (pr **Léonie Clhrtine Détavie** **rue St Lazare 82**, à **Honoré Bouscaratets**

LOC LOC  
Jeune, a **ntelaux** ( **Anepon** età **Cchille Pareau**, **rue Guyot 17** debontique aud nlier

PER PER  
**de lecanda y Mondicta** (t **Joaquina Polores Conceptionde Orbeta y Atguirre**

PER LOC LOC  
V° Mannel) décédée à **Bilbas**. **23 celle de Ptodebavriéta** le 25 Mai1900

PER PER LOC PER  
**Théonle** (par **Alfred Slain Geordes** ) dt à **Paris**, **rue Brag. 4**, à1o nee

PER PER LOC LOC  
**Deeuit** (par **Jaéol Guillaume** ) dtà **Paris**, **rue Falévy 6**, deonbeusdtales appt à

PER  
**Mme de Boissieu**

PER PER  
**de Mrangfiès** (de **Marie Gertaide bout**, épouse de **Clodoinis CMort** ) dt à

PER  
**Paris**, **rue du Frantouze 22**

FIGURE H.4 – Page d'annotation de la plate-forme Doccano.

hugo-scheithauer: Lectaurep-template/entry\_FRAN\_0025\_5038\_L-0.txt 1-16 / 16 lines [doc 5 / 6]

Layer: Named entity

Annotation: No annotation selected

1 An 1911, mois de Mai~  
 PER PER LOC LOC LOC PER

2 Gauvin-des reprises de M^me Suzanne Louise farnet spe au Mans (Sarthe) rue dus chalets 31 épouse séparée de corps de M. Auguste Aleandre~  
 PER

3 Gauvin de l'Etat Liquidatif sur enonce~

4 Tricon (au prêt consenti par le Crédit foncier à Benjamin Joseph Maurice 1 rue de Paris boulevard de Courcelles 50 d'un prêt de 90000~  
 PER ORG PER LOCATION LOC LOC PER

5 Roge par M^me Charlotte Amelia Félinde Buron proprésitaire a Paris rue Leon Coqniel 11 veuf Joeph Wilfrid Arnand en bl. pour interveur a uacte secession et detransport~  
 PER PER LOCATION LOC LOC PER

6 Dalenne (par M^me Syanne Nathalie Rogé rentière a Paris rue Léon Cogniet 13, veuve de M. le Baron Jean Baptiste Alfred) en bl. pour ceder et transférer créance~  
 PER PER LOCATION LOC LOC PER

7 Levyy par M^me Ronne Bloch épouseauguste Salomon p^t a Paris rue Beri 24 a son mari pour bendtre terrain a Paris avenue de Suffren et rue~  
 PER PER LOCATION LOC LOC LOCATION LOC LOC

8 Fribre (a Mme Jeanne Estelle Séraphine Borget) époux de M. Joseph Paul dessinateur a Paris 19 rue Montchanin 19 par le Credit foncier de 500000 ^ pour 75 ans, hypothèque sur un immeuble à Paris 19 rue Montchanin~  
 LOCATION LOC LOC ORG

FIGURE H.5 – Page d'annotation de la plate-forme Inception, avec les suggestions du recommandeur, en gris.

The screenshot shows the INCEPTION platform interface. At the top, there is a navigation bar with links for 'Projects' and 'Dashboard'. On the right side of the top bar, there are user information and a 'Logout' button. Below the top bar, the main title of the project is 'Lectaurep-template'. To the right of the title is a red button labeled 'Delete project'. On the far left, there is a vertical sidebar with various icons representing different project components. The 'Layers' icon is highlighted. The main content area is divided into several sections: 'Layers' (with a dropdown menu for 'Import' and 'Create'), 'Layer Details' (containing 'Properties' like Name and Description, and 'Technical Properties' like Internal Name and Type), and 'Features' (a list of key-value pairs). The 'Named entity' layer is currently selected, as indicated by a red highlight around its name in the 'Layers' list and its details in the central 'Layer Details' section.

FIGURE H.6 – Page des *layers* de la plate-forme Inception.

The screenshot shows the INCEPTION platform interface. At the top, there is a navigation bar with links for 'Projects' and 'Dashboard'. On the right side of the top bar, there are user information and logout links. The main title of the page is 'Lectaurep-template'. On the left, there is a sidebar with various icons representing different features like projects, dashboard, and help. The main content area is titled 'Recommenders' and contains a list with one item: '[Named entity@value] String Matcher'. This item is highlighted with a red background. To the right of the list is a 'Details' panel. The 'Name' field is set to '[Named entity@value] String Matcher' with an 'auto-generate' checkbox checked. The 'Enabled' checkbox is checked. Under 'Layer', it is set to 'Named entity'. Under 'Feature', it is set to 'value'. Under 'Tool', it is set to 'String Matcher'. The 'Activation strategy' section has a checked checkbox for 'Always active (no evaluation)'. The 'Max. recommendations' field contains the number '3'. The 'States used for training' section lists several options: 'Annotation not started yet (new)', 'Annotation in progress', 'Annotation finished', and 'Document not available for annotation (locked)'. The 'Case insensitive' checkbox is unchecked. The 'Gazeteers' section includes a 'Import gazeteers ...' button with a folder icon.

Technische Universität Darmstadt -- Computer Science Department – INCEPTION – 0.19.6 (2021-06-08 16:59:33, build 36a1032b)

FIGURE H.7 – Page des recommandeurs de la plate-forme Inception.

INCEPTION Projects Dashboard

Help hugo-scheithauer Log out 29 min

# Lectaurep-template

Delete project

Tagsets Import Create

Coreference mentions  
Coreference relations  
Dependency flavors  
Named Entity tags  
Operation  
**Tagset\_Lectaurep**  
UD Universal Dependencies (v2)  
UD Universal POS tags (v2)

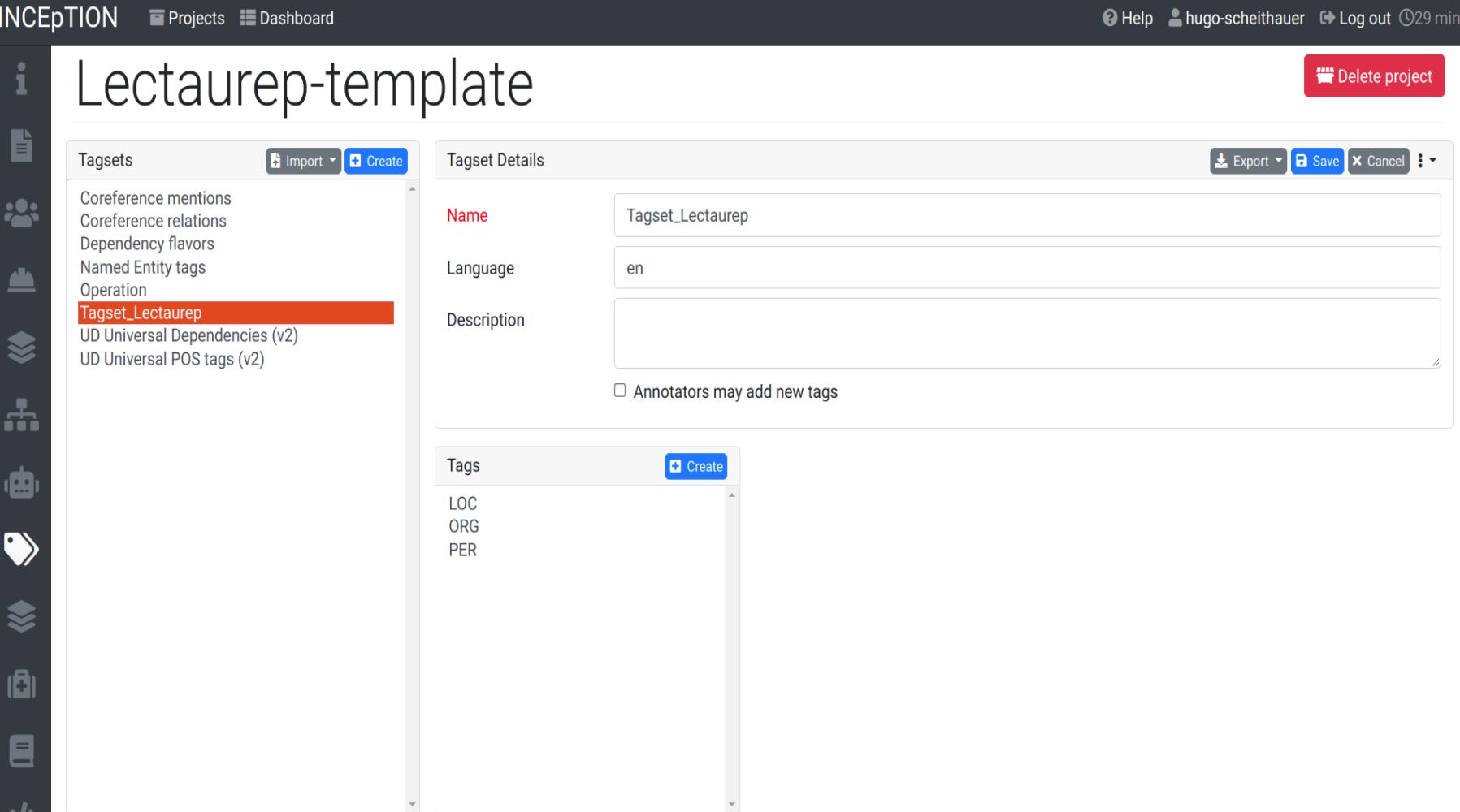
Tagset Details

Name Tagset\_Lectaurep  
Language en  
Description

Annotators may add new tags

Tags Create

LOC  
ORG  
PER



213

Technische Universität Darmstadt -- Computer Science Department -- INCEPTION -- 0.19.6 (2021-06-08 16:59:33, build 36a1032b)

FIGURE H.8 – Page des tagsets de la plate-forme Inception.

INCEPTION Projects Dashboard Help hugo-scheithauer Log out 29 min

hugo-scheithauer: Lectaurep-template/entry\_FRAN\_0025\_5038\_L-0.txt 1-16 / 16 lines [doc 5 / 6]

**Layer:** Named entity

**Annotation:** No annotation selected

**Export**

**Format:**

- CoNLL 2000
- CoNLL 2002
- CoNLL 2003
- CoNLL 2006
- CoNLL 2009
- CoNLL 2012
- CoNLL CoreNLP
- CoNLL-U
- Inline XML
- LAPPS Interchange Format

**Buttons:** Export, Cancel

Text content (lines 1-8):

- An 1911, mois de Mai~
- Gauvin-des reprises de M^me Suzanne Louise farnet sp~
- Gauvin de l'Etat Liquidatif sur enonce~
- Tricon (au prêt consenti par le Crédit foncier à Benjamin
- Roge par M^me Charlotte Amelia Féline Buron propres secession et detransport~
- Dalenne (par M^me Syanne Nathalie Rogé rentière a transférer créance~
- Levy par M^me Ronne Bloch épouseauguste Salomon rue~
- Fribre (a Mme Jeanne Estelle Séraphine Borget) époux de M. Joseph Paul dessinateur a Paris 19 rue Montchanin 19 par le Credit foncier de 500000 ^ pour 75 ans, hypothèque sur un immeuble à Paris 19 rue Montchanin~

FIGURE H.9 – Fenêtre présentant les formats d'exports de la plate-forme Inception.

Listing H.1 – Extrait d'une annotation au format CoNLL 2002 d'une page pré-traitée d'un répertoire de notaire.

1	Gauvin B-PER	1	Roge B-PER
2	de O	2	par O
3	l'Etat O	3	M B-PER
4	Liquidatif O	4	^ I-PER
5	sur O	5	me I-PER
6	enonce O	6	Charlotte I-PER
7	— O	7	Amelia I-PER
8	Tricon B-PER	8	Felinte I-PER
9	( O	9	Buron I-PER
10	au O	10	proprietaire O
11	pret O	11	a O
12	consenti O	12	Paris B-LOC
13	par O	13	rue B-LOC
14	le O	14	Leon I-LOC
15	Credit B-ORG	15	Coqniel I-LOC
16	foncier I-ORG	16	11 I-LOC
17	a O	17	veuf O
18	Benjamin B-PER	18	Joseph B-PER
19	Joseph I-PER	19	Wilfrid I-PER
20	Maurice I-PER	20	Arnand I-PER
21	1 B-LOC	21	en O
22	rue I-LOC	22	bl O
23	de I-LOC	23	. O
24	Paris I-LOC	24	pour O
25	boulevard B-LOC	25	interveur O
26	de I-LOC	26	a O
27	Courcelles I-LOC	27	uacte O
28	50 I-LOC	28	secession O
29	d'un O	29	et O
30	pret O	30	detransport O
31	de O	31	— O
32	90000 O		
33	— O		

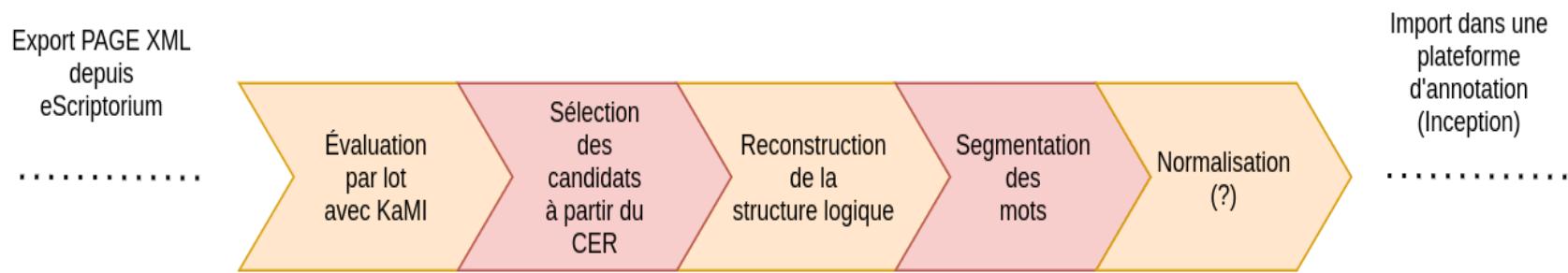


FIGURE H.10 – Chaîne de pré-traitement proposée pour annoter des données en vue de l'entraînement/affinage d'un modèle de REN.

## **Annexe I**

### **Évaluer par lots les transcription automatiques du projet LECTAUREP avec la librairie KaMI**

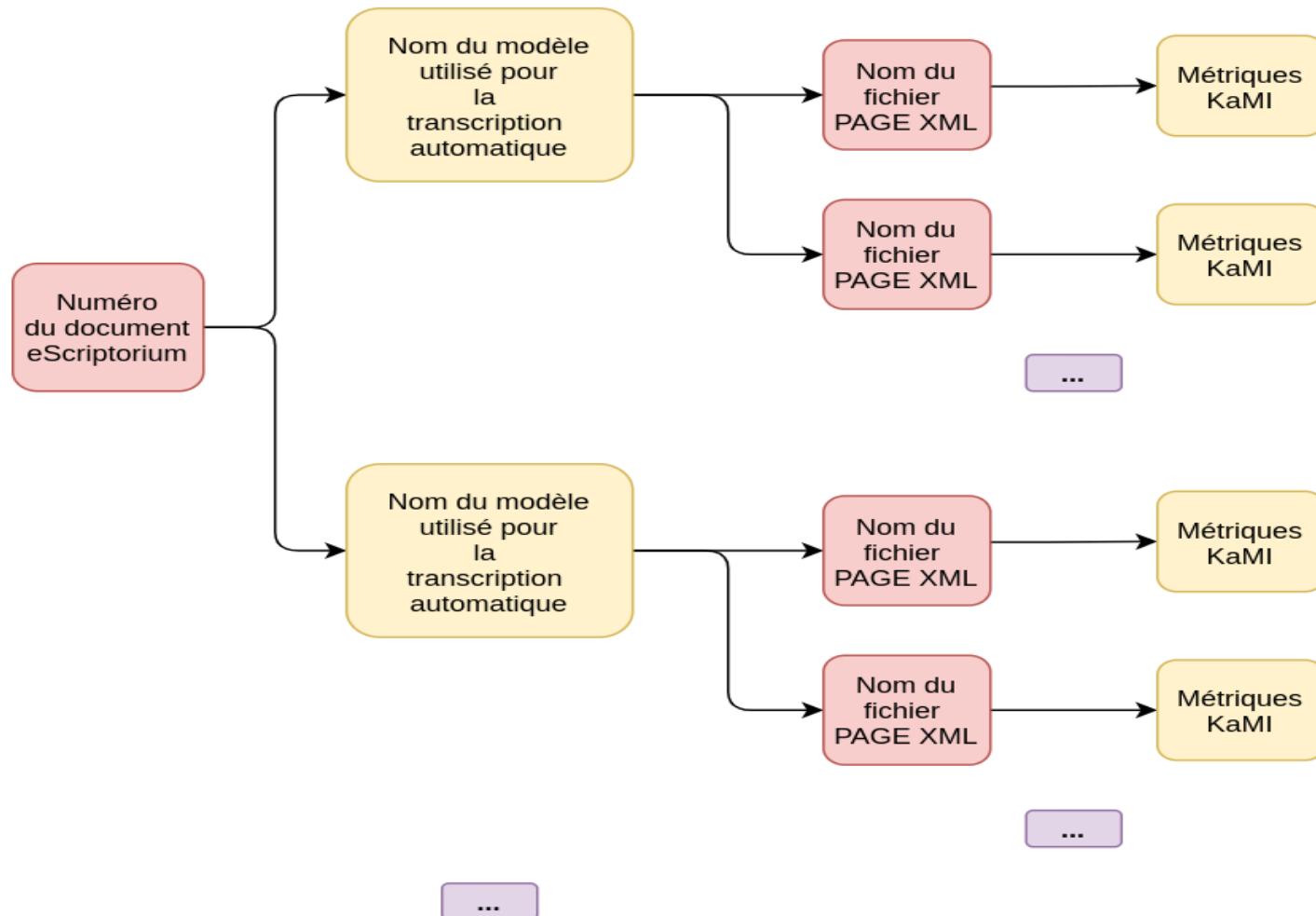


FIGURE I.1 – Schéma du JSON résultant du CLI développé pour évaluer par lots les transcription automatiques du projet LECTAUREP avec la librairie KaMI.

## **Annexe J**

### **Modéliser les répertoires des notaires en TEI**

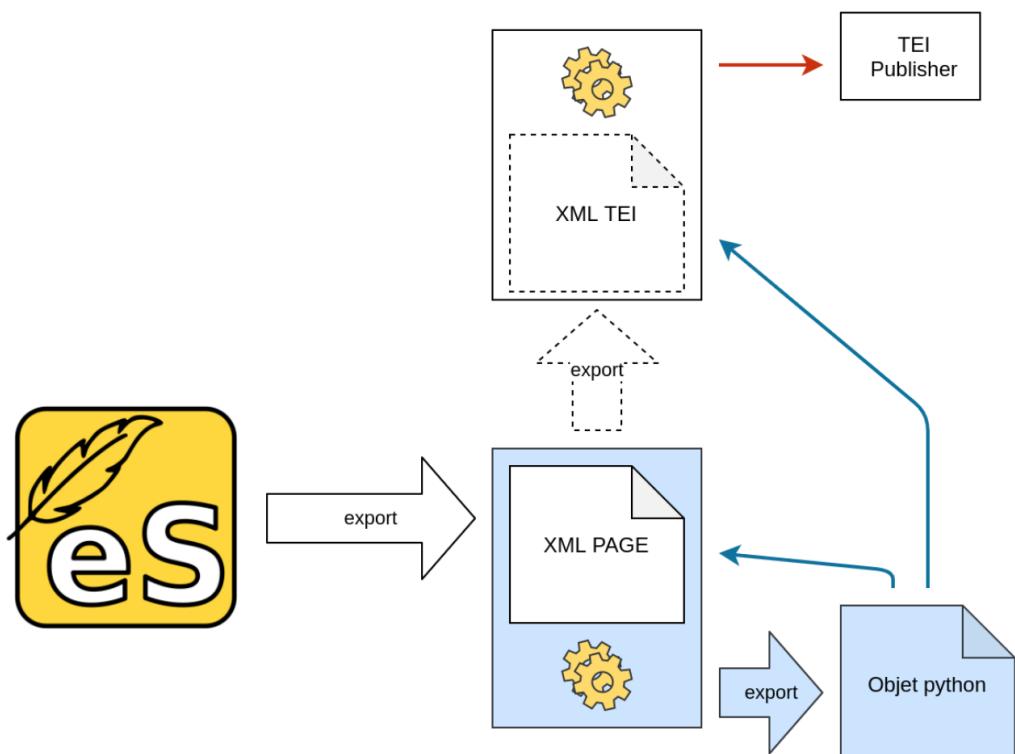


FIGURE J.1 – Chaîne de traitement de la transformation de l'export PAGE XML issu de la transcription et de sa transformation XML TEI. Modifications apportés au schéma originalelement créé par Alix Chagué, source : [https://gitlab.inria.fr/almanach/lectaurep/documentation/-/issues/17#note\\_528453](https://gitlab.inria.fr/almanach/lectaurep/documentation/-/issues/17#note_528453) (consulté le 16/08/21).

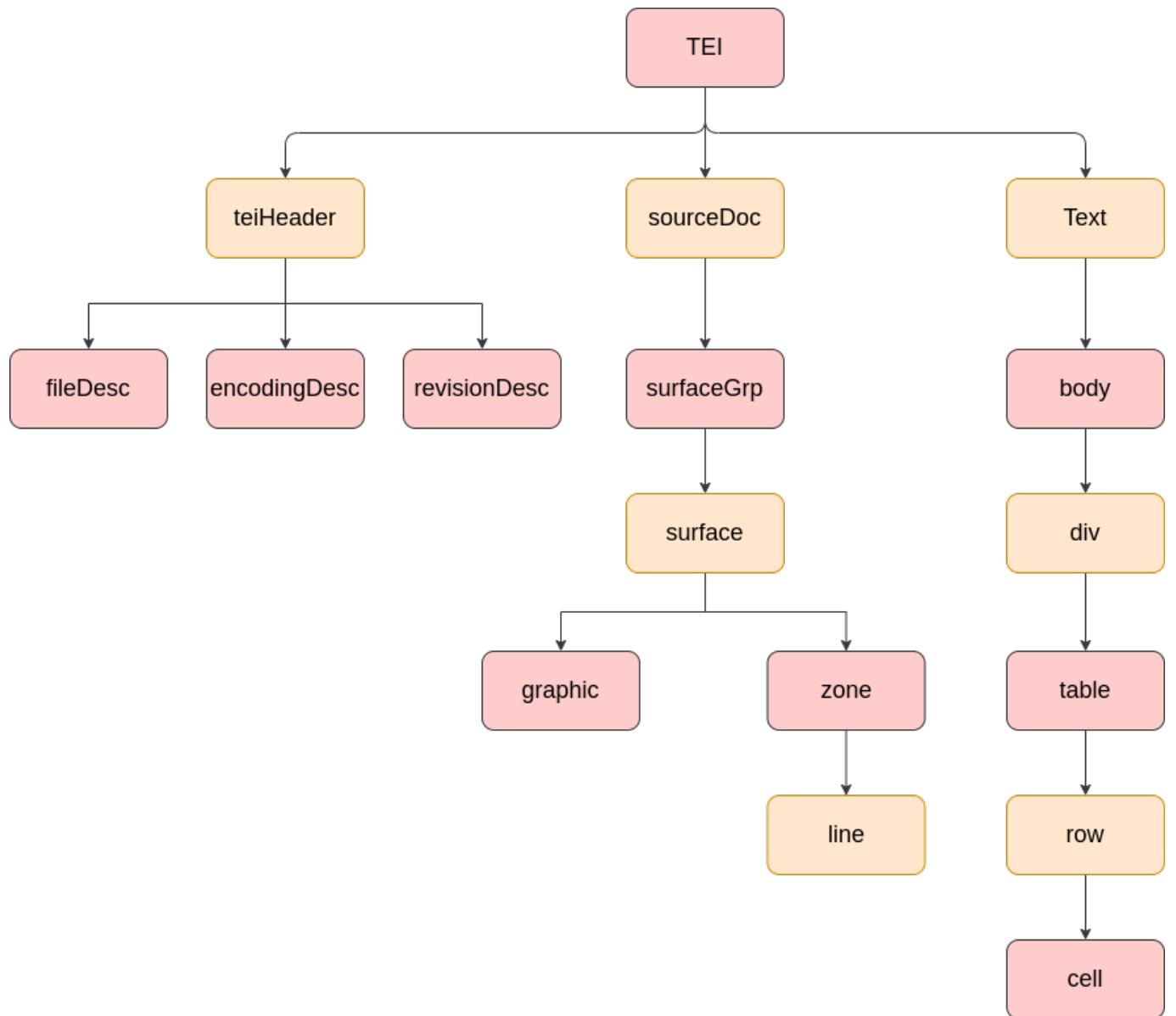


FIGURE J.2 – Arbre XML TEI de la modélisation des répertoires des notaires.



## **Annexe K**

**Publier la transcription des réertoires  
des notaires sur la plate-forme  
*open-source* TEI Publisher**

exist

File Edit Navigate Application XQuery XML Help | Logged in as admin.

New New XQuery Open Save Close Eval Run

Current app: tei-publisher File Type: XML

outline directory

db

- apps
  - dashboard
  - doc
  - eXide
  - fundocs
  - lectaurep-demo
  - lectaurep\_demo
  - markdown
  - monex
  - packageservice
  - shared-resources
- tei-publisher
  - .devcontainer
  - data
    - doc
    - playground
      - DAFANCH96\_023MIC07633\_L-0-tei.xml
      - DAFANCH96\_023MIC07645\_L-0-tei.xml
      - FRAN\_0025\_0227\_L-0-tei.xml
      - FRAN\_0025\_1290\_L-1-tei.xml
      - FRAN\_0025\_3056\_L-0-tei.xml
      - FRAN\_0025\_3657\_L-1-tei.xml
      - FRAN\_0025\_4648\_L-1-tei.xml
      - FRAN\_0025\_5094\_L-1-tei.xml
      - FRAN\_0025\_5795\_L-1-tei.xml
      - FRAN\_0025\_6067\_L-1-tei.xml
    - temp
    - test
    - about.md
    - collection.html
    - demo.png
    - documentation.png
    - playground.png
    - taxonomy.xml
  - modules
  - odd
  - resources
  - templates
  - test

Filter by...

FRAN\_0025\_1290\_L-1-tei.xml\* lectaurep\_tem... DAFANCH96\_023... lectaurep\_cel...

```

1270    <lb facs="#eSc_line_1820d578" n="3"/>
1271    Montigny
1272    </cell>
1273    <cell n="6" role="col6">
1274      <lb facs="#eSc_line_79f940c3"/>
1275      <date when-iso="22">
1276        22
1277      </date>
1278      </cell>
1279      <cell n="7" role="col7">
1280        <lb facs="#eSc_line_187af570"/>
1281        35-
1282      </cell>
1283      <cell n="8" role="misc"/>
1284    </row>
1285    <row>
1286      <cell n="1" role="col1">
1287        <lb facs="#eSc_line_64947134"/>
1288        294
1289      </cell>
1290      <cell n="2" role="col2"/>
1291      <cell n="3" role="col3">
1292        <lb facs="#eSc_line_4ddec319"/>
1293        16
1294      </cell>
1295      <cell n="4" role="col4">
1296        <lb facs="#eSc_line_e2d2d4b4"/>
1297        Vente
1298      </cell>
1299      <cell n="5" role="col5">
1300        <lb facs="#eSc_line_920c6d23" n="1"/>
1301        Pierrat et Anderson à Jules Menny à Drancy 11 rue Daisy
1302        <lb facs="#eSc_line_3b56b155" n="2"/>
1303        de 297f 02 terrain étoile Drancy n°203 du plan XXX A. P n°110
1304        <lb facs="#eSc_line_4e45652e" n="3"/>
1305        prix 9756,65
1306      </cell>
1307      <cell n="6" role="col6">
1308        <lb facs="#eSc_line_12ccdae2"/>
1309        <date when-iso="23">
1310          23
1311        </date>
1312      </cell>
1313      <cell n="7" role="col7">

```

/db/apps/tei-publisher/data/playground/FRAN\_0025\_1290\_L-1-tei.xml

An invalid XML character (Unicode: 0x1) was found in the element content of the document.

FIGURE K.1 – Capture d'écran de la base de données exist-db, et d'un encodage TEI d'une transcription d'une page d'un répertoire de notaire.

The screenshot shows the tei Publisher interface for editing an ODD (Open Document Definition) file named "Lectaurep - demo.odd". The top navigation bar includes links for Start, Documentation, News, and a user icon. On the right, there are Language (English) and Login options. The main workspace displays a hierarchical structure of XML elements: teiHeader, div, and graphic. The "graphic" element is currently selected, indicated by a blue underline. Below this, a detailed view of the "graphic" element spec is shown, including fields for Output (Description: [Document the model]), Predicate (1 [Define further conditions that have to be met (in xquery)]), behaviour (inline or Custom Behaviour), CSS Class ([Define CSS class name (for external CSS)]), and Template (1 <pb-facs-link facs="[[url]]" emit="transcription"/>). A sidebar on the left lists other element specs: div, graphic, and teiHeader. The number 225 is visible near the bottom left corner.

FIGURE K.2 – Page de l’éditeur d’ODD dans TEI Publisher

The screenshot shows the tei Publisher interface. At the top, there is a navigation bar with links for "Start", "Documentation", "News", and "Enter Query". There is also a search icon and a language selection dropdown set to "English". On the far right, there is a "Login" button.

The main content area displays a list of items. At the top left, there are "Sort by" and "Filter by" dropdown menus, both currently set to "Title". Below these, a page navigation bar shows "1" of 10 items found. A "GO TO PARENT" link is available above the first item.

The first item in the list is "DAFANCH96\_023MIC07633\_L-0-tei", which includes links for "Lectaurep" and "Licence", and a "DOWNLOAD" button.

The second item is "DAFANCH96\_023MIC07645\_L-0-tei", also with "Lectaurep" and "Licence" links and a "DOWNLOAD" button.

The third item is "FRAN\_0025\_0227\_L-0-tei", also with "Lectaurep" and "Licence" links and a "DOWNLOAD" button.

To the right of the list, there is a sidebar with several entries, each with a link and a "↔" icon:

- cell width
- Viewer\_metadata
- lectaurep\_custom.css
- Van Gogh Letters
- DOCX Import
- Docbook v5
- Processing of Docbook format
- Shakespeare Plays
- Serafin Letters
- 15th c. manuscript correspondence with parallel translation
- DOCX Output Preview
- A test ODD to preview TEI imported from Word docx
- Standard Default
- TEI-C processing defaults; this ODD is a starting point for TEI Publisher ODD chaining
- Dantiscus' Letters

FIGURE K.3 – Parcourir les pages des répertoires des notaires : navigation à travers une collection.

FRAN\_0025\_1290\_L-1-tei

Numéros du répertoire	Dates des actes brevets	Actes en P. V. ouv. Cof.	Actes en minutes	Noms, prénoms et domiciles des parties ; indication, situations et prix des biens	Date de l'enregistrement	Droits de l'enregistrement	Autres
292	15			Bajul au décès de Désiré Alexandre Marius à Romainville 7 et 9 rue des Oseraies arrivé le 10 8 <sup>bre</sup> 1938	22	- -	
293	15		Inventaire	Jadin au décès de Dieudonné Joseph Adolphe à Paris 6 rue des Batigolles y arrivé le 22-12-1938 ep <sup>^x</sup> de Henriette Eugénie Montigny	22	35 -	
294	16	Vente		Pierrat et Anderson à Jules Menné à Drancy 11 rue Daisy de 297 <sup>^f</sup> 02 terrain étoile Drancy n°203 du plan XXX A. P n°110 prix 9756,65	23	1747 04	
295		17 Certif Sign.		Baer concé Berthe Isabelle Henriette Bourdin ep <sup>^se</sup> de Richard	23	35 -	

FIGURE K.4 – Visualisation d'une page transcrise d'un répertoire des notaires, sans feuille de style (1).

tei Publisher Start Documentation News Download GitHub Enter Query Language English Login

≡ ☰ 🔍 🔍 ≡

plan de vente  
prix 9756,65

295	17	Certif Sign.	Baer conc Berthe Isabelle Henriette Bourdin ep^se de Richard Philippe André à Paris 20 B^d de Courcelles	23	35 -
296	17	Procuration	Cote p. Raymond à Paris 6 rue d'Amsterdam en bl. p^r vendre	23	35 -
297	17	Certi ppté	Hultz conc prorat. div. pens. au nom de Louis -		- -
298	17	Dépôt	Seba de grosse env. en poss. SS^ion de Suzanne Rebecca Rachel Carvallo à Paris 200 fg S^t Denis V^e de Fernand	22	35 -
299	17	Déliv de Legs	Raer p. Berthe Isabelle Henriette Bourdin epse de Richard Philippe André à Paris 20 B^d de Courcelles à Angele Miller  à Paris 81 rue de la Pompe Villa Hérau Janson entre Georges Marcel Marie à Orbec	23	275 55

Contenu Calendrier 125

FIGURE K.5 – Visualisation d'une page transcrise d'un répertoire des notaires, sans feuille de style (2).

```
table {
    border-collapse: collapse;
    font-size: smaller; background-color: #F0F0F0;
}

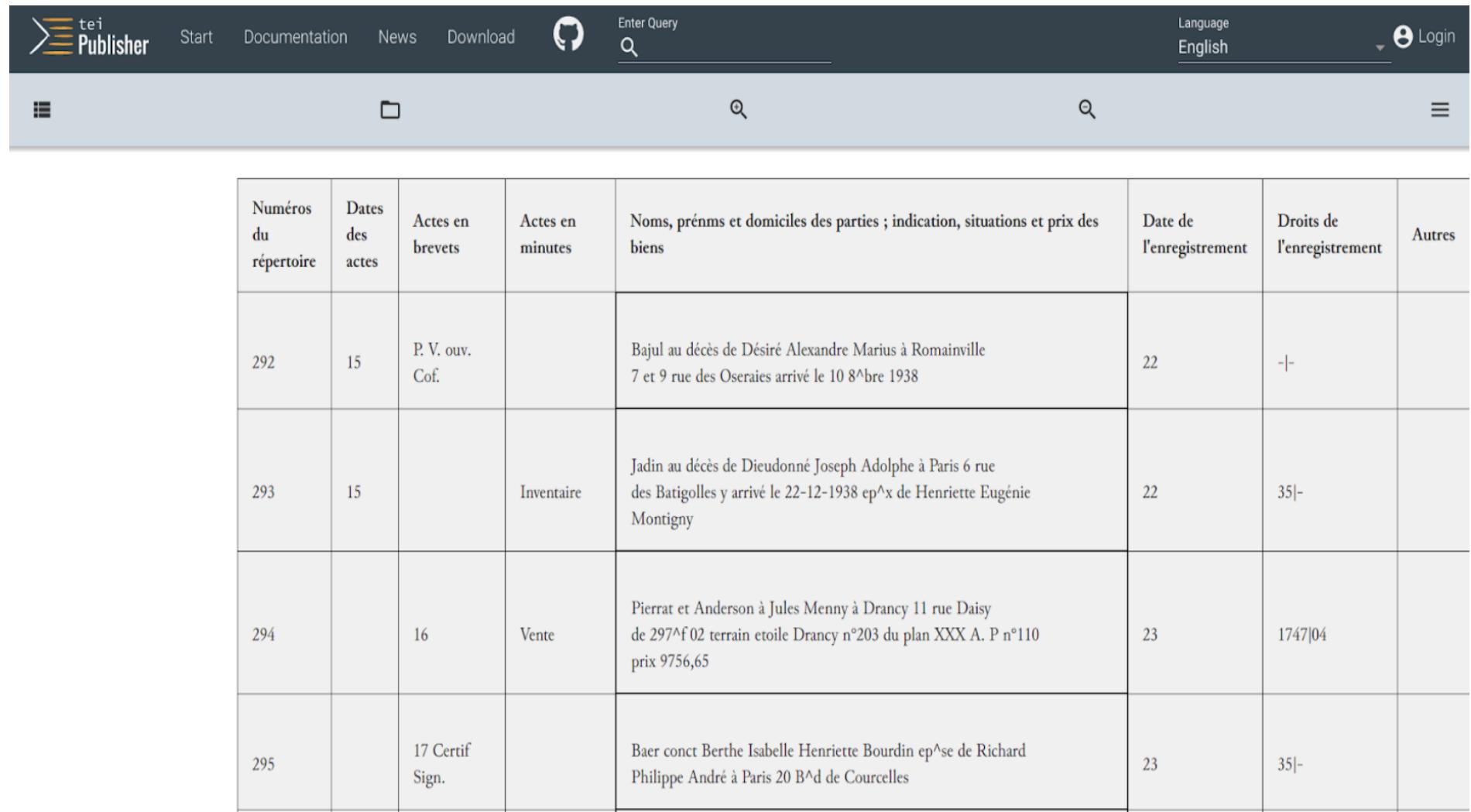
table, th, td {
    border: 1px solid black;
}

tr:hover {
    background-color: #f5f5f5;
}

th, td {
    padding: 15px;
    text-align: left;
}
```

FIGURE K.6 – Feuille CSS pour afficher une transcription d'une page d'un répertoire de notaire sous forme de tableau

230



The screenshot shows the tei Publisher interface with a dark header bar. The header includes the logo 'tei Publisher' with three orange bars, navigation links 'Start', 'Documentation', 'News', 'Download', a GitHub icon, 'Enter Query' with a magnifying glass icon, 'Language English', and a 'Login' button.

The main area features a table with the following columns:

- Numéros du répertoire
- Dates des actes
- Actes en brevets
- Actes en minutes
- Noms, prénoms et domiciles des parties ; indication, situations et prix des biens
- Date de l'enregistrement
- Droits de l'enregistrement
- Autres

The table contains four rows of data:

292	15	P. V. ouv. Cof.		Bajul au décès de Désiré Alexandre Marius à Romainville 7 et 9 rue des Oseraies arrivé le 10 8 <sup>bre</sup> 1938	22	- -	
293	15		Inventaire	Jadin au décès de Dieudonné Joseph Adolphe à Paris 6 rue des Batigolles y arrivé le 22-12-1938 ep <sup>x</sup> de Henriette Eugénie Montigny	22	35 -	
294		16	Vente	Pierrat et Anderson à Jules Menné à Drancy 11 rue Daisy de 297 <sup>f</sup> 02 terrain étoile Drancy n°203 du plan XXX A. P n°110 prix 9756,65	23	1747 04	
295		17 Certif Sign.		Baer conc't Berthe Isabelle Henriette Bourdin ep <sup>se</sup> de Richard Philippe André à Paris 20 B <sup>d</sup> de Courcelles	23	35 -	

FIGURE K.7 – Visualisation d'une page transcrise d'un répertoire des notaires, avec une feuille de style basique.

tei Publisher Start Documentation News Download Enter Query Language English Login

FRAN\_0025\_1290\_I-1-tei

292	15	P. V. ouv. Cof.	Bajul au décès de Désiré Alexandre Marius à Romainville 22 7 et 9 rue des Oseraies arrivé le 10 8 <sup>bre</sup> 1938	- -
231			Jadin au décès de Dieudonné Joseph Adolphe à Paris 6 rue des Batignolles y arrivé le 22-12- 1938 ep <sup>x</sup> de Henriette Eugénie Montigny	35 -
293	15	Inventaire	Pierrat et Anderson à Jules	

**An 1939 , mois de Mars**

Bajul au décès de Désiré Alexandre Marius à Romainville  
et 9 rue des Oseraies arrivé le 10 8<sup>bre</sup> 1938 22

Martin au décès de Dieudonné Joseph Adolphe à Paris 6 rue  
des Batignolles y arrivé le 22-12-1938 épouse Henriette Eugénie  
Montigny 22

Pierrat et Anderson à Jules Merny à Drancy 11 rue Saïy  
de 197<sup>e</sup> 12 terrains, étoile Drancy n° 20 3 du plan S<sup>ur</sup> A. P. H. 110  
mis 9756,65 23 17

Raer court Berthe Isabelle Henriette Bourdin épouse Richard  
Philippe André à Paris 20<sup>e</sup> de Courcelles 23

Côte f. Raymond à Paris 6<sup>e</sup> 5 Amsterdam en bl.  
française 23

Holtz court prorat dir. spes au nom de Louis  
Selba de grosse eau en post. 11<sup>e</sup> de Suzanne Rebecca 23

Rachel Carvalho à Paris 20<sup>e</sup> 1<sup>e</sup> Denis 7<sup>e</sup> de Fernand 22

Raer p. Berthe Isabelle Henriette Bourdin, épouse Richard  
Philippe André à Paris 20<sup>e</sup> de Courcelles à Angèle Nilla  
à Paris 81 rue de la Pompe Villa Heran 23 2

Depôt

éch. de legs

FIGURE K.8 – Visualisation confrontant la transcription d'une page d'un répertoire de notaires avec sa numérisation grâce à l'utilisation du protocole IIIF.

The screenshot shows the tei Publisher application interface. At the top, there is a navigation bar with links for 'Start' and 'Documentation'. On the right side of the bar are 'Language English' and a 'Login' button. Below the navigation bar, a search bar contains the query 'Dépôt'. The main content area displays search results across four documents:

- 1 DAFANCH96\_023MIC07645\_L-0-tei**
  - ... Suite du 8 juillet 1869 [Dépôt](#) An 1875 mois de Mai ...
  - ... 29 [Dépôt](#) Carmona (deprocuration donn...)
  - ... Suite du 19 avril 75 [Dépôt](#) Bertaux (depièces de publicic...)
  - ... 1 [Dépôt](#) n°33, à (en blanc) à l'effe...
- 2 FRAN\_0025\_5795\_L-1-tei**
  - ... 13 [Dépôt](#) du testament ...
  - ... Contrat creud^er [dépôt](#) à veuve deXXXtaire ...
- 3 FRAN\_0025\_4648\_L-1-tei**
  - ... 27 [Dépôt](#) frais préalables + loyers a...
  - ... Acceptation Command [Dépôt](#) de Décharge Breuilheid (con...
  - ... 4 [Dépôt](#) état Matériel Mainlevée ...
- 4 DAFANCH96\_023MIC07633\_L-0-tei**
  - ... 8 [Dépôt](#) Guell y Baro (de Son donnée...

FIGURE K.9 – Résultats d'une recherche du mot « dépôt » parmi les documents ajoutés à l'instance locale installée sur mon ordinateur.

tei Publisher		Start	Documentation	News	Download	 Enter Query	 Dépôt	Language English	Login
									
441	29	suite des 3 & 4 mai 1875	Arrêté de compte de tutelle	de la Cie des) Poutrel (entre francoise Eugénie Deton veuve de Louis Denis) demt	1	3.75			
442	29	Dépôt		à Paris, rue Oberkampf n°42 etses trois enfants		16 50			
443	29	Procuration		Carmona (deprocuration donnée par Dolorès Arriago epouse de Jorge) demeurant ) Mexico à Abaroa Uriarren & Goguel	29	375			
				demeurant à Paris rue de Richelieu 102, pour Vendre de rentes & valeurs					
444	29	Rectification		Manoury (par Ferdinand) demeurant à Paris, boulevart des Capucines. n°6, & demelles Eugénie & Joséphine Delasalle, demeurant					
				à Levallois Perret, rue Valentin, n°10, Prudhon (entre hippolyte Eudamidas)					
				Rectification demeurant à Neuilly avenue de Villiers n°21 et Eugène Louis Duhamel, demeurant à	1	375			

FIGURE K.10 – Signalement du résultat d'une recherche du mot « dépôt » parmi les documents ajoutés à l'instance locale installée sur mon ordinateur.

 Start Documentation  Enter Query

Language English 

1 Found 15 items

... sur prix de vente du 22 octobre 1822 et qui visiter ... [Search](#) Josephine de Beauharnais ...

<b>6</b>	DAFANCH96_023MIC07633_L-0-tei
...t à Paris rue Caumartin n°75 et Adèle) ve de <a href="#">Jules</a> Laurence dt au même lieu à l'effet de tou...	
...rue Delaborde n°11 le 29 9bre 1874 de Pierre <a href="#">Jules</a> ) en son vivant architecte ...	
...nderest (par jean Joseph marie) et Augustine <a href="#">Julie</a> Hauté sa fe dt à Paris rue Lecourbe 84 à Fçois ...	
... Paviset et Cie (requête de Etienne <a href="#">jules</a> Giraudeau dt à Paris rue de Londres n°56 ...	
<b>7</b>	FRAN_0025_6067_L-1-tei
...vie à Villemomble av. du Raincy 37 et V <sup>e</sup> de <a href="#">Jules</a> Joseph Guilhem Cros legataire universelle de Alfred) de ...	
... Cession de bail Lerminez (par <a href="#">Jules</a> ) et son ép. à Paris rue de Charonne 71 et ...	
...ges) à Paris aven. Philippe Auguste 52 et de <a href="#">Julie</a> Archimbault à S <sup>t</sup> Mandé G <sup>de</sup> rue N° 110 ...	
<b>8</b>	FRAN_0025_5094_L-1-tei
... Bapst, (par M. Louis Auguste <a href="#">Julien</a> ) cap. en retraite chev. dela leg. d'honne...	
...f <sup>os</sup> de C <sup>ie</sup> de boulangerie Pablis appar <sup>t</sup> a <a href="#">Julienne</a> Deschamps, veuve de M. Lazare) ayant demeuré dans ...	
...ration de corps et d'entre M Charles Gustave <a href="#">Jules</a> Simon dit Jules) attaché au Sénat à Paris 19 rue Turgot ...	
...et d'entre M Charles Gustave Jules Simon dit <a href="#">Jules</a> ) attaché au Sénat à Paris 19 rue Turgot et ...	

Enter Query

Search sections  Search headings

**Genre**  Documentation 4  Letters and Correspondence 1  Show top 50

**Language**  French 1  Show top 50

FIGURE K.11 – Résultats d'une recherche à troncature « Jul\* » parmi les documents ajoutés à l'instance locale installée sur mon ordinateur.

# Bibliographie

- ADAK (Chandranath), CHAUDHURI (Bidyut B.) et BLUMENSTEIN (Michael), « Named Entity Recognition from Unstructured Handwritten Document Images », dans *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, 2016, p. 375-380, DOI : 10.1109/DAS.2016.15.
- ADNAN (Kiran) et AKBAR (Rehan), « Limitations of information extraction methods and techniques for heterogeneous unstructured big data », *International Journal of Engineering Business Management*, 11 (1<sup>er</sup> janv. 2019), Publisher : SAGE Publications Ltd STM, p. 1847979019890771, DOI : 10.1177/1847979019890771.
- AKBIK (Alan), BERGMANN (Tanja), BLYTHE (Duncan), RASUL (Kashif), SCHWETER (Stefan) et VOLLGRAF (Roland), « FLAIR : An Easy-to-Use Framework for State-of-the-Art NLP », dans *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Minneapolis, Minnesota, 2019, p. 54-59, DOI : 10.18653/v1/N19-4010.
- BATISTA (David), *Named-Entity evaluation metrics based on entity-level*, 9 mai 2018, URL : [http://www.davidsbatista.net/blog/2018/05/09/Named\\_Entity\\_Evaluation/](http://www.davidsbatista.net/blog/2018/05/09/Named_Entity_Evaluation/) (visité le 23/04/2021).
- BENESTY (Michaël), *NER algo benchmark : spaCy, Flair, m-BERT and camemBERT on anonymizing French commercial legal cases*, Medium, 10 déc. 2019, URL : <https://towardsdatascience.com/benchmark-ner-algorithm-d4ab01b2d4c3> (visité le 23/04/2021).
- *Why we switched from Spacy to Flair to anonymize French legal cases*, Medium, 29 sept. 2019, URL : <https://towardsdatascience.com/why-we-switched-from-spacy-to-flair-to-anonymize-french-legal-cases-e7588566825f> (visité le 23/04/2021).
- BERMÈS (Emmanuelle) et MOIRAGHI (Eleonora), « Le patrimoine numérique national à l'heure de l'intelligence artificielle. Le programme de recherche Corpus comme espace d'expérimentation pour les humanités numériques », *Revue des Sciences et Technologies de l'Information - Série RIA : Revue d'Intelligence Artificielle* (, 2019), URL : <https://hal-bnf.archives-ouvertes.fr/hal-02122073> (visité le 10/08/2021).

BERTINO (Andrea), FOPPIANO (Luca), ROMARY (Laurent) et MOUNIER (Pierre), « Leveraging Concepts in Open Access Publications », *Journal of Data Mining and Digital Humanities*, 2019 (15 juin 2020), URL : <https://hal.inria.fr/hal-01981922> (visité le 30/04/2021).

BLUCHE (Théodore), HAMEL (Sébastien), KERMORVANT (Christopher), PUIGCERVER (Joan), STUTZMANN (Dominique), TOSSELLI (Alejandro J.) et VIDAL (Enrique), « Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript Collection in the HIMANIS Project », dans *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Kyoto, France, 2017, DOI : 10.1109/ICDAR.2017.59.

BOCHET (Charles), *How to Train your Own Model with NLTK and Stanford NER Tagger ? (for English, French, German...)* Medium, 10 mai 2018, URL : <https://medium.com/sicara/train-ner-model-with-nltk-stanford-tagger-english-french-german-6d90573a9486> (visité le 12/07/2021).

BONHOMME (Marie-Laurence), *Défis et opportunités de la reconnaissance automatique d'écriture manuscrite pour les documents d'archives : l'exemple des répertoires des notaires de Paris, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Jean-Baptiste Camps*, 2018.

BRAU (Pierre-Frédéric), BOUDAUD (Sylvie), CHARPENTIER (Alix), CHARBONNIER (Pauline), CLAVAUD (Florence), VIGNAUD (Louis), CHENARD (Gaël), CROS (Céline), FOLLET-CLAVREUL (Sylviane), GRIMONT (Alex-Adriana), et al., *Guide d'indexation pour le web*, 2021, URL : [https://francearchives.fr/file/6686af73e52bd3dd7d56cbad92228977cbe576f5/GuideIndexation\\_Web\\_v202108.pdf](https://francearchives.fr/file/6686af73e52bd3dd7d56cbad92228977cbe576f5/GuideIndexation_Web_v202108.pdf).

CALVO (Miguel Romero), *Dissecting BERT Appendix : The Decoder*, Medium, 27 nov. 2018, URL : <https://medium.com/dissecting-bert/dissecting-bert-appendix-the-decoder-3b86f66b0e5f> (visité le 12/07/2021).

— *Dissecting BERT Part 1 : The Encoder*, Medium, 7 mai 2019, URL : <https://medium.com/dissecting-bert/dissecting-bert-part-1-d3c3d495cdb3> (visité le 12/07/2021).

CANDITO (Marie) et SEDDAH (Djamé), « Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical », dans, 2012, URL : <https://hal.inria.fr/hal-00698938> (visité le 24/08/2021).

CARBONELL (Manuel), VILLEGAS (Mauricio), FORNÉS (Alicia) et LLADÓS (Josep), « Joint Recognition of Handwritten Text and Named Entities with a Neural End-to-end Model » (, 16 avr. 2018).

- CAVALIÉ (DIR.) (Étienne), *L'indexation matière en transition : de la réforme de Rameau à l'indexation automatique*, ISSN : 0184-0886, Paris, France, 2019.
- CHAGUÉ (Alix), *Comment faire lire des gribouillis à mon ordinateur ?*, 2021, URL : <https://hal.archives-ouvertes.fr/hal-03170345> (visité le 06/04/2021).
- *Création de modèles de transcription pour le projet LECTAUREP #1*, LECTAUREP, 6 sept. 2021, URL : <https://lectaurep.hypotheses.org/475> (visité le 28/08/2021).
  - *Traces6 : notre serveur principal*, LECTAUREP, URL : <https://lectaurep.hypotheses.org/402> (visité le 09/08/2021).
- CHAGUÉ (Alix) et CHIFFOLEAU (Floriane), *An accessible and transparent pipeline for publishing historical egodocuments*, mars 2021, URL : <https://hal.archives-ouvertes.fr/hal-03180669> (visité le 11/08/2021).
- CHAGUÉ (Alix) et ROSTAING (Aurélia), « Présentation du projet Lectaurep (Lecture automatique de répertoires) », dans *Atelier sur la transcription des écritures manuscrites - BnF DataLab*, Paris, France, 2021, URL : <https://hal.archives-ouvertes.fr/hal-03122019> (visité le 06/04/2021).
- CHARNOCK (Ross), « Les langues de spécialité et le langage technique : considérations didactiques », *ASp. la revue du GERAS*-23 (1<sup>er</sup> déc. 1999), Number : 23-26 Publisher : Groupe d'étude et de recherche en anglais de spécialité, p. 281-302, DOI : 10.4000/asp.2566.
- CHIFFOLEAU (Floriane), *Publication of my digital edition – Working with TEI Publisher*, Digital Intellectuals, déc. 2020, URL : <https://digitalintellectuals.hypotheses.org/3912> (visité le 19/08/2021).
- *Starting a new project – Discovering its source material*, Digital Intellectuals, mars 2021, URL : <https://digitalintellectuals.hypotheses.org/3398> (visité le 19/08/2021).
- CLARK (Alexander) et ISSCO (Alexander Clark), « Pre-Processing Very Noisy Text », dans *Proc. of Workshop on Shallow Processing of Large Corpora*, 2003.
- CLÉRICE (Thibault), « Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Texts in Old French and Latin », *Journal of Data Mining & Digital Humanities*, 2020 (Towards a Digital Ecosystem :...[ 2020]), p. 5581, DOI : 10.46298/jdmdh.5581.
- CURRIE (David), *Creating a Spell Checker with TensorFlow*, Medium, 18 mai 2017, URL : <https://towardsdatascience.com/creating-a-spell-checker-with-tensorflow-d35b23939f60> (visité le 12/07/2021).
- DAHL (Christian M.), JOHANSEN (Torben), SØRENSEN (Emil N.) et WITTROCK (Simon), « HANA : A HAndwritten NAme Database for Offline Handwritten Text Recognition », *arXiv :2101.10862 [cs, econ]* (, 22 janv. 2021), arXiv : 2101 . 10862, URL : <http://arxiv.org/abs/2101.10862> (visité le 14/04/2021).

- DARAEE (Fatemeh), MOZAFFARI (Saeed) et RAZAVI (Seyyed Mohammad), « Handwritten keyword spotting using deep neural networks and certainty prediction », *Computers & Electrical Engineering*, 92 (1<sup>er</sup> juin 2021), p. 107111, DOI : 10.1016/j.compeleceng.2021.107111.
- DEVLIN (Jacob), CHANG (Ming-Wei), LEE (Kenton) et TOUTANOVA (Kristina), « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding » (, 11 oct. 2018), URL : <https://arxiv.org/abs/1810.04805v2> (visité le 30/08/2021).
- DOVAL (Yerai) et GÓMEZ-RODRÍGUEZ (Carlos), « Comparing Neural- and N-Gram-Based Language Models for Word Segmentation », *Journal of the Association for Information Science and Technology*, 70–2 (févr. 2019), p. 187-197, DOI : 10.1002/asi.24082, arXiv : 1812.00815.
- DRUCKER (Johanna), « Humanistic Theory and Digital Scholarship », dans *Debates in the Digital Humanities*, dir. Matthew K. Gold, NED - New edition, 2012, p. 85-95, URL : <https://www.jstor.org/stable/10.5749/j.ctttv8hq.9> (visité le 10/08/2021).
- DUPONT (Yoann), « Exploration de traits pour la reconnaissance d'entités nommées du Français par apprentissage automatique », dans *TALN 2017*, Orléans, France, 2017, URL : <https://hal.archives-ouvertes.fr/hal-02448614> (visité le 13/04/2021).
- « La structuration dans les entités nommées », thèse de doct., Université Sorbonne Paris Cité, 2017, URL : <https://tel.archives-ouvertes.fr/tel-01772268> (visité le 13/07/2021).
- EHRMANN (Maud), *Les Entités Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguisation*, Theses, Paris Diderot University, 2008, URL : <https://hal.archives-ouvertes.fr/tel-01639190> (visité le 25/08/2021).
- Maud Ehrmann, et al. (éd.), « Extended Overview of CLEF HIPE 2020 : Named Entity Processing on Historical Newspapers », *CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings ( oct. 2020), Meeting Name : 11th Conference and Labs of the Evaluation Forum (CLEF 2020) Num Pages : 38 Publisher : CEUR-WS Series Number : 2696, DOI : 10.5281/zenodo.4117566.
- ESTEVES (Diego), MARCELINO (José), CHAWLA (Piyush), FISCHER (Asja) et LEHMANN (Jens), « HORUS-NER : A Multimodal Named Entity Recognition Framework for Noisy Data », dans *IDA 2021 : Advances in Intelligent Data Analysis XIX*, 2021, pp. 89-100, URL : <https://openreview.net/forum?id=eoWnVtxS1su> (visité le 10/08/2021).
- Etalab - Pseudonymiser des documents grâce à l'IA*, URL : <https://guides.etalab.gouv.fr/pseudonymisation/> (visité le 12/07/2021).
- FOPPIANO (Luca) et ROMARY (Laurent), « Entity-fishing : a DARIAH entity recognition and disambiguation service », *Journal of the Japanese Association for Digital Humanities*, 5–

1 (nov. 2020), Publisher : Japanese Association for Digital Humanities, p. 22-60, DOI : 10.17928/jjadh.5.1\_22.

FRANÇAIS (Association des archivistes), *Abrégé d'archivistique : principes et pratiques du métier d'archiviste*, Paris, France, 2020.

FRONTINI (Francesca), BRANDO (Carmen), BYSZUK (Joanna), GALLERON (Ioana), SANTOS (Diana) et STANKOVIC (Ranka), « Named Entity Recognition for Distant Reading in ELTeC », dans *CLARIN Annual Conference 2020*, Virtual Event, France, 2020, URL : <https://hal.archives-ouvertes.fr/hal-03160438> (visité le 10/08/2021).

FUCHS (Catherine) et HABERT (Benoit), « le traitement automatique des langues : des modèles aux ressources », dans *Le Français Moderne - Revue de linguistique Française*, 2004, t. LXXII : 1, URL : <https://halshs.archives-ouvertes.fr/halshs-00067884> (visité le 08/08/2021).

GABAY (Simon), CLÉRICE (Thibault) et REUL (Christian), *OCR17 : Ground Truth and Models for 17th c. French Prints (and hopefully more)*, mai 2020, URL : <https://hal.archives-ouvertes.fr/hal-02577236> (visité le 19/07/2021).

GARBE (Wolf), *Fast Word Segmentation of Noisy Text*, Medium, 23 sept. 2020, URL : <https://towardsdatascience.com/fast-word-segmentation-for-noisy-text-2c2c41f9e8da> (visité le 12/07/2021).

GATOS (Basilis), LOULoudis (Georgios), CAUSER (Tim), GRINT (Kris), ROMERO (Veronica), SANCHEZ (Joan Andreu), TOSELLI (Alejandro H.) et VIDAL (Enrique), « Ground-Truth Production in the Transcriptorium Project », dans *2014 11th IAPR International Workshop on Document Analysis Systems*, Tours, France, 2014, p. 237-241, DOI : 10.1109/DAS.2014.23.

GIOTIS (Angelos P.), SFIKAS (Giorgos), GATOS (Basilis) et NIKOU (Christophoros), « A survey of document image word spotting techniques », *Pattern Recognition*, 68 (1<sup>er</sup> août 2017), p. 310-332, DOI : 10.1016/j.patcog.2017.02.023.

GTURRET (PSEUDONYME), *gturret/frenchngrams*, original-date : 2015-12-06T12:18:04Z, 9 avr. 2021, URL : <https://github.com/gturret/frenchngrams> (visité le 26/07/2021).

HAMDI (Ahmed), JEAN-CAURANT (Axel), SIDÈRE (Nicolas), COUSTATY (Mickaël) et DOUCET (Antoine), « An Analysis of the Performance of Named Entity Recognition over OCRed Documents », dans *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Issue : 1, Champaign, United States, 2019, t. 24, p. 333-334, DOI : 10.1109/JCDL.2019.00057.

HAMMADI (Nouha), BELAÏD (Abdel) et BELAÏD (Yolande), « Named Entity Recognition by Neural Prediction », dans, 2018, URL : <https://hal.inria.fr/hal-01981613> (visité le 27/04/2021).

- HENGCHEN (Simon), Hooland (Seth van), VERBORGH (Ruben) et WILDE (Max De), « L'extraction d'entités nommées : une opportunité pour le secteur culturel ? », *I2D - Information, donnees documents*, Volume 52–2 (7 juil. 2015), p. 70-79, URL : <https://www.cairn.info/revue-i2d-information-donnees-et-documents-2015-2-page-70.htm> (visité le 13/07/2021).
- HENRY (Cyprien) et CLAVAUD (Florence), *Vers un référentiel national des notaires ?*, fdocuments.fr, URL : <https://fdocuments.fr/document/relier-donnees-referentielnotaireschenryfclavaud-final.html> (visité le 13/07/2021).
- HILL (Mark J) et HENGCHEN (Simon), « Quantifying the impact of dirty OCR on historical text analysis : Eighteenth Century Collections Online as a case study », *Digital Scholarship in the Humanities*, 34–4 (1<sup>er</sup> déc. 2019), p. 825-843, DOI : 10.1093/llc/fqz024.
- Home | Transkriptorium*, URL : <http://www.transkriptorium.com> (visité le 24/07/2021).
- HUSEBY (Kristin H.), *How to improve the performance of a machine learning model with post processing employing Levenshtein distance*, Medium, 28 mai 2020, URL : <https://towardsdatascience.com/how-to-improve-the-performance-of-a-machine-learning-model-with-post-processing-employing-b8559d2d670a> (visité le 16/07/2021).
- INGHAM (Francisco), *Dissecting BERT Part 2 : BERT Specifics*, Medium, 27 nov. 2018, URL : <https://medium.com/dissecting-bert/dissecting-bert-part2-335ff2ed9c73> (visité le 12/07/2021).
- INRIA, ARCHIVES NATIONALES et MINISTÈRE DE LA CULTURE, *Convention de recherche particulière relative au projet : LECTAUREP (phase 3) (LECTure Automatique de RE-Pertoires)*, 2019.
- J2KUN, *Word Segmentation, or Makingsenseofthis*, Math Programming, 15 janv. 2012, URL : <https://jeremykun.com/2012/01/15/word-segmentation/> (visité le 16/07/2021).
- JURAFSKY (Daniel) et MARTIN (James), *Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, t. 2, 2020.
- KERMORVANT (Christopher), *La reconnaissance d'écriture manuscrite*, Data Analytics Post, 8 oct. 2019, URL : <https://dataanalyticspost.com/la-reconnaissance-decriture-manuscrite-de-nouvelles-applications-pour-un-des-plus-vieux-problemes-dia/> (visité le 08/08/2021).
- KIESSLING (Benjamin), *Kraken - an Universal Text Recognizer for the Humanities*, 2019, URL : <https://dev.clariah.nl/files/dh2019/boa/0673.html> (visité le 19/07/2021).

- KIESSLING (Benjamin), TISSOT (Robin), STOKES (Peter) et STÖKL BEN EZRA (Daniel), « eScriptorium : An Open Source Platform for Historical Document Analysis », dans *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2019, t. 2, p. 19-19, DOI : 10.1109/ICDARW.2019.90032.
- KLIE (Jan-Christoph), ECKART DE CASTILHO (Richard) et GUREVYCH (Iryna), « From Zero to Hero : Human-In-The-Loop Entity Linking in Low Resource Domains », dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020, p. 6982-6993, DOI : 10.18653/v1/2020.acl-main.624.
- KRISHNAN (Praveen), *POS Tagging and Named Entity Recognition on Handwritten Documents*, Praveen Krishnan, 25 déc. 2018, URL : <https://kris314.github.io/publication/pos-ner/> (visité le 27/04/2021).
- KRISTANTI (Tanti) et ROMARY (Laurent), « DeLFT and entity-fishing : Tools for CLEF HIPE 2020 Shared Task », dans, 2020, t. 2696, URL : <https://hal.inria.fr/hal-02974946> (visité le 30/08/2021).
- LE PEVEDIC (Solenn) et MAUREL (Denis), « Retour sur les annotations des entités nommées dans les campagnes d'évaluation françaises et comparaison avec la TEI », *Corela. Cognition, représentation, langage*–14 (22 nov. 2016), Number : 14-2 Publisher : Université de Poitiers, DOI : 10.4000/corela.4644.
- LIMON-BONNET (Marie-Françoise), MOUFFLET (Jean-François) et PIRAINO (Gaetano), « L'innovation numérique : un cercle vertueux pour l'archivistique », *La Gazette des archives*, 2 (numéro 254[ 2019]), p. 247-252.
- LIMON-BONNET (Marie-Françoise) et PIRAINO (Gaetano), « Préparer l'innovation : l'informatisation des ressources du minutier central », *La Gazette des archives*, 2 (numéro 254[ 2019]), p. 252-266.
- MARTIN (Louis), MULLER (Benjamin), ORTIZ SUÁREZ (Pedro Javier), DUPONT (Yoann), ROMARY (Laurent), CLERGERIE (Éric de la), SEDDAH (Djamé) et SAGOT (Benoît), « CamemBERT : a Tasty French Language Model », dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020, p. 7203-7219, DOI : 10.18653/v1/2020.acl-main.645.
- « Les modèles de langue contextuels Camembert pour le français : impact de la taille et de l'hétérogénéité des données d'entraînement », dans *JEP-TALN-RECITAL 2020 - 33ème Journées d'Études sur la Parole, 27ème Conférence sur le Traitement Automatique des Langues Naturelles, 22ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, dir. Christophe Benzitoun, et al., Nancy / Virtuel, France, 2020, p. 54-65, URL : <https://hal.archives-ouvertes.fr/hal-02784755> (visité le 16/04/2021).

MASINI (Francesca), *Multi-Word Expressions and Morphology*, Oxford Research Encyclopedia of Linguistics, ISBN : 9780199384655, 30 sept. 2019, DOI : 10.1093/acrefore/9780199384655.013.611.

MASSOT (Marie-Laure), MOREUX (Jean-Philippe) et VENTRESQUE (Vincent), « Expérimenter Transkribus sur les fiches de lecture de Michel Foucault », dans *Colloque de clôture du projet ANR Foucault Fiches de lecture Seconde partie “Editer Michel Foucault (1994-2021)”*, Paris, France, 2020, URL : <https://hal.archives-ouvertes.fr/hal-02974811> (visité le 07/04/2021).

MASSOT (Marie-Laure), SFORZINI (Arianna) et VENTRESQUE (Vincent), « Transcribing Foucault's handwriting with Transkribus », *Journal of Data Mining and Digital Humanities*, Atelier Digit\_Hum (mars 2019), Publisher : Episciences.org, URL : <https://hal.archives-ouvertes.fr/hal-01913435> (visité le 07/04/2021).

MCCALLUM (Andrew), « Information Extraction : Distilling structured data from unstructured text », *Queue*, 3–9 (1<sup>er</sup> nov. 2005), p. 48-57, DOI : 10.1145/1105664.1105679.

MIRET (Blanche), *Teklia - NERVAL : A NER evaluation python package for noisy text*, 2021, URL : <https://teklia.com/blog/202104-nerval/> (visité le 23/04/2021).

MISHRA (Shubhanshu) et DIESNER (Jana), « Semi-supervised Named Entity Recognition in noisy-text », dans *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, Osaka, Japan, 2016, p. 203-212, URL : <https://aclanthology.org/W16-3927> (visité le 10/08/2021).

MORETTI (Franco), « Conjectures on World Literature », *New Left Review*–1 (1<sup>er</sup> févr. 2000), p. 54-68.

MOUFFLET (Jean-François) et PIRAINO (Gaetano), « Au coeur du document d'archives : le projet Himanis », *La Gazette des archives*, 2 (numéro 254[ 2019]), p. 267-281.

MUEHLBERGER (Guenter), SEWARD (Louise), TERRAS (Melissa), ARES (Oliveira Sofia), BOSCH (Vicente), BRYAN (Maximilian), COLUTTO (Sebastian), DÉJEAN (Hervé), DIEM (Markus), FIEL (Stefan), *et al.*, « Transforming scholarship in the archives through handwritten text recognition : Transkribus as a case study », *Journal of Documentation*, 75–5 (1<sup>er</sup> janv. 2019), Publisher : Emerald Publishing Limited, p. 954-976, DOI : 10.1108/JD-07-2018-0114.

MÜHLBERGER (Günter), « Preprint : Handwritten Text Recognition (HTR) of Historical Documents as a Shared Task for Archivists, Computer Scientists and Humanities Scholars. The Model of a Transcription & Recognition Platform (TRP) » (, 2019), URL : [https://www.academia.edu/8601748/Preprint\\_Handwritten\\_Text\\_Recognition\\_HTR\\_of\\_Historical\\_Documents\\_as\\_a\\_Shared\\_Task\\_for\\_Archivists\\_Computer\\_Scientists\\_and\\_Humanities\\_Scholars\\_The\\_Model\\_of\\_a\\_Transcription\\_and\\_Recognition\\_Platform\\_TRP\\_](https://www.academia.edu/8601748/Preprint_Handwritten_Text_Recognition_HTR_of_Historical_Documents_as_a_Shared_Task_for_Archivists_Computer_Scientists_and_Humanities_Scholars_The_Model_of_a_Transcription_and_Recognition_Platform_TRP_) (visité le 21/07/2021).

- N (Nishanth), *Train NER with Custom training data using spaCy*. Medium, 29 juil. 2020, URL : <https://towardsdatascience.com/train-ner-with-custom-training-data-using-spacy-525ce748fab7> (visité le 12/07/2021).
- NETO (Arthur Flor de Sousa), BEZERRA (Byron Leite Dantas) et TOSELLI (Alejandro Héctor), « Towards the Natural Language Processing as Spelling Correction for Offline Handwritten Text Recognition Systems », *Applied Sciences*, 10–21 (janv. 2020), Number : 21 Publisher : Multidisciplinary Digital Publishing Institute, p. 7711, DOI : 10.3390/app10217711.
- NGUYEN (T.), JATOWT (A.), COUSTATY (M.), NGUYEN (N.) et DOUCET (A.), « Post-OCR Error Detection by Generating Plausible Candidates », dans *2019 International Conference on Document Analysis and Recognition (ICDAR)*, ISSN : 2379-2140, 2019, p. 876-881, DOI : 10.1109/ICDAR.2019.00145.
- NGUYEN (Thi-Tuyet-Hai), JATOWT (Adam), COUSTATY (Mickaël), NGUYEN (Nhu-Van) et DOUCET (Antoine), « Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing », dans *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Champaign, France, 2019, p. 29-38, DOI : 10.1109/jcdl.2019.00015.
- NGUYEN HONG (Vu), NGUYEN (Hien) et SNASEL (Vaclav), « Text normalization for named entity recognition in Vietnamese tweets », *Computational Social Networks*, 3 (1<sup>er</sup> déc. 2016), DOI : 10.1186/s40649-016-0032-0.
- NORVIG (Peter), « Natural Language Corpus Data », dans *Beautiful Data*, 2009, p. 219-242.
- ORTIZ SUÁREZ (Pedro Javier), DUPONT (Yoann), MULLER (Benjamin), ROMARY (Laurent) et SAGOT (Benoît), « Establishing a New State-of-the-Art for French Named Entity Recognition », dans *LREC 2020 - 12th Language Resources and Evaluation Conference*, Marseille, France, 2020, URL : <https://hal.inria.fr/hal-02617950> (visité le 13/04/2021).
- PASQUER (Caroline), « Expressions polylexicales verbales : étude de la variabilité en corpus », dans *TALN-RECITAL 2017*, Orléans, France, 2017, URL : <https://hal.archives-ouvertes.fr/hal-01637355> (visité le 27/07/2021).
- PINCHE (Ariane), CAMPS (Jean-Baptiste) et CLÉRICE (Thibault), « Styliometry for Noisy Medieval Data : Evaluating Paul Meyer's Hagiographic Hypothesis », dans *Digital Humanities Conference 2019 - DH2019*, Utrecht, Netherlands, 2019, URL : <https://hal.archives-ouvertes.fr/hal-02182737> (visité le 16/07/2021).
- PLETSCHACHER (Stefan) et ANTONACOPOULOS (Apostolos), « The PAGE (Page Analysis and Ground-Truth Elements) Format Framework », dans *2010 20th International Conference on Pattern Recognition*, Istanbul, Turkey, 2010, p. 257-260, DOI : 10.1109/ICPR.2010.72.

- PRASAD (Animesh), DÉJEAN (Hervé), MEUNIER (Jean-Luc), WEIDEMANN (Max), MICHAEL (Johannes) et LEIFERT (Gundram), *Bench-Marking Information Extraction in Semi-Structured Historical Handwritten Records*, 2018.
- Python Word Segmentation — Word Segment 1.3.1 documentation*, URL : <http://www.gantjenks.com/docs/wordsegment/index.html#tutorial> (visité le 16/07/2021).
- QI (Peng), ZHANG (Yuhao), ZHANG (Yuhui), BOLTON (Jason) et MANNING (Christopher D.), « Stanza : A Python Natural Language Processing Toolkit for Many Human Languages », *arXiv :2003.07082 [cs]* (, 23 avr. 2020), arXiv : 2003.07082, URL : <http://arxiv.org/abs/2003.07082> (visité le 22/04/2021).
- RAJAPAKSE (Thilina), *Named Entity Recognition Specifics*, Simple Transformers, URL : <https://simpletransformers.ai/docs/ner-specifics/> (visité le 12/07/2021).
- REBORA (Simone), « A Digital Edition between Stylometry and OCR : The Klagenfurter Ausgabe of Robert Musil », *Textual Cultures*, 12–2 (2019), Publisher : [Society for Textual Scholarship, Indiana University Press], p. 71-90, URL : <https://www.jstor.org/stable/26821537> (visité le 08/04/2021).
- RIEGEL (Martin), PELLAT (Jean-Christophe) et RIOUL (René), *Grammaire méthodique du français*, ISSN : 0291-0489, Paris, France, 2004.
- RIONDET (Charles) et FOPPIANO (Luca), « History Fishing When engineering meets History », dans *Text as a Resource. Text Mining in Historical Science #dhiha7*, Paris, France, 2017, URL : <https://hal.inria.fr/hal-01830713> (visité le 07/04/2021).
- ROSSET (Sophie), GROUIN (Cyril) et ZWEIGENBAUM (Pierre), *Entités nommées structurées : guide d'annotation Quaero*, Google-Books-ID : il9bMwEACAAJ, LIMSI, 2011.
- ROSTAING (Aurélia), *Les archives notariales aux Archives nationales*, Formation, URL : <https://fr.slideshare.net/AR2012/les-archives-notariales-aux-archives-nationales> (visité le 07/04/2021).
- *Méthodologie de recherche dans les archives notariales des Archives n...* Formation, URL : <https://fr.slideshare.net/AR2012/mthodologie-de-recherche-dans-les-archives-notariales-des-archives-nationales> (visité le 07/04/2021).
- RUSAKOV (E.), ROTHACKER (L.), MO (H.) et FINK (G. A.), « A Probabilistic Retrieval Model for Word Spotting Based on Direct Attribute Prediction », dans *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2018, p. 38-43, DOI : 10.1109/ICFHR-2018.2018.00016.
- SÁNCHEZ (Joan Andreu), MÜHLBERGER (Günter), GATOS (Basilis), SCHOFIELD (Philip), DEPUYDT (Katrien), DAVIS (Richard M.), VIDAL (Enrique) et DOES (Jesse de), « transCriptorium : a european project on handwritten text recognition », dans *Proceedings of the 2013 ACM symposium on Document engineering*, Florence Italy, 2013, p. 227-228, DOI : 10.1145/2494266.2494294.

SANG (Erik F. Tjong Kim) et DE MEULDER (Fien), « Introduction to the CoNLL-2003 Shared Task : Language-Independent Named Entity Recognition », *Proceedings of CoNLL-2003* (, 12 juin 2003), arXiv : cs/0306050, URL : <http://arxiv.org/abs/cs/0306050> (visité le 19/07/2021).

SCHLAGDENHAUFFEN (Régis), « Optical Recognition Assisted Transcription with Transkribus : The Experiment concerning Eugène Wilhelm's Personal Diary (1885-1951) », *Journal of Data Mining and Digital Humanities*, Atelier Digit\_Hum (août 2020), Publisher : Episciences.org, URL : <https://hal.archives-ouvertes.fr/hal-02520508> (visité le 07/04/2021).

SCHMITT (Xavier), KUBLER (Sylvain), ROBERT (Jérémy), PAPADAKIS (Mike) et LE TRAON (Yves), « A replicable comparison study of NER software : StanfordNLP, NLTK, OpenNLP, SpaCy, Gate », dans *Sixth International Conference on Social Networks Analysis, Management and Security, SNAMS 2019*, Grenada, Spain, 2019, DOI : 10.1109/SNAMS.2019.8931850.

SIMEONOVA (Lilia), SIMOV (Kiril), OSENOVA (Petya) et NAKOV (Preslav), « A Morpho-Syntactically Informed LSTM-CRF Model for Named Entity Recognition », *arXiv :1908.10261 [cs]* (, 27 août 2019), arXiv : 1908.10261, URL : <http://arxiv.org/abs/1908.10261> (visité le 12/07/2021).

STERN (Rosa), *Identification automatique d'entités pour l'enrichissement de contenus textuels*, Theses, Université Paris-Diderot - Paris VII, 2013, URL : <https://tel.archives-ouvertes.fr/tel-00939420> (visité le 12/07/2021).

STOKES (Peter A.), KIESSLING (Benjamin), STÖKL BEN EZRA (Daniel), TISSOT (Robin) et GARGEM (El Hassane), « The eScriptorium VRE for Manuscript Cultures, in Ancient Manuscripts and Virtual Research Environments », *Classic @ Journal-18* (2021), URL : <https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/>.

STÖKL BEN EZRA (Daniel), *L'infrastructure eScriptorium de reconnaissance automatique d'écriture manuscrite (HTR)*, Biblissima, 24 mars 2021, URL : <https://projet.biblissima.fr/fr/infrastructure-escriptorium-reconnaissance-automatique-ecriture-manuscrite-htr> (visité le 27/04/2021).

STORK (Lise), WEBER (Andreas), VERBEEK (Fons) et WOLSTENCROFT (Katherine), « From Historical Handwritten Manuscripts to Linked Data », *Digital Libraries for Open Knowledge - 22nd International Conference on Theory and Practice of Digital Libraries, TPDL 2018, Proceedings*, 11057 (2018).

STUTZMANN (Dominique), MOUFFLET (Jean-François) et HAMEL (Sébastien), « Full Text Search in Medieval Manuscripts : Issues and Perspectives of the HIMANIS Project for

Electronic Publishing », *Medievales -Paris-*, 73–73 (déc. 2017), Publisher : Puv, p. 67-96, DOI : 10.4000/medievales.8198.

SUÁREZ (Pedro Javier Ortiz), SAGOT (Benoît) et ROMARY (Laurent), « Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures », dans, 2019, DOI : 10.14618/IDS-PUB-9021.

TERRIEL (Lucas), *Représenter et évaluer les données issues du traitement automatique d'un corpus de documents historiques. L'exemple de la reconnaissance des écritures manuscrites dans les répertoires de notaires du projet LectAuRep.*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Alix Chagué et Thibault Clérice, 2020.

TOSELLI (Alejandro H.) et VIDAL (Enrique), *Transkribus User Conference. Keyword Spotting in Large Scale Documents*, nov. 2017, URL : [https://readcoop.eu/wp-content/uploads/2017/07/Toselli\\_Keyword\\_Spotting.pdf](https://readcoop.eu/wp-content/uploads/2017/07/Toselli_Keyword_Spotting.pdf) (visité le 08/09/2021).

VIDAL (E.), TOSELLI (A. H.) et PUIGCERVER (J.), « A Probabilistic Framework for Lexicon-based Keyword Spotting in Handwritten Text Images », *arXiv :2104.04556 [cs]* (, 9 avr. 2021), arXiv : 2104.04556, URL : <http://arxiv.org/abs/2104.04556> (visité le 23/04/2021).

# Table des figures

1	Exemple d'un enregistrement au sein d'une page d'un répertoire de notaire, avec souligné en rouge, l'adresse de la personne concernée. . . . .	5
2	Exemple d'une page d'un répertoire de notaire parisien, datée du mois de février 1901. . . . .	11
1.1	Exemple d'un enregistrement situé dans la colonne 5 (noms, prénoms et domiciles des parties) d'une page d'un répertoire de notaire (1). . . . .	20
1.2	Illustration des index de l'entité nommée « Dupuis », étant un nom de personne.	20
1.3	Exemple d'un enregistrement situé dans la colonne 5 (noms, prénoms et domiciles des parties) d'une page d'un répertoire de notaire (2). . . . .	21
2.1	Exemple d'un enregistrement complet dans une page d'un répertoire de notaire.	32
2.2	Exemple d'un enregistrement avec l'abréviation « St// Ouen » . . . . .	41
2.3	Exemple d'un enregistrement avec l'abréviation « St Antoine » . . . . .	42
3.1	Chaînes de traitement pour le KWS et la REM avec eScriptorium . . . . .	49
4.1	Schéma simplifié illustrant le découpage horizontal des enregistrements. . . . .	56
4.2	Exemple d'une ligne indiquant la date d'une page d'un répertoires de notaire.	59
4.3	Exemple générique des caractéristiques visuelles des premières lignes de chaque enregistrement. Ce cas de figure est rencontré fréquemment, avec parfois de faibles variations. . . . .	59
4.4	Exemple d'une page où les premières lignes de chaque enregistrement sont très distinctes, du fait d'un alinéa prononcé et d'un nom de famille mis en avant.	60
4.5	Exemple de l'indice visuel minimum existant pour le début d'un enregistrement : l'alinéa. . . . .	61
4.6	Mean Intersection-Over-Union . . . . .	62
4.7	Frequency Intersection-Over-Union . . . . .	63
4.8	Mean accuracy . . . . .	64

5.1	Test d' <i>accuracy</i> pour l'extraction des premiers tokens PER et ORG de chaque enregistrement sur une page aléatoire d'un répertoire de notaire selon taux de CER (1) . . . . .	90
5.2	Test d' <i>accuracy</i> pour l'extraction des premiers tokens PER et ORG de chaque enregistrement sur une page aléatoire d'un répertoire de notaire selon taux de CER (2) . . . . .	90
5.3	Nombre d'entités extraites sur une page aléatoire d'un répertoire de notaire selon taux de CER (1) . . . . .	92
5.4	Nombre d'entités extraites sur une page aléatoire d'un répertoire de notaire selon taux de CER (2) . . . . .	93
6.1	Résultats de la classification des tokens d'un enregistrement transcrit automatique (9% de CER) avec un modèle de REN basé sur camemBERT. Source : <a href="https://huggingface.co/Jean-Baptiste/camembert-ner">https://huggingface.co/Jean-Baptiste/camembert-ner</a> (consulté le 23/08/21). . . . .	109
7.1	Exemple d'un enregistrement issu d'une page d'un répertoire de notaire, avec une indication d'adresse introduite par la préposition « à ». . . . .	112
7.2	Exemple de syntaxe pour les enregistrements concernant des organisations. . . . .	112
7.3	Exemple de syntaxe pour les enregistrements concernant des personnes. . . . .	113
7.4	Indications de substitutions dans la troisième et quatrième colonne. . . . .	113
8.1	Arbre XML-TEI simplifié de la modélisation des pages des répertoires des notaires. . . . .	122
9.1	Chaîne de traitement pour la REN dans le cadre du projet LECTAUREP . . . . .	147
A.1	Exemple de recherches des actes passés en minute et en brevet dans la SIV pour le notaire Ernest Legay. . . . .	152
A.2	Liste des documents produits par le notaire Ernest Legay, inventoriés dans la SIV. . . . .	153
A.3	Liste des numérisations des actes passés en minutes et en brevet dans les répertoires des notaires parisiens du notaire Ernest Legay . . . . .	154
B.1	Structure syntaxique pour le type d'acte « inventaire ». . . . .	156
B.2	Structure syntaxique pour le type d'acte « continuation d'inventaire ». . . . .	156
B.3	Structure syntaxique pour le type d'acte « dépôt de testament ». . . . .	157
B.4	Structure syntaxique pour l'indication d'un « idem ». . . . .	157
B.5	Structure syntaxique pour le type d'acte « procuration » et autres. . . . .	158

D.1	Exemple d'un résultat de recherche donné par le système de KWS du projet HIMANIS pour le mot « mensis », avec un taux de confiance de 86% . . . . .	166
D.2	Exemple d'un résultat de recherche donné par le système de KWS du projet HIMANIS pour le mot « mensis », avec un taux de confiance de 63% . . . . .	167
E.1	Illustration du numérotage des lignes sur eScriptorium sans segmentation des régions, et donc de l'ordre dans lequel les lignes transcrivent apparaissent dans un fichier texte. . . . .	170
E.2	Exemple d'un enregistrement présent dans une page d'un répertoire de notaire.	176
E.3	Illustration du numérotage des lignes sur eScriptorium avec segmentation des régions, et donc de l'ordre dans lequel les lignes transcrivent apparaissent dans un fichier texte. . . . .	177
E.4	Schéma illustrant le découpage horizontal des enregistrements présents sur une page d'un répertoire de notaire en s'appuyant sur l'annotation de leurs premières lignes. Schéma créé par Alix Chagué, voir <a href="https://gitlab.inria.fr/almanach/lectaurep/documentation/-/issues/17#note_525626">https://gitlab.inria.fr/almanach/lectaurep/documentation/-/issues/17#note_525626</a> (consulté le 11/08/21). . . . .	178
F.1	Exemple d'une page d'un répertoire de notaire avec un texte dense, pouvant potentiellement troubler une segmentation horizontale des différents enregistrements. Les traits de couleurs correspondent à la segmentation des colonnes sur eScriptorium. . . . .	182
F.2	Attribution d'une nouvelle étiquette pour les colonnes pour les régions pré-annotée par le modèle de segmentation entraîné par Alix Chagué. Chaque couleur représente une étiquette et une région différente. . . . .	183
F.3	Annotation des lignes indiquant la date sur une page d'un répertoire de notaire. L'attribution d'un label à cette <i>baseline</i> est symbolisé par le trait vertical rose au début de celle-ci. . . . .	184
F.4	Annotation des premières lignes de chaque enregistrement dans une page d'un répertoire de notaire. L'attribution d'un label à ces <i>baseline</i> est symbolisé par le trait vertical rose au début de celles-ci. Les <i>baselines</i> non annotées possèdent un trait vertical violet. . . . .	185
F.5	Exemple de visualisation de l'intersection de deux cadre de délimitation, le vert étant la vérité de terrain et le rouge étant la prédiction d'un modèle. Source : <a href="https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/">https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/</a> (consulté le 10/08/21). . . . .	191
F.6	Exemple d'une annotation vérité de terrain des régions et des <i>baselines</i> dans une page d'un répertoire de notaire. . . . .	192

F.7	Exemple d'une prédiction de l'annotation des régions et des <i>baselines</i> dans une page d'un répertoire de notaire. . . . .	193
F.8	Calcul de la métrique <i>Intersection over Union</i> , ou indice de Jaccard. Source : <a href="https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/">https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/</a> (consulté le 10/08/21). . . . .	194
F.9	Visualisation des scores de la métrique <i>Intersection over Union</i> , ou indice de Jaccard. Source : <a href="https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/">https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/</a> (consulté le 10/08/21). . . . .	194
F.10	Résultats d'une prédiction faite par le modèle affiné de segmentation sémantique. Sur cet échantillon, toutes les premières ligne de chaque enregistrement ont été correctement identifiés, et la segmentation des régions est également bonne. La ligne indiquant la date de la page est cependant mal segmentée. . . . .	195
F.11	Résultats d'une prédiction faite par le modèle affiné de segmentation sémantique, similaire à l'exemple précédent. . . . .	196
F.12	Résultats d'une prédiction faite par le modèle affiné de segmentation sémantique, similaire à l'exemple précédent. . . . .	197
F.13	Exemple de problèmes rencontrés dans une prédiction faites par le modèle affiné de segmentation sur une page de répertoire de notaire. Ici, la segmentation des régions a été fortement perturbée, mais l'annotation des premières lignes dans la colonne 5 reste régulière. . . . .	198
G.1	Visualisation du résultat de la REN avec la librairie DisplaCy d'une lettre de Paul d'Estournelles de Constant, octobre 1919, <a href="https://github.com/FloChiff/DAHNProject/blob/master/Correspondence/Paul_d_Estournelles_de_Constant/Corpus/Lettre569_3octobre1919.xml">https://github.com/FloChiff/DAHNProject/blob/master/Correspondence/Paul_d_Estournelles_de_Constant/Corpus/Lettre569_3octobre1919.xml</a> , corpus du projet DAHN (1). Visualisation issue du <i>notebook</i> suivant : <a href="https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_spacy/test_dahn_spaCy.ipynb">https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_spacy/test_dahn_spaCy.ipynb</a> . . . . .	200
G.2	Visualisation du résultat de la REN avec la librairie DisplaCy d'une lettre de Paul d'Estournelles de Constant, octobre 1919, <a href="https://github.com/FloChiff/DAHNProject/blob/master/Correspondence/Paul_d_Estournelles_de_Constant/Corpus/Lettre569_3octobre1919.xml">https://github.com/FloChiff/DAHNProject/blob/master/Correspondence/Paul_d_Estournelles_de_Constant/Corpus/Lettre569_3octobre1919.xml</a> , corpus du projet DAHN (2). Visualisation issue du <i>notebook</i> suivant : <a href="https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_spacy/test_dahn_spaCy.ipynb">https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_spacy/test_dahn_spaCy.ipynb</a> . . . . .	201

G.3	Visualisation avec la librairie DisplaCy du résultat de la REN avec la librairie DisplaCy d'une lettre de Paul d'Estournelles de Constant, octobre 1919, <a href="https://github.com/FloChiff/DAHNProject/blob/master/Correspondence/Paul_d_Estournelles_de_Constant/Corpus/Lettre569_3octobre1919.xml">https://github.com/FloChiff/DAHNProject/blob/master/Correspondence/Paul_d_Estournelles_de_Constant/Corpus/Lettre569_3octobre1919.xml</a> , corpus du projet DAHN (3). Visualisation issue du <i>notebook</i> suivant : <a href="https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_spacy/test_dahn_spacy.ipynb">https://gitlab.inria.fr/almanach/lectaurep/ner/-/blob/master/ner_python/ner_spacy/test_dahn_spacy.ipynb</a> . . . . .	202
H.1	Page des projets de la plate-forme Doccano. . . . .	206
H.2	Page du corpus de la plate-forme Doccano. . . . .	207
H.3	Page des étiquettes pour les EN de la plate-forme Doccano. . . . .	208
H.4	Page d'annotation de la plate-forme Doccano. . . . .	209
H.5	Page d'annotation de la plate-forme Inception, avec les suggestions du recommandeur, en gris. . . . .	210
H.6	Page des <i>layers</i> de la plate-forme Inception. . . . .	211
H.7	Page des recommandeurs de la plate-forme Inception. . . . .	212
H.8	Page des <i>tagsets</i> de la plate-forme Inception. . . . .	213
H.9	Fenêtre présentant les formats d'exports de la plate-forme Inception. . . . .	214
H.10	Chaîne de pré-traitement proposée pour annoter des données en vue de l'entraînement/affinage d'un modèle de REN. . . . .	216
I.1	Schéma du JSON résultant du CLI développé pour évaluer par lots les transcription automatiques du projet LECTAUREP avec la librairie KaMI. . . . .	218
J.1	Chaîne de traitement de la transformation de l'export PAGE XML issu de la transcription et de sa transformation XML TEI. Modifications apportées au schéma originalement créé par Alix Chagué, source : <a href="https://gitlab.inria.fr/almanach/lectaurep/documentation/-/issues/17#note_528453">https://gitlab.inria.fr/almanach/lectaurep/documentation/-/issues/17#note_528453</a> (consulté le 16/08/21). . . . .	220
J.2	Arbre XML TEI de la modélisation des répertoires des notaires. . . . .	221
K.1	Capture d'écran de la base de données exist-db, et d'un encodage TEI d'une transcription d'une page d'un répertoire de notaire. . . . .	224
K.2	Page de l'éditeur d'ODD dans TEI Publisher . . . . .	225
K.3	Parcourir les pages des répertoires des notaires : navigation à travers une collection. . . . .	226
K.4	Visualisation d'une page transcrise d'un répertoire des notaires, sans feuille de style (1). . . . .	227

K.5	Visualisation d'une page transcrive d'un répertoire des notaires, sans feuille de style (2) . . . . .	228
K.6	Feuille CSS pour afficher une transcription d'une page d'un répertoire de notaire sous forme de tableau . . . . .	229
K.7	Visualisation d'une page transcrive d'un répertoire des notaires, avec une feuille de style basique. . . . .	230
K.8	Visualisation confrontant la transcription d'une page d'un répertoire de notaires avec sa numérisation grâce à l'utilisation du protocole IIIF. . . . .	231
K.9	Résultats d'une recherche du mot « dépôt » parmi les documents ajoutés à l'instance locale installée sur mon ordinateur. . . . .	232
K.10	Signalement du résultat d'une recherche du mot « dépôt » parmi les documents ajoutés à l'instance locale installée sur mon ordinateur. . . . .	233
K.11	Résultats d'une recherche à troncature « Jul* » parmi les documents ajoutés à l'instance locale installée sur mon ordinateur. . . . .	234

# Liste des tableaux

1.1 Exemple d'une matrice de confusion pour un modèle de classification binaire. Source : <a href="https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/">https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/</a> (consulté le 10/08/21) . . . . .	25
5.1 Tableaux des scores du modèle générique de REN pour la chaîne de traitement <i>fr_core_news_lg</i> de SpaCy sur une transcription automatique (CER de 21%) d'une page aléatoire d'un répertoire de notaire sans reconstitution de la structure logique. . . . .	82
5.2 Tableaux des scores du modèle générique de REN pour la chaîne de traitement de Stanza sur une transcription automatique (CER de 9%) d'une page aléatoire d'un répertoire de notaire avec reconstitution de la structure logique, obtenus avec la librairie NERVAL, avec un seuil de distance d'édition réglé à 0.30. . .	85
5.3 Tableaux des scores du modèle générique de REN pour la chaîne de traitement de Stanza sur une transcription automatique (CER de 9%) d'une page aléatoire d'un répertoire de notaire avec reconstitution de la structure logique, obtenus avec la librairie NERVAL, avec un seuil de distance d'édition réglé à 0.20. . .	86
5.4 Tableaux des scores du modèle générique de REN pour la chaîne de traitement de Stanza sur une transcription automatique (CER de 9%) d'une page aléatoire d'un répertoire de notaire avec reconstitution de la structure logique, obtenus avec la librairie NERVAL, avec un seuil de distance d'édition réglé à 0. . .	86
5.5 Tableaux des scores du modèle générique de REN pour la chaîne de traitement de Stanza sur une transcription automatique (CER de 9%) d'une page aléatoire d'un répertoire de notaire avec reconstitution de la structure logique, obtenus avec la librairie NERVAL, avec un seuil de distance d'édition réglé à 0.40. . .	86
5.6 Tableaux des scores du modèle générique de REN pour la chaîne de traitement <i>fr_core_news_lg</i> de SpaCy sur une transcription manuelle (vérité de terrain) d'une page aléatoire d'un répertoire de notaire sans reconstitution de la structure logique. . . . .	87

5.7	Tableaux des scores du modèle générique de REN pour la chaîne de traitement <i>fr_core_news_lg</i> de SpaCy sur une transcription manuelle (vérité de terrain) d'une page aléatoire d'un répertoire de notaire avec reconstitution de la structure logique . . . . .	87
5.8	Tableaux des scores du modèle générique de REN pour la chaîne de traitement <i>fr_core_news_lg</i> de SpaCy sur une transcription automatique corrigée d'un document du corpus du projet DAHN, une lettre de Paul d'Estournelles de Constant. . . . .	94

# Table des matières

Résumé	i
Remerciements	iii
Liste des sigles et abréviations	v
Introduction	3
<b>I Exploiter les transcriptions automatiques avec la reconnaissance d'entités nommées : présentation des répertoires des notaires et des données textuelles produites avec la REM dans le cadre du projet LECTAUREP</b>	<b>13</b>
<b>1 La reconnaissance d'entités nommées, une sous-tâche de l'extraction d'information dans un texte : définitions</b>	<b>15</b>
1.1 la reconnaissance d'entités nommées, une tâche d'extraction d'information . . . . .	15
1.1.1 Le traitement automatique des langues . . . . .	15
1.1.2 L'extraction d'information . . . . .	16
1.1.3 La reconnaissance d'entités nommées et le concept d'entité nommée	18
1.1.4 La désambiguïsation des EN . . . . .	21
1.2 Les systèmes de reconnaissance par règles . . . . .	22
1.3 La reconnaissance d'entités nommées et l'IA . . . . .	23
1.4 Évaluer un modèle de reconnaissance d'entités nommées : présentation des métriques . . . . .	24
1.5 Un exemple de l'implémentation de la reconnaissance d'entités nommées dans des plate-formes de publications numériques . . . . .	27
1.5.1 "A machine learning tool for fishing entities" : présentation d' <i>Entity-fishing</i> . . . . .	27

1.5.2 Utiliser <i>Entity-Fishing</i> pour améliorer le processus de publication et la recherche dans des documents publiés . . . . .	28
<b>2 Des modèles de reconnaissance d'écriture manuscrite visant la perfection : étude du matériau source et de la production des données du projet LECTAUREP</b>	<b>31</b>
2.1 La réforme du notariat de 1803 : le début d'un enregistrement standardisé de l'information des activités des notaires . . . . .	31
2.2 La nature des données textuelles des répertoires de notaires : un langage spécialisé ? . . . . .	32
2.2.1 Les types d'actes : mots et mots complexes . . . . .	32
2.2.2 Une syntaxe particulière induite par le notariat et la structure tabulaire des répertoires . . . . .	34
2.3 Les objectifs du projet LECTAUREP pour la reconnaissance d'entités nommées	35
2.4 Quelles solutions technologiques pour une campagne de reconnaissance d'écriture manuscrite réalisée dans le cadre d'un projet d'humanités numériques ? .	36
2.4.1 La reconnaissance d'écriture manuscrite : définition . . . . .	36
2.4.2 La démocratisation de la REM avec Transkribus . . . . .	37
2.4.3 Une interface open-source pour la transcription et l'entraînement de modèles de REM : eScriptorium . . . . .	38
2.5 Transcrire pour entraîner des modèles de reconnaissance d'écriture manuscrite	40
2.6 Analyse des types d'erreurs générées par la reconnaissance d'écriture manuscrite	42
2.7 Quel outil pour évaluer les données produites dans le cadre de LECTAUREP ?	43
2.7.1 Un outil développé dans l'équipe ALMAnaCH : la librairie python KaMI	43
2.7.2 Définition de trois métriques d'évaluation de la REM : la distance de Levenshtein, le CER, et le WER. . . . .	43
<b>3 Autres méthodes de recherche dans un document patrimonial</b>	<b>45</b>
3.1 Les moteurs de recherche : recherches plein texte et recherches floues . . . . .	45
3.1.1 Une première exploration du texte par la recherche plein texte . . . . .	45
3.1.2 Un système de recherche flexible : les recherches floues . . . . .	46
3.2 Le <i>keyword spotting</i> : une affaire de vision par ordinateur . . . . .	46
3.3 Déetecter les entités nommées sans utiliser de données textuelles grâce à la vision par ordinateur . . . . .	48

## **II La reconnaissance d'entités nommées appliquée à des données bruitées issues de la transcription automatique de documents pa-**

# **trimoniaux : expérimentations à partir des données du projet LECTAUREP**

## **51**

<b>4 Quels pré-traitements des données pour la reconnaissance d'entités nommées ?</b>	<b>53</b>
4.1 Reconstruire un document déconstruit par la reconnaissance d'écriture manuscrite . . . . .	54
4.1.1 Reconstituer la structure logique des répertoires des notaires . . . . .	54
4.1.2 Description de l'ontologie utilisée pour l'annotation des régions et des <i>baselines</i> . . . . .	55
4.1.3 Méthodologie . . . . .	56
4.2 Entraîner un modèle de segmentation sémantique pour reconstruire automatiquement la structure logique . . . . .	58
4.2.1 Objectifs et motivations du modèle de segmentation affiné . . . . .	58
4.2.2 Description du processus d'annotation du corpus d'entraînement . . . . .	59
4.2.3 Affinage du modèle de segmentation et évaluation . . . . .	60
4.2.4 Explication des métriques d'évaluation du modèle de segmentation . . . . .	65
4.2.5 Analyse des scores et pistes d'améliorations pour le modèle . . . . .	66
4.3 Vers un « débruitage » des données issues de la reconnaissance d'écriture manuscrite en vue de la tokenisation . . . . .	67
4.3.1 La segmentation des mots, une étape essentielle pour l'exploitation automatique des données du projet LECTAUREP . . . . .	67
4.3.2 Segmenter les mots avec des dictionnaires de fréquence de n-grammes . . . . .	68
4.3.3 Utiliser des expressions régulières pour segmenter les mots ? Le risque de la non-exhaustivité. . . . .	71
4.3.4 Segmenter les mots avec une approche neuronale . . . . .	72
4.3.5 La tokenisation des données textuelles . . . . .	73
4.4 Une étape optionnelle : la normalisation des abréviations . . . . .	73
4.5 La correction orthographique post transcription automatique . . . . .	74
<b>5 Les modèles génériques de reconnaissance d'entités nommées : une solution clé en main ?</b>	<b>77</b>
5.1 Tour d'horizon des modèles génériques de reconnaissance d'entités nommées disponibles avec les librairies de TAL python . . . . .	78
5.1.1 SpaCy . . . . .	78
5.1.2 Stanza . . . . .	79
5.1.3 Autres librairies de TAL n'ayant pas de modèle de REN pour le français	79

5.2	Une étude des performances des modèles génériques par seuil de taux d'erreur de caractères selon une performance idéale . . . . .	80
5.2.1	Évaluation de la chaîne de traitement de SpaCy . . . . .	80
5.2.2	Évaluation de la chaîne de traitement de Stanza . . . . .	82
5.2.3	Observation des performance des modèles génériques sur une vérité de terrain . . . . .	86
5.3	Des performances aléatoires pour les modèles génériques de reconnaissance d'entités nommées? . . . . .	88
5.3.1	Cas d'usage de l'annotation des noms de personnes et des noms d'organisation en début d'enregistrement . . . . .	88
5.3.2	Une variation dans le nombre d'entités extraites selon les taux de CER	91
5.4	Quelles performances pour des textes propres présentant des caractéristiques linguistiques différentes des répertoires de notaires ? . . . . .	93
<b>6</b>	<b>Une solution spécifique à un corpus : entraîner un modèle de reconnaissance d'entités nommées</b>	<b>97</b>
6.1	Entraîner un modèle de classification avec le modèle de langue camemBERT	97
6.1.1	Présentation du modèle de langue camemBERT . . . . .	97
6.1.2	Entraînément d'un modèle de classification avec camemBERT . . . . .	99
6.2	Quels objectifs peut-on donner à un modèle affiné ? . . . . .	101
6.3	Rassemblement et préparation de données d'entraînement . . . . .	102
6.3.1	Sélectionner des candidats pour l'annotation de données d'entraînement en évaluant par lot . . . . .	103
6.3.2	Un CLI pour débruiter et normaliser les candidats sélectionnés . . . . .	104
6.3.3	Retour d'expérience sur l'utilisation de la plate-forme d'annotation <i>open-source</i> Inception . . . . .	105
6.3.4	Un format d'annotation pour des données d'entraînement : CoNLL 2002	106
6.4	Préconisations pour l'entraînement d'un modèle de reconnaissance d'entités nommées à partir des données de LECTAUREP . . . . .	107
<b>7</b>	<b>Une solution hétérodoxe : la reconnaissance d'entités nommées basée sur un système de règles</b>	<b>111</b>
7.1	Une solution spécifique et non généralisable, profitant de la nature des données exploitées par le projet LECTAUREP . . . . .	111
7.2	De l'utilité des référentiels pour de la reconnaissance d'entités nommées avec un système de règles . . . . .	113

### **III Exploiter les entités nommées dans le contexte des métiers du patrimoine** 115

<b>8 Le signalement des entités nommées au sein d'un encodage en XML-TEI : l'après NER</b>	<b>117</b>
8.1 Modélisation de la structure logique des répertoires des notaires en TEI . . . . .	117
8.1.1 Transformer le PAGE XML en XML-TEI . . . . .	117
8.1.2 Modéliser en TEI la structure tabulaire des répertoires des notaires dans la balise <text> . . . . .	120
8.2 Automatisation de la transformation du PAGE XML résultant de la REM en XML-TEI . . . . .	121
8.3 Signalement des entités nommées dans un encodage TEI . . . . .	122
8.4 Une plate-forme de publication pour les transcriptions des répertoires des notaires : visualisation et recherches dans les pages des répertoires de notaires transcrits avec TEI Publisher. . . . .	125
8.4.1 Une interface configurable pour la publication des transcriptions des répertoires des notaires : l'application <i>open-source</i> TEI Publisher. . . . .	125
8.4.2 TEI Publisher comme plate-forme de <i>crowdsourcing</i> pour la REN ? . . . . .	128
8.4.3 Parcourir les transcriptions des répertoires des notaires avec TEI Publisher . . . . .	128
<b>9 Indexer les répertoires des notaires grâce à la reconnaissance d'entités nommées</b>	<b>131</b>
9.1 Indexer des documents patrimoniaux : définition . . . . .	131
9.1.1 L'indexation : définition . . . . .	131
9.1.2 Automatiser l'indexation pour réaliser cette tâche à grande échelle . . . . .	132
9.2 Indexer automatiquement grâce à la reconnaissance d'entités nommées . . . . .	132
9.2.1 Retrouver la structure logique des documents issus de la REM pour améliorer la recherche d'information . . . . .	132
9.2.2 Les EN comme matériau d'indexation . . . . .	133
9.2.3 Application de l'indexation automatique à partir des entités nommées pour les répertoires des notaires . . . . .	134
9.2.4 Présentation d'autres méthodes d'indexation automatique applicable aux transcriptions des répertoires des notaires . . . . .	136
9.3 L'indexation automatique dans le secteur culturel . . . . .	137
9.3.1 Quels impacts métiers pour l'indexation automatique ? . . . . .	137
9.3.2 Un guide du service interministériel des Archives de France pour indexer	138

<b>Conclusion</b>	<b>141</b>
<b>Annexes</b>	<b>151</b>
<b>A Chercher un acte dans les répertoires de notaires dans la SIV</b>	<b>151</b>
<b>B Structuration de l'information dans la cinquième colonne : exemples</b>	<b>155</b>
<b>C Les formats d'export XML d'eScriptorium : ALTO et PAGE</b>	<b>159</b>
<b>D Le Keyword spotting : exemple de mise en oeuvre avec le projet HIMANIS</b>	<b>165</b>
<b>E Reconstruire la structure logique des des répertoires des notaires après la REM</b>	<b>169</b>
<b>F Affinage d'un modèle de segmentation en vue de la reconstitution de la structure logique des pages des répertoires de notaire</b>	<b>181</b>
<b>G Visualiser la reconnaissance d'entités nommées</b>	<b>199</b>
<b>H Annoter des entités nommées : les plate-formes d'annotation Doccano et Inception</b>	<b>205</b>
<b>I Évaluer par lots les transcription automatiques du projet LECTAUREP avec la librairie KaMI</b>	<b>217</b>
<b>J Modéliser les répertoires des notaires en TEI</b>	<b>219</b>
<b>K Publier la transcription des répertoires des notaires sur la plate-forme <i>open-source</i> TEI Publisher</b>	<b>223</b>