Hugo Scheithauer

21/04/21

Fiche de lecture
HTR, NER, et le projet LECTAUREP

1 Introduction

La présente fiche de lecture se présente comme un outil de travail rendant compte de la synthèse de plusieurs articles lus dans le cadre du début de mon stage au sein de l'équipe ALMAnaCH - INRIA, pour le projet LECTAUREP (LECTure Automatique des REPertoires). [1] Les références lues se trouvent en bibliographie.

2 LECTAUREP - présentation et avancées du projet

2.1 Intérêts et chiffres du projet

Commencé au premier semestre 2018, le projet LECTAUREP est entré dans sa troisième phase au mois de novembre 2019. A terme, le projet vise à enrichir l'utilisation faite des répertoires d'actes de notaires conservés au Minutier dans les Archives Nationales, par les technologies de reconnaissance automatique d'écriture manuscrite (REM - Reconnaissance automatique d'Ecriture Manuscrite ; HTR - Handwritten Text Recognition). Se faisant, les usagers et les chercheurs auraient la possibilité d'effectuer des recherches rapides, exhaustives, plein texte et à facettes, au sein de cet imposant corpus, tout en faisant gagner à ce public un gain de temps non négligeable en raccourcissant, voire en supprimant, la phase de ddépouillement. [2] La convention de recherche particulière de la phase 3 suggère également de nouvelles exploitations du corpus rendues possibles par cette exploitation à grande échelle, comme par exemple de nouvelles lectures sémantiques favorisant de potentielles

1. Voir `https://lectaurep.hypotheses.org/`

2. Voir 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) pour plus de détails sur la méthodologie de recherche, les procédures aux Archives Nationales, et les difficultés rencontrées pour les recherches dans le Minutier.

analyses quantitatives : nombre d'actes, nombre de clients selon la chronologie, périodicité de l'activité professionnelle, taxation et fiscalité sur les différents types d'actes, etc. Des liens avec des référentiels patronymiques, géographiques, etc., ainsi que des mises en relation avec des bases de données externes seraient également réalisables, avec Wikidata, par exemple.

Le corpus contient plus de 917 notaires ayant exercé entre 1803 et 1940 à Paris. Il existe 26 kml de documents produits par les notaires de la capitale. Au total, on estime que ce corps professionnel a produit entre 1800 et 2000 registres de répertoires, contenant chacun 300 à 500 pages. Un simple et rapide calcul montre qu'il y aurait ainsi plus ou moins 760 000 pages à traiter. [3] Ce seul chiffre justifie l'intérêt d'un traitement automatique.

Les minutes des notaires de Paris représentent en moyenne 65% des communications journalières de la salle de lecture aux Archives Nationales de Paris. Le projet LECTAUREP aurait ainsi un impact positif sur l'expérience des usagers des Archives Nationales, du moins pour les recherches de documents compris dans la fourchette chronologique du projet.

2.2 Présentation du corpus

Le projet LECTAUREP se concentre donc sur les répertoires chronologiques d'actes de notaires conservés aux département du Minutier central des notaires de Paris, aux Archives Nationales.

Les minutes concernent plusieurs types d'actes (liste non exhaustive) [4] :

- Les droits des personnes physiques et des familles :
  - Contrats de mariage
  - Donations entre époux
  - Consentements ou opposition à mariage
  - Testaments
  - Pièces relatives au règlement des successions
  - Inventaire après décès
  - Etc.
- Les transactions relatives aux biens meubles ou immeubles :
  - Contrats de vente
  - Cahiers des charges de copropriété

3. 1900 registres de 400 pages contiendraient 760 000 pages au total.

4. Voir Id., *Les archives notariales aux Archives nationales*, Formation, URL : `https://fr.slideshare.net/AR2012/les-archives-notariales-aux-archives-nationales` (visité le 07/04/2021) pour un exposé plus complet.

- Baux
- Actes et transactions relatives aux sociétés commerciales
- Vente de fonds de commerce
- Etc.

Le corpus traité par LECTAUREP débute en 1803. Cette année marque en effet la formalisation des actes de notaires effectuée lors de la réorganisation du notariat par le Premier Consul Bonaparte (loi du 25 Ventôse an XI, art. 29 et 301). Les actes deviennent des formulaires pré-imprimés dotés de 6 colonnes de classement des informations.

1. La colonne 1 donne le numéro de l'acte.
2. La colonne 2 donne le quantième dans le mois de la date de l'acte.
3. La colonne 3 est divisée en deux sous-colonnes : brevet et minute. Elle donne la nature de l'acte.
4. La colonne 4 donne le patronyme principal de l'acte, les noms et prénoms des parties, leurs domiciles, leurs relations d'ordre familial, ainsi que l'indication des biens, de leur situation et de leur prix.
5. Les colonnes 5 et 6 donnent la relation de l'enregistrement.
   - En colonne 5 on retrouve la date (quantième du jour d'enregistrement).
   - En colonne 6 on retrouve les droits payés aux services fiscaux, en francs et centimes.

2.3 Objectifs techniques

Les objectifs techniques, tels que stipulés sur la convention de recherche, sont les suivants :

- Rechercher une chaîne de caractère à l'aide d'un moteur de recherche et la repérer visuellement sur les images du corpus.
- Possibilité d'entreprendre des recherches floues.
- Possibilité d'exporter dans plusieurs formats différents :
  - XML-EAD
  - XML-EAC(CPF)
  - XML-TEI
  - RDF (sérialisation XML, JSON-LD, NT, N3)
  - CSV/TSV
  - JSON
- Possibilité de limiter la transcription à telle ou telle colonne du registre.

— Possibilité d'enrichir les données depuis des entrepôts de données (API de géocodage, comme Nominatim, OpenStreetMap, Etalab, Dicotopo.

2.4 Pipeline actuelle

Le projet LECTAUREP s'appuie actuellement sur le développement de la plate-forme de transcription manuelle et automatique et d'annotation eScriptorium (PSL, SCRIPTA). [5] eScriptorium permet d'agglomérer les documents du projet LECTAUREP et de proposer un outil de crowdsourcing : les images peuvent être annotées et transcrites pour établir une vérité de terrain qui est réutilisée pour entraîner des modèles. Ces modèles servent ensuite à la transcription automatique. Deux pistes sont actuellement étudiées :

— un modèle générique, qui pourrait être appliqué à tous les documents. L'objectif pour ce modèle est d'avoir un character error rate (CER) [6] de 20% maximum. [7]
— un modèle par main d'écriture. L'objectif ici est d'avoir un CER de 10% maximum. [8]

La pipeline prend en input des images, importées depuis la Salle des Inventaires ou par IIIF. Les images sont traitées avec eScriptorium, où est faite la transcription automatique. La segmentation découpe le document en zones et identifie par conséquent les zones de textes. Elle prépare donc le document pour la transcription. Elle permet d'identifier depuis quels endroits du document sont tirés des morceaux de textes : les prix dans la colonne des prix, par exemple. La transcription y est également faite. Celle-ci peut être corrigée manuellement directement sur la plate-forme. Les documents, à ce stade, peuvent être consultés à l'aide d'une recherche floue ou exacte. Ensuite, les documents sont exportables en XML PAGE, ALTO, et/ou TEI. Cette étape permet de constituer la vérité de terrain, et la transcription mise à disposition des publics. Ces documents numériques sont ensuite transformés en XML EAD ou texte simple, puis versés dans la Salle des Inventaires Virtuels des Archives Nationales. [9]

La vérité de terrain sert à constituer un set de données utilisé pour mesurer l'efficacité et la précision du modèle de transcription automatique, en comparant les données produites automatiquement et les données créées lors de l'étape de transcription manuelle. [10]

5. Voir `https://escripta.hypotheses.org/`
6. (taux d'erreur par caractère transcrit automatiquement)
7. Alix Chagué 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)
8. *Ibid.*
9. Voir *Ibid.*
10. Voir 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*

De manière générale, la vérité terrain peut également être augmentée par des référentiels externes : dictionnaires linguistiques, référentiels d'entités nommées (noms de lieux, noms de personnes, concepts, etc.) et des outils de lemmatisation. [11]

2.5 L'implémentation d'outils de NER dans la pipeline

Mon stage consiste à faire une prospections des outils de Named-Entity Recognition (NER) existants et sur la façon dont ils pourraient être intégrés dans le projet. Il s'agit de trouver une solution généralisable en s'appuyant sur le cas précis de LECTAUREP.

Les technologies NER s'inscrivent dans le contexte de l'extraction automatique d'informations dans un texte. Elles ont pour but d'identifier des morceaux de textes référant à des personnes, des localisations, des noms d'organisations. Jurafsky et al. définissent le NER en ces termes : « [it] is a subtask of information extraction that seeks to locate and classify named entities (a real-world object, such as persons, locations, organizations, products, etc, that can be denoted with a proper name) in text (...). » [12] Les morceaux extraits sont nommés "named entity mentions." [13] Le NER peut aussi être étendu à la détection de texte qui ne suit pas la grammaire générale du langage analysé, permettant ainsi de détecter les URL, les noms d'entreprises, de marque, les dates, les montants monétaires, et autres chiffres. [14]

Elles rendent également possible la désambiguïsation des termes, selon le contexte de la phrase dans laquelle se trouve le mot extrait. Ainsi, dans la phrase suivante "Je m'appelle Lucien, je travaille pour Orange et je suis intervenu chez les Dupont à Orange, en France, le 16 avril 2021." il serait possible d'extraire et de classifier les mots comme suit :

Je m'appelle Lucien**(name)**, je travaille pour Orange**(organization)**, et je suis intervenu chez les Dupont**(name)** à Orange**(location)**, en France**(location)**, le 16 avril 2021**(date)**.

Ortiz Suárez et al. rappellent que le NER proposent deux types d'analyse : l'analyse de type intrinsèque - dans la phrase précédente, "France" est toujours une localisation -, et l'analyse de type contextuelle - Orange pouvant se référer à l'entreprise française (travailler

-Paris-, 73–73 (déc. 2017), Publisher : Puv, p. 67-96, DOI : 10.4000/medievales.8198, la vérité de terrain est « un échantillon représentatif de textes tels quils sont en réalité sur le terrain, par opposition au « modèle ». »

11. *Ibid.*

12. 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), page.3. Voir également Jurafsky, Dan. Speech language processing. Pearson Education India, 2000.

13. Pedro Javier Ortiz Suárez, Yoann Dupont, Benjamin Muller, Laurent Romary et Benoît 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.2

14. *Ibid.*

à Orange), et une ville française (habiter à Orange). [15]

On comprend ainsi l'intérêt de ce type de technologies dans le cadre du projet LECTAUREP. Ainsi implémenté, le NER permettrait de proposer des recherches par type (par nom par exemple), de récupérer automatiquement les différentes adresses, de repérer les sommes payées, etc.

3 Panorama des projets d'HTR en Humanités Numériques : méthodologies et problématiques

3.1 La bibliothèque Foucaldienne

Le projet ANR de la Bibliothèque Foucaldienne, clôturé en 2020, visait à rendre accessible aux chercheurs les carnets de note du philosophe Michel Foucault, en proposant une transcription automatique réalisée avec le logiciel Transkribus. [16] Un modèle de transcription automatique a été réalisé après une campagne de transcription manuelle. Les documents transcrits ont été exportés en TEI, un instrument de recherche en EAD a été produit, et une interface permet aujourd'hui de les consulter. Un moteur de recherche à été mis en place pour rechercher des mots-clés, des titres de fiches, des références bibliographiques, ainsi que des noms de personne.

3.2 Les journaux d'Eugène Wilhelm (1885-1951)

Régis Schlagdenhauffen a conduit une expérience de transcription automatique sur les carnets du juriste Eugène Wilhelm. La tâche comprenait deux principaux défis : les documents représentent 66 ans de vie, l'écriture du juriste a considérablement évolué, devenant difficilement lisible à la fin de sa vie ; le juriste a utilisé de façon concomitante l'alphabet latin pour les aspects concernant sa vie quotidienne, écrits en français, et l'alphabet grec pour les aspects privés. [17]

15. *Ibid.*

16. 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.

17. Voir 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).

Un dictionnaire propre à Eugène Wilhelm a été créé pour essayer d'atteindre une transcription automatique en français la plus précise. En effet, les transcriptions utilisées ont été réalisées à l'aide de Transkribus, qui n'intègre pas de dictionnaire. La transcription automatique se faisant caractère par caractère, et non mot par mot. Schlagdenhauffen explique qu'il n'était pas utile de récupérer un dictionnaire de français *général*, car apportant trop d'informations superflues. Un dictionnaire propre aux mots employés régulièrement par Wilhelm s'est avéré plus utile, car ne contenant que des mots issus des sphères de la vie du juriste. [18]

Le chercheur arrive à la conclusion que, bien que les transcriptions ne soient pas parfaites, elles permettent tout de même la recherche plein texte. Schlagdenhauffen insiste cependant, à la date de rédaction de son article et dans sa conclusion, sur le manque de dictionnaires adaptés à la reconnaissance et au traitement automatique de la langue française. [19]

3.3 L'analyse stylométrique de l'Oeuvre de Robert Musil grâce à l'OCR

Simone Rebora a entrepris une analyse stylométrique de textes attribués à Robert Musil avec des outils de transcriptions automatique, de l'OCR (optical character recognition) en l'occurrence. L'objectif était notamment de réétudier l'attribution qu'avait été faite d'articles écrits entre 1916 et 1917 à l'auteur au sein de la revue de propagande autrichienne *Tiroler Soldaten-Zeitung*. Simone Rebora ne remet pas en question son rôle d'éditeur, historiquement prouvé, mais bien des attributions jugées infondées. [20]

Le corpus de la revue *Tiroler Soldaten-Zeitung* représente 43 numéros. Le chercheur a décidé de mener une campagne d'océrisation sur ces numéros, afin de pouvoir procéder à l'analyse stylométrique. Les données produites ont pu être comparées avec d'anciennes transcriptions manuelles de l'Oeuvre complète de Musil, comprenant les articles de la revue qui lui étaient attribués (36 articles au total, sur 38 attribués à Musil appartenant à la revue). Le CER était de 10%. Ce problème était causé par une mauvaise segmentation réalisé par le logiciel d'OCR.

18. Voir *Ibid.*, p.11. "Limitations encountered by the absence of a dictionary integrated into Transkribus should be considered. To correct this, we opted for the creation of a dictionary of the words used by Eugène Wilhelm (see section 1.3) rather than using a dictionary of the French language. Indeed, as each individuals vocabulary remains limited - or circumscribed - to certain spheres, we saw no point in backing up transcriptions with a general dictionary. For example, vocabularies from the fields of natural sciences, medicine or crafts, are all unused, and therefore unnecessary, words in the context of writing a lawyers diary."

19. Est-ce que des outils développés par l'équipe ALMAnaCH auraient pu être utiles ? Notamment OSCAR et camemBERT.

20. Simone Rebora, "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), p.73

Pour y remédier, le chercheur a implémenté une pipeline logicielle de segmentation visant à rendre cette étape plus efficace. [21] Celle-ci se résume à :

— Binarisation des images par seuillage (transformation d'une image en niveau de gris en une image binaire, noir et blanc, où les pixels ne peuvent avoir qu'une valeur de 1 ou de 0) avec OCRopus/OCRopy. Les anomalies, telles que les ombres et les variations de lumière, sont également compensées automatiquement.
— Segmentation des zones de chaque page avec Transkribus.
— Le lissage et le redressage (de-skewing [22]) des pages avec ScanTailor.

Le tout lié avec des scripts rédigés en R.

Une fois les pages segmentées, des modèles de transcription automatique avec OCRopus/OCRopy et Transkribus ont été produits. Le taux d'erreur final, après l'entraînement d'un réseau de neurones récurrents sur Transkribus, était de 0,11%, ce qui est un résultat très satisfaisant. [23]

3.4 Le projet HIMANIS

Le projet HIMANIS « vise à l'indexation du texte des registres de la chancellerie royale française des années 1302-1483, conservés aux Archives nationales, à partir des images produites par leur numérisation. » [24]

A bien des égards, les objectifs du projet HIMANIS sont semblables à ceux du projet LECTAUREP : simplifier la recherche d'information dans les registres de la chancellerie en supprimant l'étape complexe du dépouillement et en améliorant les solutions préalablement mise en place (des inventaires existent déjà, produits à l'issue d'opérations de dépouillement).

Pour transcrire automatiquement les textes, le choix de Transkribus a été fait. La vérité de terrain se base sur l'édition électronique des *Actes royaux du Poitou*. La concordance entre le texte et les images a été réalisée par des renvois faits aux zones des différents manuscrits directement dans l'encodage en XML-TEI des transcriptions. [25] La vérité de terrain a été enrichie par des transcriptions supplémentaires mettant l'accent sur les abréviations.

21. *Ibid.*, p.76
22. Voir `https://fsix.github.io/mnist/Deskewing.html`
23. *Ibid.*, p.83
24. D. Stutzmann, J.F. Moufflet et S. Hamel, "Full Text Search in Medieval Manuscripts…", abstract.
25. *Ibid.*, « Ces renvois sont enregistrés par un attribut @facs inséré dans lélément <pb/>. Lorsque lacte ne couvre pas toute la page, lattribut @facs renvoie à une <zone> déclarée dans une section <facsimile> entre le <teiHeader> et le <text>. Ensuite, pour faciliter lapprentissage de la lecture, la transcription a été enrichie des indications de passage de ligne par lélément <lb/> dans près de 300 actes, parmi les 474 donnés par Paul Guérin dans ses trois premiers tomes et correspondant à la tranche chronologique des registres JJ35-JJ91, déjà numérisée par les Archives nationales en début de projet et sur laquelle sest fondé lapprentissage. »

L'indexation des registres de la chancellerie royale française se base sur cette opération d'extraction automatique. L'index « vise [...] à retrouver les occurrences dun terme choisi dans limmense continent des registres de la chancellerie royale. »[26] Pour le constituer, il a été nécessaire de s'appuyer sur des données d'autorité, des référentiels, et des outils linguistiques.

— Les outils linguistiques améliorent « la reconnaissance et lindexation du texte en régularisant et normalisant des termes que les outils automatiques auraient mal lus et dont ils auraient mutilé lun ou lautre caractère. »[27] Grâce à ces outils, il est possible de calculer un indice de confiance pour chaque mot afin d'affiner les recherches possibles dans le document. La lemmatisation permet des recherches par radical, laissant envisager des analyses lexicométriques. Un référentiel a été constitué à partir des textes des éditions, et se résume à « une liste de mots avec des statistiques de fréquence. »[28] En ce qui concerne les référentiels externes au corpus, des dictionnaires ont été utilisés, permettant de définir les termes, ainsi que des outils de lemmatisation pour le latin et le français.[29]
— La détection d'entités nommées (NER) - noms de lieux et de personnes, concepts et matières - permettrait également d'améliorer l'indexation. Le projet HIMANIS fait appel à des référentiels constitués lors des travaux des Archives Nationales sur les documents originaux, les index publiés en appendice des inventaires. Ceux-ci ont été océrisés et rétroconvertis en EAD. Ils traitent des noms de personnes et de lieux et ont pu être utilisés pour mesurer la précision de l'indexation automatique. Des référentiels externes ont également été sélectionnés :
  — Données universelles : GeoNames, Wikipedia/DBpedia : ces référentiels permettent d'obtenir des coordonnées géographiques.
  — Dictionnaires topographiques, comme par exemple Dico-topo.[30]

Dominique Stutzmann et al. mentionnent à juste titre la nécessité de s'interroger sur la nature des renvois de l'index. En effet, l'extraction d'entités nommées est d'abord localisée par le script sur une image, une page dans le document originel.[31] Cependant, ils précisent que ces renvois peuvent être problématiques, certaines parties de l'image étant ambiguës quant à leur appartenance à tel ou tel acte. Il a donc été décidé que l'indexation devrait renvoyer à des coordonnées sur l'image de l'entité extraite. Cette image renvoie elle-même à une page. Ensuite, les coordonnées « doivent être comparées à celles des actes pour

26. *Ibid.*
27. *Ibid.*
28. *Ibid.*
29. A la date de rédaction de l'article, les auteurs précisent que ces outils n'ont pas encore été implémentés.
30. A la date de rédaction de l'article, les auteurs précisent que ces outils n'ont pas encore été implémentés.
31. *Ibid.*

savoir auquel de ceux-ci attribuer le mot. » Une réflexion semblable serait potentiellement à réaliser pour le projet LECTAUREP dans le cadre de l'extraction des entités nommées (recherche d'un nom en fonction d'un type d'acte, recherche d'un nom dans un enregistrement sur une fin de page qui déborde sur la page suivante, recherche d'un nom en fonction d'une date, etc.).

Les chercheurs mettent également en avant les avantages d'un « modèle de négociation de contenu. »[32] Celui-ci permet à l'utilisateur-ice de formuler une requête en choisissant le degré de bruit informationnel toléré, et permet ainsi de brasser plus ou moins large, tout en reconnaissant que la machine puisse se tromper dans son traitement automatique.

Enfin, ils proposent une réflexion sur le degré d'incertitude généré par la transcription et l'extraction automatique. Celui-ci est mesuré en termes de précision et de rappel.

— La précision « en matière dindexation, est le pourcentage doccurrences correctes parmi celles identifiées comme répondant à une requête. »[33]
— Le rappel est « le rappel est le pourcentage doccurrences correctement identifiées parmi lensemble des occurrences dans le corpus. »[34]

Ils soutiennent que la précision est un facteur dont l'importance doit être relativisée, un-e chercheur profiterait en effet de dizaines de milliers de pages traitées automatiquement et gagnerait du temps même si un travail de correction manuelle devrait être réalisé. Le rappel est, à l'opposé, un facteur essentiel. Il permet d'atteindre l'exhaustivité et évite un silence généré accidentellement (imaginons qu'une recherche pour un nom apparaissant une seule fois dans les actes des notaires du projet LECTAUREP soit faite, et qu'elle ne renvoie rien à cause d'une erreur d'extraction, cela serait problématique).

L'indice de confiance, autre métrique, permet d'attribuer un score à chaque indexation, et laisse la liberté aux usagers de trier les résultats (voir plus haut, recherche par négociation).

D'un point de vue technique, les étapes de transcription et l'indexation automatique sont réalisés grâce à du keyword spotting (KWS). Les probabilités de pertinence des mots sont calculés en utilisant des treillis ("lattices"[35]) de caractère produits par un réseau de neurones récurrents et un modèle de langage n-gram.[36] Le keyword spotting se réalise-

32. *Ibid.*
33. *Ibid.*
34. *Ibid.*
35. Voir `https://fr.wikipedia.org/wiki/Treillis_(ensemble_ordonn%C3%A9)` , `https://en.wikipedia.org/wiki/Lattice_(order)`
36. 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.

rait, dans ce cas de figure, avec des « query-by-string » (requêtes par chaînes de caractères). Le KWS est un type de recherche différent des recherches NER. Le KWS ne nécessite pas la transcription de tous les documents, et permet d'effectuer des recherches approximatives sur l'ensemble des images du corpus.

Les résultats ont été jugés concluants. Les images du corpus ont ainsi toutes été indexées de manière probabiliste. Le nombre moyen d'entrées d'index par page a été optimisé de sorte à pouvoir réduire les contraintes de stockage, tout maintenant de bonnes performances en ce qui concerne le score de précision et de rappel. [37]

3.5 Le traitement automatique des annuaires de propriétaires de Paris

Carmen Brando et Gabriela Elgarrista ont récemment présenté, au cours de la cinquième session du séminaire bimensuel des Sources aux Systèmes dInformation Géographique, le 6 avril 2021, une chaîne de traitement destinée à l'analyse des annuaires de propriétaire de Paris. [38]

La chaîne de traitement correspond aujourd'hui à traiter des images, et en proposer une transcription automatique à l'aide d'OCR et d'HTR (anciennement Transkribus, et passage actuellement à eScriptorium). La transcription est structurée en XML/TEI, et ensuite intervient l'extraction et la spatialisation des adresses et la géolocalisation des entités récupérées.

4 Les outils de NER

4.1 Définitions avancées du NER

Dans le cadre d'un cas d'usage proposant l'analyse des mentions des acteurs de la seconde guerre mondiale dans des écrits personnels de soldats français, Luca Foppiano

37. *Ibid.*, p.316-317. « Finally the whole collection has been actually indexed. The process required about 1 month of intensive multi-core computation and the resulting probabilistic index contains about 266 million entries and requires about 10 gigabytes of storage. During this process, about three million lattices were generated, then used to compute the probabilistic index entries, and finally discarded. All in all, this workflow involved handling about 250 gigabytes of data during the whole process time span of about two months. A beta version of the query and search system for the 67,282 page images of the full Chancery collection is available at prhlt-kws.prhlt.upv.es/himanis. »Voir le site actuel du projet HIMANIS `http://himanis.huma-num.fr/himanis/`

38. DYPAC UVSQ, *cinquième session du séminaire bimensuel des Sources aux Systèmes dInformation Géographique,* URL : `https://www.youtube.com/watch?v=4xkZHdy88DU&ab_channel=DYPACUVSQ` (visité le 07/04/2021)

et Charles Riondet ont proposé des explications éclairantes sur les principes avancées des technologies NER. [39]

Afin d'analyser ces mentions, il était nécessaire de premièrement les extraire, et deuxièmement de replacer l'extraction dans la structure du discours. En effet, certaines expressions utilisés par les soldats français pour qualifier les soldats allemands relèvent du langage oral, des expressions communément partagés par les individus, qu'un programme de NER ne peut qualifier efficacement qu'en analysant le contexte syntaxique dans lequel est situé le mot. Pour chaque mention extraite, il est ainsi importer de collecter chaque token [40] aidant à comprendre le sens de la mention extraites. Ces tokens transportant l'information peuvent être des verbes et des adjectifs, et renseignent dans la phrase le sentiment attribué à la mention, mais aussi le jugement, le point de vue du soldat qui écrit, etc.

Pour analyser la structure du discours, les chercheurs s'appuient sur le « part of speech tagging (POST) », qui est défini comme étant le processus permettant d'assigner une catégorie morphosyntaxique (part-of-speech) ou une autre classe lexicale à un mot. [41] Pour cela, on utilise des classificateurs grammaticaux (grammar classifiers) basés sur 8 « parts of speech » :

1. Verbe
2. Nom
3. Adjectif
4. Pronom
5. Préposition
6. Adverbe
7. Conjonction
8. Interjection

Se faisant, on obtient des informations sur le mot extrait et sur les mots qui lui sont voisins dans le discours.

Les chercheurs ont réalisé un schéma synthétique et explicite du workflow, aidant considérablement la compréhension de cette tâche, voir figure 1. [42]

39. C. Riondet et L. Foppiano, "History Fishing When engineering meets History"…
40. Une chaîne de caractère pouvant être analysée par un ordinateur et correspondant à un mot.
41. *Ibid.*, p.4, « The approach for the recognition of structural discourse is based on the part of speech tagging (POST), which is is the process of assigning a part-of-speech or other lexical class marker to each word in a corpus. »
42. *Ibid.*

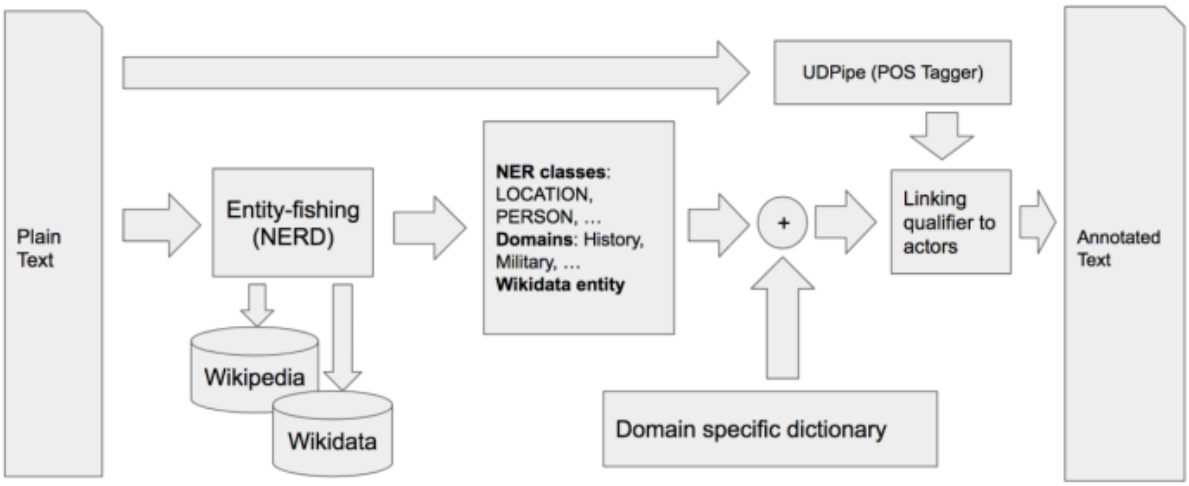

FIGURE 1 – Workflow NER

Deux outils ont été utilisés : Entity-Fishing et UDPipe. Entity-fishing est un outil de NER et de désambiguïsation se basant sur Wikipedia et Wikidata, par comparaison. [43] Nous reviendrons en détail sur cet outil. UDPipe [44] est un outil permettant de parser selon les Universal Dependencies et permet de faire le POST. (UD) [45]

Entity-fishing permet de récupérer automatiquement des « mentions », qui ne sont pas nécessairement des entités *nommées*, et de les mettre en relation avec Wikipedia et Wikidata pour trouver, ou non, un match. Cette tâche est nommée « disambiguation » ou « entity-linking ». [46]

UDPipe est une pipeline pour tokenizer, tagger, lemmatiser et pour parser les dépendances (« depencency parsing » [47]. La pipeline est construite selon un modèle entraîné. Il permet d'obtenir un graphe des dépendances de chaque phrase, avec le POST de chaque token. Chaque phrase est représentée comme un arbre, dont la racine part du verbe. Chaque token est connecté à une « head » [48] - un token plus près de la racine que le token taggé - ainsi qu'à un ou plusieurs tokens descendants. Le résultat permet d'analyser les relations grammaticales entre les mots d'une phrase. Voir figure 2. [49]

Le cas d'usage des journaux de soldats français posent des problèmes d'identifi-

43. `https://github.com/kermitt2/entity-fishing`
44. `https://ufal.mff.cuni.cz/udpipe`
45. `https://universaldependencies.org/`. Framework permettant l'annotation de la grammaire de différents langages.
46. Tanti Kristanti et L. Romary, "DeLFT and entity-fishing : Tools for CLEF HIPE 2020 Shared Task", dans *CLEF 2020 - Conference and Labs of the Evaluation Forum*, dir. Linda Cappellato, Carsten Eickhoff, Nicola Ferro et Aurélie Névéol, Thessaloniki / Virtual, Greece, 2020 (CLEF 2020 Working Notes), t. 2696, URL : `https://hal.inria.fr/hal-02974946` (visité le 09/04/2021), p.2
47. C. Riondet et L. Foppiano, "History Fishing When engineering meets History"...
48. *Ibid.*, p.5
49. *Ibid.*

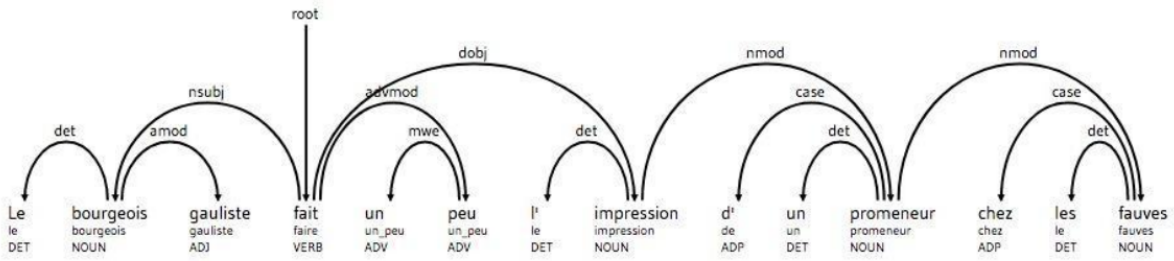

FIGURE 2 – Dependency graph

cation de vocabulaire. En effet, comme énoncé précédemment, certaines mentions sont issues d'un vocabulaire spécifique, utilisé par les locuteurs du français de cette époque. Par exemple, le terme "souris grises", désignant les auxiliaires de l'armée allemande qui étaient des femmes. [50] D'autres mentions sont issues d'une terminologie propre aux personnes ayant écrit ces journaux. On retrouve par exemple, dans un carnet, le terme "cocs", étant un diminutif péjoratif pour "coco", désignant le parti communiste français. [51] Ces particularités entraînent la nécessité de ne pas se contenter des bases de données de connaissances générales telles que Wikipedia et Wikidata, et de créer des dictionnaires spécifiques. Dans le cas d'usage discuté ici, le dictionnaire a été encodé au format XML TEI-TBX. Il rassemble les mentions désignant un concept et établit des liens entre les termes.

Le workflow se résume ainsi à ces trois étapes [52] :

1. Extraction automatique des mentions et désambiguïsation avec Wikipedia et Wikidata.
   - Entity-fishing catégorise chaque mention extraite avec des classes (jusqu'à 27) : « location », « person », « title », « organisation », « institution », etc.
   - Entity-fishing permet également d'obtenir pour chaque mention l'article/les articles Wikipedia/Wikidata lui correspondant. Le processus de désambiguïsation permet d'obtenir une prédiction sur l'article le plus à même de correspondre à la mention extraite.

2. Extraction automatique des mentions spécifiques au texte avec les dictionnaires

3. Analyse des dépendances et POST

50. *Ibid.*, p.6
51. *Ibid.*
52. *Ibid.*, p.7

4.2 Entity-fishing

4.2.1 Description de l'outil

Entity-fishing est un « generic NERD system against Wikidata. »[53] NERD pour Named Entity Recognition and Disambiguation. Cet outil est un service du projet européen DARIAH (Digital Research Infrastructure for the Arts and Humanities) et est déployé par la TGIR Huma-Num. Il accessible via une interface, permettant le traitement d'annotations standardisées (en TEI, par exemple) jusqu'à l'intégration de l'outil dans des secteurs professionnels.[54] Le système se veut également fiable et scalable, de plus, la plate-forme a été développée de façon durable.[55] Entity-fishing n'utilise pas de dictionnaire de « domaine », et laisse ainsi le choix à ses usagers de choisir l'outil de désambiguïsation des entités. L'outil peut être utilisé avec 5 langages, dont le français.

Tanti Kristanti et Laurent Romary, dans un article datant de septembre 2020, explique qu'Entity-fishing procède généralement selon trois étapes :

1. Identification du langage (« language identification »). Cette étape est nécessaire pour sélectionner les outils adéquats pour le traitement du texte (tokenizer, segmentation des phrases, etc.) et pour sélectionner un Wikipedia dans telle ou telle langue.[56]

2. Reconnaissance des mentions (« mention recognition ») à partir du document en input. Pour cela, Entity-fishing utilise Wikipedia, en comparant la mention à ce qui existe sur la base. Wikipedia permet notamment d'identifier les acronymes. Pour ce qui relève du NER, Entity-fishing se base sur la librairie GROBID-NER[57] qui permet de traiter le texte, d'extraire les entités nommées et de les classifier dans 27 classes différentes, dont nous avons déjà parlé précédemment. Pour ce faire, GROBID-NER utilise un modèle statistique de Conditional Random Field (CRF). La comparaison avec Wikipedia est complémentaire à l'étape de machine learning. Elle est réalisée avec une analyse des n-gram. Le résultat de la reconnaissance des mentions prend la forme d'une aggregated list of objects containing raw values from the original text, their actual positions, and NER classes (within the 27 classes). »[58]

53. T. Kristanti et L. Romary, "DeLFT and entity-fishing…", p.3

54. Voir 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 et `http://nerd.huma-num.fr/nerd/`, `https://github.com/kermitt2/entity-fishing`, et `https://nerd.readthedocs.io/en/latest/`.

55. *Ibid.*, p.4

56. *Ibid.*, p.11-12, T. Kristanti et L. Romary, "DeLFT and entity-fishing…", p.3

57. `https://github.com/kermitt2/grobid-ner`

58. *Ibid.*, p.4, L. Foppiano et L. Romary, "Entity-fishing…", p.12

3. Identification des entités (« entity resolution »). Cette étape associe les mentions aux entrées de Wikidata. Trois phases sont nécessaires :
   (a) La génération de candidats à partir des entrées de Wikidata. Chaque mention est liée à une liste de concepts, qui sont autant de candidats potentiellement sélectionnés par le processus de désambiguïsation.
4. Le classement des candidats par l'attribution d'un score. [59]
5. La sélection du candidat.

Le processus de désambiguïsation est l'étape où les entités sont liés avec les mentions. Une mention est définie comme étant un « textual segment, one or a combination of words, that can be identified in the text. » Les entités sont « something that exists as itself, as a subject of as an object, actually or potentially, concretely or abstractly, physically or not. » [60] La liaison (« linking ») est l'étape où une knowledge base (KB) est sélectionnée en fonction de ce à quoi la mention se réfère dans son contexte. Aujourd'hui, Wikipedia, avec Wikidata, grâce aux à leur open licence, sont les knowledge bases de référence. Entity-fishing exploite Wikipedia sur la base de son réseau de graphes de concepts, et non sur le sens des articles. [61]

Le service est accessible via une interface web mettant à disposition une REST API, permettant une intégration facilitées dans d'autres projets. [62]

Entity-fishing prend en input une requête structurée en JSON et la retourne enrichie en output avec une liste des entités identifiées et désambiguïsées. [63] L'outil propose aussi un output standardisé en XML-TEI. Il peut traiter du texte brut avec des identifications de mentions ou d'identités préalablement réalisées, pour faciliter la désambiguïsation. Il peut également traiter des documents PDF.

4.2.2 Entity-Fishing et DeLFT

Cet outil peut être utilisé en concordance avec un framework open-source nommé Deep Learning Framework for Text (DeLFT). [64] Ce framework utilise les librairies python Keras et TensorFlow pour réaliser du traitement de texte avec du deep learning. DeLFT peut être employé dans plusieurs architectures de deep learning [65] :

— Bidirectional LSTMs and Conditional Random Fields

59. T. Kristanti et L. Romary, "DeLFT and entity-fishing…", p.4, « Then, in the candidate ranking, each candidate is assigned a confidence score calculated as regression probability using various features. » Voir également L. Foppiano et L. Romary, "Entity-fishing…", p.13-14
60. *Ibid.*, p.7-8
61. *Ibid.*, p.25
62. Voir *Ibid.*, p.17-19
63. T. Kristanti et L. Romary, "DeLFT and entity-fishing…", p.7
64. `https://github.com/kermitt2/delft`
65. *Ibid.*, p.3

— Bidirectional LSTM and Convolutional Networks
— Bidirectional Gated Recurrent Unit
— Contextualized embeddings (ELMo, BERT).

DeLFT utilise des embeddings [66] depuis une source extérieure, et les gère depuis une base de données.

Dans le contexte du HIPE 2020 [67], Tanti Kristanti et Laurent Romary ont utilisé plusieurs datasets pour construire et évaluer un modèle de NER pour la langue française : le corpus annoté French TreeBank (FTB) et le dataset fourni par la campagne HIPE.

La création de DeLFT a permis de réimplémenter les architectures neuronales (DL - Deep Learning) pour les outils de NER. De plus, celles-ci présentent de meilleures performances que des systèmes de machine learning. [68]

Les performances d'Entity-fishing et de DeLFT ont été jugées satisfaisantes, mais une importante quantité de données est nécessaire. Celles-ci doivent également être de qualité. De plus, la performance peut être impactée négativement par des données bruitées. [69]

4.3 La question des datasets

Dahl et al. ont publié un article au mois de janvier 2021 partageant leur initiative de créer une base de données noms et prénoms à destination des projets de reconnaissance automatique d'écriture manuscrite : HANA (HAndwritten NAme Database for Offline Handwritten Text Recognition). [70] Grâce à la récupération des noms propres présents dans les colonnes des formulaires pré-imprimés des registres de police danois, partagés par les archives de Copenhague, HANA concentre 3,355,388 noms, distribués sur 1,106,020 images. Tous ces noms ont été récupérés depuis les registres de police danois, partagés au projet par les archives de Copenhague. [71]

Une des principales sources pour l'analyse morphosyntaxique et syntaxique des

66. `https://en.wikipedia.org/wiki/Word_embedding`

67. « HIPE (Identifying Historical People, Places and other Entities) is a evaluation campaign on named entity processing on historical newspapers in French, German and English, which was organized in the context of the impresso project and run as a CLEF 2020 Evaluation Lab. » Voir `https://impresso.github.io/CLEF-HIPE-2020/`

68. *Ibid.*, p.6, « The French model trained with the FTB corpus within the BiLSTM-CRF architecture and French Wikipedia fastText reaches an 87.45 F1-score. Meanwhile, with the use of French ELMo, the score is improving into 89.23. »

69. *Ibid.*, p.9-10

70. 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]* (, janv. 2021), arXiv : 2101.10862, URL : `http://arxiv.org/abs/2101.10862` (visité le 14/04/2021)

71. *Ibid.*, p.7

textes en français est la French TreeBank. [72] Cependant, cet ensemble ne comprend pas d'informations sur les entités nommées. Pour pallier cela, Ortiz Suárez et al. ont introduit une nouvelle version annotée de la French TreeBank au mois de mai 2020, permettant de l'utiliser plus efficacement pour les tâches de NER.

4.4 Évaluation des performances de l'OCR et des résultats post-OCR

Les données produites par l'OCR et l'HTR doivent être mesurées selon leur précision. Il en va de même pour les performances des extractions réalisées par les outils de NER, dont les résultats sont les principales portes d'entrées pour les usagers d'une bibliothèque en ligne, par exemple. [73] Pouvoir mesurer la performance des outils de NER est donc une tâche importante à ne pas négliger, afin de ne pas entraver la consultation des documents par les usagers.

Hill et Hengchen, en travaillant sur l'impact d'un mauvais OCR sur le corpus de textes Eighteenth Century Collections Online (ECCO), ont prouvé que des textes possédant un score de précision en dessous de 70% - 75% ont un impact négatif fort sur les analyses pouvant être menées sur ces documents, du moins sur le corpus utilisé pour réaliser leur études. [74] Au niveau des tokens, les meilleurs résultats qualitatifs et quantitatifs sont obtenus au-dessus de 80% de précision. [75] Ils mettent également en avant les particularités des textes du XVIIIe siècles posant problème à l'OCR, aidant à aller traiter le problèmes. Les lettres les plus difficiles à traiter s'avèrent être le s long et les ligatures. Ils préconisent également de ne pas se fier aux mesures de diversité lexicale, pouvant être trompeuses et ne pas indiquer efficacement la précision de l'OCR. Un échantillonnage aléatoire serait plus efficace pour mesurer la qualité. [76] Nous comprenons ainsi la nécessité d'avoir conscience de l'impact d'un mauvais OCR sur le corpus de LECTAUREP, et de savoir d'où les problèmes peuvent émerger dans les textes.

Hamdi et al. ont proposé dans un article de décembre 2017 une méthodologie d'analyse des performances de NER sur des documents océrisés. [77] Ils ont montré que les outils

72. P. J. Ortiz Suárez, Y. Dupont, B. Muller, *et al.*, "Establishing a New State-of-the-Art for French Named Entity Recognition"…, p.2

73. 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.1. 80% du top 500 des requêtes d'une librairie en ligne contiennent des entités nommées.

74. Mark J Hill et Simon Hengchen, "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 (déc. 2019), p. 825-843, DOI : 10.1093/llc/fqz024, p.837

75. *Ibid.*, p.840

76. *Ibid.*

77. A. Hamdi, A. Jean-Caurant, N. Sidère, *et al.*, "An Analysis of the Performance of Named Entity Re-

de NER peuvent surmonter les documents bruités océrisés (character degradation, phantom degradation, bleed-through, blurring effect) où les entités nommées ont donc été mal transcrites. Dans ce cas, les outils de NER peuvent potentiellement arriver à reconnaître des parties de l'entité et à l'identifier.

Par exemple, pour des documents bruités par du flou, ils mesurent à 21,5 % le taux de transcription erronée des entités nommées extraites par l'OCR. Cependant, dans celles-ci, 8,1% ont été justement identifiées. [78]

Cependant, d'après leurs tests, la précision des outils de NER descend de 9O% à 60% quand le word error rate et le character error rate de l'output OCR monte de 1% à 7%, et de 8% à 20%. Ils concluent que les algorithmes de NER doivent ainsi s'appuyer sur des documents océrisés avec précision pour avoir les meilleurs résultats possibles. [79]

Une analyse semblable pourrait être appliquée au corpus de LECTAUREP. Il serait intéressant de calculer le taux d'erreur du NER selon la qualité de l'HTR.

4.5 Les traitements post-OCR

Savoir analyser un texte post-OCR permet de la corriger et de le rendre le plus exploitable possible. [80]

Il existe également plusieurs types communs d'erreurs [81] :

— Single-errors typos avec une « edit distance » de 1 : schopl pour school, par exemple. »
— Multi-errors tokens avec une « edit distance » plus importante : schopi pour school par exemple.

Les erreurs d'orthographes peuvent être classifiées comme suit [82] :

— First-position errors. En moyenne, elles représentent 11% des erreurs d'OCR.
— Non-word errors (quand le token n'est pas dans le lexique) / real-word errors (quand le token correspond à un mot dans un lexique mais qu'il est dans le mauvais contexte). En moyenne, elles représentent 67,5% des erreurs d'OCR.
— Word boundaries. Quand un espace a été ajouté dans un mot (split errors, elles

cognition over OCRed Documents"...

78. *Ibid.*, p.2
79. *Ibid.*
80. 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
81. *Ibid.*, p.2
82. *Ibid.*, p.2-3

représentent en moyenne 13% des erreurs), ou quand un espace a été supprimé dans un mot (run-on errors, elles représentent en moyenne 2% des erreurs).

Ces erreurs sont causées par des problèmes de capture d'image, des symboles similaires difficiles à différencier pour le logiciel d'OCR, par la ponctuation et par la typographie. [83]

Il existe deux approches principales pour le traitement post-OCR :

1. Une approchée basée sur les dictionnaires permet de corriger les termes isolés, en ne prenant pas compte du contexte dans lequel ceux-ci s'insèrent. Cette approche ne peut traiter que des non-word errors.
2. Une approche basée sur le contexte se base sur le contexte grammaticale et sémantique des erreurs, et cherche à les résoudre en fonction. Pour cette approche, les solutions techniques se basent sur le « noisy channel model » [84], sur des modèles de langages, et sur des solutions de machine learning.

Pour corriger ces erreurs, quatre types d'opérations peuvent être réalisées à l'échelle du token [85] :

1. Suppression
2. Insertion
3. Substitution
4. Transposition

De manière générale, Nguyen et al. ont montré que la majorité des erreurs d'OCR, 58,92%, sont des single-errors tokens. Par conséquent, pour commencer à éliminer ces erreurs, il convient d'adopter une approche de traitement se concentrant premièrement sur les edit distance de 1 et de 2. [86]

En ce qui concerne la relation entre la longueur des tokens et le taux d'erreurs, 85,27% des erreurs se trouvent dans les token d'une longueur de 2 à 9. Seulement 42,1% des erreurs se trouvent dans des tokens courts (short-word errors). En termes de ratio, il est préconisé d'utiliser un seuil d'edit distance de 2 pour les tokens d'une longueur de 4, de 3 pour 10, et de 4 pour 13. [87]

83. *Ibid.*, p.3
84. `https://en.wikipedia.org/wiki/Noisy_channel_model#:~:text=From%20Wikipedia%2C%20the%20free%20encyclopedia,been%20scrambled%20in%20some%20manner.`
85. *Ibid.*
86. *Ibid.*, p.4
87. *Ibid.*

A l'intérieur des tokens même, 27,37% des erreurs se trouvent en dernière position, et 28,39% se trouvent au milieu. Les statistiques montrent également que moins de 10% des erreurs correspondent à une combinaison d'un caractère erronée au début et d'un caractère erroné à la fin du token. Nguyen et al. préconisent ainsi que le traitement post-OCR se concentre en priorité sur les erreurs solitaires, ou sur des erreurs apparaissant à la fois au début ou à la fin et après le début du token. [88]

Les erreurs post-OCR contiennent en moyenne 59,21% d'erreurs liées à des real-word errors. Il est également important de souligner qu'un lexique de taille insuffisante risque d'entraîner des faux négatifs car ignorant des tokens valides, tandis qu'un dictionnaire trop important pourrait valider des tokens invalides en les associant à des des domaines spécifiques à certains domaines, augmentant donc le nombre de faux positifs. [89]

Enfin, les statistiques concernant les erreurs liées à la limite des mots montrent que l'importante majorité de celles-ci concernent les mots qui ne possèdent pas de problèmes de limite, représentant environ 82,85% de celles-ci. [90]

5 Le Keyword Spotting (KWS)

Rusakov et al. définissent le processus de word spotting comme processus permettant de « retrieve a list containing word images that are relevant with respect to the [user] query. » De plus, « word spotting methods rank all retrieved word images from a given document collection by a certain criterion and sort them by their similarities. » [91]

Les queries, ou requêtes, peuvent être de deux natures [92] :

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 les plus 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 caractère. Ce dernier se charge de récupérer les occurrences similaires à la chaîne de caractère dans le document soumis à la requête. Cette requête ne renvoie pas forcément des résultats en image.

88. *Ibid.*, p.8
89. *Ibid.*
90. *Ibid.*, p.9
91. 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
92. *Ibid.*

Gurjar et al. ont mis en avant avec un article publié au mois de janvier 2018 une méthodologie pour accélérer l'entraînement des réseaux de neurones convolutifs pour les tâches de KWS. En effet, bien que présentant de très bonnes performances, l'entraînement de ces modèles prend beaucoup de temps de par la nécessité d'avoir à disposition un nombre conséquent de données annotées. Pour cela, les chercheurs ont entraîné un modèle sous « weak supervision » - avec des documents bruités -, en utilisant une combinaison de données d'entraînement créées artificiellement et une petite part de données d'entraînement directement issues des images d'un corpus soumis à l'HTR. [93] Les résultats obtenus par l'équipe ont été jugés satisfaisants.

6 Outils envisagés après ces premières lectures

1. Entity-fishing...
   — `https://github.com/kermitt2/entity-fishing`
   — `https://nerd.readthedocs.io/en/latest/`
2. ...accompagné de DeLFT
   — `https://github.com/kermitt2/delft`
3. spaCy
   — `https://spacy.io/`
4. T-NER
   — `https://github.com/asahi417/tner`
5. (Neural network system based LSTM-CRF) [94]

Des recherches restent à faire sur :

— CamemBERT
   — `https://camembert-model.fr/`
— OSCAR
   — `https://oscar-corpus.com/`
— "Sequential tagging methods : Hidden Markov Models" [95]

93. Neha Gurjar, Sebastian Sudholt et Gernot A. Fink, "Learning Deep Representations for Word Spotting Under Weak Supervision", *arXiv :1712.00250 [cs]* (, janv. 2018), arXiv : 1712.00250, URL : `http://arxiv.org/abs/1712.00250` (visité le 13/04/2021)

94. A. Hamdi, A. Jean-Caurant, N. Sidère, *et al.*, "An Analysis of the Performance of Named Entity Recognition over OCRed Documents"..., p.1. "It generates the most probable sequence of predicted labels from surrounding words. Long Short-Term Memory (LSTM) networks compute a representation of the context of each word. A CRF layer allows generating the most probable sequence of predicted labels from surrounding words." Voir Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. arXiv preprint arXiv :1603.01360 (2016).

95. *Ibid.* Voir Daniel M Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1998. Nymble : a high-performance learning name-finder. arXiv preprint cmp-lg/9803003 (1998).

— Conditional Random Fields (CRFs). [96]

— SEM

  — `https://github.com/YoannDupont/SEM`

— NERVAL - Teklia

  — `https://teklia.com/blog/202104-nerval/`

96. *Ibid.* Voir Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. Journal of Machine Learning Research 12, Aug (2011), 24932537.

Bibliographie

ARTAUD (Chloé), SIDÈRE (Nicolas), DOUCET (Antoine), OGIER (Jean-Marc) et POULAIN D'ANDECY (Vincent), "Find it! Fraud Detection Contest Report", dans *24th International Conference on Pattern Recognition (ICPR 2018)*, Beijing, China, 2018, pp. 13-18, URL : https://hal.archives-ouvertes.fr/hal-02316399 (visité le 09/04/2021).

asahi417/tner, en, URL : https://github.com/asahi417/tner (visité le 09/04/2021).

BLUCHE (Théodore), HAMEL (Sébastien), KERMORVANT (Christopher), PUIGCERVER (Joan), STUTZMANN (Dominique), TOSELLI (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.

CHAGUÉ (Alix), *Comment faire lire des gribouillis à mon ordinateur?*, 2021, URL : https://hal.archives-ouvertes.fr/hal-03170345 (visité le 06/04/2021).

CHAGUÉ (Alix) et AURÉLIA (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).

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]* (, janv. 2021), arXiv : 2101.10862, URL : http://arxiv.org/abs/2101.10862 (visité le 14/04/2021).

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).

DYPAC UVSQ, *cinquième session du séminaire bimensuel des Sources aux Systèmes dInformation Géographique*, URL : https://www.youtube.com/watch?v=4xkZHdy88DU&ab_channel=DYPACUVSQ (visité le 07/04/2021).

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.

GURJAR (Neha), SUDHOLT (Sebastian) et FINK (Gernot A.), "Learning Deep Representations for Word Spotting Under Weak Supervision", *arXiv :1712.00250 [cs]* (, janv. 2018), arXiv : 1712.00250, URL : http://arxiv.org/abs/1712.00250 (visité le 13/04/2021).

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)*, Is-

sue : 1, Champaign, United States, 2019, t. 24, p. 333-334, DOI : 10.1109/JCDL.2019.00057.

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 (déc. 2019), p. 825-843, DOI : 10.1093/llc/fqz024.

KRISTANTI (Tanti) et ROMARY (Laurent), "DeLFT and entity-fishing : Tools for CLEF HIPE 2020 Shared Task", dans *CLEF 2020 - Conference and Labs of the Evaluation Forum*, dir. Linda Cappellato, Carsten Eickhoff, Nicola Ferro et Aurélie Névéol, Thessaloniki / Virtual, Greece, 2020 (CLEF 2020 Working Notes), t. 2696, URL : https://hal.inria.fr/hal-02974946 (visité le 09/04/2021).

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).

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.

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).

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).

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).

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.

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).

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.