De la variation linguistique et de son influence sur l'application de méthodes de Traitement Automatique des Langues

Gaël Lejeune

SORBONNE UNIVERSITÉ

École Doctorale 130, EDITE (Informatique, Tlcommunications et Electronique)

De la variation linguistique et de son influence sur l'application de méthodes de Traitement Automatique des Langues

HABILITATION

présentée et soutenue publiquement le 18 décembre 2023

pour l'obtention d'une

Habilitation de Sorbonne Université

(mention informatique)

par

Gaël Lejeune

Composition du jury

Président :	Franck Neveu, Professeur des Universités, STIH, Sorbonne Université
Rapporteurs :	Cécile Fabre, Professeure des Universités, CLLE, Université de Toulouse, Éric Gaussier, Professeur des Universités, LIG, Université Grenoble Alpes Laurent Romary, Directeur de Recherches INRIA, Almanach, INRIA Paris
Examinateurs :	Franck Neveu, Professeur des Universités, STIH, Sorbonne Université Franois Rioult Matre de Conférences HDR, GREYC, Université de Caen Virginie Julliard, Professeure des Universités, GRIPIC, Sorbonne Université (Garante)

Sens Texte Informatique Histoire (STIH)

Mis en page avec la classe thesul.

À mes parents, mon frère, ma sur

À Alice, Oscar, Marianne et Jacques

Remerciements

Je tiens tout d'abord à remercier mes rapporteurs Cécile Fabre, Éric Gaussier et Laurent Romary de m'avoir fait l'honneur de relire ce tapuscrit et de siéger dans le jury. Votre présence dans ce processus m'a rassuré, vos remarques m'ont éveillé, et je ne doute pas que vos questions vont m'inspirer. Mon jury est aussi constitué de personnes qui ont fortement contribué à la construction de l'enseignant et du chercheur que je suis. Je remercie Franck Neveu d'avoir accepté de faire partie de ce jury. C'est un immense plaisir de pouvoir présenter mon travail devant le chercheur qui avait suscité auprès de l'étudiant de Sciences du Langage que j'étais, une passion jamais démentie depuis pour l'étude des faits linguistiques. Je remercie ensuite François Rioult qui a élargi les horizons de ma recherche grâce à nos discussions sur la fouille de données, l'apprentissage automatique, l'algorithmique du texte . . . et aussi provoqué mon intérêt pour de multiples objets de recherche. Ta compétence et ta bienveillance ont beaucoup compté pour moi. Enfin, Virginie Julliard. Virginie, en regardant dans le rétroviseur j'ai peine à croire que cela ne fait que 4 ans que nous nous connaissons. Toutes les idées échangées, les projets, les questions de recherche suffiraient sans difficulté à nourrir un deuxième tome d'habilitation. Je remercie les personnes engagées dans le lourd travail de relecture, mes relecteurs et relectrices : Karine Abiven, Karën Fort, Charlotte Lecluze, Alice Millour, Jordane Raisin-Dardre, Lichao Zhu.

Quand j'ai commencé à écrire ce tapuscrit, 10 années s'étaient écoulées depuis ma transition des sciences du langage vers la science informatique. Avant de démarrer cette aventure pour de vrai, j'ai fait une rencontre qui a marqué ma vie, à savoir Monsieur le Professeur des Universités Antoine Doucet. Notre passion commune pour le ballon rond et les céréales qui font des bulles nous a un peu poussé dans les bras l'un de l'autre. Antoine, tu m'as accordé ta confiance depuis le *very beginning* et tu m'as aidé pendant toute ma carrière jusqu'ici, en espérant que cela continue, ou que je devienne pleinement autonome un jour !

Quand j'ai commencé à écrire cette habilitation, 6 ans avaient passé depuis ma soutenance de thèse. Ma dette, contractée durant la thèse (voir [Lejeune, 2013]), envers Romain Brixtel n'est sans doute pas tout à fait soldée, nos conférences ACM inspirent encore nombre de mes travaux. Je n'oublie pas bien sûr Charlotte Lecluze qui fut là dans les mauvais moments de la rédaction pour me pousser vers l'avant. J'ai ensuite eu la chance de faire un certain nombre de très belles rencontres scientifiques. En premier lieu, lors de mon passage à Nantes au LINA (aujourd'hui LS2N) où j'ai beaucoup appris, notamment auprès de Béatrice Daille et EMmanuel Morin. De retour sur Caen, au-delà du cas François Rioult évoqué plus haut, j'ai eu la chance de trouver au sein de l'équipe CODAG et du projet Nareca un creuset scientifique fantastique avec des personnes comme Bruno Crémilleux, Jean-Philippe Métivier ou encore Alexandre Pauchet. Je remercie Emmanuel Cartier qui m'a offert un deuxième et passionnant contrat post-doctorat en m'accueillant au sein du LIPN. En plus d'éveiller mon intérêt pour la néologie de toute sorte, cette expérience m'a permis de rencontrer des collègues de grande valeur tels que Davide Buscaldi, Thierry Charnois, Audre Grezka et Joseph Le Roux. Et, par le hasard des tubes à poster, j'ai pu aussi grâce à cette expérience au LIPN faire la connaissance de l'excellent Caio Corro .

Dans l'étape suivante de ma carrière j'ai eu la chance de connaître, en plus de Virginie Julliard, quatre femmes exceptionnelles qui ont beaucoup marqué mon séjour à la Faculté des Lettres. Chronologiquement, Karën Fort qui, grâce à nos discussions régulières sur nos intersections et nos complémentaires nous amené (enfin !) à publier ensemble. Merci Karën, pour tout ce que tu as apporté à ma pratique professionnelle, pour nos discussions animées et pour

ton soutien indéfectible [1].

La faculté des Lettres de la Sorbonne reste évidemment marqué pour moi par la personne de Karine Abiven. Nous avons partagé les premières idées sur les données bruitées, les premiers projets, la garde des jeunes collègues de la 212, des articles, des corpus, des projets de thèse, n'en jetez plus. Merci à toi Karine, d'avoir un jour de janvier 2018 coché sur ton petit carnet l'idée d'écrire à ce néo-MCF en Informatique qui fut tout heureux de découvrir les tourbillons de la vie XVIIèmiste.

La faculté des Lettres, c'est aussi la maison de la recherche de la rue Serpente et son groupe de recherche et d'animation sociologique : Alice Millour et les petit.e.s serpentin.e.s. Que de beaux moments partagés grâce à toi Alice, avant COVID dans nos repaires de St Michel, pendant COVID dans la folie des visio-conférences régulières pour rompre la monotonie du confinement et après COVID dans tes réguliers et étincelants retours à la maison-mère. Dans la troupe des serpentins, nombre d'évènements ont rythmé notre vie de bureau toutes ces années (`Kobra, Sibon, Thamnophis` et consorts), du Karaoké de juillet à la Raclette de janvier en passant par les bars corses et les pots de thèse, les rencontres en 206, les déjeuners en 205, en 212, dans le couloir sur la terrasse ... Merci pour cette belle ambiance Théophile Bagur, Hana Bouchired, Solenn Caroff, Pierre-Marie Chauvin, Nicolas Cheplagin, Romain Davière, Margot Déage, Renaud Debailly, Sylvain Fontaine, Laurie-Anne Galiby, Camille Gillet, Cyril Jayet, Hugo Jeaningros, Sébastien Mosbah-Natanson, Gaëlle Meslay, Hugo Mestayer, Margot Lenouvel, Raphaël Piters, Lucas Sage, Zara Salzmann, Melchior Simioni, Marie Trespeuch, Hugo Touzet et Elise Verley. Merci également à Gianluca Manzo et Thierry Tirbois pour leurs sympathiques encouragements.

Un grand merci évidemment à Shehrazad Lakaf pour ses réponses à mes 1001 questions arrivant dès 8h05, à peine le café coulé. Un spécial à Anastaz ja Gavr lov pour les emplois du temps et les fous rires et merci aussi à ma compatriote manchotte Amélie Renault d'avoir supporté tout ce bruit. J'ai eu le plaisir de beaucoup échanger pédagogiquement avec Pascal Boldini et Christian Vincent, merci pour tout ce que vous m'avez apporté ! Un grande pensée pour Françoise Guérin, ne peut plus pouvoir partager avec toi sur les théories linguistiques et leur applicabilité informatique va être un vrai déchirement.

Serpente depuis 2017, ce fut aussi des vagues successives de jeunes chercheurs et chercheuses en informatique qui par leur activité ont permis de voir arriver ces rares jours où les informaticien.ne.s dominèrent numériquement les sociologues, une pensée particulière à Fouad Aouinti , Alexandre Bartz , Amina Boublenza , Toufik Bouheziz , Andrea Briglia , Corina Chutaux , Pauline Delahaye , Yoann Dupont , Dhaou Ghoul , Carlos Gonzalez Eleni Kogkitsidou, , Imed Laaridh , Caroline Langlet ,Luce Lefeuvre, Luis Moreno , Yingsi Liang ,Vincent Lully , Olga Seminck, Tian Tian et Lichao Zhu bien sûr puisque nous avons enfin finit par partager, sur un temps certes bien trop court, le même bureau, bouclant ainsi une série de colocations estudiantines, scientifiques familiales et administratives.

Serpente c'est aussi tout une équipe de collègues qui nous rendent la vie plus facile et avec qui il est toujours agréable de partager des petits moments de bonne humeur, Erwan Alcan, Andres Angulo Mora, Patrice Badin, Olivier Crouillebois, Guy Delaporte, Marie Delaporte, Anne-Claire Garcia, Daouda Guiro, Sylvie Lamain, Olivier Lemire, Laure Quiefin, Maxime Quiefin et aussi Olivier Lessoud grâce à qui la raclette de bureau est devenue une activité autorisable ;-) C'est aussi les facilitateurs et facilitatrices qui se décarcassent pour environner la rechercher et sans qui rien ne serait possible, je pense en particulier à Elena Billi-Rizza, Julie Legangneux, Joan Owangal, Marina Salvetti, Giovanni

1. Like skyscrapers rising up Floor by floor, I'm not giving up [...] No, I'm not down, I'm not down

Tona et Barbara van Doosselaere. Serpente c'est encore, le CERES et cette formidable dynamique de recherche initiée par Virginie Julliard et Caroline Marti. La team 007 m'a occasionné tellement de moments forts, irracontables hors contexte, avec Félix Alié, Léa Andolfi, Kenza Benabdelouhab, Julien Bezançon, Edouard Bouté, david Gödicke, Thibault Grison, Marceau Hernandez, Rimane Karam, Adélie Laruncet, Marine Tiger et merci à Thomas Bottini à qui je dois aussi ma présence dans cette belle aventure.

Serpente c'est une grande famille mais dans cette famille, il y eut aussi celles et ceux qui nous ont quittées, Philippe Laublet, Joseph Chaulleau, Didier Lapeyronnie. Et bien sûr Beate Collet, merci à toi Beate d'avoir lors de notre toute dernière rencontre relancé ma motivation et réveillé le feu qui s'éteignait.

Si le feu sacré de la recherche ne s'est pas tu, je le dois aussi à mes étudiants et étudiantes, en particulier celles et ceux qui ont eu pour une période plus ou moins longue à supporter mon encadrement en mémoire ou en thèse. D'Anaëlle Baledent, mon tout premier padawan, à Julien Bezançon le défigeur fou, en passant par Jean-Baptiste Tanguy, Caroline Koudoro-Parfait, HeeSoo Choi, Nicolas Hiebel, Tiago André De Carvalho Bene, Richy Buth, Solveig Poder, Morgann Sabatier, Zijian Wang, Amélie Hip, Steve Mutuvi et Khoa Nguyen.

Et il reste bien d'autres personnes auxquelles je dois de beaux moments intellectuels comme amicaux, que ce soit à STIH autour de Annna Arzoumanov, Romain Benini, Anne Carlier, Gilles Couffignal, Antoine Gautier, Philippe Monneret, Gilles Siouffi, dans des travaux avec des co-auteurs de génie tels que Alexandre Bartz, Adrien Barbaresi, Emanuela Boros, Frédéric Dumonceaux, Adam Jatowt, Yves Lepage, Pedro Ortiz Suarez, Yannick Toussaint, Manuela Yapomo. Ce sont aussi bien des projets pédagogiques et scientifiques qui m'ont amené à travailler avec Motasem Alrahabi, Claire Badiou, Denisa Bumba, Gabriella Parussa, Glenn Roe, Richard Walter ou Karell Bertet, Fabrice Issac et Xavier-Laurent Salvador.

Finir cette habilitation en retravaillant cette si importante section des remerciements, c'est se rappeler que l'on n'est pas grand chose tout seul et que tout le sel de l'enseignement et de la recherche ce ne sont ni des institutions, ni des politiques publiques mais ce sont des humains qui décident ou non de collaborer, de se remettre en question, de discuter et qui permettent, par leur compétence et leur bienveillance, à l'université française et à ses acteurs et actrices de traverser bien des tempêtes. Merci donc à tous les collègues pré-cités de faire de mon univers pro un univers formidable.

Table des matières

Avant-propos

Les recherches présentées dans cette habilitation forment un résumé de mon parcours de recherche post-thèse, débuté par deux contrats d'ATER, un premier à l'IUT de Cherbourg (tout en continuant ma recherche au GREYC) puis à la faculté des Sciences de Nantes (au sein de l'équipe TALN du laboratoire LINA), ensuite en tant que post-doctorant au sein de l'équipe de fouille de données (CODAG) du GREYC et enfin dans l'équipe RCLN du LIPN (Université Paris XIII, Villetaneuse). Ces recherches se sont poursuivies depuis ma prise de poste en 2017 dans l'équipe de Linguistique Computationnelle (laboratoire STIH) de Paris-Sorbonne (aujourd'hui Faculté des Lettres de Sorbonne Université). Depuis cette prise de poste, deux projets d'ampleur importante m'ont permis de développer ma vision du TAL, au sens large, en m'intéressant à de nouveaux objets de recherche. D'une part, à partir de janvier 2018 dans le projet Antonomaz (Analyse Automatique et Numérisation des Mazarinades) [2] sur l'analyse de données textuelles du 17ème siècle avec Karine Abiven. D'autre part, à travers la fondation en février 2021, par Virginie Julliard et Caroline Marti, de l'unité de service CERES (Centre d'Expérimentations en Méthodes Numériques pour les SHS) au sein de laquelle je suis depuis fortement impliqué et dont j'assure la direction adjointe.

Je voudrais ici mettre en avant l'aspect collectif, central dans la recherche universitaire telle que je la conçois. Bien entendu, ces recherches ne suivent pas une droite ligne, en raison de l'influence sur mon historique de chercheur d'une succession de rencontres qui ont façonné mon profil et ont contribué à motiver mon intérêt pour les problématiques épistémologiques et pratiques en TAL.

C'est aussi une analyse personnelle de mon parcours de recherche à ce jour. Je n'ai pas cherché à intégrer tous mes travaux, ni toutes les publications associées, de manière à conserver une certaine cohérence. Le choix de ne pas procéder à un inventaire à la Prévert des réalisations auxquelles j'ai participé a amené à quelques décisions douloureuses, j'espère que les collègues dont les collaborations ne sont pas intégrées ici ne m'en tiendront pas rigueur. Enfin, je chercherai à conserver une dimension expérimentale importante dans cette synthèse de manière à raccrocher plus facilement mes interrogations sur les paradigmes du TAL à des applications concrètes.

Dans une première partie, je propose une réflexion épistémologique sur le réductionnisme en Traitement Automatique des Langues, à savoir ce paradigme méthodologique consistant à découper les problèmes complexes non résolus en sous-problèmes pour les simplifier. Ce paradigme favorise la factorisation et à la modularisation qui sont des objectifs majeurs de la science et du développement informatique. Je défends dans cette partie l'idée que le paradigme réductionniste se heurte à différents types de variation inhérentes aux données langagières, rendant difficile la définition d'observables intangibles, de chaînes de traitement véritablement ré-utilisables et posant finalement la question de la pertinence de chercher à standardiser les données dans des contextes expérimentaux réels. Ceci m'amène à chercher comment la souplesse méthodologique, permet de ne pas rester coincé dans des schémas expérimentaux trop contraignants, des « recettes » toutes faites. Je propose notamment de poser la question d'exploiter des observables de granularité différentes, à la fois pour (I) explorer pleinement les situations où la richesse des données autorise la variation des observables et (II) les cas où le bruitage ou l'hétérogénéité des données imposent de fait d'explorer des pistes originales.

La suite de cette habilitation est plus expérimentale, elle débute par des questionnements autour de la notion de granularité dans les analyses automatiques et en particulier aux alternatives existant à une vision centrée autour de la notion de mot dont je souhaite interroger la

2. `http://antonomaz.huma-num.fr/` consulté le 2 octobre 2023

systématicité ou disons la centralité.

Cette systématicité amène avec elle un certain nombre de paradigmes dont il me semble important d'examiner la pertinence. Dans une première partie, j'interroge donc le paradigme qui amène à voir avant tout dans la langue des mots et des phrases et je propose quelques résultats et analyses pour enrichir les interrogations sur la pertinence de considérer le mot graphique (ou ses dérivés) comme un descripteur incontournable. Je m'intéresse tout d'abord à deux problèmes d'attribution d'auteur dans le chapitre 2. D'une part, en examinant l'intérêt de redescriptions très simples telles qu'une approche en patrons syntaxiques pour rivaliser avec des méthodes purement lexicales, en termes de classification non supervisée, sur un corpus littéraire. Il s'agit d'un travail mené avec Anaëlle Baledent alors étudiante à Sorbonne Université dans le Master Langue et Informatique. Ensuite, en classification supervisée cette fois, en s'intéressant aux n-grammes de caractère (mots ou non mots), cette fois sur un corpus de presse. Ce travail est une collaboration avec mes collègues de l'Université de Caen, Romain Brixtel et Charlotte Lecluze. Enfin, dans le chapitre 3 je m'intéresse cette fois aux aspects du grain document et aux observables qui peuvent être utilisés lorsqu'une structure riche de document est disponible. Ces travaux ont démarré lors de mon contrat d'ATER à l'Université de Nantes dans le cadre d'une collaboration fructueuse avec Béatrice Daille.

Cette question des observables a trouvé un terrain d'expression fécond dans le cadre du défi fouille de textes (DEFT) auquel je participe depuis 2011. Une série d'éditions du DEFT où la tâche à réaliser était de la classification supervisée, travaux réalisés notamment lors de mon post-doctorat dans l'équipe RCLN de l'Université Paris XIII (aujourd'hui Sorbonne Paris Nord) notamment avec Davide Buscaldi, Aude Grezka et Joseph Le Roux. Il s'agissait en l'espèce de l'analyse des tweets en français : Deft 2015 [Lejeune and Dumonceaux, 2015] sur l'analyse de sentiments, Deft 2017 sur le langage figuratif et l'analyse d'opinions [Buscaldi et al., 2017] et Deft 2018 consacré à l'analyse d'opinions et à l'analyse thématique. Dans ce dernier article, les résultats d'approches au grain caractère ont été particulièrement spectaculaires puisque nous avons atteint la première place sur une des tâches du Défi. Notamment, parce que la souplesse des chaînes de caractères permet de corriger les limites du grain mot en particulier dans des contextes où l'orthographe est moins normée que dans des écrits plus institutionnalisés. Enfin, compléter des approches en mots par des approches en caractères s'avère assez fécond en terme de résultats de classification de même que pour des tâches de régression. À ce titre , voir par exemple [Dupont et al., 2021] et [Ben Ltaifa et al., 2022] pour des travaux menés avec des collègues de STIH sur la notation automatique de copies d'étudiants, toujours dans le cadre du DEFT.

L'utilisation de chaînes de caractères ne saurait être, plus que la tokenisation systématique, la panacée du Traitement Automatique des Langues. Toutefois, en particulier sur une langue non standard ou bruitée, il peut être fructueux de traiter la donnée telle quelle sans nécessiter une préparation, une correction en amont, ce que permettent ces observables. Sur des données réelles, nous sommes dans une situation différente du corpus Xml bien formé, issu d'un processus de redescription des données, dont les observables ont été polis. Les données dont je vais parler ici ne sont pas considérées comme prêtes à être traitées par des méthodes de TAL et par conséquent, il est d'usage de considérer que ces données doivent impérativement être préparées, pré-traitées (en enlevant la ponctuation, les fautes d'orthographe ou encore les mots outils. . .). Toutefois, le concept de pré-traitement ne soit lui même pas complètement satisfaisant. Selon [Millour, 2020], ce sont en fait des traitements à part entière puisqu'ils ont un impact non nul sur les opérations subséquentes réalisées (la tokenisation en tout premier lieu). Dissocier ces deux étapes amène, à mon sens, à conférer au premier un statut inférieur aux seconds et à dénier aux pré-traitements une influence, positive ou non, sur les résultats. Ainsi, les pré-traitements sont souvent à la fois

peu décrits, et peu justifiés expérimentalement. De plus, à travers ces altérations du matériau d'origine il y a un risque faire disparaître des observables qui pourraient s'avérer utiles. Par exemple, les variations orthographiques et le bruit qui peuvent être des caractéristiques idiosyncratiques d'un thème, d'un auteur ou encore d'une période temporelle (voir [Baledent et al., 2020] pour une étude sur des corpus anciens bruités). Je pense donc qu'il ne faut modifier/supprimer les objets textuels traités que si l'on a l'assurance que cela améliore les résultats, les rend plus interprétables ou éventuellement plus rapides à obtenir.

La troisième partie de mon travail s'intéresse justement à la double question de la conservation des observables selon les modalités de constitution des données, et de l'applicabilité en aval de méthodes de traitement automatique des langues. La question transversale de ce chapitre est celle de l'évaluation de la qualité des données et notamment la faculté qu'ont les mesures d'agrégation (la moyenne arithmétique en particulier) de masquer les sauts qualitatifs dans les corpus.

Je présenterai tout d'abord cette problématique de conservation des observables lors de la collecte de données issues du web via un processus de *text scraping* (chapitre 4), travaux menés avec Lichao Zhu (à l'époque post-doctorant à Paris XIII) puis avec Adrien Barbaresi (BBAW, Berlin) et Emmanuel Giguet (GREYC, Caen). Puis, dans le chapitre 5, je m'intéresse aux données obtenues avec des outils d'OCR (*Optical Character Recognition* ou reconnaissance optique de caractères) qui ont fait l'objet de deux thèses que je co-encadre. D'une part, la thèse de Jean-Baptiste Tanguy sur l'applicabilité des techniques de TAL à des données bruitées (co-encadrement Glenn Roe et Karine Abiven, soutenue le 16/09/2022), thèse qui a été financée par le Domaine D'intérêt majeur (DIM) Sciences du texte et Connaissances Nouvelles (STCN [3]) et qui était liée au projet Antonomaz susmentionné. D'autre part la thèse de Caroline Koudoro-Parfait (co-encadrement Glenn Roe et Motasem Alrahabi) sur la comparaison de systèmes d'extraction d'entités nommées dans des textes littéraires, thèse financée par l'initiative SCAI de Sorbonne Université [4]. Cette orientation de mes recherches m'a amené à participer à la création, à l'initiative de Caroline Koudoro-Parfait, de Christophe Kermorvant de la société Teklia [5] et de Joseph Chazalon d'EPITA, d'un groupe d'intérêt sur l'extraction d'entités sur des documents historiques numérisés [6]. Enfin, ceci rejoint d'autres préoccupations qui émergent dans la communauté TAL, visibles par exemple à travers l'organisation avec Caio Corro du LISN d'une journée d'études ATALA sur la robustesse des systèmes de TAL [7] qui a eu lieu à la Sorbonne le 25 novembre 2022. Ces problématiques sont au centre des recherches que je compte mener sur les prochaines années et que je présenterai dans la conclusion de cette synthèse.

3. `https://www.dim-humanites-numeriques.fr/` consulté le 2 octobre 2023
4. `https://scai.sorbonne-universite.fr/` consulté le 2 octobre 2023
5. `https://teklia.com/` consulté le 2 octobre 2023
6. `https://ner-for-historical-docs.github.io/` consulté le 2 octobre 2023
7. `https://www.atala.org/content/robustesse-des-systemes-de-tal` consulté le 2 octobre 2023

Première partie

Le paradigme réductionniste en Traitement Automatique des Langues face au défi de la variation

Chapitre 1

Qu'est-ce que le TAL sait dire du langage naturel ?

Sommaire

Je pars d'un double questionnement sur l'épistémologie de ma discipline qui est le suivant : *in fine* que souhaite-t-on observer dans des textes en langue naturelle ? Et, que sait-on observer dans ces mêmes données ? Ceci m'a amené à m'interroger sur la pertinence des différents éléments classiques, des différents chaînons « incontournables », d'une chaîne de traitement de TAL de « recettes » traditionnelles incluant, explicitement ou non, nettoyage, tokenisation, étiquetage et lemmatisation ou encore passage par des plongements lexicaux, des modèles de langues contextuels ou des architectures neuronales :

— Quel le fondement linguistique de ces méthodologies ?
— Leur efficacité est-elle empiriquement démontrée ?
— Ou survivent-elles par la force de l'habitude ?
— Quelle est leur robustesse face à la variation des tâches et des données à traiter ?

On peut considérer que le TAL contribue à définir une série de résultats méthodologiques visant à réaliser des tâches particulières efficacement sur différents types de données. Il s'agit

donc d'articuler des données et des tâches avec des méthodes que l'on cherche à améliorer graduellement pour donner de meilleurs résultats sur des données connues ou sur de nouvelles données, sur des cas bein documentés ou des cas peu documentés ... C'est ici que la notion de variation me paraît très importante. Les données textuelles sont rarement disponibles dans un état standard, unique ou même figés, les corpus vont varier :

— en langue, on ne traite pas seulement de l'anglais ni même seulement des états de langue contemporains ou standards ;
— en genre textuel, avec un style décriture variable tant au niveau lexical ou syntaxique qu'au niveau stylistique ;
— en qualité, les documents traités peuvent être bruités ou incomplets, notamment du fait du processus de collecte ;
— en homogénéité, au sein même d'un corpus ou d'un document la qualité ne sera pas constante, les corpus sont souvent traversés par des sauts qualitatifs.

Le traitement séquentiel et ascendant généralement impliqué par les chaînes de traitement classique de TAL pose question sur sa capacité à embrasser ces variations. Le fait d'appliquer régulièrement cette recette à des données linguistiques indépendamment de leurs caractéristiques propres (langue et genre textuel notamment) est sans doute discutable d'un point de vue méthodologique comme d'un point de vue d'efficacité pure. Si je reviendrai dans les pages qui suivent sur mes premières expériences dans le domaine de la veille épidémiologique multilingue, c'est que c'est la question du coût marginal de traitement d'une nouvelle langue par cette méthodologie ou ses dérivés qui avait fait naître des interrogations épistémologiques dans mon esprit [Lejeune, 2013]. En économie, schématiquement, le coût marginal est le coût de production de l'unité $N + 1$ quand on a déjà produit N unités. Le coût marginal est faible, et décroissant, tant que l'on ne doit pas investir dans de nouvelles machines et que l'on a donc une économie d'échelle. Inversement, il y a des effets de seuils important quand on atteint le maximum de sa capacité de production et que l'on doit par exemple investir dans un nouveau bâtiment ou une nouvelle machine.

Appliquée au TAL, la notion de coût marginal permet de réfléchir aux coûts impliqués par le traitement de nouvelles données. Un cas intéressant est le traitement de nouvelles langues. Étant donnée une chaîne de traitement qui réalise une tâche sur une langue A très bien dotée (au hasard l'anglais), que doit-on mettre en uvre pour traiter une langue B moins dotée quoique bien dotée (disons le chinois) puis une langue C très peu dotée (par exemple le mayalam [8]). Dans le cas d'une chaîne de traitement classique comme d'une chaîne de traitement « neuronale », on voit par exemple que si la tokenisation a longtemps posé problème en chinois, on peut désormais trouver « sur l'étagère » des tokeniseurs tout à fait raisonnables même si sans doute moins efficaces que pour l'anglais. Toutefois, il n'est pas certain que l'on en trouve un de qualité raisonnable pour la langue C. La tokenisation, en particulier dans des langues peu dotées et typologiquement éloignées de l'anglais [Ogundepo et al., 2022], s'avère difficile et est évidemment un frein à l'exploitation de tout produit dérivé de la tokenisation [Rust et al., 2020]. Le coût de traitement d'une langue peu dotée peut donc être significatif, amenant cette situation fréquente dans la litérature du domaine où l'expérimentation mutlilingue figure plus souvent dans les perspectives que dans les contributions. Le coût marginal va augmenter selon différents aspects :

la disponibilité : l'existence même des pré-requis (outils et ressources) pour déployer une technique donnée sur une nouvelle langue, l'absence potentielle de ces pré-requis augmentant de fait le coût marginal ;

8. Langue dravidienne parlée dans le sud de l'Inde, plus de 30 millions de locuteurs

l'efficacité : ces outils et ressources n'ont pas forcément la même qualité dans la nouvelle langue à traiter, invitant donc à les améliorer ou à prévoir des infrastructures logicielles spécfifiques ;

la faisabilité technique : avec la prise en main d'outils non maintenus ou aux documentations parcellaires voire pour des tâches plus complexes la disponibilité même d'une infrastructure matérielle permettant leur exécution [9]

Concernant les deux premiers aspects on voit bien que la tendance générale étant de traiter les langues bien dotées puis les langues moins dotées, le coût marginal est naturellement croissant. Si le troisième aspect est d'un certain point de vue du ressort de l'ingénierie, il pose aussi de vraies questions sur la capacité de méthodes état de l'art peu frugales à traiter beaucoup de langues. Ces aspects incitent à mon sens à questionner la capacité des paradigmes traditionnels du TAL, notamment ceux où la tokenisation occupe un rôle central, à traiter des situations variées. Bien souvent, l'analyse des textes se fait de façon ascendante dans un processus où des informations sur les mots on tire une représentation des textes en passant éventuellement par une représeantation des phrases. La pertinence de cette représentation « locale », sensible aux variations linguistiques au grain mot, sera discutée dans la prochaine section.

1.1 Un paradigme du TAL : modéliser une intuition linguistique sur les phrases

Un des paradigmes centraux du TAL est l'utilisation de mots graphiques dans la représentation, il est visible par exemple dans la prééminence des approches en sacs de mots (*Bag of Word models*) qui, bien que ne représentant plus l'état de l'art, constituent souvent la *baseline* par défaut. Une des manières d'aller plus loin que ce paradigme a été de travailler dans le cadre de la phrase pour aller au-delà des formes brutes et ainsi « ne pas considérer implicitement les occurrences de mots comme des faits indépendants » [Mathet et al., 2008]. Si l'on cherche donc à représenter des phénomènes linguistiques au-delà du niveau des formes pures, on va s'intéresser à travailler la syntaxe voire à la sémantique. C'est l'objet des tâches d'Extraction d'Information (EI).

1.1.1 Des principes de la modélisation au niveau phrastique

Usuellement, le terrain d'étude en EI c'est la phrase, ainsi traiter différents scénarios consiste souvent à modéliser des phrases sous formes de patrons, soit de manière très explicite (voir par exemple [Huttunen et al., 2002, Cellier et al., 2015]) dans la lignée des *Message Understanding Conference* (MUC) ou de façon plus implicite en intégrant à un réseau de neurones la position des mots dans les phrases [Zeng et al., 2015, Wu and He, 2019]. Si l'approche est élégante, on voit bien l'idée de jouer sur des paradigmes (verbaux notamment) afin de dégager une série de structures syntaxiques relativement couvrantes où l'on ferait varier notamment les verbes et les noms. La difficulté est que le nombre de patrons possibles est trop grand pour être décrit en extension, trop grand même pour être couvert par les exemples observés sur un échantillon raisonnable [Zouaq et al., 2012]. La génération automatique des patrons via des méthodes d'apprentissage n'a pas d'autre part apporté les résultats escomptés [Chambers and Jurafsky, 2011, Patel et al., 2018] d'autant qu'il a été montré que l'excès de généralisation introduit du bruit [Jijkoun et al., 2004].

9. On songe ici évidemment à la gourmandise matérielle des *Large Language Models*, par LLAMA dont l'utilisabilité hors serveur de calcul est toujours problématique.

En résumé, la variation des structures est beaucoup plus forte que l'on pourrait l'imaginer de prime abord de sorte que la couverture des patrons syntaxiques est asymptotique. On trouvera un certain nombre de patrons relativement fréquents mais il n'est pas toujours évident de définir en intension l'ensemble des structures de phrases ou de propositions décrivant un évènement. Ceci a amené très récemment à redéfinir l'extraction des déclencheurs (*triggers*) décrivant l'évènement comme une tâche d'étiquetage de séquences [Du and Cardie, 2020] voire véritablement comme une sous-tâche d'un système de questions-réponses [Li et al., 2019,Boros et al., 2021]. On retrouve un peu ici l'idée de Chambers et Jurafsky de *template without template* [Chambers and Jurafsky, 2011].

Ces acquis de l'état de l'art sont par certains aspects plus fragiles que dans d'autres domaines puisque l'on voit souvent les même jeux de données (ACE-2005 [10] par exemple) ou la même langue (l'anglais), néanmoins la perspective me semble intéressante puisque cela questionne la manière dont on construit une représentation efficace. d'un phénomène (ici un type d'évènement) dont les manifestations en corpus, les instances, vont être soumises à une forte variation. Je pense qu'il ne s'agit pas simplement d'un exemple, parmi d'autres, de tâche pour laquelle des approches d'apprentissage (profond ou non) ont surpassé des approches basée sur des règles mais d'une limite forte rencontrée par les approches fondées sur des motifs ou des *templates*.

1.1.2 Des limites de la modélisation au niveau phrastique

Afin d'illustrer ce propos, nous allons présenter une série de tableaux présentant les occurrences des termes associés à *flu* dans le corpus Daniel ou DANIEL-DATASET ([Lejeune et al., 2013, Mutuvi et al., 2020b]). Ce corpus pour la veille épidémiologique a notamment été créé pour l'étude [Brixtel et al., 2013], il comporte 2089 documents en 5 langues dont 475 pour le sous-corpus anglais. Le corpus ayant été préalablement filtré sur le domaine médical il permet d'avoir une idée relativement fiable de la manière dont les épidémies sont exposées dans des articles de presse. Nous avons dans ce sous-corpus 61 phrases qui contiennent ce terme ou ses variantes (*avian flu*, *influenza* ...) dans le sous-corpus anglais. Le tableau 1.1a présente les occurrences où l'on peut repérer un déclencheur verbal, que nous définissons ici comme un verbe fortement marqueur du domaine étudié (en l'espèce l'épidémiologie).

On peut ici voir que la richesse de la syntaxe et des verbes associés est assez grande puisque seuls deux déclencheurs verbaux sont présents plusieurs fois, à savoir *spread* (dont une occurrence sous forme négative) et *reported* (3 occurrences) [11]. Si l'on observe les occurrences de *spread* dans le sous-corpus anglais, on se rend compte que les cas où le nom de la maladie est antéposé sont rares, on va trouver : *cancer, snake's poison, parasitic infection, dengue fever* (tous avec une seule occurrence) ainsi que *virus* (3 occurrences) et donc *flu* (2 occurrences). Cette variété importante dans les lexèmes comme dans la syntaxe limite de fait les capacités de systèmes d'apprentissage à généraliser efficacement et à dépasser les 85/90% de F-Mesure [Mutuvi et al., 2020a]

On retrouve *spread*, sous sa forme nominale cette fois, dans le tableau 1.1b qui recense d'autres types de déclencheurs que l'on peut trouver dans le voisinage de *flu*. On retrouve aussi des termes formés à partir de cases (*confirmed cases* et *number of cases*) qui sont sûrement de bons indicateurs pour savoir s'il est véritablement question d'une épidémie dans la phrase considérée. Si l'on observe les occurrences de *spread of* dans tout le corpus, on va trouver d'autres noms de maladies : *bird flu, wild polio, winter sickness, the virus*, mais aussi [*infectious*] *disease*[*s*]

10. `https://catalog.ldc.upenn.edu/LDC2006T06` consulté le 2 octobre 2023

11. On pourrait considérer que le verbe *reported* n'est pas un déclencheur verbal, dans la mesure où il est précédé de *case[s]* mais cela correspond à un des exemples de motif répertoriés dans l'article fondateur de Ralph Grishman [Grishman et al., 2002], j'ai donc considéré qu'il méritait d'être inclus ici

Contexte gauche	Entité(s)	Contexte droit
The first	H1N1 flu	**case** of the season **has been reported** at[...]
[...] in Khurda where the first	avian influenza	**positive cases were reported** on Sunday.
Although the	flu	season often begins in the fall, Newfoundland and Labrador **is now reporting** its **first cases**[...]
[...] other La Nina events have not seen novel	flu strains	**spread** around the world, they caution.
"Our business is concentrated in Cuttack and Bhubaneswar, where the	flu	**has not spread** so far," said B Soundararajan[...]
[...] live wildlife trade can also **expose humans to** infectious diseases such as	monkeypox, bird flu[...]	
The regions bordering Bangladesh **have been susceptible to**	avian flu	.
[...]when the Indian government **confirmed**	bird flu	.
[...]whether they rushed into mass prescriptions of Tamiflu when	swine flu	**hit** the nation in 2009.
Doctors recommend a	flu	shot to prevent **catching the virus.**
[...]**to detect** any possible outbreak of	avian influenza	
Kidney disease and	pneumonia/ influenza	switched places [...]
	Pneumonia and influenza	**have really dropped** a lot [...]
[...] officials said the hospital **treated** its first	influenza A/H1N1	, patient over the weekend.
Death rates also **declined** for	influenza and pneumonia	(by 8.5 percent)[...]

(a) Phrases avec des déclencheurs verbaux

[...] Cover and Contain to help prevent the **spread of**	influenza	.
[...] to help prevent **the spread of**	influenza	during its peak season.
[...]We are beginning to see some initial **confirmed cases** of	influenza B	in the province
While the **number of cases** is low and symptoms are mild, receiving the seasonal	flu	shot [...], along with practising regular hand washing, is the most effective way to **protect** [...]
	Flu	symptoms include rapid onset of **cough, fever, headache, chills**[...]
[...] alterations in migration patterns promote the development of dangerous **new strains of**	influenza	[...]
Yet hasnt a	swine flu	**victim** also ingested (or at least inhaled) the **virus** one way or another ?
High-risk adults and children with chronic disease[...]are at increased **risk** for	influenza	.
[...]some passengers had	stomach flu	and the islands are ill-equipped to handle a **contagious virus**.

(b) Phrases avec des déclencheur nominaux

TABLE 1.1 – Phrases contenant des termes liés à *flu* dans le sous corpus anglais du DANIEL-DATASET

(3) et *influenza* (2). On pourrait penser qu'il y a là matière à apprendre quelque chose, et c'est sans doute vrai, il faut toujours penser aux contre-exemples et ici nous avons *what this was* qui complémente *spread of*[12].

On peut bien sûr penser à des déclencheurs plus évidents tels que *outbreak* ou *pandemic*. Nous en donnons des exemples, toujours autour de *flu* et des termes associés, respectivement

12. *B.C. Public Health suggested that based on the evidence that we were sharing with them and the fast spread of what this was, that we're looking at norovirus* Source CBC.ca (`https://www.cbc.ca/news/canada/british-columbia/norovirus-outbreak-suspected-at-b-c-student-conference-1.1208696`) consulté le 2 octobre 2023

Poultry exports hit after	flu	**outbreak** in Odhisa, Meghalaya
[. . .] whether they prescribed Tamiflu on a mass scale too quickly in the wake of the 2009	swine flu	**outbreak**.
Professor Seto Wing-Hong, director of[. . .] in Hong Kong, said that during the 1997	avian flu	**outbreak**, health workers there were required to report any contact made with patients.
Until now surveillance of	seasonal flu	**outbreaks** and potential **pandemics** such as bird flu and swine flu[. . .]
[. . .] such as an	influenza	**outbreak** during World Youth Day in 2008 in Australia.
Poultry companies are exercising caution with the **outbreak** of	bird flu	in parts of Odisha and Meghalaya[. . .]
"In the past seven months, poultry exports from India have been banned due to the **outbreak** of	bird flu	in West Bengal and the Northeast.
But the monitoring of birds, pigs, people and the genetics of the	influenza virus	have all been stepped up in response to recent **outbreaks** of both swine flu and bird flu.
But the monitoring of birds, pigs, people and the genetics of the influenza virus have all been stepped up in response to recent **outbreaks** of	both swine flu and bird flu	.

(a) Phrases avec le déclencheur « *outbreak* »

If the	swine flu	**pandemic** of 2009-10 was part of this pattern[. . .]
La Nina linked to killer	flu	**pandemics**
The La Niña weather pattern could be triggering	flu	**pandemics** such as 2009's swine flu, scientists suggest.
The researchers studied ocean temperatures records [. . .]before the four most recent	flu	**pandemics** emerged.
[. . .] that's believed to have triggered the 2009	swine flu	**pandemic**.
[. . .] every allegedly responsible health agency[. . .] warned of the imminent onset of an	avian flu	**pandemic** [. . .] comparable to the Spanish flu of 1918, which claimed up to 100 million lives.
La Nina **'may abet'**	flu	**pandemics**
La Nina events may **make**	flu	**pandemics more likely**, research suggests.
We know that **pandemics** arise from dramatic changes in the	influenza	genome [. . .]
Nevertheless, the last four **pandemics** - the	Spanish Flu and [. . .]	- were all preceded by periods of La Nina conditions.

(b) Phrases avec le déclencheur « *pandemic* »

TABLE 1.2 – Phrases contenant des termes liés à *Flu* dans le sous corpus anglais du DANIEL-DATASET

dans les tableaux 1.2a et 1.2b. On voit ici que les *patterns* sont plus évidents, en particulier [maladie]+[pandemic/outbreak]. Il faudrait toutefois régler les cas, très fréquents, où ni la syntaxe ni les déclencheurs ne semblent aisément modélisables (Tableau 1.3). Si plusieurs des phrases recensées dans ce tableau ne semblent pas décrire de nouveaux cas, plusieurs sont tirés de textes étiquetés comme pertinents dans DANIEL-DATASET.

Cette étude, évidemment trop succincte, vise à montrer que la modélisation au niveau phrastique atteint rapidement ses limites. Ceci peut expliquer que les approches d'EI classique, agrémentée ou non d'apprentissage automatique ou de modèles de langues, n'aient pas permis à ce jour de résoudre le problème de la veille épidémiologique multilingue. Cette typologie, limitée, a simplement une vocation illustrative : il sera difficile (I) d'énumérer les structures possibles ou (II) d'en déduire un nombre significatif d'un corpus d'entraînement. Dans [Mutuvi et al., 2020a] nous avons, résultat attendu, montré que la qualité des résultats était directement liée à la dotation en ressources des langues, en particulier pour la langue polonaise où il fut compliqué de trouver une architecture un tant soit peu efficace.

Que faire dès lors lorsque l'on voudra passer à l'échelle, c'est-à-dire traiter plus de textes dans

[...]when a nice Jewish boy *came down with*	swine flu	.
The	flu	shot can be obtained through a [...] physician[...]
While the number of cases is low and symptoms are mild, receiving the seasonal	flu	shot[...] is the most effective way to *protect* against influenza viruses.
A two-year review [...]reveal findings about the effectiveness and side-effects of the	swine flu	'wonder-drug'.
The	flu	is an airborne virus, easily spread through the nose, throat and lungs.
To receive the	flu	shot, contact your family physician or public health office.
No need for panic over	bird flu	, say experts
After reviewing studies of Tamiflu during the	avian flu	scare, Dr. Tom Jefferson of [...]had concluded in a 2006 report that the drug was effective.
[...]there was no proof that Tamiflu reduced serious	flu	complications like pneumonia or death.
The Department of Health and Community Services is reminding residents that	seasonal flu	vaccinations are still available throughout the province.
[...] The Department of Health and Community Services sent out a reminder to residents on Tuesday that the seasonal	influenza	vaccination is still available throughout the province.
[...] The	influenza	vaccine is provided at no cost to high-risk individuals
Household contacts of people at high risk of	influenza	complications ; Pregnant women [...]
[...] the La Niña pattern is known to alter the migratory patterns of birds - believed to be a primary reservoir of human	influenza	
[...] migratory birds [...] are a major reservoir for	influenza	.
[...] These conditions could favor the kind of gene swapping that creates new variations of the	influenza	virus, say the scientists.
The	seasonal influenza	vaccination is still available throughout the province[...]
The	influenza	vaccine is provided at no cost for individuals at high risk for complications[...]
[...] but this affirms the national priority on immunization, both	influenza and pneumococcal	.

TABLE 1.3 – Phrases contenant des termes liés à *Flu* dans le sous corpus anglais du DANIEL-DATASET sans déclencheur évident

plus de langues ? À quel moment la modélisation est prête à être exploitée sur des corpus ? Quid du rappel quand on sait que les *patterns* ajoutés risquent d'offrir une couverture de plus en plus limitée ? Que faire surtout lorsque la donnée va varier, en particulier en langue ? Si l'on en revient au coût marginal, on voit que l'exigence en terme de ressources (lexiques, ontologies, plongements ...) ou de modules d'analyse (tokeniseur, lemmatiseur, étiqueteur morpho-syntaxique...) est liée à la variété. On peut aussi renverser la question, à quel point le travail déjà réalisé sur une langue est réellement réutilisable pour en traiter une puis plusieurs autres. C'est là je crois que deux des grands projets historiques de veille épidémiologique automatique ont rencontré leurs limites. PULS [13], système de veille auquel j'ai contribué [Lejeune, 2009], n'a finalement affiché que deux langues (anglais et russe) [14] tandis que Biocaster [Collier, 2011] en a traité jusqu'à 11 (anglais, arabe, chinois, coréen, espagnol, français, japonais, portugais, russe, thaï et vietnamien) sans toutefois que la qualité des résultats ait été mesuré dans la majorité de ces langues [Lyon et al., 2011]. Les approches fondées sur l'apprentissage éprouvent sans doute moins de difficulté à négocier ce passage, même si se pose la question de la stabilité multilingue

13. `http://puls.cs.helsinki.fi/static/index.html` consulté le 2 octobre 2023

14. Le système que j'avais développé en 2009 pour le français n'est toutefois resté que quelques mois en ligne, ce qui invite probablement à la modestie.

des architectures mises en uvre qu'il s'agisse du choix des algorithmes de classification ou de leur paramétrage. De plus, les performances sont évidemment limitées par la taille des données disponibles, introduisant une forte variation entre langues dans la qualité des résultats [Mutuvi et al., 2021b]. Si les techniques évoluent, et définissent de nouveaux résultats « état de l'art », j'observe certaines constantes dans la manière dont les problèmes sont modélisés, dans la manière dont on observe la langue et donc dans les limites de ce que l'on peut tirer de ces observations.

1.2 Quels sont en réalité les observables du TAL ?

La question de ce que les techniques de TAL permettent d'observer dans les textes, les énoncés, est très rapidement devenue une question centrale dans mon travail. Il me semble indispensable de se représenter ce que telle ou telle méthode existante cherche à représenter, quelle est l'intuition linguistique ou algorithmique sous-jacente. La manière de décrire ou de modéliser des données textuelles est en fait suspendue à la possibilité d'isoler algorithmiquement les observables pertinents, les unités constitutives de la modélisation en question. Il y a de mon point de vue un certain intérêt à renverser la question : que cherche-t-on dans les données textuelles et en quoi cela a une influence sur les méthodes que l'on va utiliser ?

1.2.1 Des mots pour quoi lire ?

On peut se demander si une analyse fine des données textuelles, comme peut le réclamer une tâche comme l'extraction d'information, implique nécessairement une modélisation séquentielle mettant en jeu des analyses locales fortement dépendantes de la langue. Ce paradigme a pourtant longtemps prédominé [15]. Si l'on s'intéresse au dernier élément du triptyque données – méthode – tâche évoqué plus haut, on peut se demander dans quelle mesure une représentation puisse se concevoir indépendamment des données à traiter d'une part et de la tâche visée d'autre part. Il est assez naturel bien sûr de chercher une méthode un peu tout terrain qui puisse s'adapter facilement à différents types de données comme à différentes tâches. D'une part, c'est avec ce genre de méthodes que l'on construit des *baselines* et d'autre part, plus pragmatiquement encore, cela correspond à un certain besoin de factoriser un travail de modélisation et une implantation informatique déjà réalisés, et partant de limiter le coût marginal. Il est intéressant de noter que les *baselines* sont un reflet de l'évolution des paradigmes du domaine dans une période donnée. Si l'on observe la littérature en TAL de part et d'autre du changement de paradigme impliqué par les modèles de langue en particulier à la suite de Word2vec [Mikolov et al., 2013] [16], le phénomène est très visible. Ainsi, on verrait probablement que la *baseline* standard en classification supervisée était la représentation en sac de mots (BOW pour *Bag Of Words*) couplée avec un classifieur linéaire par exemple. Aujourd'hui, il semble de plus en plus incongru de lire ou de relire un article qui ne présenterait pas de *baseline* exploitant des plongements lexicaux. On a donc bien ici un changement de paradigme (*paradigm shift*) au sens de [Kuhn, 1962] puisque les plongements de mots sont passés de « découverte extraordinaire » au stade de nouveau paradigme, Kuhn parle aussi de changement de matrice disciplinaire (*disciplinary matrix*). Les plongements de mots associés à des réseaux de neurones sont donc la nouvelle matrice du TAL, la manière naturelle de représenter les problèmes. Un champ de recherche important

15. Un relecteur d'une célèbre conférence du domaine : « Je ne crois pas dans ces idées brillantes qui évitent le recours à la chaîne de traitement classique de TAL qui démarre avec la tokenisation, la lemmatisation et l'étiquetage en parties du discours »

16. Des travaux plus anciens tels que [Bengio et al., 2003] voire [Miikkulainen and Dyer, 1991] ont l'antériorité sur ce sujet mais sans avoir toutefois produit de semblable révolution dans l'état de l'art.

s'est donc constitué autour de l'idée de pousser dans leurs retranchements ces représentations. En ce sens les modèles de *transformers* type BERT [Devlin et al., 2019] constituent un élément de cette nouvelle matrice disciplinaire, un *paradigm shift* de moindre degré que WORD2VEC qu'il convient d'explorer afin d'en tirer tout le sel. C'est l'objet de la BERT-ology (terme à ma connaissance mentionné tout premier lieu dans la littérature par [Ravishankar et al., 2019]). Si le changement de paradigme est évident il n'en reste pas moins que l'observation de la donnée reste fondamentalement liée à la notion de mot. On l'observe assez facilement puisque les mots hors vocabulaire (OOV pour *Out Of Vocabulary*) restent un sujet important et que l'amélioration de leur traitement a présidé à des développements ultérieurs marquants pour le domaine tels FASTTEXT [Bojanowski et al., 2017] et plus généralement une arrivée au premier plan des approches utilisant des chaînes de caractères non-mots (*subwords* par exemple). Ceci montre une prise en considération des limites de l'utilisation des formes brutes des mots comme des limites de processus favorisant un rapprochement de formes apparentées telles que la racinisation ou la lemmatisation. Toutefois, la représentation des textes reste cantonnée à un travail sur ces formes locales, la notion de document ou celle de genre textuel reste peu exploitée dans les approches état de l'art.

1.2.2 Des documents et des corpus pour quoi faire ?

J'ai cherché à explorer les notions de répétition et de position dans un genre textuel particulier qui m'avaient intéressées dans mes travaux initiaux sur la presse ([Lejeune et al., 2010]) et que j'ai proposé de généraliser en travaillant sur les articles scientifiques ([Lejeune et al., 2011, Doualan et al., 2012, Lejeune and Daille, 2015, Daille et al., 2016]). Ces deux genres textuels présentent un style collectif, dans le sens utilisé par [Lucas, 2009], une pratique discursive assez normée d'organisation de l'information qui apparaît comme une pratique très régulière pour les scripteurs opérant, notamment, dans un genre textuel particulier. Une propriété régulière du style collectif des écrits académiques est par exemple constituée par le modèle IMRAD (*Introduction, Material and Methods, Results Ands Discussion*) [Sollaci and Pereira, 2004], dans la presse on trouvera notamment le style dit en pyramide inversé (*inverted pyramid style*) dans la presse [Pöttker, 2003]. Ces principes d'organisation textuelle, typiques de sous-corpus homogènes du point de vue du genre textuel, sont destinés à faciliter la lecture par l'humain et l'extraction des informations essentielles. Exploiter ces indices d'organisation textuelle permet d'éviter de tomber dans le piège formé par un paradigme de résolution fondé sur une structuration ascendante (voir [Giguet, 2011, P.58]).

Ceci ne signifie pas que des propriétés de style collectif soient identifiables, et traduisibles sous forme d'algorithme, pour tous les genres textuels. Toutefois, réfléchir aux spécificités des documents et des corpus auxquels ils appartiennent peut permettre de s'affranchir du carcan d'une structuration ascendante de la représentation du texte. On observe dans l'analyse de *tweets* qu'il est prometteur de ne pas chercher à standardiser la langue mais à tenir compte de la variabilité locale en exploitant par exemple des représentations en chaînes répétées maximales (*Repeated Character Substrings*) [Buscaldi et al., 2018, Ghoul and Lejeune, 2019]. Dès lors, on se trouve dans une approche expérimentale où on n'appliquera des « pré-traitements » que s'ils s'avèrent absolument nécessaires [Ghoul and Lejeune, 2020], considérant que la variation locale est inhérente au genre des *tweets* et mérite d'échapper à la standardisation. Ma réflexion vise donc à exploiter les spécificités des corpus comme une plus-value que comme une contrainte, dans l'esprit de François Rastier quand il écrit que « le texte est pour une linguistique évoluée l'unité minimale [d'analyse] [Rastier, 2002]. À l'autre extrémité du spectre, c'est accepter que la représentation en mots n'est qu'une commodité méthodologique et que considérer un ensemble

moins contraint des sous-chaînes de caractères d'un texte est souvent productif. Ce qui dans mon esprit est très liée à une maxime de Umemura et Church *Anything you can do with words, we ought to be able to do with substrings*[17] [Umemura and Church, 2009]. *In fine* le découpage en mots n'est qu'une commodité de lecture qui a eu pour but notamment de faciliter la lecture silencieuse et s'est véritablement généralisé au moyen-âge [Saenger, 2001] pour mieux adapter l'écriture à l'il humain qui la recevra. Il n'est pas certain que le découpage en mots graphiques soit nécessairement le choix le plus pertinent quand il est question d'un il électronique.

Exploiter des propriétés globales des documents, des corpus et assouplir la définition des observables au niveau local est une voie productive et originale pour traiter des corpus variés avec des résultats tangibles. Il ne s'agit toutefois pas de définir une martingale mais de systématiser l'adaptation aux spécificités des données, notamment en cherchant dans les données des observables utiles pour une tâche particulière. La recherche des observables doit être de mon point de vue liée à l'observatoire, au corpus, au sein duquel ces observables peuvent être identifiés, analysés et interprétés. Dit autrement, on ne doit pas sacrifier la réflexion sur les données sur l'autel de l'application d'une modélisation qui peut être inadaptée et écarter des observables potentiellement intéressants. Le choix des observables est donc intimement lié à l'observatoire. Damon Mayaffre [Mayaffre, 2005] oppose deux usages de la notion d'observatoire selon qu'ils déterminent *a priori* une théorie ou bien qu'ils participent à la construction du modèle *a posteriori*. Je me range plutôt dans la seconde catégorie dans la mesure où ma réflexion se veut avant tout pragmatique, sinon opportuniste, et vise à s'adapter à ce que les données disponibles permettent.

Ceci n'exclut pas naturellement d'avoir des fondements épistémologiques forts dans son travail, une matrice de pensée de la donnée textuelle mais de s'autoriser à la questionner, à la modeler, à en douter[18]. En plus de questionner le caractère central de la notion de *token* dans l'épistémologie du TAL, il me semble donc que les notions de document et de corpus sont elles largement sous exploitées. Bien sûr les plongements lexicaux exploitent dans une certaine mesure la notion de corpus puisque ces représentations sont tirées de calculs effectués sur de grands ensembles documentaires. Toutefois, il ne s'agit pas de corpus au sens linguistique du terme, l'unité et l'homogénéité des grands jeux de données tirées par exemple du *Common Crawl* est discutable. On voit d'ailleurs qu'il est fréquemment nécessaire de re-spécialiser (via du *fine-tuning*) les représentations pour les rendre moins tout-terrain et plus adaptées à un usage spécifique.

De plus la notion de document, ou plus encore de genre textuel, est absente de ces représentations. Ceci n'empêche pas ces représentations d'être utiles, je ne suis pas certain que l'on doive chercher à tout prix, ni que l'on puisse trouver, des méthodes qui soient à la fois parfaitement et solidement fondées linguistiquement parlant et efficaces. Néanmoins, je pense que l'on peut trouver des compromis entre ces deux aspects d'une autre manière et que la notion de corpus, dans le sens d'ensemble documentaire homogène avec des traits identifiés, pourrait être mieux investie informatiquement parlant et qu'il ne faut pas négliger la relation entre les corpus et les observables. Le corpus c'est en somme l'éco-système qui régit les relations entre les observables dans les documents qui le composent. Ceci définit un contexte, au-delà du simple contexte phrastique ou sous-phrastique, d'apparition des observables qui est essentiel à la compréhension des relations entre ces mêmes observables. Ces observables peuvent varier à la fois selon la langue, certaines langues s'analysent très bien en mots, comme selon le type de documents traités ou

17. Tout ce que vous êtes capable de faire avec des mots, nous devrions pouvoir le faire avec des chaînes de caractères (traduction personnelle)

18. Mon professeur d'échecs disait souvent pour questionner les paradigmes trop bien établis de l'analyse des parties, une phrase qui je trouve s'adapte bien ici : « Aux échecs, comme dans tout domaine, il faut avoir des certitudes . . . et douter de tout »

encore la tâche à réaliser. Les identifier correctement impose de se placer dans la perspective du corpus afin de déterminer aussi précisément que possible la manière optimale de structurer les documents.

1.3 Observer les textes dans leur éco-système

Certaines propriétés intrinsèques d'un texte, en particulier la relation implicite qu'il trace entre le message, l'émetteur et le récepteur, sont des éléments qui favorisent l'émission ou la réception dudit message. Faciliter l'émission du message se produit lorsque dans la situation de communication l'émetteur du message s'autorise, sur invitation ou non du récepteur, à contrevenir à certaines règles par désir de fluidifier/favoriser la communication, ou parfois par simple paresse. Le problème de cette facilitation est qu'avec elle viennent des difficultés de compréhension du message, difficultés liés à une moindre prise en compte du profil du récepteur, à des ambiguïtés mal anticipées qui risquent de rendre plus incertaine la réception pleine et entière du message. Si le récepteur est un système automatique on peut imaginer un grand nombre de spécificités discursives de ces corpus qui rejaillissent sur tout ou partie des documents qui les composent :

— Texte mal retranscrit avec comme résultante un bruitage des données avec des caractères surnuméraires (dans le cas de l'OCR par exemple), des mots incorrects (en particulier avec de l'ASR [19]) ou encore des empans plus larges de texte manquants ou superflus (par exemple en *web scraping*, voir par exemple [Barbaresi, 2019])
— Fautes d'orthographes, et plus généralement inventivité orthographique, amenant dans des mots mal reconnus (le cas le plus emblématique restant le cas de corpus tirés de réseaux sociaux numériques)
— Autres mots hors vocabulaire liés par exemple à la variation diachronique, à la variation diatopique ou encore à la créativité lexicale
— Inconsistance grammaticale, entrainant des difficultés pour l'analyse syntaxique (par exemple pour la Reconnaissance d'entités nommées)
— Inconsistance de la ponctuation, impliquant des difficultés pour le découpage en phrases (par exemple dans le cas de PDF nativement numériques [Giguet and Lejeune, 2021a])

Ces catégories ne sont pas mutuellement exclusives puisque dans des processus d'analyse ascendante, les erreurs à un niveau provoquent naturellement des erreurs plus graves au niveau supérieur du fait des erreurs en cascade.

1.3.1 Contraindre les données ou adapter les traitements ?

On voit ici que traiter des sources de données présentant de la variation affecte les paradigmes du TAL en plusieurs points. Dès lors que le document émis est peu contraint, peu standardisé, les observables traditionnels sont moins aisés à identifier. D'un autre côté, faciliter la réception lorsque l'on produit un énoncé consiste à exercer des contraintes sur l'acceptabilité d'un message, contraintes qui amènent à s'attarder sur la construction de l'énoncé afin de l'adapter à son interlocuteur et à la situation d'énonciation. Chercher à contrôler l'entrée, l'*input* vise alors à faciliter l'analyse ultérieure.

Faut-il dès lors adapter les données en fonction d'un usage par une machine ? Ceci suppose que l'on ne cherche pas à traiter des textes faits pour des humains, que l'on conditionne la ro-

19. Automatic Speech Recognition pour reconnaissance automatique de la parole

bustesse des systèmes . . . à l'adaptation des données en entrée. L'adaptation au destinataire, au récepteur du message est une évidence d'un point de vue linguistique, comme on le retrouve par exemple dans la notion de *Lector in Fabula* d'Umberto Eco [Eco, 1985]. Les contraintes mises sur l'émission du message vont constituer les bonnes propriétés communicationnelles du message : les locuteurs sont incités, socialement notamment, à bien encoder, à bien préparer leur message pour en faciliter le décodage. Un bon message respecte donc un certain nombre principes, possède des caractéristiques attendues des destinataires, sur différents niveaux linguistiques, qui vont faciliter son intelligibilité et créer l'effet attendu. Ces bonnes propriétés ne sont pas situées uniquement dans la norme « phrastique ». La grammaticalité, entendue ici dans le cadre de la phrase, est une caractéristique importante mais l'on voit bien qu'il faut un certain degré de dégradation ou une certaine complexité d'énoncé pour que la grammaticalité soit véritablement problématique pour la compréhension d'un document par un locuteur humain raisonnablement compétent. C'est bien plus les faiblesses dans la structuration d'un document qui peuvent altérer fortement la compréhensibilité d'un document, que l'on songe par exemple à une affiche annonçant un évènement et qui respecterait pas la règle des 5W [20] de sorte que le nombre d'inférences pour combler les informations manquantes devienne trop grand ou impossible à réaliser pour le récepteur. On voit bien que la difficulté est plus grande si l'on est dans le cas d'une langue moins bien maîtrisée. Il me semble que l'un des verrous du traitement automatique des langues se situe ici : la machine est un locuteur aux capacités d'adaptation limitées. C'est un locuteur faiblement compétent et le fait d'avoir trop d'attendus sur la conformité des énoncés limite les capacités de généralisation. Cette limite peut-être liée, par exemple à des choix méthodologique consistant à se placer dans un cas (trop) bien déterminé : corpus monolingue, langue standard, tâche très spécifique . . . La faible représentativité des données sur lesquelles la machine a été testée ou entraînée nourrit cette limite. Mais il s'agit sans doute aussi d'un problème de vision globale dans le développement de modèles par la communauté TAL. La plupart des modèles sont fondés sur une approche ascendante, où la représentation du texte est reconstruite par combinaison des résultats d'analyses locales (la tokenisation), négligeant ainsi des dimensions linguistiques de plus haut niveau telles que la stylistique ou la pragmatique qui sont difficilement encodables à partir d'unités de base comme les tokens. Pourtant, je considère que, sans pour autant mener une analyse stylistique de grande envergure, il y a beaucoup de choses à exploiter de ce côté là car il y a des éléments de modélisation derrière certaines normes, certaines propriétés stylistiques ou pragmatiques visibles dans les corpus. Les normes d'encodage du message liées à la pragmatique sont essentielles dans un texte, bien organiser son message c'est faire que son essence soit correctement transmise en exploitant les outils, les propriétés propres au contexte d'énonciation et aux compétences du récepteur. On peut observer que le TAL a plutôt privilégié les aspects lexicaux et syntaxiques, et plus globalement des observables repérables au niveau phrastique. Ce penchant s'est développé pour différentes raisons. Tout d'abord, l'analyse lexicale et l'analyse syntaxique sont deux tâches assez autonomes et utiles par elles mêmes. On peut tirer de l'analyse syntaxique, des observations de faits de langue à cet échelon linguistique, mais on peut aussi intégrer l'analyse syntaxique comme maillon d'une chaîne de traitement de TAL. Et en amont de cette analyse syntaxique on aura généralement une analyse lexicale, avec *a minima* une tokenisation.

20. Principe d'écriture impliquant d'inclure dans son énoncé une réponse explicite aux questions *Who, What, Where, When, Why* ou dans sa version étendue en français QQOQCCP pour Qui ? Quoi ? Où ? Quand ? Comment ? Combien ? Pourquoi ?

1.3.2 Analyser des sacs de phrases ?

La, relative, restriction de l'analyse automatique au cadre de la phrase introduit un fort biais conceptuel. L'analyse syntaxique automatique est, par exemple, souvent menée hors du contexte du document, phrase par phrase, sans que le contexte d'énonciation n'intervienne. On sait pourtant que si la syntaxe informe le texte, le texte informe également la syntaxe [Rastier, 2005,Boré and Bosredon, 2016]. Le contexte d'énonciation, et notamment ce qui résulte du genre des textes, intervient de manière plutôt anecdotique dans la conception des modèles de TAL. Les genres textuels sont souvent mis en avant comme des cadres applicatifs, traiter un corpus du genre Z avec une méthode qui a déjà été éprouvée sur les genres X et Y, plutôt que comme des questionnements (ou des solutions) méthodologiques. Le « corpus » est ici plutôt un ensemble de textes, et en ce sens synonyme de *dataset*, et les textes définis par leur appartenance à telle ou telle classe plutôt qu'à des propriétés intrinsèques des documents qui serviraient de base à une structuration plus riche du jeu de données.

Le corpus donne au texte des propriétés essentielles qui rejaillissent sur son interprétabilité par l'humain, il est donc dommageable de ne pas chercher davantage à développer des algorithmes pour exploiter ces propriétés. Les propriétés textuelles inhérentes à un corpus, à un éco-système de textes, facilitent son analyse, et permettent de ne pas voir un texte comme un simple sac de phrases, et le corpus comme un simple sac de textes (éventuellement enrichis par un étiquetage). L'organisation de l'information dans le texte est largement influencée par le corpus dans lequel le texte est baigné et facilite l'analyse des éléments de plus bas niveau (le paragraphe, la phrase, le token ...). Analyser une phrase hors-sol présente bien sûr un intérêt puisque cela favorise une certaine forme de généralisation et la factorisation des opérations réalisées. On peut tout à fait de ce point de vue considérer que l'on va ramener tout problème d'analyse de texte à une succession de problèmes d'analyse de phrase. Mais, cette restriction apparente dans la complexité du matériau analysé masque le fait que l'on a perdu de l'information en réduisant le texte aux phrases le composant.

D'un autre côté, il n'y a pas de raison objective de traiter les textes uniquement par le truchement des mots qui les composent hors le fait que c'est une unité commode à manipuler pour l'humain (après tout il faut bien segmenter, chercher des observables) et que cette unité constitue un point d'entrée important dans l'analyse linguistique. Il est selon moi essentiel de faire la liaison entre le questionnement sur la définition du contexte d'analyse d'un texte et sur l'unité de base qui doit être manipulée. Ceci amène à n'opérer des pré-traitements que s'ils sont strictement nécessaires, de manière à conserver les propriétés des textes, de ne pas écraser les observables en conservant les indices laissés, consciemment ou non, par l'émetteur du texte. Une question que l'on pourrait se poser sur les modèles de TAL est la suivante : dans quelle mesure un système va se comporter de la même manière si on lui présente les phrases ou les mots dans un ordre différent. On sait bien qu'un modèle BOW sur les unigrammes sera parfaitement insensible à un *shuffle* sur les *tokens*, mais dans quelle mesure un modèle fondé sur BERT est affecté par un changement de l'ordre des phrases dans le texte ? [Hessel and Schofield, 2021] ont montré qu'un BERT qui n'aurait d'autres connaissances que le comptage des unigrammes (modèle que les auteurs nomment Bow-BERT) obtient des résultats très proches d'un BERT classique sur des tâches de compréhension du langage du *benchmark* GLUE[21]. Ces cas d'insensibilité des modèles sont d'un certain point de vue synonymes de robustesse, mais ce sont aussi peut-être des signes d'une relative pauvreté conceptuelle de nos approches et d'une trop grande volonté d'uniformisation de l'appréhension du matériau et des observables que l'on y identifie. On peut en effet se demander dans quelle mesure la structure de la langue a une importance s'il est si

21. *General Language Understanding Evaluation*

simple d'obtenir des *baselines* de qualité.

Bien sûr les approches tout-terrain sont séduisantes, et par bien des aspects efficaces. Je pense qu'il ne faut toutefois pas s'interdire de chercher à varier les observables, à prendre en compte les spécificités des corpus, à modéliser un contexte plus large que celui du niveau local, lexical ou phrastique. Certains principes communicationnels qui pilotent la collaboration entre l'émission et la réception du message véhiculé par un énoncé sont à l'uvre dans la plupart des genres textuels bien déterminés et donc repérables, utilisables dans des corpus bien construits. Ils sont à mon avis à rapprocher du *handshake* dans les communications réseaux non pas en tant que série de règles impérativement respectées mais en tant que série de vérifications opératoires qui gouvernent le style collectif des locuteurs produisants et assurent la bonne compréhension par les locuteurs entendants et détermine la bonne réussite de la communication. Il me semble important de montrer que ces principes permettent d'appréhender de façon efficace un certain nombre de problématiques de TAL même s'ils ne peuvent être constitutifs d'une quelconque « nouvelle recette ». Les règles communicationnelles permettent de faire face à la variation des données, en particulier la variation en langues tant la notion de style collectif transcende les frontières linguistiques, mais aussi des tâches. En effet, j'ai eu l'occasion de montrer que certaines propriétés du style collectif journalistique étaient très robustes à la variation en langue ([Lejeune et al., 2015a] de même du côté des articles scientifiques ([Lejeune, 2013]. Quand j'écris que cela ne constitue pas une quelconque recette c'est qu'il ne s'agit pas pour le chercheur d'avoir la même méthode, le même raisonnement, quelle que soit la tâche ou le corpus mais de ne pas oublier que dans la liste des caractéristiques disponibles existent aussi des éléments liés au discours : la répétition, la position, les isotopies de toute forme même s'ils doivent n'apparaître que sous forme d'heuristiques [Huttunen et al., 2011]. Toutefois, ces caractéristiques discursives ne sont pas si simples à détecter, il ne s'agit pas de formes de surface, ni à exploiter, en particulier si la tâche visée d'une manière ou d'une autre. En effet, l'évaluation amène avec elle une représentation finale très normée. Par exemple, la tâche d'extraction d'évènements (*Event Extraction* ou EE) implique de livrer un certain *template* généralement composé de mots appartenant à des classes sémantiques définies. Avec ce genre d'*output* attendu, il faut donc à un moment ou à un autre revenir à une représentation sous forme de mots qui de fait favorise les approches qui travaillent déjà au grain mot. C'est un problème que nous avons rencontré dans un travail très récent sur l'extraction d'évènements épidémiologiques ([Mutuvi et al., 2021b]) pour lequel nous avons dû mettre à jour le Daniel-dataset pour le présenter sous la forme généralement attendue dans le domaine à savoir un format de type IOB (*Inside Outside Beginning* [Ramshaw and Marcus, 1999]). Ce format impliquait la disparition d'un certain nombre de propriétés, l'écrasement de certains observables du corpus : disparition de la notion de document et de paragraphes via un découpage en un *token* par ligne et l'assignation des étiquettes de document à des « entités » identifiées au niveau phrastique. [22]

Les entités à identifier étaient des noms de maladies et des noms de lieux. Pour ce qui est des noms de maladie, un double problème s'est posé : d'une part celui de « tagguer » différentes formes plus ou moins équivalentes présentes au long de chaque document (Grippe, Grippe A, H1N1 ...) et d'autre part de traiter le cas des maladies mentionnées sans qu'il soit question d'une quelconque épidémie de cette maladie. Ce dernier cas est intéressant car une phrase telle que *Doctors recommended a Flu shot to prevent catching the virus* peut aussi bien se trouver

22. L'annotation initiale était réalisée au grain document, dans le double but de tenir compte du genre textuel, considérant que dans les articles expositifs était respecté le principe un document = un évènement rendant inutile l'annotation de plusieurs phrases décrivant la même information, et de diminuer le coût des annotations multilingues (il y avait peu d'annotateurs pour la langue grecque notamment) et partant d'augmenter le nombre de documents annotés.

dans un article décrivant l'épidémie, l'aspect vaccinal étant alors secondaire, comme dans un article historique n'ayant rien à voir avec de nouveaux cas. On doit aussi parler des problèmes de co-référence (cette maladie, *this disease* ...). Le cas des noms de lieux est plus intéressant encore. Les problèmes rencontrés pour les noms de maladie (polysémie, mots inconnus ...) se posaient de la même manière mais, se rajoutaient les cas où le nom de lieu (le pays) n'était pas mentionné du tout dans le document. Ce phénomène, que nous avions nommé localisation implicite ([Brixtel et al., 2013]), est le plus souvent lié au fait que la localisation de la source implique implicitement la localisation de l'évènement : si je lis le *Süddeutsche Zeitung*, en l'absence d'information géographiques identifiables je vais considérer que l'évènement dont il est question a lieu en Allemagne. On voit bien ici la limite d'une vision où la définition d'un évènement est circonscrite au grain de la phrase. La cohérence entre les phrases disparaît, et si l'information sur la source n'est pas disponible avec une phrase détachée du document dont elle provient, la localisation implicite ne peut plus être inférée.

Un grand nombre de tâches peuvent certainement être résolues via une analyse locale, phrastique et simplement avec des représentations en mots sans utilisation de propriétés stylistiques ou discursives. Dans certains cas, le simple cadre de la phrase suffit amplement à l'interprétation. Mais c'est aussi parfois la conséquence, quelque peu fataliste, de la disparition d'informations sur la structure (disparition du balisage consécutive au « nettoyage ») du document ou de la perte de ses méta-données. Je n'ai donc pas l'ambition de proposer un quelconque modèle de raisonnement mais plutôt d'interroger le modèle existant et de montrer que l'on peut interroger un certain nombres de paradigmes, au sens épistémologique du terme, du TAL et en montrer les limites de manière expérimentale. À rebours de la fameuse boutade attribuée à Frederik Jelinek, dont la formulation la plus célèbre est je crois « chaque fois que j'enlève un linguiste de mon équipe, mon système fonctionne mieux », je pense que la linguistique peut toujours apporter une connaissance du matériau, pour peu que l'on ne réduise pas la linguistique à la morpho-syntaxe ou à des aspects peu susceptibles d'être traduites sous la forme d'algorithmes. Il ne s'agit pas de dire que les seuls théories linguistiques qui présentent de l'intérêt sont celles que l'on peut traduire mais plutôt de dire qu'il y a encore matière à parler de linguistique dans un domaine qui se donne encore le nom de *Computational Linguistics*.

1.3.3 De l'analyse locale à la robustesse

Je souhaite poser la question de la valeur ajoutée offerte par ce cadre de travail face à un défi à mon sens central du TAL qui est la robustesse face à la variation : variation en langue, en genre textuel, ou en bruitage. Pour traiter la variation, on peut évidemment chercher à se replonger dans un cas que l'on sait traiter et donc chercher dans de nouvelles données des observables tels qu'ils avaient pu être identifiés sur d'autres données plus « standard ». Ainsi, si ces données sont bruitées, il est commode de chercher à les corriger pour pouvoir obtenir des mots de la langue standard et ainsi utiliser le *pipeline*, la chaîne de traitement existante. Dans ce paradigme, on adapte les données à la méthode, et non pas l'inverse. C'est le cas par exemple lorsque l'on pratique de la correction ou de la normalisation sur un corpus de tweets ou de SMS ou bien encore quand on va éprouver le besoin de retranscrire en français contemporain un texte écrit en français classique par exemple. Ceci a aussi un impact sur les tâches que le TAL va traiter : il y a une certaine tendance à redéfinir de nombreux objets d'étude comme des tâches de classification ou d'étiquetage. Autrement dit, des cas où l'on ramène un nouveau problème à un problème que l'on sait résoudre par des algorithmes bien connus. On peut sans doute alors parler de robustesse à la variation dans les tâches, c'est-à-dire que des méthodologies de TAL vont pouvoir être ré-exploitées pour de nouveaux problèmes préalablement redéfinis sous forme

de problèmes de classification. En l'espèce, ce sont les données et les méthodes vont exercer des contraintes sur les tâches que l'on peut réaliser de manière automatique. La tâche prend alors elle aussi le pas sur les données, les données doivent être adaptées pour la méthode choisie et dans une certaine mesure, il faudra aussi les adapter aussi aux tâches : on définit des tâches en fonction des méthodes que l'on sait pouvoir appliquer. Ceci n'est pas sans rappeler le concept du marteau d'Abraham Maslow [Maslow, 1966] (ou loi de l'instrument) : « J'imagine qu'il est tentant, si le seul outil dont vous disposez est un marteau, de tout considérer comme étant un clou » [23]. On peut en effet se dire que, lorsque l'on a de bons outils de vectorisation de documents, il est tentant de vouloir transformer toute donnée en une matrice termes-documents. On peut imaginer aussi qu'il est tentant quand on a de bonnes méthodes de classification de redéfinir tout problème comme l'assignation d'instances dans des classes prédéfinies par une annotation humaine. Cette tendance à adapter le travail à l'outil est assez naturelle, et correspond d'ailleurs assez bien à ce que l'on fait en informatique lorsque l'on cherche à factoriser le travail en réutilisant des composants existants [Tomkins, 1963]. Ceci peut amener des effets vertueux en permettant d'aborder des domaines originaux, constituant des sortes de détournement d'usage (voire par exemple [Giannetti et al., 2019] pour une application de la classification supervisée à la détection des changements de température sur des objets) comme les détournements d'usage.

Mais, ce penchant naturel peut amener des biais, on perçoit bien par exemple les limites de réduire des problèmes concrets à de la classification supervisée : classes potentiellement recouvrantes, classes multiples, variation du poids des erreurs ... Par ailleurs ceci limite les possibilités que l'on a d'aborder un problème nouveau ou des données nouvelles puisque l'imagination expérimentale est bornée par les outils que l'on a à sa disposition ou inversement. On pourrait traduire cela comme le fait Robert Kagan [Kagan, 2004] : « En l'absence de marteau, vous ne voulez rien qui ressemble à un clou » [24]. Si le TAL exploite des algorithmes, de *machine learning* ou de *deep learning*, capables, étant donné des données étiquetés, de classifier de nouvelles données, que sommes nous en mesure de proposer en l'absence de ces données étiquetés ?

C'est ce qui nous amène nous informaticiens à reformuler naturellement les problèmes comme des problèmes de classification : c'est ce que l'ordinateur sait bien faire. Donc, étiqueter des données est (trop) souvent un prérequis pour faire du TAL. Néanmoins, je pense que tout ceci n'interdit pas de modifier, ou ne pas s'interdire de modifier, l'approche des données textuelles que l'on transmet naturellement à une machine. C'est à dire s'interroger sur la pertinence en contexte de tel ou tel paradigme de l'informatique en général ou du TAL en particulier. Ceci impose un certain nombre de questionnements. Le premier est de savoir ce qui est visible dans un texte, en tant que locuteur, et comment on peut le faire voir à une machine ? Ensuite, comment peut-on expliquer suffisamment précisément les phénomènes langagiers utiles à l'interprétation de manière à pouvoir les traduire en algorithmes ? Enfin, comment rendre accessible, représentable, le sens d'un énoncé avec des observables qui soient calculables afin de les rendre accessibles à ce locuteur peu compétent qu'est la machine. Les propriétés, les caractéristiques dont l'identification et l'interprétation sont évidentes pour le locuteur humain ne sont pas les mêmes que celle d'un « locuteur artificiel » mais on doit pouvoir modéliser ce que montrent les données textuelles.

23. *I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail*
24. Adaptation personnelle de *When you don't have a hammer, you don't want anything to look like a nail.*

1.4 Que montrent les données textuelles ?

Nous avons vu qu'il existe une certaine obsession à démarrer tout processus de TAL par une segmentation en mots (avec ou sans sous-mots), éventuellement précédée d'une segmentation en phrases, qui nécessite un nettoyage des données (ces pré-traitements qui sont d'ailleurs déjà des traitements selon [Millour, 2020]). Ceci est très visible dans les tutoriels de TAL disponibles en ligne. Les tutoriels sont, pédagogiquement et scientifiquement, un outil très précieux pour connaître avec un léger décalage temporel les paradigmes majoritaires en uvre dans un domaine. C'est une image un peu déformée sans doute de la réalité mais c'est aussi la trace de ce que la vulgarisation d'un domaine produit. Mon observation régulière de ces tutoriels est que l'application d'une phase de nettoyage du texte, avec différentes formes (suppression des accents, des mots vides, de la ponctuation. . .) est quasi-systématique. Il est bien sûr difficile de mener un analyse exhaustive sur les nombreux sites de tutoriels. Une étude du corpus d'articles tirés de la plateforme *medium* disponible sur KAGGLE [25] permet de se faire une idée de la question. On observe que 93% des approches de TAL présentées impliquent un « nettoyage » des données dans le cadre d'une approche fondamentalement ascendante. L'idée que les différents éléments intervenant dans la formation du sens des énoncés exprimés en langue naturelle se combinent de manière ascendante est un paradigme central dans l'histoire du TAL.

1.4.1 Approches ascendantes et réductionnisme

Selon Monique Slodzian [Slodzian, 2014] la « tripartition syntaxe – sémantique – pragmatique [...] est un paradigme peu contesté par le Traitement Automatique des Langues ». Sachant que les aspects pragmatiques sont relativement peu traités en TAL et que les aspects morphologiques ont eux reçu beaucoup d'attention, on serait tenté de dire que la tripartition centrale en TAL serait plutôt morphologie – syntaxe – sémantique. De mon point de vue, ceci se ressent assez naturellement dans la pratique du TAL puisque la communauté a beaucoup travaillé sur l'analyse morpho-syntaxique d'une part et sur l'interface syntaxe-sémantique d'autre part. L'aspect central de la syntaxe se ressent dans des approches très formelles, par exemple quand Montague [Montague, 1970] considère que l'on peut formaliser la langue syntaxiquement parlant puisqu'il n'existe selon lui que peu de différences entre langages formels et naturels [26]. On le voit de manière beaucoup plus « pratique » quand la tâche d'extraction d'information se concentrait sur l'utilisation de patrons syntaxiques pour affecter du sens aux unités de la phrase et poursuivre la construction ascendante du sens du texte par addition ou composition du sens de ses phrases.

Épistémologiquement parlant, dans cette vision le locuteur et la situation de communication sont comme détachables des autres faits de langue puisqu'il existerait une manière standard d'analyser les énoncés à travers leur construction sous-jacente, structure qu'il s'agit de retrouver afin d'obtenir une représentation du sens. Ceci implique souvent, informatiquement parlant, de traiter de façon successive et relativement étanche différents étages de la description linguistique, par exemple la tripartition dont parle Monique Slodzian. Cette séparation explicite des niveaux de représentation linguistique est quelque part déjà contestée par le côté ubiquitaire des modèles de type BERT même s'il reste commode de chercher à rattacher les performances de ces modèles aux niveaux de représentation linguistique. [Martin et al., 2020] choisissent par exemple de présenter les performances du modèle Camembert sur différentes tâches en s'efforçant d'attribuer ces performances à une qualité de modélisation de la morphologie, de la syntaxe ou encore de

25. `https://www.kaggle.com/datasets/dorianlazar/medium-articles-dataset` consulté le 2 octobre 2023
26. *I reject the contention that an important theoretical difference exists between formal and natural languages*

la sémantique. Ceci masque le fait que ces niveaux sont bien évidemment fortement entrelacés et que leur séparation n'est pas nécessairement explicite dans la pratique des locuteurs, même s'il est parfois nécessaire de les identifier par exemple lorsque l'on fait de la remédiation sur l'apprentissage d'une langue ([Sparks and Ganschow, 1993]). Lorsque le TAL cherche à décrire la langue cette séparation a du sens mais lorsque le TAL cherche à reproduire la faculté de langage (en compréhension ou en expression pour parler à grands traits) alors il est sans doute trop artificiel de masquer cette intrication des différents niveaux de représentation. L'intrication en linguistique postulerait que le système de la langue ou même simplement le système de la phrase ne peut être réduit à une somme d'état individuels que seraient des caractéristiques extraites directement des mots, visibles directement dans le matériau linguistique (des caractéristiques endogènes en somme) ou recomposées à partir d'une redescription ou d'un enrichissement linguistique (souvent issues d'informations exogènes). En ce sens l'enrichissement des formes de surface, les chaînes de caractères, par le truchement de différents niveaux d'analyse automatique, dont la tokenisation, ne permet de capturer qu'une partie des informations linguistiques servant au décodage de l'énoncé.

Les variables influençant l'interprétation d'un énoncé ne sont pas situées dans une sphère déterminée de manière finie. Le sens d'un « mot » dans le contexte de la $N^{\text{ième}}$ phrase d'un texte peut être influencé par une information provenant d'une phrase placée à d'autres positions dans le texte, ou dans d'autres textes du corpus. Le niveau textuel renseigne, précise, actualise la façon dont la phrase doit être interprétée. Or, il est souvent plus confortable, et d'une certaine manière plus efficace, de travailler hors-sol, en sacrifiant la complétude des interactions entre les différentes strates de représentation linguistique au profit d'une représentation plus commode en tokens dont on essaie d'exploiter les relations de manière plus ou moins fine (les approches en sac de mots permettent ainsi de représenter la co-présence de tokens). Ceci s'entend tout à fait puisque cela favorise la généralisation, la modélisation des énoncés qui seront vus avant tout comme de simples séquences de tokens. Une représentation unifiée favorise grandement la réutilisation des modèles. On va en effet pouvoir considérer que le découpage en tokens est l'étape primordiale dans le traitement des données textuelles. Ceci permet de réduire le nombre de variables à considérer et favorise la généralisation et la factorisation. Traiter des corpus de phrases en traduction automatique est ainsi une bonne manière de s'affranchir d'un certain nombre de particularités liées aux différentes données à traiter. Ainsi l'utilisation du BTEC, corpus multilingue dédié aux conversations touristiques aligné en phrases, a considérablement contribué à la recherche en Traduction Automatique et a favorisé la conception et la comparaison des approches de TA. Mais de fait ceci introduit un biais sur ce qu'est une traduction puisque l'on limite la représentation au cadre restrictif de la phase. La cohérence inter-prastique n'est donc pas véritablement prise en compte et de fait la manière de formaliser la tâche a une influence sur les observables que l'on va considérer. Comment dans un cadre aussi restrictif peut-on sortir des phrases et des mots ?.

Une autre illustration de l'écrasement des observables dans les corpus est le *Million Song Dataset* [Bertin-mahieux et al., 2011], corpus dédié donc aux paroles de chansons. Il est intéressant d'observer déjà que le nom du corpus met particulièrement en avant la problématique de la taille, au détriment par exemple d'informations peut être plus essentielles comme la langue. Il faut noter aussi que l'on parle de jeu de données plutôt que de corpus et c'est peut-être justifié. En effet, un peu plus des trois quart du corpus (77% exactement) dispose des données textuelles proprement dites (le *musiXmatch Dataset*) mais sous la forme d'un sac de mots. On a donc ici une taille qui est trompeuse et d'autre part on voit mal comment exploiter les propriétés spécifiques d'un corpus de chansons réduites à des vecteurs de mots, la différence avec un corpus de presse par exemple étant en partie écrasée par le passage au format matriciel. Il est même

impossible de retrouver le grain phrase, dont on connaît l'importance en TAL. Il est bien évident que des considérations de droits d'auteurs ont présidé à ce choix, il faut garantir que l'on ne puisse pas « reconstruire » les chansons et mettre ainsi à disposition publiquement ces données. Mais cela limite de fait la possibilité d'analyse fine des phénomènes langagiers. Ce simple fait montre bien la prédominance du mot en tant que base de l'analyse automatique, en effet on aurait aussi bien pu donner une matrice de trigrammes de caractères, une table de compression ...

La nécessaire segmentation des phénomènes langagiers en des unités objectivables amène à des appauvrissements des données traitées : on peine à intégrer des informations intégrant différents niveaux ou des propriétés qui ne recouvrent que partiellement la segmentation en tokens. La segmentation en mots semble rester une nécessité absolue en l'état actuel du TAL [27] que cette segmentation soit faite de manière explicite ou non. Notons que c'est la segmentation par défaut lors de la vectorisation par la bibliothèque Scikit-learn par exemple. Ceci conduit donc à considérer que la seule segment valide, le seul grain d'analyse, est celui du mot et que les autres grains en découlent plus ou moins directement. On aurait bien des peines à chercher par exemple des tutoriels de TAL ne comprenant pas la phase de tokenisation (sans parler du nettoyage de la ponctuation, de la suppression des mots outils ...) ce qui n'a pas d'autres justifications que la discrétisation de la donnée textuelle. Il existe pourtant d'autres manières de discrétiser, les chaînes de caractères mots ou non-mots par exemple.

Je pense que ce paradigme du TAL introduit un certain nombre de verrous épistémologiques et pratiques. Au premier rang des verrous je souhaiterais citer la reproductibilité des chaînes de traitement, et des résultats d'expériences. En effet, les techniques utilisées pour la tokenisation sont souvent peu ou pas documentées, ceci est assez dommageable car dans la littérature récente des conférences ACL par exemple, la tokenisation semble parfois être la seule partie du travail où l'on va considérer que l'on traite des textes et non pas n'importe quel autre type de données séquentielles. Encore une fois, ceci est pleinement explicable par des nécessités de factorisation des procédés mais cela n'en soulève pas moins des questions sur ce paradigme central du TAL.

1.4.2 Des difficultés liées à la tokenisation à l'écrasement des observables.

Le problème derrière cette obsession pour la tokenisation est la difficulté de se mettre d'accord sur ce qu'est un mot dans des langues aussi différentes que le français, le chinois et le finnois par exemple. Des modèles de type Fastext cherchent à s'affranchir de ce carcan en modélisant les mots à l'aide des n-grammes de caractères les composant [Joulin et al., 2016] mais l'on reste coincé au grain mot. Quelques approches ont toutefois cherché à s'affranchir plus explicitement du grain mot. Par exemple pour modéliser des morphèmes [Luong et al., 2013], ou des syntagmes [Socher et al., 2012]. Ceci reste toutefois minoritaire, au vu notamment de la capacité, des modèles BERT, encore eux, à apporter des performances remarquables sans chercher à détacher des observables linguistiques de plus haut niveau. Le grain document va éventuellement être présent (par exemple) en complément de la représentation en mots (on pourrait même parler d'heuristique) par exemple dans [Mutuvi et al., 2020a] où il est montré qu'exploiter le « chapô » des articles de presse améliore les résultats, conséquence du modèle de la pyramide inversée ([Pöttker, 2003,Piskorski et al., 2011]) à l'uvre dans les articles de presse, et que la combinaison des débuts et des fins permet d'améliorer les résultats de classification en améliorant le rappel. Enfin, les approches de type Doc2Vec ([Le and Mikolov, 2014]) reconstruisent une représentation du document en représentant le paragraphe comme une construction à partir des mots puis

27. Des modèles de *Deep Learning* apprennent certainement des caractéristiques utiles de la langue de façon « non-supervisée » mais ce sont avant tout des caractéristiques liées à la notion de mot telles que les *subwords*.

le document commune construction à partir des paragraphes. La modélisation du document dans cette représentation semble être restreinte à la prise en compte de la longueur et de la séquentialité des paragraphes. Les mots restent donc les unités, les briques de base de l'analyse des données textuelles et toute représentation, toute construction se fonde nécessairement sur eux. On pourrait arguer que les différences entre deux tokeniseurs pour la même langue ne sont sans doute pas suffisamment significatives pour être dignes d'intérêt. Si les exemples ne sont pas légion sur le sujet, on pourra lire avec intérêt un travail récent où les auteurs montrent au travers d'une étude sur 5 tokeniseurs et 10 langues que les différences de tokenisation ont un impact significatif sur les résultats en traduction automatique ([Domingo et al., 2018]). Autrement dit, la tokenisation n'est pas une opération innocente qui mérite d'être laissée de côté.

Ce sont donc deux aspects, l'un épistémologique et l'autre pratique (ou disons expérimental), qui questionnent la pertinence du recours systématique à des représentations en mots. Le troisième aspect est plutôt de l'ordre de l'éthique, il s'agit de la possibilité effective de calculer cette représentation en mots pour toutes les langues que l'on peut avoir à traiter. En effet, nous avons ici une représentation gourmande en ressources qui exclut de fait les langues pour lesquelles on dispose de peu de documents [28], mais aussi de langues pour lesquelles encore une fois le concept de mot est beaucoup moins pertinent qu'il ne l'est pour la langue anglaise ou des langues proches. C'est le cas notamment des langues agglutinantes telles que le finnois ou le japonais ou encore de langues où la séparation des mots sera moins marquée visuellement, comme l'arabe, ou pas du tout marquée, comme le chinois. Ainsi que l'a observé Emily Bender [Bender, 2019], il serait faux de penser que l'anglais puisse d'une quelconque manière être une langue représentative de la variété des langues du monde et il faux donc que la communauté assume de dire explicitement la langue des corpus traités plutôt que de considérer l'anglais comme la langue par défaut du TAL (voir [Ducel et al., 2022a, Ducel et al., 2022b] pour des études statistiques sur l'absence de mention de la langue traitée dans les conférences ACL, LREC et TALN).

L'écrasement des observables ([Lucas, 2009, Lecluze and Lejeune, 2014]) induit par les représentations centrées autour du mot laisse tout de même une place pour des représentations plus globales, oserais-je dire « plus ambitieuses ». Certaines problématiques linguistiques relativement éloignées des turpitudes des variations locales permettent d'exploiter des propriétés des textes qui ne sont pas simplement héritées des mots qui les composent. Le mot n'est pas la seule unité d'analyse possible mais on voit bien que les caractéristiques langagières liées à la pragmatique et au style ont été plutôt ignorées dans le TAL. Sans doute du fait qu'elles s'accommodent mal d'une analyse centrée sur les mots ce qui rappelle encore la forte dépendance du TAL au « mot ». La langue ce n'est pas simplement des unités détachables, analysables hors-sol, c'est aussi, et peut être surtout, du contexte. Hors du contexte, la transmission de l'information est incertaine. Quand le contexte est là, à travers la dimension du document ou mieux encore du corpus, il convient sans doute de se poser la question de son utilisabilité. En effet, ce n'est pas seulement intéressant *per se* de disposer du contexte, et d'en tirer partie, c'est utile aussi pour la tâche. Cela implique de trouver le moyen de représenter ce contexte d'une manière calculable [29].

1.4.3 Comment le global détermine le local

Les représentations fondées uniquement sur des critères locaux limitent la stabilité de l'interprétation. En effet, on ne peut garantir que le même énoncé, dans d'autres conditions d'énoncia-

28. Ou encore une fois on les condamne à disposer des corpus les plus opportunistes, constituées de l'agrégation de sources diverses et variées.

29. Sans doute trouve-t-on ici une limite de ce que l'on sait faire en terme de calculabilité de la langue naturelle : *Within a computer natural language is unnatural* ([Perlis, 1982]

tion, admette exactement la même ou les mêmes interprétations. Dès lors, la prise en compte de paramètres tels que le genre de texte, le public visé et le contexte d'énonciation est fondamental. Toutes ces propriétés, dont la prise en considération est naturelle pour le linguiste, inconsciente pour le locuteur, ont tout intérêt à être prises en compte pour l'analyse automatique mais si elles sont problématiques pour le programmeur. On pourrait les appeler paramètres « externes », par opposition on paramètres internes, visibles, segmentables du contenu langagier. Mais c'est bien évidemment une simplification majeure de ce qu'est la langue, de ce qu'est la communication. Dans une présentation invitée à l'atelier Ethique et TRaitemeNt Automatique des Langues (ETeRNAL) en 2020 Dirk Hovy dans une présentation invitée intitulée « Layers, Biases, and Responsibility » [30] donne quelques éléments qui montrent que ces considérations ont aussi une importance d'un point de vue expérimental. Hovy montre comment le fait de considérer que l'on traite des langues standardisées est un biais majeur. Mais, encore une fois pas simplement d'un point de vue éthique mais d'un point de vue purement pratique. Par exemple, il montre comment la connaissance de l'âge des locuteurs ayant produit les énoncés influence le résultat de classifieurs [Hovy, 2015]. D'un point de vue plus général Dirk Hovy montre à quel point il est biaisé de considérer que l'on travaille sur des langues standardisées et que notre manière de modéliser est un biais qui peut générer des handicaps dans les systèmes que nous créons.

Réduire l'analyse de données langagières à une série de mécanismes de transformation « rationnels »qui permettent d'extraire le sens est donc une idée élégante mais par certains aspects réductrice et moyennement opérante. En ce sens cette approche amène une limite, un plafond de verre je dirais, à la faculté de compréhension de la langue que l'on peut transmettre à une machine ; Mais on peut légitimement se demander bien sûr, quelle faculté on cherche à transmettre. Dans un *pipeline* traditionnel de TAL, la représentation ressemble assez à un automate, une série d'états et de transitions censés représenter les différents paliers de la représentation linguistique et donc *in fine* de capturer l'essence des énoncés à traiter.

En ce sens, les modèles de types réseaux de neurones profonds sont peut être plus proches de la non-localité puisqu'ils permettent de modéliser le non-modélisable ou en tout cas de le modéliser implicitement plutôt qu'explicitement. On pourra objecter que la représentation initiale de la langue, une séquence de mots et de sous-mots, est déjà une réduction importante de la réalité du matériau.

Ce qui va être traité par un programme de TAL va correspondre à un état de langue qui sera décrit comme traitable automatiquement, puisque standard, état de langue à opposer à un état de langue non traitable par la recette automatique standard, état de langue dégradé, bruité ... Les données en entrée doivent donc être nettoyées, débruitées, expurgées en un mot standardisées. On peut alors parler d'une tendance répandue en traitement Automatique des Langues à l'écrasement des observables. Certaines dimensions du texte ont tendance à être escamotées de manière à faciliter la factorisation et la réutilisation.

Ceci n'est pas sans rapport avec l'indétermination de la traduction, ou l'impossibilité de la traduction radicale, telle qu'elle est décrite chez Quine [Quine, 1960]. Quine postule que l'on ne peut garantir que l'on traduise l'intégralité d'un énoncé linguistique. Il imagine un linguiste observant une peuplade et qui remarque que le passage d'un lapin devant ses yeux et ceux de la peuplade qu'il observe amène l'utilisation du mot « Gavagai ». Ici, on peut selon Quine circonscrire le mot *gavagai* à quelques significations probables parmi lesquelles *Gavagai* = `lapin`. Pour autant on ne peut pas dire avec certitude que Gavagai n'a pas une signification plus large (*Gavagai* = `lapinidé` par exemple) ou plus étroite (« ah tiens, le même animal que l'autre fois »). Un nombre supplémentaire d'observations ne garantirait pas une certitude totale. L'observation

30. `https://jep-taln2020.loria.fr/conference-virtuelle/programme/keynote/` consulté le 2 octobre 2023

est limitée, l'information est incomplète de sorte que le point de vue du linguiste de terrain n'est pas exempt de biais et on doit donc faire appel à l'interprétation. Ce que Quine décrit, au travers d'un exemple de linguistique de terrain, c'est la problématique de lister avec certitude les implications d'un acte langagier, de pouvoir énumérer ce qui est observable et utile à la compréhension de l'acte langagier. Ce faisant, il montre une démarcation importante entre le courant pragmatique, dont Quine est issu au même titre que des philosophes plus connus dans la sphère des sciences du langage tels que Peirce, et le courant de l'empirisme logique qu'il critique. Tous les observables langagiers n'étant pas présents au même niveau, choisir un niveau d'observation c'est donc simplifier pour, *a priori*, mieux généraliser, mais aussi sacrifier un ou plusieurs observables.

1.4.4 Gestion de la variation : entre généralisation et adaptation

Ceci nous amène au problème central de TAL dont il sera question ici, la question des observables (les unités de bases) et des observatoires (les textes et des corpus) dans lequel ils pourront être identifiés ou non selon la variation observée dans les corpus traités. Comment est-ce que l'on applique des modèles appris sur des données X à des données Y ? Quelle fiabilité peut-on attendre ? Quel est le degré de confiance des modèles ? Dans quelle mesure ont-ils une grande couverture ?

Est-ce qu'un texte, un corpus sont des accumulations des enchevêtrements, des réseaux d'observables ? Qu'est-ce qu'un corpus prêt à l'usage pour le TAL pour l'ADT [31] ? Je présenterai une série de réflexions appuyées par différentes expérimentations menées autour de l'usage des observables et des observatoires. Tout d'abord, dans la partie II je présenterai des cas où les corpus traités ne présentent pas de variation problématique au grain local, laissant plus de liberté dans la détermination des méthodes à mettre en uvre. J'examinerai trois genres textuels : un corpus d'uvres littéraire, un corpus d'articles de Presse et un corpus d'articles scientifiques. Ces corpus sont des corpus prêts à l'emploi dans la mesure où ils ne présentent pas de variation inattendue :

— retranscriptions d'uvres littéraires retouchées à la main ;
— articles de presse provenant d'une seule source et permettant donc une approche *ad hoc* de l'extraction du contenu textuel ;
— articles scientifiques avec une retranscription corrigée et présentant en plus une structure XML riche.

Dans ce dernier cas, je montrerai comment la richesse structurelle d'un corpus constitue un observatoire fécond pour la définition d'observables pertinents. Il se trouve que ces situations où les corpus sont non bruités, globalement homogènes et où les documents sont explicitement structurés ne constituent pas la situation la plus commune ni, je crois, la plus intéressante. Dans la partie III, je décrirai des situations différente avec des expériences menées sur des données hors « conditions de laboratoire » avec là aussi l'utilisation de données de genres textuels différents et présentant différents types de variation :

— Un corpus de presse multilingue, provenant de plusieurs dizaine de sources et examiné au prisme de la variation induite par cette multiplicité ;
— Un corpus de textes littéraires obtenus par océrisation mais comparé à une vérité de terrain disponible en ligne ;
— Un corpus de documents historiques plus bruités et n'offrant pas de vérité de terrain.

31. Analyse statistique des Données Textuelles

Deuxième partie

Renouveler les paradigmes du TAL en faisant varier les observables

Chapitre 2

Un auteur se définit-il autrement que par ses mots ?

Sommaire

Dans ce chapitre je m'intéresse à la question de la systématicité du recours au mot comme unité d'analyse minimale en prenant pour exemple des tâches liées à la reconnaissance d'auteurs avec une finalité de classification non-supervisée puis supervisée. Il faut entendre ici la notion d'auteur au sens de scripteur c'est-à-dire de créateur principal d'une donnée textuelle à analyser. Dans les deux cas, l'objectif est de caractériser des styles individuels, autrement dit de chercher des descripteurs calculables automatiquement qui soient susceptibles de discriminer des auteurs entre eux.

Dans un premier temps (Section 2.1), il sera question de la manière de caractériser des auteurs autrement que par le simple lexique. Il s'agira ensuite (Section 2.2), de s'attaquer à une tâche de classification non supervisée dans le domaine littéraire en ayant recours à une redescription du texte en motifs syntaxiques pour discriminer des auteurs. Enfin, (Section 2.3), sera abordée une question de classification supervisée où l'on aura cette fois recours à des motifs sous forme

de chaînes de caractères, mots ou non-mots [32], pour une tâche d'attribution d'auteur dans le domaine journalistique.

2.1 Comment caractériser le style d'un auteur autrement qu'avec des mots ?

La caractérisation du style d'un auteur est un sujet intéressant du point de vue de l'épistémologie du TAL puisque l'on peut tout aussi bien voir le style comme un moyen, les hypothèses sur le style favorisant la réalisation d'une tâche donnée, ou comme une fin c'est-à-dire que la réalisation efficace de la tâche permet de valider la pertinence des descripteurs. Un exemple de la première approche peut être fourni par l'utilisation d'hypothèses sur le style collectif [33] pour réaliser une tâche de classification en contexte multilingue [Lejeune et al., 2015a], ou une caractérisation du style individuel exploitant des chaînes de caractères récurrentes, incluant la ponctuation en particulier, [Brixtel, 2015,Stamatatos et al., 2016] pour discriminer ou regrouper des auteurs. Dans [Legallois, 2016], les auteurs abordent la seconde approche en s'intéressant à l'identification automatique de traits saillants du style individuel qui puissent être analysés pour eux-mêmes. Bien entendu, ces traits sont aussi amenés à nourrir des approches de classification automatique et donc d'être des supports pour une évaluation quantitative.

Dans le domaine des humanités numériques, il semble particulièrement important de ne pas simplement rechercher l'efficacité, mesurée avec des mesures d'agrégation (telles que précision, rappel et F-mesure), mais d'avoir des résultats qui soient interprétables notamment en ayant recours à des observables appropriés. L'idée est de chercher dans la définition des observables et leur exploitation algorithmique un équilibre entre efficacité et interprétabilité. Ceci rejoint une préoccupation historique de l'intelligence artificielle en général, particulièrement prégnante avec l'essor des approches fondées sur de l'apprentissage automatique ou profond : à quel point l'amélioration de la qualité des résultats peut suffire à justifier la complexification des approches ? Cette question n'est pas anodine et dépasse le simple cadre de l'application de méthodes d'IA pour résoudre une tâche donnée, schéma un peu classique de la collaboration entre l'informaticien agissant en qualité d'ingénieur et le chercheur en humanités agissant en qualité de client qui fournit des données et choisit dans un catalogue les « sorties » que l'on pourrait lui fournir. Un des objets de la création du CERES est justement d'intégrer la question des utilisateurs et de leur *desiderata* comme une partie intégrante de l'épistémologie de l'informatique appliquée [34].

Ce terrain nous permet d'explorer la question de l'adaptation de la méthode aux propriétés intrinsèques des données traitées et aux besoins des utilisateurs. Mathieu Valette [Valette, 2016] parle ainsi d'une tendance forte du TAL à se poser peu de questions épistémologiques et à utiliser des sortes de recettes qui tiennent peu compte du matériau textuel étudié ou des besoins réels des utilisateurs finaux. Or, les utilisateurs ne se satisfont pas toujours d'avoir une boîte noire dont on ne peut exploiter que les résultats en sortie, dont le fonctionnement est difficile à comprendre et qui amène souvent à devoir redéfinir ses objets de recherche en fonction des tâches que la boîte noire peut réaliser. Au-delà du choix des algorithmes eux mêmes, une façon de favoriser l'interprétabilité serait de choisir des descripteurs moins nombreux mais plus à

32. Dans le sens que l'on peut avoir à la fois des mots, des sous-mots (*subwords*) mais aussi plusieurs mots ou sous-mots, en somme on a des chaînes non-restreintes par les frontières des mots.

33. Dans le sens défini dans la section 1.2.2 page 15 de pratique discursive régulièrement observée dans un genre textuel particulier

34. Il s'agit de ne pas réduire au seul domaine du TAL cette question de l'implication des utilisateurs dans le processus de conception des « outils informatiques » . Les même questions se posent, ou devraient se poser, en vision par ordinateur ou plus généralement en apprentissage automatique

même d'être analysés qualitativement. Autrement dit, réduire la taille de l'espace de description quitte à accepter une efficacité moindre et donc choisir les observables pour eux-mêmes et non pour leur capacité à optimiser des métriques d'évaluation. Ces descripteurs, que l'on va utiliser en entrée d'un système de classification par exemple, doivent à la fois pouvoir être identifiés par une machine, qui n'est pas un locuteur expert et se limite à des opérations somme toutes sommaires, et être pertinents pour l'expert. Or, la difficulté est que les observables faciles à identifier ne sont pas toujours les plus intéressants pour l'analyse qualitative, quand bien même ils offriraient de bons résultats, et que les stylèmes dont le choix apparaîtrait vraiment judicieux aux yeux de l'expert semblent parfois très difficiles à repérer algorithmiquement (voir par exemple ceux décrits dans [Molinié and Viala, 1993]) [35]. Les inférences nécessaires à l'extraction de stylèmes sont difficiles à énumérer exhaustivement, rendant sans doute leur caractérisation automatique illusoire. Il s'agit donc de trouver un équilibre entre des représentations exploitant des observables pauvres, telles que approches en sacs de mots, mais qui sent faciles à obtenir, et des représentations avec des observables plus riches, proches de ce que « voit » un expert en analyse de style, mais ardues à décrire formellement.

2.1.1 Quels observables pour quels usages ?

Les observables de bas niveau vont être des sous-parties de l'ensemble des sous-chaînes de caractères d'un texte. Les mots, les *subwords*, les n-grammes de caractères (avec n fixe ou compris dans un intervalle) qu'ils soient mots ou non-mots sont donc simplement des sous-ensembles particuliers de formes de surface. On peut aussi utiliser des observables obtenus via une redescription de ces éléments de surface : descripteurs syntaxiques (POS tags [36]), lemmes, entités nommées ou encore représentations « sémantiques » provenant de plongements de mots par exemple. J'utilise ici le terme de redescription des données au sens donné en fouille de données (voir par exemple [Rioult, 2017]) c'est-à-dire :

« *une transformation de l'espace dentrée, peu structuré conceptuellement ou temporellement, vers un espace des régularités sur lequel lalgorithme générique de décision peut sappuyer.* »

La redescription s'entend donc ici dans le sens d'une transformation d'observables de premier niveau en des observables plus riches en information ou des représentations plus concises et à même de produire des résultats de qualité mais qui laissent la possibilité à l'expert d'exercer des interprétations. Sur ce sujet, voir par exemple [Lejeune et al., 2016] pour l'utilisation de redescriptions de dialogues avec la taxonomie DIT^{++} [Bunt, 2009] pour la prédiction de réactions à des stimuli ou [Rioult, 2017] pour une vue plus générale des applications dans des domaines tels que l'analyse de trajectoires dans le sport.

Dans ce processus de redescription, les observables peuvent être redéfinis de manière exogène, par exemples avec des listes finies telles que les listes de POS tags et de signes de ponctuations considérés comme les plus utiles pour l'AA énumérés par [Stamatatos, 2009]. À l'opposé, on trouve dans la littérature l'utilisation d'observables choisis de façon plus endogène et donc spécifiques au corpus (voir [Quiniou et al., 2012] pour l'utilisation de motifs syntaxiques). On pourrait enfin utiliser des modèles de langue, ce qui rentre dans la catégorie exogène pour les modèles pré-entraînés mais devient plus endogène lorsque l'on opère un *fine-tuning* sur le corpus de travail.

Qualifier le style d'un auteur peut être vu avec un objectif de tâche de classification supervisée, ce qui est le cas le plus fréquent en informatique. [El Bouanani and Kassou, 2014] définissent un ensemble d'observables (de traits si l'on conserve la terminologie des auteurs) qui ont des

35. J'utilise ici le terme *stylème* dans le sens très générique d'observable ayant pour but de caractériser un style.
36. *Part of Speech* tags ou étiquettes en parties du discours

distributions relativement constantes pour un auteur donné et sont suffisamment distinctifs de son style d'écriture par opposition à celui d'autres auteurs. [Koppel et al., 2011] ont exploité des informations telles que l'utilisation de certaines suites particulières de mots et de caractères pour capturer des traits stylistiques personnels. Si l'exploitation des mots et des lemmes nécessite des ressources *a priori*, l'exploitation des chaînes de caractères non-mots est indépendante de la langue de ce texte. Un profil d'auteur est alors construit à partir des n-grammes contenus dans les textes qui lui sont associés. Des techniques d'apprentissage automatique supervisé sont utilisées pour apprendre à partir de ces profils, en fonction d'un corpus d'entraînement où les paires ([texte, auteur]) sont connues. Les recherches en AA peuvent se concentrer sur certains de ces problèmes : le passage à l'échelle quand un grand nombre d'auteurs candidats est considéré ou l'indépendance vis-à-vis de la langue lorsque les ressources linguistiques sont rares ou manquantes.

Dans ces travaux, l'indépendance vis-à-vis de la langue est abordée avec des méthodes fondées sur les caractères mais le calcul et l'exploitation de toutes les chaînes de caractères d'un texte peut s'avérer coûteux, en particulier d'un point de vue d'espace mémoire. La contribution principale de l'expérience qui va suivre consiste en l'utilisation d'un nouvel algorithme pour manipuler des chaînes de caractères, en vue de réduire l'espace de description et ainsi le temps et le coût d'entraînement. L'approche classique fondée sur les n-grammes de caractères de longueurs variables est comparée à une approche exploitant des *répétitions maximales* ainsi que des *répétitions maximales du* $2^{ème}$ *ordre*. Les expériences ont été menées grâce à la constitution de trois corpus : un en anglais, un en français et un correspondant à la concaténation des deux autres. Pour un état de l'art complet, se référer aux travaux de [Koppel et al., 2009], de [Stamatatos, 2009] et de [El Bouanani and Kassou, 2014].

L'AA est dans cette étude vue comme une tâche de catégorisation multi-classe de textes à étiquette (*label*) unique. Comme détaillé dans [Sun et al., 2012], trois éléments principaux doivent être définis en amont : (I) la nature des traits exploités, (II) le mode de calcul de l'ensemble des traits représentant un texte et (III) la façon de manipuler ces représentations pour relier un texte à un auteur.

2.1.2 Définitions de caractéristiques pour l'attribution d'auteur

Les observables utilisés en AA peuvent être séparés en différents groupes [Abbasi and Chen, 2008] :

- **observables lexicaux** : des valeurs associées à des statistiques sur les mots (nombre de mots dans les textes, nombre de caractères par mot, nombre de bi-grammes/tri-grammes de caractères au sein de ces mots) ;
- **observables syntaxiques** : des valeurs associées à la syntaxe des phrases (effectif des mots outils, des uni-grammes/bi-grammes/tri-grammes de ces mots outils ou des séquences de parties du discours) ;
- **observables structurels** : des valeurs numériques associées à des unités plus grandes (nombre de paragraphes ou encore longueur moyenne des paragraphes) ;
- **observables thématiques** : des valeurs associées avec le contenu thématique (des sacs de mots, des n-grammes de mots clés) ;
- **observables morphologiques** : des particularités en rapport avec des pratiques idiosyncrasiques (telles que les fautes d'orthographe ou de frappe) [37].

37. Idée que l'on retrouve aussi en datation automatique : le bruit présent dans des textes issus d'OCR peut constituer un indice concernant la période d'impression [Baledent et al., 2020]

	T	O	I	TO	OT	TI	IT	TOT	OTO	OTI	TIT	ITI
$1 \leq n \leq 1$	4	2	2									
$2 \leq n \leq 2$				2	2	2	1					
$3 \leq n \leq 3$								2	1	1	1	1
$1 \leq n \leq 2$	4	2	2	2	2	2	1					
$1 \leq n \leq 3$	4	2	2	2	2	2	1	2	1	1	1	1

TABLE 2.1 – Vectorisation en n-grammes de caractères de la chaîne $TOTOTITI$ selon l'intervalle de valeur de n considéré

Parmi ces caractéristiques, certaines sont spécifiques à des types de langue et de graphie. Si découper un texte en mots est aisé dans certains cas (en définissant un mot comme une chaîne de caractères entourée d'espaces), ce n'est pas une tâche triviale en chinois ou en japonais par exemple. Les approches exploitant les n-grammes de caractères apparaissent alors comme étant les plus simples pour traiter n'importe quelle langue naturelle [Grieve, 2007, Stamatatos, 2006] ou non [Burrows et al., 2014]), ainsi que les plus performantes.

Il convient toutefois d'être prudent sur le qualificatif de « multilingue »appliquée à ce genre de méthodes, car ainsi que le souligne Émilie Bender ([Bender, 2009]), une méthode indépendante des langues ne doit pas forcément être dépourvue de considérations linguistiques. Si l'extraction de n-grammes est réalisée indépendamment de la langue traitée, le choix du paramètre n doit être fait en fonction des langues abordées. Étant donné les différences morphologiques des langues (flexionnelles, agglutinantes, *etc.*), ce paramètre ne pourra pas amener les mêmes résultats selon la langue.

[Sun et al., 2012] défendent l'idée qu'utiliser une valeur fixe de n ne peut mener qu'à l'extraction d'informations lexicales (pour de petites valeurs de n), contextuelles ou thématiques (pour des plus grandes valeurs), mais n'expliquent pas si cette propriété est valide pour le chinois ou pour toute langue. Les auteurs soutiennent que l'on peut escamoter ce problème du choix de la longueur en exploitant des n-grammes de longueurs variables (des sous-chaînes de longueur entre 1 et n), donc en capturant des informations de types différents (lexicales, contextuelles et thématiques). On observe une redondance dans la représentation présente lorsque l'on analyse des n-grammes avec une valeur de n comprise dans un intervalle d'empan supérieur à 1. En effet, l'augmentation de la taille des n-grammes induit rapidement une diminution de la fréquence des n-grammes considérés et donc une augmentation de la taille de la représentation (voir tableau 2.1 pour un exemple). On voit que TO et TOT sont redondants, ayant la même fréquence sur l'intervalle $1 \leq n \leq 3$.

On observe régulièrement qu'il n'y a plus de gain en termes d'efficacités à utiliser des intervalles d'empan supérieur à 3 ou 4. Ceci s'est avéré vrai que ce soit pour des tâches de « bas niveau » comme l'identification de langue (voir [Ghoul and Lejeune, 2019, Ghoul and Lejeune, 2020] pour un travail sur les dialectes de l'arabe) ou sur des tâches de classification mettant en jeu plus de sémantique telle que l'analyse de sentiments [Buscaldi et al., 2017], de polarité [Buscaldi et al., 2018] ou encore la détection de sarcasmes [Ghoul and Lejeune, 2021].

Un même trait peut être attribué à plusieurs paires ([texte, auteur]) mais chaque texte et auteur ne partagent pas pour autant un grand ensemble de traits caractéristiques. Différents ensembles d'observables peuvent être définis pour représenter des textes (et par extension, pour représenter des auteurs). Considérant les méthodes d'AA existantes, deux catégories principales d'observables peuvent être définies :

— les observables dits *hors-ligne* : considérés *a priori* pertinents pour une tâche donnée, comme ceux largement décrits par [Chaski, 2001]. C'est le cas par exemple si l'on considère que le n-gramme *ment* ou le patron syntaxique $DET - ADJ - NOM$ seront inclus dans la représentation quoi qu'il arrive.
— les observables *en-ligne* : calculés *a posteriori* à partir de la collecte du corpus, ils sont définis pendant le traitement de celui-ci (dans le cas de méthodes supervisées, en fonction des corpus d'entraînement et de test, voir par exemple le modèle de langue en caractères décrit par [Peng et al., 2003]). Ils ne peuvent être complètement définis que lorsque le corpus à traiter est complet.

Différentes techniques pour exploiter ces observable ont été proposées. SVM (*Support Vector Machine* ou Séparateur à Vaste Marge) et les réseaux de neurones (*neural network*) sont des approches efficaces pour mener la tâche d'AA suivant le paradigme d'apprentissage automatique supervisé [Tweedie et al., 1996, Kacmarcik and Gamon, 2006]. Quand l'ensemble des auteurs candidats est extrêmement grand ou incomplet, rendant la représentation trop coûteuse à calculer, d'autres approches comparent les textes comme des ensembles d'observables avec des fonctions spécifiques pour calculer les similarités entre ces ensembles [Koppel et al., 2011]. D'autres approches utilisent des ensembles de caractéristiques individuelles pour construire un classifieur par auteur.

Nous exploitons dans les deux expériences qui suivent des observables *en-ligne* mais avec une différence : dans la première expérience il s'agira d'observables extraits via une méthode exogène, on utilise une ressource externe pour redécrire les données, alors que dans le second cas il s'agit d'une méthode purement endogène, la représentation étant entièrement calculée en fonction des données elles mêmes.

2.2 Classification non-supervisée : distinguer Dumas et Féval

Ici nous nous intéressons tout d'abord aux patrons syntaxiques (séquences d'étiquettes POS en l'espèce) que l'on pouvait trouver chez les deux auteurs puisque ce niveau de redescription présente l'avantage d'être raisonnablement efficace (les performances des POS *tagger* dépassant allègrement les 90% d'exactitude) et d'être intéressants pour l'expert comme l'a montré [Quiniou et al., 2012]. Ce travail prend la suite d'expérimentations réalisées dans la foulée du mémoire de Master Langue et Informatique d'Anaëlle Baledent, il a donné lieu à une communication orale en colloque et à la publication d'un chapitre de livre [Baledent and Lejeune, 2020]. Les expériences qui sont présentées ici ont été intégralement refaites pour les besoins de ce tapuscrit, le code originel permettant de reproduire les expériences est disponible en ligne [38].

2.2.1 Le Corpus Dumas-Féval (CDF) et sa redescription

Ce corpus comprend des uvres de deux auteurs comparables par bien des aspects afin de mettre en lumière au mieux les spécificités de chacun. L'objectif initial était de trouver s'il existait des spécificités du style de Dumas mais bien entendu l'usage peut être renversé pour trouver celles de Féval.

Nous avons sélectionné 9 uvres pour Dumas et 10 pour Féval en combinant romans et feuilletons de manière à obtenir un corpus qui soit équilibré en nombre de tokens. Il s'agit des retranscriptions de livres disponibles sur la plateforme Gutenberg [39]. Le tableau 2.2 présente les

38. `https://github.com/rundimeco/analyse-stylometrique_dumas-feval` consulté le 2 octobre 2023
39. `https://gutenberg.org/` consulté le 2 octobre 2023

statistiques sur ce corpus. Les données sur les genres sont issues de nos propres annotations, la taille en tokens a été obtenue avec TXM [Heiden et al., 2010]. En gras nous présentons les livres du corpus qui constituent des extrema en terme de taille et introduisent quelques différences qui auront leur importance : (I) parmi les livres qui comportent plus de 200 000 tokens, trois sont de Dumas et un seul de Féval ; (II) nous avons sur chaque sous-corpus un livre beaucoup plus court que les autres *Othon l'archer* pour Dumas (38 000 tokens) et *Le médecin bleu* pour Féval (16 000 tokens).

Auteur	**Titr**	**Année**	**Genre**	**# tokens**	**# chapitres**
DUMAS	*Le capitaine Paul*	1838	Aventure	76 787	19
DUMAS	*Acté*	1839	Historique	81 972	19
DUMAS	*Othon larcher*	1840	Historique, Aventure	**38 619**	11
DUMAS	*Le chevalier dHarmental*	1843	Historique, Aventure	164 631	48
DUMAS	*Les trois mousquetaires*	1844	Historique, Aventure	**284 527**	68
DUMAS	*La reine Margot*	1845	Historique	**258 859**	31 + 35
DUMAS	*Le chevalier de Maison-Rouge*	1846	Historique	**201 815**	56
DUMAS	*La tulipe noire*	1850	Historique, Aventure	89 471	33
DUMAS	*La femme au collier de velours*	1851	Fantastique, Historique	69 970	17
FEVAL	*Le médecin bleu*	1842	Historique	**16 556**	9
FEVAL	*Les fanfarons du roi*	1843	Historique, Aventure	97 080	25
FEVAL	*Le loup blanc*	1843	Historique, Aventure	92 780	24
FEVAL	*La fée des grèves*	1850	Historique	95 336	34
FEVAL	*La reine des épées*	1852	Aventure	113 056	23
FEVAL	*Le bossu*	1857	Historique, Aventure	**270 531**	62
FEVAL	*Les errants de la nuit*	1857	Aventure	121 054	30
FEVAL	*Le roi des gueux*	1859	Historique, Aventure	162 748	25
FEVAL	*La vampire*	1865	Fantastique	109 227	27
FEVAL	*Le cavalier fortune*	1868	Aventure	158 413	59

TABLE 2.2 – Description du CDF, pour chaque auteur les livres sont présentés dans l'ordre chronologique

Ici l'espace d'entrée est tiré une segmentation en mots, que l'on cherche à redécrire pour obtenir des formes plus génériques, facilitant idéalement l'apprentissage automatique comme l'interprétation ultérieure des résultats. L'approche que nous avons choisie est de chercher si des suites d'étiquette syntaxiques, ci-après séquences, peuvent être spécifiques à l'un ou l'autre des auteurs, [Béchet et al., 2012] utilisait en pareil cas le terme de « motifs » mais ceci peut introduire une confusion avec le sens que motifs a en stylistique. Contrairement à [Legallois et al., 2016], nous n'exploitons donc pas du tout le lexique ; D'autre part, nous nous intéressons aux séquences sans trous, c'est-à-dire aux séquences d'éléments strictement consécutifs (ce qui différencie ce travail de celui de [Longrée and Mellet, 2013]). L'étiquetage en partie du discours a été réalisé avec STANFORD-TAGGER [Manning, 2011] qui présente de bons résultats pour le français [Falk et al., 2014] et dispose d'un *wrapper* PYTHON très pratique. Parmi les jeux d'étiquettes possibles pour le français nous avons choisi celui développé par [Crabbé and Candito, 2008] plutôt que celui du *French TreeBank*[40] car ceci nous permettait d'avoir une description plus détaillée (cf. Tableau en annexe page 136). Ces séquences sont utilisées pour entraîner un classifieur non supervisé[41]. Il s'agit de retrouver des relations connues au sein du corpus, notamment l'auctorialité pour valider des descripteurs potentiellement pertinents. L'objectif de ces

40. `https://universaldependencies.org/fr/pos/index.html` consulté le 2 octobre 2023

41. L'implantation du K-means dans SCIKIT-LEARN en l'espèce cf. `https://scikit-learn.org/stable/` consulté le 2 octobre 2023

Original :	D'Artagnan	raconte	sa	première	visite	à	Mr	de	Tréville
Redescription :	NPP	V	DET	ADJ	NC	P	NC	P	NP
Séquences :	NPP-V-DET	(fréquence = 1 longueur = 3)							
	NC-P	(fréquence = 2, longueur = 2)...							

FIGURE 2.1 – Exemple de traitement appliqué à une phrase du corpus

expériences n'est pas simplement la réalisation d'une tâche, somme toute simple, mais de s'intéresser aux différents choix méthodologiques qui peuvent présenter des perspectives équivalentes en terme de résultats.

2.2.2 Chaîne de traitement pour la discrimination d'auteurs

Pour chaque livre du corpus les traitements suivant sont appliqués (un exemple est fourni dans la figure 2.1) :

— Redescription du texte en une séquence d'étiquettes POS avec STANFORD-TAGGER ;
— Calcul des sous-séquences de POS avec la contrainte de longueur(len) : $1 \leq len \leq 5$;
— Représentation sous forme de matrice avec les fréquences absolues ou relatives.

Le choix de limiter la taille des séquences extraites à une taille de 5 répondait à plusieurs besoins. Tout d'abord, il s'agissait de limiter l'espace de description et donc le temps de calcul. Ensuite, les séquences longues risquent d'être trop rares, du fait de l'anti-monotonicité de la fréquence [Agrawal and Srikant, 1994] [42], et donc d'intérêt limité pour la généralisation. Les séquences ont été exploitées pour différentes expériences sur le corpus :

1. **Recherche** de séquences discriminantes avec le Taux de Croissance (*Growth Rate* ou GR [Rioult, 2005] [43]) ;
2. **Observation** des résultats du *clustering* : ;
   (a) K-MEANS en faisant varier le nombre de *clusters* ;
   (b) DENDROGRAMMES pour chercher des proximités entre des paires de livres ;
   (c) ANALYSE EN COMPOSANTES PRINCIPALES pour obtenir une autre vue du corpus.

Une première approche a été de représenter les données avec pour chaque texte un vecteur des fréquences absolues des séquences en choisissant de garder les séquences respectant la contrainte suivante : $4 \leq longueur(sequence) \leq 5$ puisque c'est avec cette contrainte que les résultats étaient les meilleurs. On peut voir dans la partie gauche du Tableau 2.3 [44] que cette représentation a le défaut de regrouper des textes du fait de leur proximité en terme de longueur en mots. Les textes de taille proche ayant globalement des signatures proches : les séquences fréquentes y sont plus nombreuses et le nombre de séquences présente est plus grand (le « vocabulaire » augmentant naturellement avec la taille du texte, ce qui est vrai pour les séquences de mots l'est aussi pour les séquences de POS). La partie droite du tableau montre la même expérience avec cette fois des vecteurs de valeur relative.

42. L'anti-monotonicité de la fréquence est le fait qu'une séquence de longueur N est au maximum aussi fréquente que les sous-séquences de longueur $N-1$ qui la composent. Cette propriété, assez évidente au demeurant, permet de justifier la fixation d'un seuil maximal de longueur puisque les descripteurs seront partiellement redondants .

43. Le GR mesure le ratio de fréquence entre deux distributions, une définition plus détaillée figure dans la section 2.2.2 page 40.

44. Il est à noter que les lettres n'ont pas dans ce tableau de valeur sémantique particulière, elles visent simplement à faciliter la reconnaissance des *clusters* qui auraient une teinte proche, l'auteur de ces lignes lui même daltonien, demande aux lecteurs qui auraient cherché une motivation derrière les choix de lettres de bien vouloir l'en excuser.

		Fréq. absolue				Fréq. relative			
	Nombre de clusters	2	3	4	5	2	3	4	5
DUMAS (1)	*Le capitaine Paul*	W	W	V	W	V	V	V	V
DUMAS (2)	*Acté*	W	W	V	W	V	V	V	V
DUMAS (3)	*Othon larcher*	W	W	Y	X	V	V	V	Z
DUMAS (4)	*Le chevalier dHarmental*	V	X	X	V	V	V	Y	Y
DUMAS (5)	*Les trois mousquetaires*	V	X	W	Y	V	V	Y	Y
DUMAS (6)	*La reine Margot*	V	X	W	Y	V	V	Y	Y
DUMAS (7)	*Le chevalier de Maison-Rouge*	V	V	X	V	V	V	Y	Y
DUMAS (8)	*La tulipe noire*	W	W	V	W	V	V	V	V
DUMAS (9)	*La femme au collier de velours*	W	W	V	W	V	V	V	V
FÉVAL (1)	*Le médecin bleu*	W	W	Y	X	W	X	X	X
FÉVAL (2)	*Les fanfarons du roi*	W	W	V	W	W	W	W	W
FÉVAL (3)	*Le loup blanc*	W	W	V	W	W	W	W	W
FÉVAL (4)	*La fée des grèves*	W	W	V	W	W	W	W	W
FÉVAL (5)	*La reine des épées*	W	V	V	W	W	W	W	W
FÉVAL (6)	*Le bossu*	V	W	W	Z	W	W	Y	Y
FÉVAL (7)	*Les errants de la nuit*	W	V	V	W	W	W	W	W
FÉVAL (8)	*Le roi des gueux*	V	V	X	V	W	W	W	W
FÉVAL (9)	*La vampire*	W	V	V	W	W	W	W	W
FÉVAL (10)	*Le cavalier fortune*	V	V	X	V	W	W	W	W

TABLE 2.3 – Comparaison des résultats du *clustering* selon que la vectorisation est réalisée en fréquence absolue (partie gauche) ou en fréquence relative (partie droite) des séquences (avec $4 \leq longueur(sequence) \leq 5$). Chaque *cluster* est décrit par une lettre et une couleur pour plus de lisibilité.

Avec les valeurs absolues, les regroupements se font surtout sur des proximités de longueur comme indiqué précédemment. Au contraire, avec les valeurs relatives on voit bien se profiler une différenciation entre les deux sous-corpus. Avec deux *clusters* on a le résultat attendu, séparer les deux auteurs et, lorsque l'on augmente le nombre de *clusters*, on a des *clusters* interne à chaque auteur. En premier lieu avec 3 *clusters*, le roman de Féval *Le médecin de bleu* est isolé, du fait qu'il offre une richesse de séquences syntaxiques plus faible que les autres livres du corpus. Lorsque l'on passe à 4 *clusters*, on voit un cluster inter-auteur avec quatre livres de Dumas (*Le chevalier d'Harmental*, *Les trois mousquetaires*, *La Reine Margot* et *Le chevalier de Maison-Rouge*) et un de Féval (*Le bossu*). Dans ces romans on trouve de nombreuses séquences spécifiques du genre épique. Avec 5 *clusters*, et pour les mêmes raisons que *Le médecin bleu* précédemment, *Othon l'archer* se retrouve lui aussi isolé. Enfin, les livres longs ont plus tendance à contenir des séquences qui n'apparaissent que chez un auteur. En d'autres termes dans une représentation creuse telle que celle d'une matrice terme-document (ce qui est le cas ici pour une représentation en séquences), les documents longs sont plus facilement rapprochés entre eux puisque leur vocabulaire est plus grand. L'augmentation continue du vocabulaire au fur et à mesure qu'un échantillon de textes grandit en taille est un corollaire de la loi de Zipf ainsi que l'a montré [Lardilleux, 2010] et semble bien devoir s'appliquer aux séquences syntaxiques qui suivent elles mêmes une loi de Zipf (Figure 2.5a page 44). Ces séquences très rares, sont souvent absentes d'un des deux sous corpus, ont souvent un GR infini (désigne par convention la valeur

du GR quand le dénominateur est nul), on parle aussi de *Jumping Emerging Patterns* [Rioult, 2004]. Autrement dit, GR infini désigne ici le cas où une séquence est présente dans les données observées mais absente des données de contraste. Si une séquence est beaucoup plus fréquente dans un sous-corpus que dans un autre, on parle simplement de motif émergent (*emerging pattern* [Dong and Li, 1999]).

Plus formellement, étant donné nos deux sous-corpus *scD* (Dumas) et *scF* (Féval), pour chaque séquence on a :

— Si Effectif dans scFeval = 0 Alors $GR = \infty$
— Sinon si Effectif dans scDumas = 0 Alors $GR = 0$
— Sinon $GR = \frac{\text{Effectifdans scDumas}}{\text{Effectif in scFeval}}$.

Visualisation avec une analyse en composantes principales

Le fait de discriminer les documents produits par chaque auteur était une première étape, nous nous sommes ensuite intéressés à la visualisation de ces résultats en ayant recours à une Analyse en Composantes Principales (ACP). Pour rappel, l'ACP vise à réduire la dimensionnalité de la représentation afin, notamment, de pouvoir en tirer une visualisation en deux dimensions. La Figure 2.2 offre ainsi une visualisation alternative de la cohérence des *clusters* obtenue par la représentation en séquences syntaxiques. Chaque texte y est représenté par une lettre (D pour Dumas et F pour Féval) et un chiffre correspondant au numéro d'ordre chronologique de chaque uvre (numéros figurant dans le tableau 2.3 39) D'un point de vue global, l'ACP montre une nette séparation entre le sous-corpus Dumas (partie droite de la représentation) et le sous-corpus Féval (partie gauche de la représentation).

Nous pouvons voir que les livres de Féval forment un ensemble plus compact duquel on peut détacher *Le médecin bleu* (F1) et *Le bossu* (F6). Le dendrogramme de la Figure 2.3 (page 42) est cohérent avec cette observation et on voit que le Bossu est moins isolé puisqu'il se rapproche des *Fanfarons du Roi* (F2) et du *Cavalier fortune* (F10). Du point de vue de Dumas, on repère des paires de documents qui sont associés : *Les trois mousquetaires* (D5) et *La reine Margot* (D6), *Le chevalier D'Harmental* (D4) et *Le chevalier de Maison-Rouge* (D7), *Le capitaine Paul* (D1) et *La tulipe noire* (D8) et enfin *Acté* (D2) et *Othon l'archer* (D2 et D3).

Si une analyse précise de ces *clusters* intra-auteur présente certainement un intérêt, elle nécessite une expertise littéraire qui nous échappe. En tout état de cause, il semble bien que les micro cohérences observées chez Dumas attestent, d'une autre manière que pour Féval où le regroupement est plus collectif, de la pertinence de la représentation choisie.

2.2.3 Extraire et situer les séquences syntaxiques discriminantes

Si la représentation est effectivement pertinente, quels sont alors les observables les plus déterminants pour différencier les deux auteurs ? Pour les trouver nous avons choisi d'utiliser le taux de croissance qui est un ratio entre deux distributions. Plus précisément, il s'agit simplement de comparer la fréquence absolue (l'effectif) de chaque séquence dans les différents sous-corpus.

Comme nombre des séquences présentant un GR infini l'étaient avant tout par ce qu'elles sont tout simplement rares dans le corpus, nous avons restreint l'analyse aux séquences présentant d'un effectif minimal de 200 (sur l'ensemble du corpus). Ceci permet d'éviter d'extraire des séquences trop peu intéressantes.

Le tableau 2.4 présente un échantillon des séquences du corpus (avec fréquence minimale de 200) triées par ordre croissant de GR Nous avons donc dans les premières lignes, les séquences typiques de Féval ($GR << 1$), dans les dernières lignes celles qui sont typiques de Dumas ($GR >> 1$). Au centre figurent des séquences distribuées de façon quasi égale chez les deux

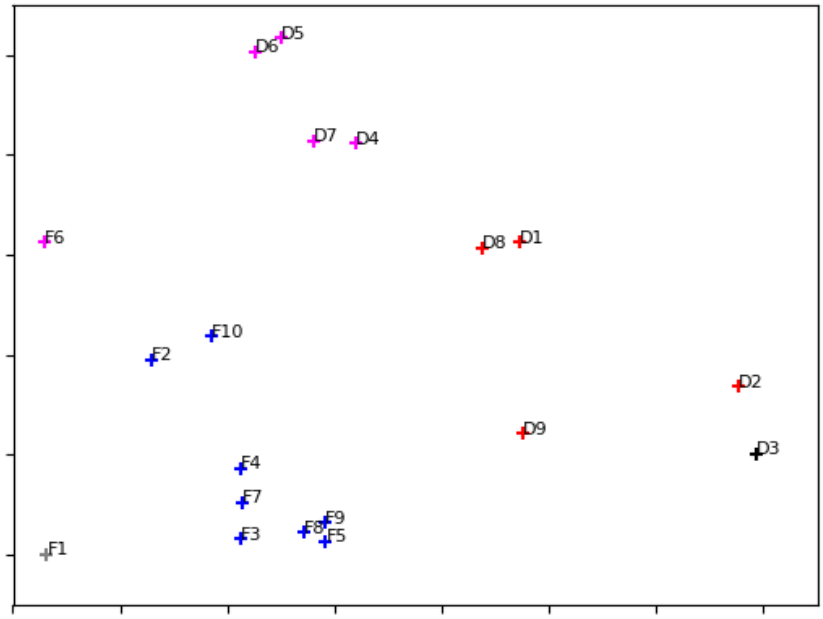


FIGURE 2.2 – ACP à partir de d'une représentation avec les séquences ($4 <= len(seq.) <= 5$), Dx représente les livres de Dumas et Fx ceux de Féval (les numéros correspondent à l'ordre chronologique pour chaque auteur, la correspondance avec les titres des uvres sont donnés dans la figure 2.3))

auteurs ($GR \simeq 1$), séquences probablement typiques du genre ou de la langue française en général.

Une manière d'utiliser les séquences extraites est d'opérer un retour au texte pour remettre les observables en contexte. En ce sens l'encodage du texte en séquences syntaxiques est une modalité de représentation, validée par la réussite de la classification, et on va pouvoir ainsi vérifier dans quelle mesure ceci permet de mettre en lumière les spécificités recherchées.

L'exemple 1, PUNC_V_NPP_PROREL (ponctuation - verbe - nom propre - pronom relatif) avec $GR = 0.33$ est spécifique à Féval. C'est une séquence syntaxique typique des dialogues dans laquelle on peut trouver des patrons lexico-syntaxiques tels que [...PUNC *Dit* NPP *qui* ...]. On trouve par exemple dans le corpus ces séquences dans *Le Médecin bleu* (Exemple 1 de la figure 2.4 tiré du chapitre 7). Avoir beaucoup de dialogues est une pratique courante chez les feuilletonistes pour favoriser le retour à la ligne, anticipant ainsi les stratégies mises en uvre par certains étudiants contemporains friands de lignes orphelines, cette stratégie se nomme tirage à la ligne [Carassus, 1970] qui s'explique notamment par une économie où le feuilletoniste est payé à la ligne. Ici, s'agissant d'une nouvelle il se pourrait que ce soit le reflet d'une uvre plus rythmée par les dialogues que les uvres de Dumas.

Comme second exemple, la séquence **NC_P_PROREL_CLS_V** (nom commun - préposition - clitique - pronom relatif) est beaucoup plus fréquente chez Dumas. Dans *Le capitaine Paul* (chapitre 6) on en trouve deux manifestations (Exemple 2 de la figure 2.4 ci-dessous).

Bien entendu, valider la pertinence de ces descripteurs nécessiterait une analyse experte plus approfondie mais ceci nous permet de valider la pertinence potentielles de redescriptions, ici sous la forme d'étiquettes syntaxiques, mais aussi l'importance de la longueur des séquences étudiées. Ainsi, les séquences de taille plus courtes (par exemple $1 \leq \text{len(seq.)} \leq 2$) ne véhiculent pas suffisamment d'information pour caractériser le style d'un auteur ce qui se reflète dans l'échec

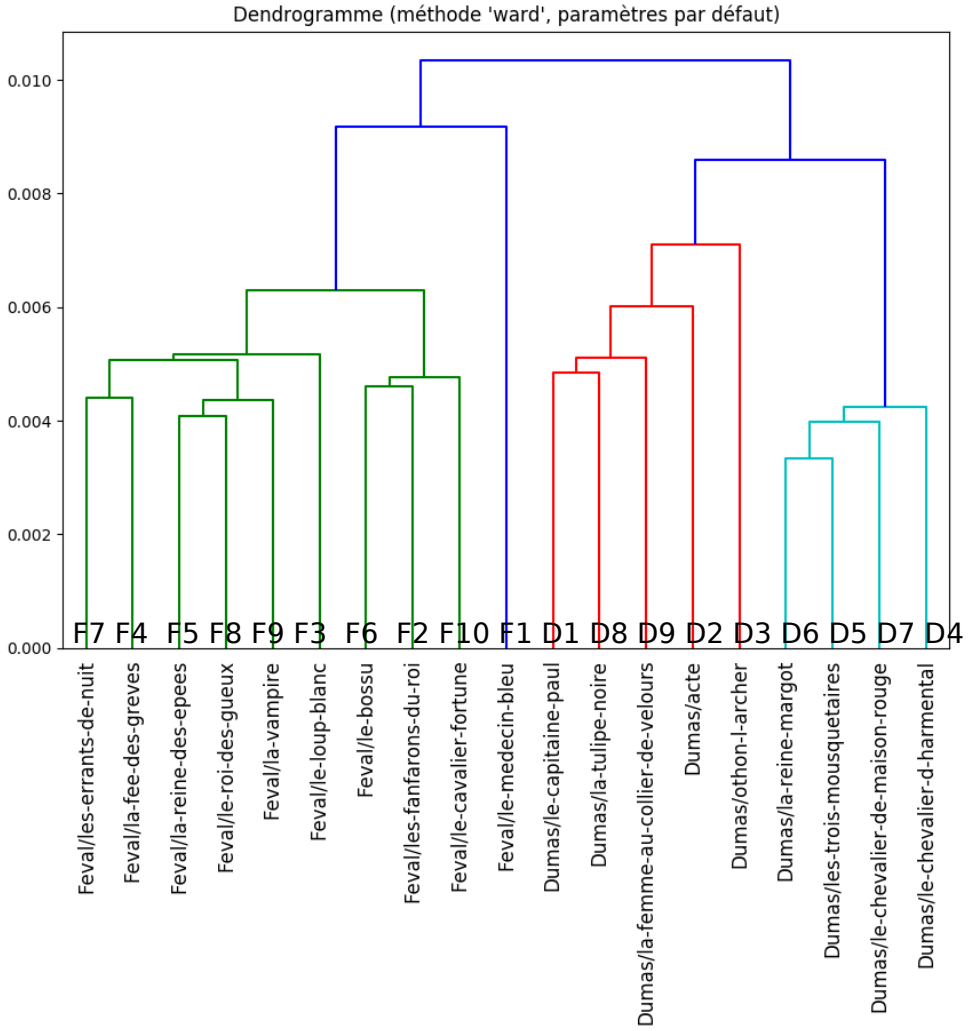


FIGURE 2.3 – Dendrogramme obtenu à partir de la représentation en séquences syntaxiques ($4 <= len(seq.) <= 5$)

Séquence	GR	Effectif dans *scDumas*	Effectif dans *scFeval*
NC_CL_CL_V_DET	0,3013	47	156
PUNC_V_NPP_PROREL	**0,3313**	55	166
CL_CL_V_DET	0,3951	162	410
DET_NC_CL_CL_V	0,4012	67	167
NC_CL_CL_V	0,4036	111	275
	...		
P_NC_P_NC	0,9822	2542	2588
NC_P_DET_NC	0,9849	10543	10705
P_NC_P_DET_NC	0,9849	1825	1853
V_P_DET_NC_P	0,9874	1566	1586
DET_NC_CC_DET_NC	0,9918	1455	1467
DET_NC_PROREL_V	1,0025	2375	2369
DET_NC_ADJ_ADJ	1,003	1007	1004
NC_P_DET_NC_P	1,0062	2287	2273
V_P_DET_NC	1,013	6071	5993
DET_DET_DET_DET	1,0151	5374	5294
	...		
PUNC_DET_NPP_PUNC	3,6667	187	51
NC_P_PROREL_CLS	3,9324	291	74
NC_P_PROREL_CLS_V	**3,9355**	244	62
DET_NC_P_PROREL_CLS	4,641	181	39
DET_I_PUNC_DET	5,9833	359	60

TABLE 2.4 – Séquences extraites (avec un effectif minimal de 200) ordonnées par ordre croissant de *GR*, nous avons donc en haut les séquences spécifique de Féval (respectivement en bas celles de Dumas) et au centre les séquences également distribuées dans les deux sous-corpus. Les séquences en gras font l'objet de commentaires plus détaillés.

Exemple 1 PUNC - V - NPP - PROREL

Voulez-vous me rendre bien content	**? dit Brand, qui**	se sentit rougir sous le cuir bronzé de sa joue.
Quoi ! sauvés tous deux ! sauvés par vous	**! dit Sainte, qui**	fondit en larmes. Que faire pour vous prouver...

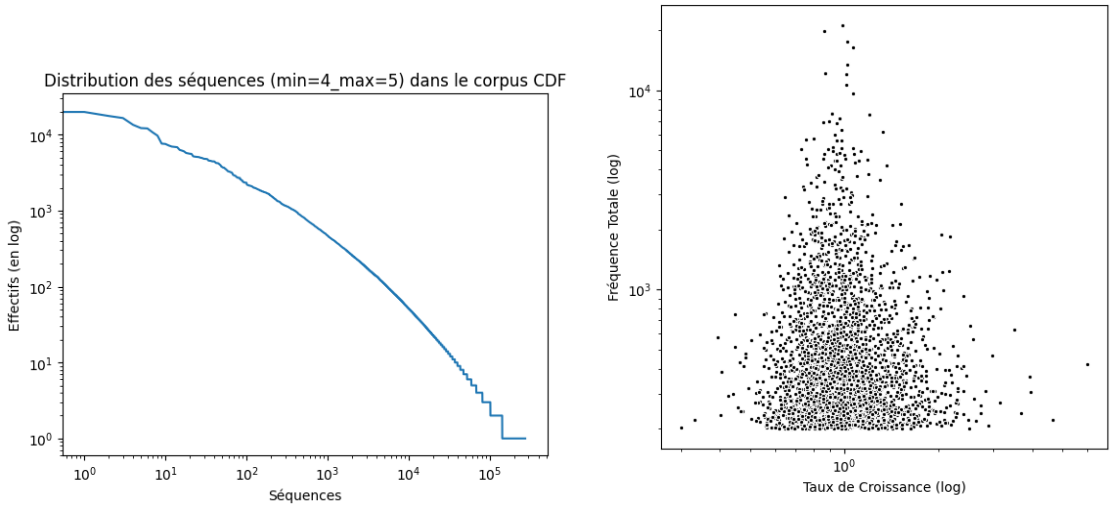
Exemple 2 NC - P - PROREL - CLS - V

car j'aurai un	**frère pour lequel je n'aurai**	plus d'amour,
et un	**mari pour lequel je n'aurai**	plus d'estime !

FIGURE 2.4 – Exemples de motifs discriminants en contexte

du clustering avec ces descripteurs alors que l'intervalle de longueur de séquences choisi ici ($4 \leq \text{len(seq.)} \leq 5$) offre de très bons résultats. On voit aussi que montrer la plus-value d'utiliser des approches automatiques ne dispense pas d'un peu de travail sur le paramétrage, il n'y a pas de recette magique, de sorte que ce n'est pas simplement le choix des descripteurs mais aussi la façon de les compter qui va permettre de faire parler les données de manière intelligible. Bien sûr, cela n'empêche pas que les livres longs soient plus facilement rapprochés puisque la probabilité que leur représentation soit moins creuse (comporte moins de zéros) est grande.

Il y a un effet de longue traîne qui est indéniable, et qui est lié au fait que de même que



(a) Loi de Zipf sur les motifs

(b) Distribution des motifs selon le GR

FIGURE 2.5 – Représentation distributionnelle des séquences d'étiquettes dans le CDF avec $4 \leq \text{len(seq.)} \leq 5$ (échelles logarithmiques). Il est à noter que dans la figure de droite pour plus de lisibilité, les séquences au taux de croissance nul (73 765 dont 59 976 hapax) ou infini (90 376 dont 71 168 hapax) ont été exclus de la représentation.

les n-grammes de mots, les n-grammes d'étiquettes suivent une loi de Zipf (Figure 2.5a). On peut raisonnablement faire l'hypothèse que les séquences représentées dans la partie gauche de la courbe sont spécifiques de la langue et que les séquences discriminantes mais fréquentes vont se situer dans le cur de la courbe. La figure 2.5b présente la distribution d'une autre manière en mettant en regard l'effectif total des séquences et leur taux de croissance. On voit sans surprise que les séquences les plus fréquentes sont aussi les mieux distribuées dans chacun des sous-corpus. De plus, les séquences présentant des taux de croissance extrêmes (nul ou infini) sont très souvent des hapax.

On le voit, il n'y a pas une seule manière d'exploiter les données textuelles pour une tâche donnée. Les deux prochaines sections présenteront un travail de classification supervisée cette fois où nous nous sommes intéressés à des observables d'un autre genre à savoir des chaînes de caractères mots ou non mots.

2.3 Classification supervisée : attribution d'auteur

Il est question ici aussi d'un travail sur les observables mais cette fois ci pour une tâche de classification supervisée, l'attribution d'auteur. La tâche d'attribution d'auteur (AA) est le plus souvent abordée sous l'angle de la stylométrie (ou étude du style). Ce domaine est aussi connu sous le nom de *writeprint*, en référence aux termes anglais « écriture » (*write*) et « empreinte digitale » (*fingerprint*). L'hypothèse sous-jacente est qu'un auteur laisse involontairement dans son message textuel des indices qui peuvent mener à son identification. Le problème d'attribution d'auteur consiste à deviner l'auteur de textes à partir d'un ensemble de candidats. Ainsi, cette tâche peut être vue comme un sous-domaine de l'apprentissage automatique supervisé.

Techniquement cela consiste à définir une nouvelle paire reliant un texte à son auteur.

Le travail présenté dans ce chapitre est issu d'une collaboration avec Romain Brixtel (contributeur principal de ce travail) et Charlotte Lecluze pendant nos contrats post-doctoraux respectifs en 2014-2015 [Brixtel et al., 2015, Brixtel, 2015]. Nous nous sommes intéressés à la tâche d'attribution d'auteur en contexte multilingue en proposant une alternative aux méthodes supervisées fondées sur les n-grammes de caractères de longueurs variables : les *répétitions maximales*. Pour un texte donné, la liste de ses n-grammes de caractères contient des informations redondantes. À contrario, les *répétitions maximales* représentent l'ensemble des répétitions de ce texte de manière condensée. Nos expériences montrent que la redondance des n-grammes, si elle est problématique pour la taille de représentation, contribue à l'efficacité des techniques d'attribution d'auteur exploitant des sous-chaînes de caractères. Ce constat posé, nous proposons une fonction de pondération sur les traits donnés en entrée aux classifieurs, en introduisant les répétitions maximales du $n^{\text{ème}}$ ordre (c'est-à-dire des répétitions maximales détectées dans un ensemble de répétitions maximales). Les résultats expérimentaux montrent de meilleures performances avec des répétitions maximales, avec moins de données que pour les approches fondées sur les n-grammes.

2.3.1 Catégorisation de textes fondée sur des observables en-ligne

L'expérience présenté ici offre plusieurs contributions :

— Présenter une alternative aux n-grammes de caractères via les répétitions maximales ;
— Montrer l'effet bénéfique de la redondance des n-grammes ;
— Proposer une nouvelle manière de prendre en compte l'interdépendance longueur-effectif pour la pondération d'une chaîne de caractères en fonction des sous-chaînes qu'elle encapsule.

Chaque classifieur agit tel un expert pour traiter un sous-ensemble de l'espace de recherche lors de la classification d'un corpus, chaque classifieur étant spécialisé dans la détection d'un auteur spécifique. Dans nos expériences nous utilisons un unique classifieur SVM, classifieur revendiqué comme le plus pertinent à l'époque par [Sun et al., 2012] et [Brennan et al., 2012], pour l'ensemble des auteurs en gardant les mêmes paramètres pour chaque expérience, en vue d'analyser finement l'influence du choix des observables sur le traitement. Les expérimentations que nous avons menées soulignent les propriétés principales des méthodes d'AA fondées sur les chaînes de caractères (en opposition avec des approches fondées sur les mots eux mêmes ou l'exploitation de leurs étiquettes morpho-syntaxiques comme vu dans la section précédente). L'étape de sélection des observables a pour but d'extraire ceux qui sont pertinents des corpus d'entraînement et de test sans *a priori* sur les langues traitées. Nous focalisons nos analyses sur l'influence de la sélection des observables en contexte multilingue.

L'ensemble des caractéristiques utilisées pour classifier correspond à l'intersection des ensembles de traits du corpus de test et du corpus d'entraînement puisque dans cette configuration celles qui seraient absentes de l'un ou l'autre des jeux de données n'auraient pas d'intérêt pour la classification. Toutefois, s'agissant des chaînes répétées maximales, les caractéristiques conservées ne seront pas les mêmes selon que l'on traite seulement le jeu d'entraînement ou l'ensemble du jeu de données. En effet, les critères de répétition de de maximalité seront affectés par l'augmentation de la taille du corpus. Nous ne pensons pas en l'espèce que cette utilisation du jeu de test constitue un biais puisque nous n'exploitons de connaissance sur les étiquettes du jeu de test mais nous extrayons des caractéristiques linguistiques de celui-ci afin de définir des observables calculés en ligne, en fonction du corpus uniquement, la segmentation des observables est donc non-supervisée [Buscaldi et al., 2017, Abiven and Lejeune, 2019].

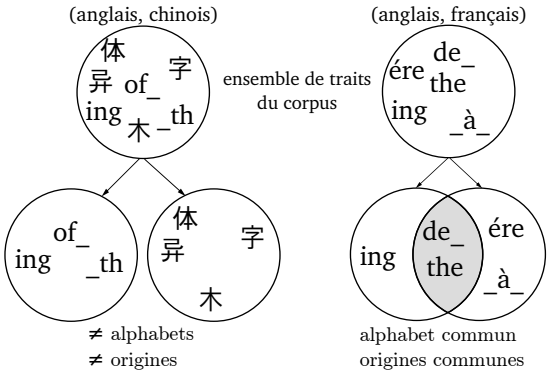


FIGURE 2.6 – Influence de la différence de nature des caractéristiques sur l'apprentissage.

	#caractères	#textes	#auteurs
corpus	$1.9\ 10^6$	631	40
auteurs (moyenne ± écart type)	$4.6\ 10^4 \pm 8075$	15.8 ± 2.6	
textes (moyenne ± écart type)	2945.1 ± 178.5		

TABLE 2.5 – Caractéristiques du corpus EBG (anglais).

Corpus utilisés Deux corpus différents ont été mobilisés, chacun constitué de textes écrits dans la même langue : un en anglais (corpus EBG), l'autre en français (corpus LIB). Ces deux langues ont été sélectionnées car elles partagent un alphabet et sont relativement proches typologiquement. Ceci rend la tâche plus difficile que lors du traitement de langues possédant plus de différences (anglais et chinois par exemple), les espaces de description au grain caractère étant, presque, totalement différents entre ces deux langues. Ainsi, une approche fondée sur SVM n'aurait aucune difficulté à séparer les textes analysés en deux sous-espaces contenant d'un coté les documents écrits en anglais, de l'autre les documents écrits en chinois (Figure 2.6).

EBG, sous-corpus de textes écrits par 40 auteurs, a été sélectionné au sein du EXTENDED BRENNAN GREENSTADT *adversarial corpus* [Brennan et al., 2012]. Ce corpus est exclusivement constitué de textes en anglais (Table 2.5). Les textes manipulés lors d'expériences menées par [Brennan et al., 2012], plagiats de styles ou de contenus, ont été exclus. La relation entre auteur et thème est ténue dans ce corpus, la plupart des auteurs ayant écrit des textes sur le même thème.

LIB, constitué pour l'occasion, comprend des articles de presse issus de la version en ligne du journal Libération, collectés pour les expériences décrites ici. Il contient des textes écrits en français par 40 auteurs qui ont écrits dans plus d'une catégorie du journal (sport, santé, politique étrangère, *etc.*). Il s'agit d'éviter des biais liés à la spécialisation de tel ou tel auteur sur une rubrique/thématique unique. Les auteurs qui écrivent exclusivement dans une seule de ces catégories ont donc été exclus afin d'éviter que le thème d'un article prime sur le style, ce qui rend ce corpus plus difficile à traiter. Les caractéristiques de ce corpus figurent dans le tableau 2.6.

Le corpus LIB contient autant d'auteurs que le corpus EBG, mais le nombre de textes pour chaque auteur est plus important (31.2 ± 4.2 textes par auteur pour le corpus LIB, 15.8 ± 2.6 pour le corpus EBG). Chaque texte, dans ces deux corpus, contient plus de 250 mots (environ 1 500 caractères), la longueur minimale nécessaire pour l'AA vue comme une tâche de classification

	#caractères	#textes	#auteurs
corpus	$5.1\ 10^6$	1247	40
auteurs (moyenne $\pm$ écart type)	$1.3\ 10^5 \pm 2.6\ 10^4$	31.2 ± 4.2	
textes (moyenne $\pm$ écart type)	4070.6 ± 1524.2		

TABLE 2.6 – Caractéristiques du corpus LIB (français).

selon [Forsyth and Holmes, 1996].

Le corpus MIXT est constitué à partir de la fusion des corpus EBG et LIB. Nous l'utilisons en vue d'éprouver le caractère multilingue des approches considérées. Des expériences sont aussi menées sur différents sous-corpus issus des corpus EBG, LIB et MIXT. Ainsi, EBG-10 (respectivement LIB-10 et MIXT-10) est un sous-ensemble de textes constitués de 10 auteurs du corpus EBG (LIB, MIXT). Aussi, les corpus MIXT-20, 40, 60 et 80 sont issus de la fusion des corpus LIB-10 + EBG-10, ... LIB-40 + EBG-40. Nous décrivons dans les sections suivantes les expérimentations menées sur ces corpus dans le but de mettre en valeur les différentes caractéristiques utilisées ainsi que les différents éléments de la chaîne de traitement.

2.3.2 Les chaînes de caractères répétées maximales d'ordre 1 et d'ordre n

Les répétitions maximales (*maximal repeats* ou *motifs* dans les travaux de [Ukkonen, 2009]) sont calculées en se fondant sur les tableaux de suffixes [Kärkkäinen et al., 2006]. Ces motifs représentent de manière condensée toutes les sous-chaînes d'un corpus. Pour la détection des chaînes *hapax* d'un corpus à partir de ses motifs, se référer aux travaux de [Ilie and Smyth, 2011]. Les motifs sont des sous-chaînes de caractères avec les caractéristiques suivantes :

répétition : les motifs apparaissent deux fois ou plus dans le corpus traité ;

maximalité : ajouter à un motif le caractère se situant sur sa gauche ou sa droite (on dit aussi « étendre » une occurrence d'un motif), donne une chaîne de caractères avec un nombre d'occurrences moindre que le motif de base.

Une description complète de ces caractéristiques et des algorithmes permettant de les vérifier figure dans [Brixtel, 2015]. La propriété intéressante de ces algorithmes est que leur complexité est quasi-linéaire puisque si n est la taille en caractères des données et k le nombre de motifs extraits, la complexité est de $O(n) + k$. Ensuite, il faut simplement retenir que ces descripteurs visent la non-redondance et que nous allons en fait chercher les redondances dans les motifs non redondants.

Soit $\mathcal{R}$ l'ensemble des répétitions maximales (ou *motifs*) détectées sur n chaînes de caractères $\mathcal{S} = \{\mathcal{S}_0, \ldots, \mathcal{S}_{n-1}\}$, avec $|\mathcal{S}| = \sum_{i=1}^{n} size(\mathcal{S}_i)$. L'ensemble de motifs $\mathcal{R}$ est calculé sur la concaténation de chaque chaîne $\mathcal{S}_i$: $c(\mathcal{S}) = \mathcal{S}_0\$_{n-1} \ldots \mathcal{S}_{n-1}\$_0$. Les répétitions maximales du deuxième ordre $\mathcal{R}^2$ dans $\mathcal{S}$ sont calculées sur la concaténation de l'ensemble des m motifs de $\mathcal{R}$, $c(\mathcal{R}) = \mathcal{R}_0\$_{m-1} \ldots \mathcal{R}_{m-1}\$_0$ avec $m < |\mathcal{S}|$, chaque $\mathcal{R}_i$ étant un motif de $\mathcal{S}$. L'ensemble des motifs du $n^{\text{ème}}$ ordre est noté $\mathcal{R}^n$.

Par exemple, soit $c(\mathcal{S}) = \texttt{HATTIV}\$_1\texttt{ATTIAA}\$_0$. L'ensemble de motifs $\mathcal{R}$ sur $c(\mathcal{S})$ est constitué des motifs suivants : $\mathcal{R} = \{\texttt{ATTI}, \texttt{A}, \texttt{T}\}$. L'ensemble des répétitions du deuxième ordre $\mathcal{R}^2$ est composé des motifs `T` (deux fois dans `ATTI` et une fois dans `T`) et `A` (une fois dans `ATTI` et une fois dans `A`).

L'ensemble des motifs dans $\mathcal{R}^n$ est donc un sous-ensemble de $\mathcal{R}^{n-1}$.

La Figure 2.7 représente le nombre de motifs différents en fonction de l'ordre des répétitions maximales. Parce que $\mathcal{R}^n \subset \mathcal{R}^{n-1} \iff |\mathcal{R}^n| < |\mathcal{R}^{n-1}|$, le nombre de motifs décroît plus l'ordre est important quelque soit le corpus. Le nombre de motifs tombe à 0 pour $n = 26$ (EBG-40, LIB-40 et MIXT-80) et $n = 25$ (MIXT-40).

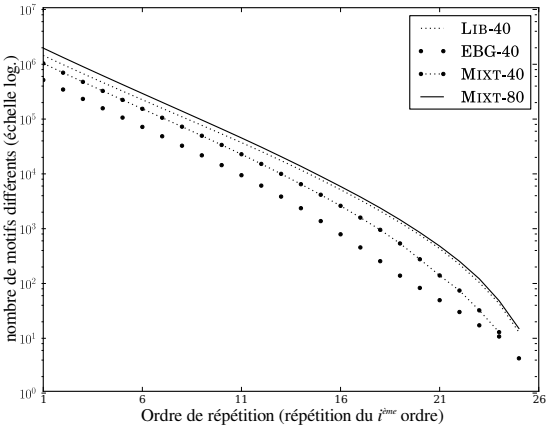


FIGURE 2.7 – Évolution du nombre de motifs (échelle logarithmique) en fonction de l'ordre des répétitions maximales (LIB-40, EBG-40, MIXT-40 et MIXT-80).

Le calcul des répétitions maximales du deuxième ordre s'effectue avec le même algorithme que celui permettant le calcul des répétitions maximales, la complexité en temps pour l'énumération de ces motifs est donc faites en $O(n)$ car la taille des ensembles de motifs décroît à chaque itération (i.e. plus l'ordre est important moins il y a des motifs puisque certains ne sont plus répétés). Par ailleurs, le calcul des répétitions maximales du deuxième ordre est utilisé pour détecter les motifs inclus dans d'autres motifs de sorte que le temps de calcul baisse graduellement.

2.3.3 Comparer *n*-grammes de caractères et répétitions maximales

Comme décrit précédemment, les répétitions maximales sont une manière condensée de représenter toutes les sous-chaînes de caractères d'un corpus. En d'autres termes, pour une valeur donnée n, l'ensemble des répétitions maximales de taille n est un sous-ensemble des n-grammes de caractères d'un corpus (et de la même manière dans le cas de chaînes de caractères de longueurs variables : les répétitions maximales ayant une longueur comprise entre $[min, max]$ et les $[min, max]$-grammes de caractères). Les sous-chaînes qui ne sont pas des répétitions maximales sont celles qui sont seulement maximales à gauche ou à droite (ou ni l'un ni l'autre, donc répétées mais non-maximales) ou des *hapax legomena*. Dans une tâche de classification supervisée, les *hapax legomena* du corpus complet n'ont alors pas d'impact sur les résultats car par définition, ces *hapax* apparaissent seulement une fois dans le corpus de test ou une fois dans le corpus d'entrainement.

Si les n-grammes de caractères peuvent capturer différentes caractéristiques sous-jacentes en fonction du choix du paramètre n (caractéristiques lexicales, contextuelles ou thématiques [Sun et al., 2012]), ces n-grammes encodent indirectement des sous-chaînes de taille supérieure à n. Par exemple, considérons `abcdef` un motif extrait d'un corpus. Ses caractères ne sont pas inclus dans d'autres sous-chaînes du corpus, parce que `abcdef` est maximale. Alors, chaque sous-chaîne de `abcdef` possède le même nombre d'occurrences que `abcdef` ($freq(abcdef) = k$). La Figure 2.8

représente comment l'usage de 3-grammes de caractères est affecté par le nombre d'occurrences du motif, et donc comment la représentation vectorielle du corpus contenant ce motif est elle aussi affectée.

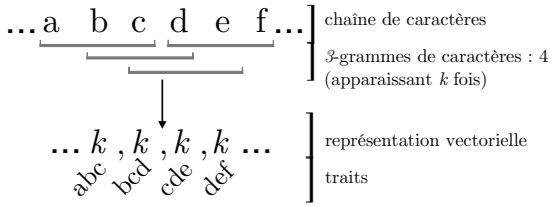


FIGURE 2.8 – Sous-chaînes d'un motif et leur influence sur la représentation vectorielle du corpus.

Ainsi, exploiter seulement les répétitions maximales de taille 3 ne permettra pas dans cet exemple d'exploiter des sous-chaînes ayant le même nombre d'occurrences que le motif `abcdef`. Considérer certaines longueurs affectera la représentation vectorielle fondées sur les occurrences des chaînesi. Pour prendre en compte ces influences, nous définissons une fonction de pondération $w_{2^{nd}}(trait)$ qui exploite les sous-chaînes qu'un trait encapsule. $w_{2^{nd}}(trait) = pot(trait) - sub(trait)$, où $pot(trait)$ correspond au nombre de sous-chaînes potentielles à l'intérieur d'un trait et $sub(trait)$ correspond au nombre de motifs de deuxième ordre qui apparaissent à l'intérieur du trait et ailleurs dans le corpus :

— cette pondération est donc fonction de la longueur du trait ;
— pour deux observables de même longueur, le facteur de pondération peut être différent.

Si un trait varie d'un caractère par rapport à un autre, cette fonction de pondération minimisera cet ajout : les produits du facteur de pondération et de l'effectif des motifs utilisés comme caractéristiques seront proches. À l'inverse, un trait qui est plus qu'une légère variation d'un autre trait sera considéré comme « consistant » et donc aura plus d'importance. En d'autres termes, là où les n-grammes pondèrent naturellement les caractéristiques grâce à la redondance, cette fonction de pondération exploite la redondance au sein de ces caractéristiques.

Deux ensembles de traits de longueur variable sont donc envisagés : les n-grammes de caractères et les répétitions maximales (motifs). Les motifs sont considérés de trois façons différentes : pondérés par leur longueur (w_{len}), par les répétitions maximales du 2ème ordre ($w_{2^{nd}}$) et sans pondération. Une validation croisée, *stratified 10-fold cross validation*, est utilisée pour valider les performances du système pour chaque trait, chaque strate (*fold*) contient la même proportion d'auteurs. L'évaluation est réalisée par une exactitude, le nombre de textes correctement classés divisé par le nombre total de textes classés puis ramené à un pourcentage.

2.3.4 Impact de la longueur des sous-chaînes et des motifs

Le score d'AA est calculé sur trois corpus : EBG-40 (Figure 2.9), LIB-40 (Figure 2.10) et MIXT-80 (Figure 2.11). Chaque figure est constituée de quatre matrices pour chaque trait : les répétitions maximales (*motifs*), les n-grammes, les répétitions maximales pondérées par leur longueur ($motifs_{len}$) et les répétitions maximales pondérées par les répétitions maximales du 2ème ordre ($motifs_{2^{nd}}$). Le score écrit aux coordonnées (i, j) de chaque matrice est donné par l'exploitation de traits de longueur comprise entre i et j.

Les traits peuvent être ordonnés en fonction de leur performance sur les corpus : $motifs \leq$

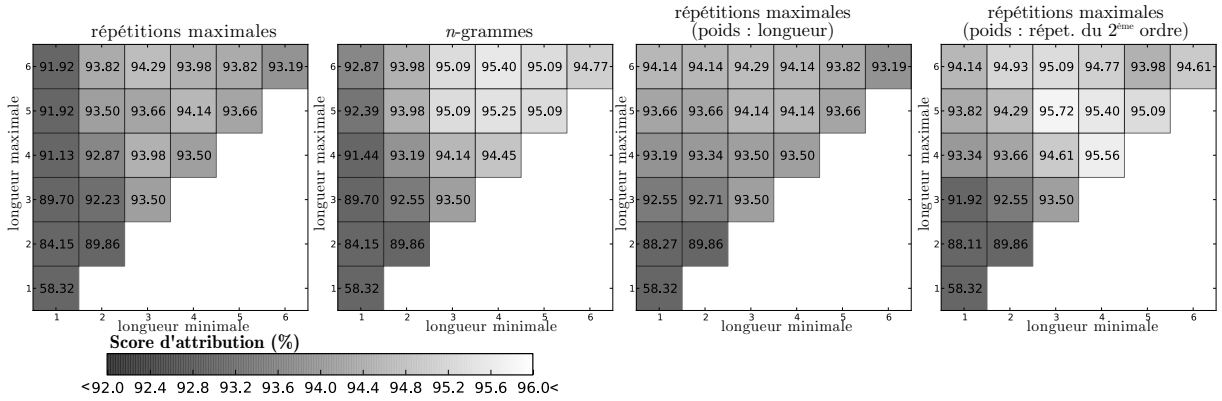


FIGURE 2.9 – Score d'attribution sur le corpus EBG-40.

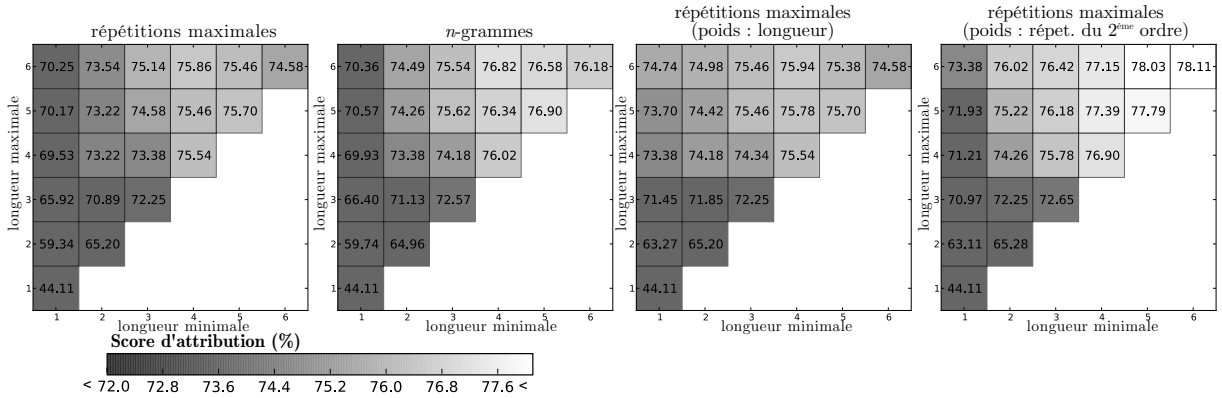


FIGURE 2.10 – Score d'attribution sur le corpus LIB-40.

$motifs_{len}$ < n-grammes < $motifs_{2^{nd}}$. Le fait que $motifs$ < n-grammes montre l'effet positif de la redondance des traits. Les scores sur les diagonales des matrices (là où les traits sont de taille fixe) utilisant $motifs$ et $motifs_{len}$ sont identiques car chaque trait est pondéré par le même facteur avec la fonction de pondération w_{len}. Les scores d'attributions sur le corpus EBG sont élevés. Cela s'explique par le lien fort entre auteur et contenu thématique (pour un auteur, chacun de ses textes appartient à la même thématique comme économie, arts ou sport). La tâche est plus difficile sur le corpus LIB car contrairement au corpus EBG, chaque auteur a été sélectionné si son ensemble de textes est constitué de textes de différents thèmes.

Le score d'attribution sur les trois corpus a aussi été calculé en utilisant $motifs_{2^{nd}}$ sans contrainte (tous les motifs sont considérés quelle que soit leur longueur) avec les scores suivants : 66,40% sur le corpus EBG-40, 48,20% sur le corpus LIB-40 et 54,21% sur le corpus MIXT-80. Ceci souligne la nécessité de la présélection des traits parmi ceux disponibles. Les meilleurs paramètres de longueur sont sélectionnés en calculant la moyenne des scores d'attribution sur chacune des matrices pour chaque intervalle de longueurs $[min, max]$ (Table 2.7).

La configuration $motifs_{2^{nd}}$ obtient les meilleurs résultats en utilisant le plus petit intervalle de longueurs. Les meilleurs paramètres de longueur calculés sur l'ensemble corpus ne constituent pas nécessairement le meilleur jeu de paramètres pour chaque corpus pris individuellement. Par exemple, $motifs_{2^{nd}}$ obtient de meilleurs résultats avec les paramètres $[6, 6]$ sur le corpus LIB-40 qu'avec les paramètres $[4, 5]$. Comme les motifs sont une représentation condensée des n-grammes, l'espace de traits est naturellement moindre en utilisant des motifs. Les expériences

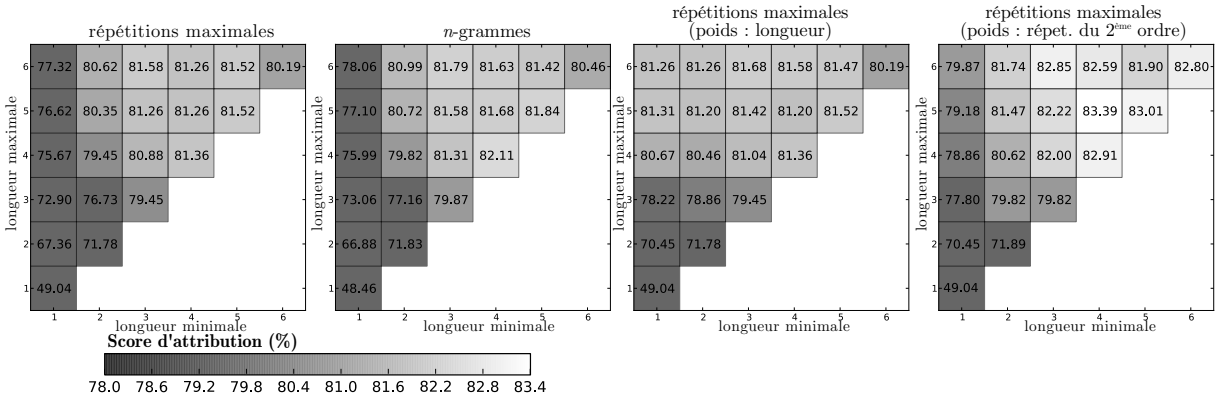


FIGURE 2.11 – Score d'attribution sur le corpus MIXT-80.

	meilleurs paramètres de longueur $[min, max]$	score moyen
n-grammes	$[4, 6]$	84,61%
$motifs$	$[4, 6]$	83,69%
$motifs_{len}$	$[4, 6]$	83,88%
$motifs_{2nd}$	$[4, 5]$	**85,39**%

TABLE 2.7 – Meilleurs paramètres en fonction du score moyen sur les corpus LIB-40, EBG-40 et MIXT-80.

montrent de meilleurs résultats avec l'utilisation de traits de longueur variable plutôt que fixe. Cependant, utiliser le plus grand intervalle de longueur lors de la sélection des traits n'est pas systématiquement un choix pertinent au regard des résultats. Par exemple, une différence de 4,01 points de pourcentage (ci-après `pp`) est observable entre l'intervalle $[1, 6]$ et l'intervalle optimal $[4, 5]$ sur les résultats du corpus LIB-40 en utilisant $motifs_{2nd}$ (Figure 2.10). Considérer le plus grand ensemble de traits disponibles permet peut être de capturer des caractéristiques utiles à cette tâche, mais surtout de capturer des traits dégradant sensiblement les résultats.

2.3.5 Influence du nombre de traits et du nombre d'auteurs

En choisissant les meilleurs paramètres pour chaque type de trait (Table 2.7), les expériences suivantes décrivent l'évolution du score d'attribution en fonction du nombre d'auteurs (Figure 2.12). Il s'agit d'illustrer à quel point les paramètres optimaux pour un certain nombre de classes seront sensibles à l'évolution de ce nombre.

Pour chaque corpus et chaque trait, le score d'attribution décroît quand le nombre d'auteurs augmente. Les résultats sont supérieurs en utilisant $motifs_{2nd}$, à l'exception des corpus EBG-30 et EBG-40. Les pires résultats sont obtenus sur le corpus LIB pour lequel le score décroît de 92,04% à 77,39% (de 89,60% à 76,82% en utilisant des n-grammes). Pondérer les motifs en fonction de la longueur ($motifs_{len}$) n'améliore pas le score de manière significative par rapport à l'utilisation des motifs sans pondération. Les évolutions du nombre de traits sont données sur la Figure 2.13. Le nombre de traits correspond à la moyenne des tailles des vecteurs représentant les textes sur chacun des dix échantillons de la validation croisée. Les résultats sont différents de ceux de la Figure 2.7 (page 48). En effet nous montrons ici le nombre de traits utilisés pour la classification et non tous ceux qui sont calculables sur les corpus.

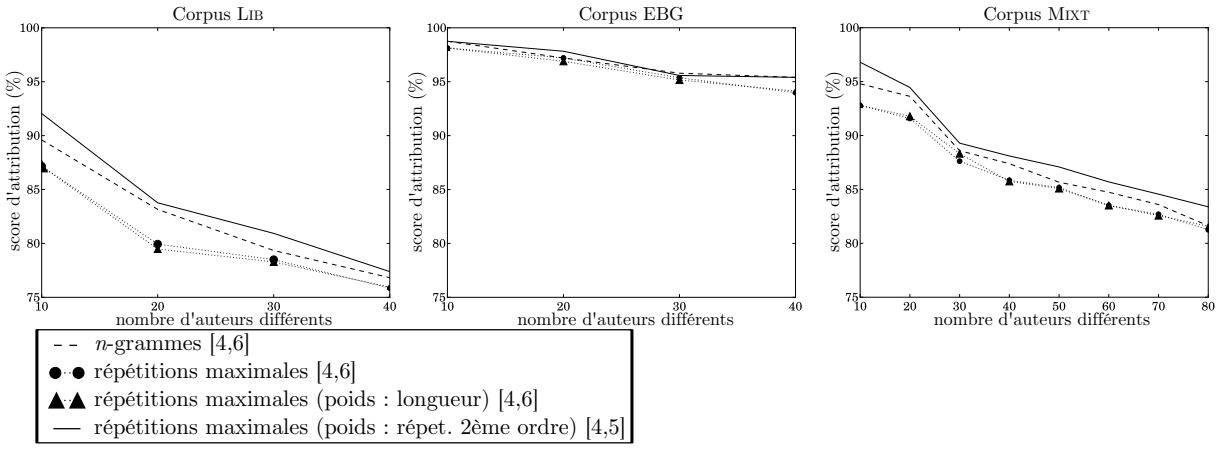


Figure 2.12 – Évolution du score d'attribution en fonction du nombre d'auteurs.

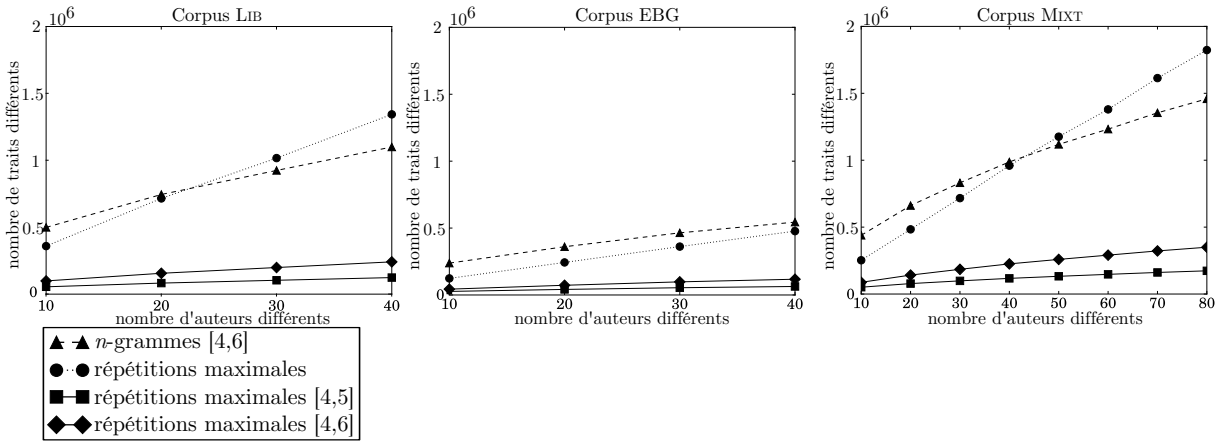


Figure 2.13 – Évolution du nombre de traits en fonction du nombre d'auteurs.

Utiliser des motifs de taille $[4, 5]$ réduit significativement le nombre de traits par rapport à l'usage de n-grammes de taille $[4, 6]$ ou de motifs sans contrainte de longueur. Le nombre de motifs augmente linéairement en fonction du nombre d'auteurs (ou en fonction de la taille du corpus). Le nombre de n-grammes de taille $[4, 6]$ est plus élevé que le nombre de motifs pour un faible nombre d'auteurs. Il est toutefois plus faible dès lors que le nombre d'auteurs augmente, du fait de la distribution sous-linéaire des n-grammes de taille $[4, 6]$. La taille de l'espace de recherche des motifs de taille $[4, 5]$ est adaptée quand la taille des données croît.

Le corpus Mixt est composé des corpus Lib en français et EBG en anglais, les deux langues partageant des traits en commun de par leurs origines communes. Dans le cadre d'une analyse multilingue, l'utilisation de deux langues proches est adaptée en AA. Cette expérience permet de vérifier si le trait choisi est efficace dans un corpus où chaque texte n'est pas écrit dans la même langue mais où les langues partagent des traits en commun. La Table 2.8 présente les scores d'attribution sur les deux corpus monolingues, Lib et Ebg, pris indépendamment puis intriqués au sein du même corpus Mixt. Le but est d'analyser comment les traits influent sur le traitement lorsque plusieurs langues sont traitées en même temps.

Les résultats sont similaires, que l'on utilise le corpus multilingue ou les corpus monolingues de manière indépendante. Quelques améliorations peuvent être notées en utilisant $motifs_{2^{nd}}$,

n-grammes (longueur $[4, 6]$)

nb. d'auteurs	EBG	EBG issu de MIXT	LIB	LIB issu de MIXT
10	**98,75**%	**98,75**%	89,60%	**91,13**%
20	**97,20**%	96,89%	**83,15**%	82,69%
30	**95,79**%	94,85%	**79,34**%	78,65%
40	**95,40**%	94,10%	**76,82**%	75,03%

$motifs_{2^{nd}}$ (longueur $[4, 5]$)

nb. d'auteurs	EBG	EBG issu de MIXT	LIB	LIB issu de MIXT
10	**98,75**%	**98,75**%	92,01%	**92,35**%
20	**97,83**%	97,52%	**83,77**%	83,46%
30	95,59%	**96,84**%	**80,93**%	80,08%
40	**95,40**%	95,09%	77,38%	**77,47**%

TABLE 2.8 – Score d'attribution sur les corpus LIB et EBG indépendamment ou issus du traitement du corpus MIXT.

où les résultats sont plus souvent meilleurs quand les deux corpus EBG et LIB sont traités ensemble. En utilisant des n-grammes de taille variable sur les corpus multilingue et monolingue, la différence de résultats augmente avec le nombre d'auteurs : la différence de score d'attribution est de -1,30 pp) sur le corpus LIB et de -1,79 pp) sur le corpus EBG. À l'inverse, l'approche fondée sur les motifs est plus adaptée à la problématique multilingue (-0,31 pp) sur le corpus LIB et +0,09 pp) sur le corpus EBG).

2.3.6 Commentaires sur les résultats de l'attribution d'auteur

Nous avons proposé une alternative efficace aux approches fondées sur les n-grammes de tailles variables via les *répétitions maximales*. Ces répétitions surpassent les approches classiques fondées sur les sous-chaînes sur deux aspects. Premièrement, les *répétitions maximales* sont, par essence et à la différence des n-grammes, non-redondantes. En effet, leur caractère maximal évite la détection et l'utilisation de nombreuses occurrences de sous-chaînes équivalentes dans le corpus. Cela réduit considérablement le nombre de traits donc l'espace de recherche et nous préconisons leur usage conjointement à des méthodes de sélection de sous-espaces de recherche (Algorithme Génétique, Recuit Simulé, sélection de Corrélations Caractéristiques, Gain d'Information). Nous avons aussi souligné dans que cette redondance avait un effet positif pour la qualité cette tâche de classification. Cette redondance a été exploitée pour proposer une méthode de pondération des traits en fonction de leur structuration interne. Deuxièmement, avec les répétitions maximales de deuxième ordre, nous réduisons davantage l'espace de recherche des traits et nous proposons une nouvelle façon d'améliorer la précision de la prédiction en AA. L'hypothèse qu'une longue chaîne répétée est plus importante si elle ne contient pas trop de sous-répétitions, est validée. Nos premières expériences montrent que l'usage des répétitions maximales permet de traiter plus aisément des flux de textes dans des langues différentes.

2.4 Conclusion intermédiaire : pourquoi et comment faire varier les observables ?

Faire varier les observables utilisés permet tout d'abord de sortir de recettes toutes faites où la même chaîne de traitement va s'appliquer à toute tâche dès lors que l'on a pu transformer un problème donné en un problème de classification. Varier les observables permet de s'adapter aux données et à la tâche que ce soit pour chercher des descripteurs alternatifs à des fins d'interprétation (c'était le cas dans le travail sur le corpus Dumas-Féval) ou pour chercher les descripteurs qui optimisent les résultats obtenus par un algorithme de classification.

Il y a une tendance certaine dans notre discipline à favoriser l'utilisation de chaînes de traitement plus sophistiquées, et notamment d'algorithmes de classification plus complexes et plus gourmands en temps de calcul. Ceci se fait au détriment, je crois, de la recherche des éléments essentiels d'une approche de classification qui peuvent, dans certains cas, être le mode de représentation des données sous forme de redescription ou de tokenisation non supervisée. Je ne prétend pas ici épuiser la liste des possibilités de variation les caractéristiques ni non plus avoir présenté la meilleure méthode dans chacun de ces cas. Par contre, ce qui me semble évident c'est que la multiplicité des représentations possibles, et les facultés d'adaptation à la variation que cela offre, constituent un éventail de solutions que notre communauté explore trop rarement.

Parmi les cas de variation qu'il me semble intéressant d'étudier par ce biais figure ce que l'on nomme parfois les données bruitées générées par des utilisateurs (*noisy user generated content*), au rang desquels figurent en bonne place les données des réseaux sociaux et en particulier de Twitter. Nous allons sortir ici du cas de l'attribution d'auteur, afin de montre la grande généricité des modèles en caractères. Nous avons montré dans une série de travaux réalisés avec Davide Buscaldi, Aude Grezka et Joseph Le Roux du LIPN que l'utilisation de motifs en caractères permettait de mettre en place des méthodes simples, efficaces et peu coûteuses en données d'entraînement pour des tâches de détection d'ironie [Buscaldi et al., 2017], des tâches de classification binaire ou encore de détection de polarité à 4 classes [Buscaldi et al., 2018]. Ces deux derniers cas concernaient des corpus en français, mais cette remarque vaut aussi pour d'autres langues. Nous avons montré aussi avec mon collègue Dhaou Ghoul (ATER à Sorbonne Université de 2018 à 2020) dans [Ghoul and Lejeune, 2021] que ces caractéristiques offraient des représentations très efficaces, cette fois pour des tâches de détection, dans des corpus en langue arabe, de sarcasme et de sentiment. Une des bonnes propriétés de ces représentations est, au contraire des représentations en mots par exemple, la capacité de tenir compte de la variation morphologique de façon endogène. Ceci nous avait aussi permis de proposer des modèles d'identification de dialectes de la langue arabe avec des résultats tout à fait satisfaisants [Ghoul and Lejeune, 2019, Ghoul and Lejeune, 2020]. Pour finir avec une application plus « linguistique », dans le cadre de mon post-doctorat avec Emmanuel Cartier en 2017 au LIPN, nous avons démontré que les motifs en caractères, étaient très efficaces pour améliorer un système de détection de néologismes en filtrant des candidats en fonction des chaînes de caractères qui les composent et des introducteurs figurant dans leurs contextes gauches et droits [Lejeune and Cartier, 2017]. Un des intérêts des méthodes en caractères, qui amène à réaliser ce que j'appellerais une tokenisation non supervisée, est de repérer en ligne des caractéristiques utiles à la classification en se détachant du schéma classique où l'on « nettoie/standardise » les textes pour tokeniser.

Après ce travail sur les observables, se pose la question, à l'autre extrémité du spectre, du cadre d'analyse, de l'observatoire dans lequel les observables sont exploités et de la manière dont il est possible de mettre à profit des caractéristiques des documents. Je parle ici des documents au sens large du terme, de leur organisation propre, du genre textuel auxquels on peut les rattacher

et aussi des corpus au sein desquels ils prennent leur sens. Le prochain chapitre sera consacré à la présentation d'une expérience visant à exploiter des propriétés spécifiques à un genre textuel pour réaliser une tâche de classification.

Chapitre 3

Peut-on exploiter la structure des documents pour identifier des classes particulières de mots ?

Sommaire

3.1 Terminologie et Stylométrie

ăDans le cadre de ma participation à l'ANR Termith TERMinologie et Indexation de Textes en sciences Humaine[45]), j'ai pu m'intéresser à la question de la valeur terminologique (non ambigüe) ou non des différentes occurrences d'un candidat terme. Ceci a été une expérience très enrichissante puisque je n'avais jamais eu l'occasion de m'intéresser à proprement parler à la problématique du lexique. Ce travail est un travail collectif qui a impliqué Béatrice Daille du LINA et nos collègues de l'ATILF (Evelyne Jacquey) et du LORIA (Luis Felipe Melo et Yannick Toussaint) dans notre article publié à LREC en 2016 [Daille et al., 2016]. J'ai pu approfondir un certain nombre des questions posées dans ce travail grâce à mes élèves de Master 2 Langue et informatique de Sorbonne Université, dans le cadre du cours « Terminologie et Stylométrie ».

Si je reprends le triptyque Méthodes/Données/Objectifs à travers lequel je structure et compare mes travaux de recherche, on a ici d'un côté une méthode (la stylométrie) et de l'autre côté

45. `https://anr.fr/Projet-ANR-12-CORD-0029` consulté le 2 octobre 2023

un objectif (la terminologie). Et c'est bien dans cette intention là que j'ai compris le titre de ce cours de M2. Mais à y regarder de plus près, la terminologie peut aussi être vue comme une donnée. En effet, connaître les termes d'un texte est une clé pour accéder à la compréhension de ce texte. La stylométrie elle-même peut être un objectif puisqu'observer la répartition des occurrences terminologiques dans des documents permet de faire des hypothèses sur les procédés par lesquels, dans un certain genre textuel, les idées sont formulées et comment la structuration du document selon les « dans les règles de l'art »facilite la transmission des informations. Ici l'hypothèse forte est de dire que les auteurs d'articles scientifiques font des efforts importants pour assurer la qualité de la transmission du message en utilisant astucieusement de la terminologie. Trop de terminologie et l'on est dans le cadre du discours spécieux, cryptique où l'on masque la faiblesse du discours derrière un jargonnage, une pyrotechnie terminologique [46]. Le manque de terminologie indique au contraire un manque de précision, du flou, voire une méconnaissance des objets étudiés. Ici nous nous plaçons dans une configuration de TAL où l'objet central de la terminologie est le terme. Au contraire, dans la distinction donnée par [LHomme, 2005], d'une configuration où l'objet central de la terminologie est le concept. Sachant que la recherche présentée ici se rattache à l'informatique appliquée, la motivation est avant tout l'indexation, et non l'analyse automatique de l'épistémologie d'une science à travers sa terminologie par exemple. La terminologie telle que traitée ici peut être vue comme l'ensemble du vocabulaire d'une science, composé en majorité, sinon totalement, de vocables possédant une signification précise (et unique) dans ce domaine. Il s'agit sans doute alors d'une vision réductrice de la notion de terminologie [Neveu, 2008] mais qui se veut adaptée à ce que l'on peut raisonnablement automatiser (voir [Kageura, 2012] pour une discussion sur les différentes définitions de terme et terminologie).

3.1.1 Le terme et ses manifestations

La question qui nous avait intéressés ici était de savoir comment l'on pouvait affiner les résultats de l'extraction terminologique automatique. Un des défauts connus des extracteurs terminologiques est que la décision est souvent globale au sein d'un corpus donné [Camacho-Collados et al., 2014], de sorte que si une forme est un terme une fois, alors toutes ses autres occurrences seront terminologiques elles aussi. Or, ceci ne résiste pas à l'analyse. Il est très facile de voir qu'une forme peut être un terme dans un domaine de spécialité et non dans un autre. Mais comment délimiter un domaine de spécialité ? Est-ce que l'on considère des domaines à très gros grain, par exemple la sociologie ? Ou bien est-ce que l'on descend encore d'un niveau pour examiner séparément la sociologie qualitative et quantitative, la sociologie de la famille et la sociologie économique ? C'est difficile à dire, sans parler bien sûr du problème consistant à imaginer une frontière étanche entre des domaines, par exemple si l'on prenait l'exemple du TAL et de la linguistique. On peut donc considérer que l'extraction terminologique consiste à opérer une sélection automatique dont la qualité dépend de l'homogénéité du corpus. À partir de là, on voit qu'il peut être difficile de séparer la question de l'évaluation elle-même, de la qualité de l'évaluation et celle de la granularité que l'on choisit comme limite pour regrouper ou non les sous-disciplines. Un extracteur qui serait en mesure de diagnostiquer le domaine pour adapter ses modalités d'analyse serait certainement un progrès. C'est un lien que l'on peut faire aussi avec la détection de néologismes où, si l'on ne s'intéresse pas qu'aux néologismes de forme, la connaissance experte du domaine est essentiel.

46. C'est une expression que j'utilise parfois pour parler de jargon, et c'est une sorte de mise en abîme puisque cette expression est elle-même une pyrotechnie formelle pour dire jargon.

Afin d'illustrer ceci, je présente un exemple (tiré de l'infolettre de France terme [47]) utilisé pour un examen d'un cours sur les néologismes [48]. L'exercice consistait à identifier (manuellement) dans un texte des termes appartenant à un même domaine technique.

> Par cette nuit si froide, le courageux furet nageait dans la piscine, éclairé par une chandelle. Le fantôme en profita pour fouiller dans la boîte à gants, mais il n'y trouva qu'un **crayon**, une chaussette et une aiguille. Et dire qu'il avait manqué de se transformer en glaçon pour un si maigre butin.

L'exercice n'est pas aisé et l'on voit bien que l'identification du domaine est un préalable indispensable à l'identification des termes et donc au repérage du sens activé. En l'espèce il s'agissait du domaine du nucléaire où « le crayon » désigne le tube en métal qui contient le combustible nucléaire et « le furet » un autre type de tube permettant d'irradier un échantillon à des fins expérimentales. On voit bien ici que considérer, sans tenir compte du domaine, que toutes les occurrences de « crayon » et de « furet » sont des termes serait une erreur. De même, n'en considérer aucune comme terminologique serait erroné. Par ailleurs, ces formes sont aussi des termes dans d'autres domaines [49]. Cette identification du domaine concerné sort du cadre de la recherche présentée ici mais nous permet de montrer l'intérêt de ne pas dissocier la catégorisation d'un élément textuel du corpus, de l'éco-système auquel il appartient.

C'est dans ce contexte que se situe la recherche que je présente ici. La question que l'on se pose est de savoir comment on peut raffiner l'extraction terminologique en s'interrogeant sur les cas où le terme candidat n'est pas utilisé en tant que terme, les cas où il n'active pas son sens terminologique. Afin de faciliter la comparaison entre différentes méthodes, nous avons choisi de voir cette tâche comme une tâche de classification, avec tout le caractère réducteur que cela implique notamment puisque le diagnostic d'ambigüité est binaire [50]. Étant donné toutes les occurrences d'un terme candidat, il s'agit donc ici de distinguer celles qui sont véritablement terminologiques, le sens en domaine de spécialité qui est activé, des autres qui sont non-terminologiques (et potentiellement ambigües). Marie-Claude L'Homme [L'Homme, 2004] et d'autres terminologues ont bien insisté sur cet aspect non universel, et aussi transitoire, du caractère terminologique d'un vocable. La plupart des candidats termes, sauf sans doute ceux dont l'usage est trop localisé ou trop peu fréquent pour leur donner un caractère polysémique, auront en effet des occurrences qui ne sont pas terminologiques. Différentes raisons peuvent bien sûr expliquer cette situation. Il y a la question des limites entre domaines par exemple. Il peut aussi y avoir le mésusage d'un terme, c'est-à-dire l'utilisation comme un mot de la langue courante d'une forme qui est attestée comme terme dans le domaine où on l'emploie. Assez peu de travaux se sont intéressés à ces problèmes particuliers, bien que le travail sur la désambiguïsation lexicale soit particulièrement actif notamment depuis [Navigli, 2009] et à l'origine de projets de grande envergure tels que BABELNET [51]. Pourtant, on sait que ceci est très important notamment pour des tâches de recherche d'information dans des contextes où les besoins sont spécifiques, par exemple les moteurs de recherche [Maeda et al., 2000] ou encore les agents conversationnels [Chen et al., 2006].

47. `http://www.culture.fr/Ressources/FranceTerme/Infolettres-parues` consulté le 2 octobre 2023
48. Je remercie chaleureusement Françoise Guérin qui partage ce cours avec moi et a trouvé ce texte.
49. En médecine, le crayon peut désigner un dispositif destiné à détruire les verrues, en plomberie le furet est une tige métallique destinée à déboucher des canalisations.
50. Même si nous aurions pu imaginer utiliser un score entre 0 et 1 par exemple.
51. `https://babelnet.org/` consulté le 2 octobre 2023

3.1.2 L'ambiguïté en TAL

La désambiguïsation sémantique est un verrou important pour le Traitement Automatique des Langues. Ce problème a souvent été abordé dans une perspective de résolution. Étant donné les sens possibles d'une unité lexicale (mot ou groupe de mots) en contexte, il s'agit de déterminer lequel de ces sens est activé pour une occurrence particulière. Ce champ de recherches a été principalement investi à la suite des travaux de Yarowsky [Yarowsky, 1992, Yarowsky, 1995] bien que les recherches dans le domaine soient bien plus anciennes avec notamment les travaux de Lesk et l'algorithme éponyme [Lesk, 1986].

Dans ce travail, nous avons abordé l'ambiguïté sémantique d'un terme, simple ou complexe, en domaine de spécialité. Nous nous intéressons au diagnostic d'ambiguïté, c'est-à-dire que nous cherchons à déterminer si, dans un contexte particulier, le sens d'un terme est difficile à appréhender. Par exemple, si l'on a le mot « classe » dans un texte relevant de la linguistique, il s'agit de savoir si l'on a un emploi terminologique (biunivoque) ou non (susceptible d'être ambigu). Pour l'unité lexicale non ambiguë, le choix du sens à activer est trivial : si son emploi relève d'un domaine de spécialité, il s'agit d'un cas de monosémie. Autrement dit, le nombre d'inférences à effectuer pour déterminer le sens est minimal pour le récepteur du texte [Sperber and Wilson, 1998, Wilson, 2006].

Si nous nous replaçons dans le domaine de la désambiguïsation sémantique, cela signifie que parmi tous les sens possibles de l'unité lexicale considérée, c'est le plus terminologique qui doit être activé. Détecter les cas d'emploi terminologique permet donc de guider le processus d'analyse. Nous pensons que le diagnostic d'ambiguïté permet de limiter la combinatoire des sens à explorer. Identifier s'il y a une réelle ambiguïté favorise alors la résolution de cette ambiguïté en permettant de savoir si l'on peut ou non se référer au domaine de spécialité concerné. Ce peut aussi être un indice pour déterminer quels sont les mots-clés pertinents pour décrire un document.

La désambiguïsation sémantique (*Word Sense Disambiguation*) vise alors à déterminer pour chaque occurrence d'un terme le sens le plus approprié parmi tous ceux que ce terme peut revêtir [52]. Le sens de l'occurrence est donc une étiquette qui selon les ressources exploitées peut revêtir différentes formes : une définition exprimée en langue naturelle, la position dans une ressource de type ontologie ou encore les traductions possibles de ce terme dans différentes langues. Résoudre l'ambiguïté des termes candidats d'un texte permet par exemple d'améliorer les performances des systèmes de traduction automatique. Pour mesurer l'intérêt de cette classification, on peut imaginer un exemple comme Linguee [53] où l'on voit en contexte les différentes acceptions d'une unité lexicale. Pouvoir classer les exemples selon que l'emploi est terminologique ou pas et de quel domaine il s'agit pourrait améliorer les résultats de la désambiguïsation.

Résoudre l'ambiguïté La désambiguïsation sémantique a suivi l'évolution du TAL en général. Les travaux répertoriés les plus anciens [Bar-Hillel, 1960, Small and Rieger, 1982] ont traité la désambiguïsation sémantique comme un problème de sélection que l'on pourrait résoudre à l'aide de systèmes experts. L'approche la plus emblématique du domaine est due comme nous l'avons dit précédemment à [Lesk, 1986] qui a exploité les premiers dictionnaires électroniques à large couverture afin d'utiliser les relations entre les définitions pour raffiner les connaissances sémantiques sur chaque mot. Puis, c'est l'apprentissage automatique qui est devenu en vogue [Gale et al., 1992] ce qui a permis d'améliorer considérablement les résultats et a autorisé l'extension

52. De manière plus ambitieuse, ou plus réaliste c'est selon, on pourrait aussi chercher à ne pas répondre dans les cas où il y a un doute sur le sens qui est activé, ce point sera abordé dans les questions sur l'évaluation.

53. `https://www.linguee.fr/` consulté le 2 octobre 2023

vers de nouveaux domaines et des langues autres que l'anglais. Ensuite, d'autres recherches ont amené de nouvelles problématiques pour le domaine comme la limitation des ressources impliquées [de Loupy and El-Bèze, 2000, Jin et al., 2009] ou l'interprétabilité des modèles générés [Navigli and Velardi, 2005]. Les traits exploités dans ces travaux et leurs successeurs sont principalement de deux ordres : classes sémantiques et étiquettes morpho-syntaxiques. Sont considérés les termes à désambiguïser ainsi que leurs voisins selon une certaine fenêtre (comportant les n termes précédents et/ou suivants). Une des principales contraintes rencontrées est la largeur de cette fenêtre, la complexité de calcul augmente avec la taille de la fenêtre. Avec une fenêtre de taille n et en moyenne m sens à observer pour chaque terme, on a une complexité exponentielle en la largeur de la fenêtre. Le choix de cette largeur ne peut donc être qu'un compromis entre efficacité et temps de calcul.

Diagnostiquer l'ambiguïté pour faciliter la résolution ? Donner un diagnostic d'ambiguïté permet, par exemple, de limiter le nombre de combinaisons à envisager. Chaque mot non ambigu permet de réduire la combinatoire pour le calcul du sens de ses voisins ou encore d'élargir à moindre coût le contexte exploré pour améliorer les résultats. Par ailleurs, le diagnostic d'ambiguïté permet d'identifier plus finement les candidats qui sont véritablement des termes pour le document considéré. Diagnostiquer l'ambiguïté revient alors à distinguer en contexte les emplois terminologiques des emplois non-terminologiques. Dans l'article dont est tiré la recherche présentée ici nous donnions pour exemple des utilisations de *aspect*, tout d'abord dans son sens terminologique en linguistique puis dans l'un de ses ses communs :

(I) *L'* `aspect` *est une catégorie qui reflète le déroulement interne d'un procès* [Cothière-Robert, 2007]

(II) Ce dernier `aspect` *est primordial* [El-Khoury, 2007]

On voit bien ici que l'on a dans un cas une utilisation proprement terminologique et dans l'autre une utilisation de la langue commune. La première occurrence est opaque pour qui ne connaît pas la terminologie linguistique, la seconde est accessible au locuteur de la langue d'usage. Bien entendu, l'ambiguïté concerne aussi les unités multi-mots, notamment quand elles sont utilisées sous une forme réduite après une première occurrence dans leur forme pleine, dans leur forme en expansion [Jacques, 2003]. La forme réduite est généralement plus encline à l'ambiguïté que la forme étendue ou la forme complète, que l'on songe par exemple au *token* « analyse » qui est souvent ambigu, est dans notre corpus désambiguïsé dans ses formes étendues, parmi lesquelles des termes candidats (CT) tels que `analyse syntaxique`, `analyse sémantique`. La forme simple `analyse` était plus fréquemment étiquetée comme non-terminologique par les annotateurs, soit que `analyse` soit véritablement utilisé dans son sens commun soit qu'il y ait un flou difficile à lever un contexte.

Pour chaque CT recensé, le diagnostic d'ambigüité d'un terme consiste alors à étiqueter chacune de ses occurrences comme terminologique (TO) ou non-terminologique (NTO). La désambiguïsation est ici une seconde étape de filtrage fondée sur l'hypothèse suivante : les occurrences véritablement terminologiques d'un terme candidat sont un sous-ensemble de ses occurrences repérées en premier lieu. Dès lors, cette phase intervient après la phase d'Acquisition Terminologique Automatique (ATA).

3.2 Méthodes pour la désambiguïsation terminologique

Nous avons proposé trois méthodes de classification supervisée qui correspondent à autant de manières de modéliser le contexte d'apparition des différentes occurrences d'un terme afin de

déduire des régularités suffisamment robustes pour automatiser la tâche. De manière à évaluer la viabilité de ses méthodes, nous avons construit deux *baselines* inspirées de l'algorithme de Lesk.

Il s'agit de mesurer un comportement particulier des occurrences terminologiques : ce que j'ai proposé de nommer leur caractère grégaire. Ce caractère sera exploité de différentes manières : la tendance d'un terme, dans un domaine de spécialité, à figurer en compagnie d'autres termes au sein de la phrase/du paragraphe et à des positions discursives inhérentes au genre textuel (ici les articles scientifiques).

3.2.1 Baselines dérivées de l'algorithme de Lesk

Nos deux *baselines* sont des versions simplifiées du célèbre algorithme de Lesk [Lesk, 1986]. Pour obtenir la classe (terminologique ou non) d'une occurrence d'un terme candidat, on compare son contexte dans le jeu de test aux contextes observés dans le jeu d'apprentissage pour ses TO et NTO. Le voisinage est ici défini comme l'ensemble des mots qui apparaissent dans le même bloc de la structure Xml : paragraphe, titre, légende .... Plus formellement :

- soit N_{cand} le voisinage d'une occurrence dans le corpus de test ;
- soit N_{term} le voisinage de ses occurrences terminologiques (TOs) et respectivement $N_{nonterm}$ le voisinage de ses occurrences non-terminologiques ($NTOs$).
- On calcule I_{TO} l'intersection de N_{cand} et N_{term} et I_{NVP} l'intersection de N_{cand} $N_{nonterm}$.
- Si $card(I_{TO}) > card(I_{NTO})$ on classifie l'occurrence comme TO
- Et inversement si $card(I_{TO}) < card(I_{NTO})$ on classifie l'occurrence comme NTO

Cette application assez simple de l'algorithme de Lesk soulève une question importante : que faire des cas d'indécision c'est-à-dire des cas où $card(I_{TO}) = card(I_{NTO})$? Nous verrons que cette question de « prise de décision »s'est posée aussi pour d'autres méthodes et amène des discussions sur la pertinence des métriques d'évaluation y compris dans des cas simples de classification (voir Section 3.3.2 page 67). Ces cas se présentent notamment si un terme candidat n'a pas été rencontré dans le jeu d'apprentissage. En effet, nous avons fait le choix de stratifier notre corpus par document et non par instance de sorte que des instances peuvent apparaître seulement dans le jeu de test (Section 3.3.1). L'autre cas de figure est celui où les intersections sont égales mais non nulles. Voici la stratégie que nous avons utilisé pour résoudre les cas d'indécision, et qui aboutit à deux *baseline* différentes :

- Lesk Orienté Précision (POL pour *Precision Oriented Lesk*) : en cas d'indécision, l'algorithme ne prend pas de décision [54] afin de favoriser la précision ;
- Lesk Orienté Rappel (ROL pour *Recall Oriented Lesk*) : en cas d'indécision, on classifie l'occurrence comme TOs de manière à favoriser le rappel.

3.2.2 Approche fondée sur les hypothèses

Dans l'approche fondée sur les hypothèses, on considère les mots et les annotations qui leur sont liées comme des marqueurs du caractère terminologique ou non terminologique de leurs voisins. La différence entre cette approche et les *baseline* est qu'il y a une restriction sur le voisinage : on ne considérera comme pertinents que les mots (et annotations associées) qui sont

54. NB : dans l'article original, nous avions écrit « *indecisiveness cases are classified as NTOs* » ce qui est inexact puisque la méthode POL affiche un taux de décision de 68,8%. Voir tableau 3.3 page 69 conforme au tableau de l'article original.

spécifiques aux occurrences terminologiques ou aux occurrences non terminologiques. [55]

L'approche fondée sur les Hypothèses [Kuznetsov, 2004, Kuznetsov, 2001] (ci-après HB) est une méthode fondée sur l'Analyse Formelle de Concept (AFC) et l'exploitation de ces concepts au profit d'une méthode d'apprentissage automatique symbolique. Cette méthode va permettre de tirer parti des *itemsets* représentant chaque mot pour représenter les hypothèses positives et négatives à partit des occurrences terminologiques (TOs) et non-terminologiques ($NTOs$) .

En Analyse Formelle de Concepts, la structure d'un jeu de données est définie par une structure conceptuelle qui exploite les objets (ici les occurrences de CT) et leurs attributs (leurs voisinages décrits sous forme d'*itemsets*. Le triplet $K = (G, M, I)$ définit un contexte formel avec :

— G, un ensemble d'objets
— M, un ensemble d'attributs
— I, la relation binaire entre les objets et les attributs : $I \subseteq G \times M$.

En conséquence, *(g, m)* $\in I$ signifie que g possède l'attribut m. Si l'on reprend l'exemple précédent sur le candidat terme `Aspect`, celui-ci serait décrit sous la forme du concept formel présenté dans le tableau 3.1. Pour plus de détails, le lecteur consultera avec profit un article précédent de Yannick Toussaint et Luis Felipe Melo sur le sujet [Melo-Mora and Toussaint, 2015].

	l'	est	une	catégorie	qui	exprime	la	séquence	interne	d'	un	procès	ce	dernier	primordial	$\mathbb{T}^{\times}$	$\mathbb{T}^{\prime}$
$\boldsymbol{S}_1$	x	x	x	x	x	x	x	x	x	x	x					x	
$\boldsymbol{S}_2$		x											x	x	x		x

TABLE 3.1 – Extraction du contexte formel de deux contextes d'apparition du CT `Aspect`

Les hypothèses sont calculées pour chaque CT séparément et regroupées pour former un ensemble d'hypothèses positives et négatives. Les hypothèses positives correspondent naturellement à l'ensemble des contextes formels tirés des occurrences terminologiques observées dans le jeu d'apprentissage et inversement pour les hypothèses négatives.

Le jeu d'entraînement est donc re-décrit sous forme de deux types de contextes, positifs $K_+ = (G_+, M, I_+)$ ou négatifs $K_- = (G_-, M, I_-)$, tandis que les instances du jeu de test correspondent à des contextes indéterminés : $K_\tau = (G_\tau, M, I_\tau)$.

La méthode va opérer un certain nombre de filtrages :

— Chaque hypothèse positive H_+ est un ensemble **non vide** d'attributs qui ne se retrouvent dans aucune description d'hypothèse négative H_- (et inversement pour les hypothèses négatives)
— Les hypothèses sont construites par regroupement des attributs, idéalement on aura une seule H_+ (respectivement H_-) pour chaque exemple de G_+ (respectivement G_-).
— Dans les faits, les combinaisons d'attributs seront souvent multiples puisque les contextes des exemples sont par définition variés ;
— Les hypothèses sont contraintes par un opérateur de fermeture [56] : on cherche pour un

55. Cette approche a été développée par nos collègues du LORIA Yannick Toussaint et Luis Felipe Melo. Je reprends ici un descriptif rapide de manière à faciliter les comparaisons avec les autres méthodes.

56. La notion de fermeture correspond à ce que l'on appelle « maximalité »en algorithmique du texte voir section 2.3.2 page 47

nombre donné d'exemples le nombre maximal d'attributs qu'ils ont en commun. Si ajouter un attribut diminue la couverture en nombre d'exemples (ou support) alors l'*itemset* est dit fermé.

Une occurrence x d'un terme candidat sera classée positive (TO) si la description de x contient au moins une hypothèse positive et aucune hypothèse négative, inversement elle sera classée négative (NTO) si la description contient une hypothèse négative et aucune hypothèse positive. Si les deux types d'hypothèses sont présentes dans la description (quelles que soient leurs cardinalités respectives) alors le statut de l'occurrence est indéterminé, le système ne donne donc pas de réponse. Pour cette méthode aucune solution de repli (*fallback*) n'a été implantée bien que cela aurait certainement amélioré le rappel ainsi que le taux de décision.

3.2.3 Spécificité de Lafon (LS)

Cette approche développée par les collègues de l'ATILF, il s'agissait d'appliquer la méthode textométrique de calcul de la spécificité [Lafon, 1980, Drouin, 2007]. La spécificité de Lafon permet d'observer les variations de fréquence des formes, des mots, au sein des différentes parties d'un corpus. Il s'agit d'observer si cette fréquence est normale, si elle correspond à l'attendu : si la spécificité est positive, la forme est sur-employée, si elle est négative, la forme est sous-employée.

Fondements de la méthode

Le calcul de spécificité s'appuie sur les variables suivantes :
— T la taille en mots du corpus
— t_i la taille dans le paragraphe[57] i
— e l'effectif (fréquence absolue) de la forme dans T
— e_i l'effectif de la forme observé dans le paragraphe i
— e'_i l'effectif de la forme attendu dans le paragraphe i

Il s'agit donc de comparer l'effectif attendu dans la partie considérée avec l'effectif effectivement observé. Pour une description plus complète de la mesure et de ses propriétés voir par exemple [Labbé and Labbé, 2013].

Réalisation

Pour chaque CT du corpus d'entraînement, on extrait les contextes lexicaux de ses occurrences terminologiques (TO) et non-terminologiques (NTO), on calcule pour les différents mots de ces contextes si leur spécificité est positive (LS_+ ou négative (LS_-). L'hypothèse est que les voisins affichant une spécificité positive dans les exemples de TO sont des indicateurs du caractère terminologique de l'occurrence examinée.

À l'issue de l'analyse du jeu d'apprentissage, on regroupe les paires (mot, spécificité) de tous les voisins d'un CT dans ses TO d'une part et dans ses NTO d'autre part. Le tableau 3.2, présente les voisins les plus spécifiques dans les contextes terminologiques et non-terminologiques du CT « aspect », triés par ordre décroissant de spécificité.

Nous pouvons voir que dans les paragraphes où `aspect` revêt son sens terminologique les voisins affichant les scores de spécificité positive les plus élevés ont à voir avec des problématiques liées aux verbes, et notamment semble-t-il à la langue anglaise. Au contraire les voisins les plus significatifs des occurrences non-terminologiques affichent moins d'unité thématique ce qui

57. Lafon parle de partie mais ici notre unité d'analyse est le paragraphe.

Voisins dans les *TO*		Voisins dans les *NTO* pairs	
Angleterre	41,16	*orientation*	19,55
passé	34,21	*communauté*	11,45
anglais	28,73	*représentation*	11,41
prétérit	19,35	*compétence*	10,34
futur	17,40	*parlant*	8,91
achevé	16,61	*familier*	8,84
durée	15,78	*esprit*	8,42
langue	14,38	*amusant*	7,86
règle	11,10	*chose*	7,77
narration	10,87	*caractéristique*	7,21
...		...	

TABLE 3.2 – Application de la spécificité de Lafon, voisins les plus spécifiques des *TO* et es *NTO* de `aspect` avec leur score de spécificité.

semble bien être en relation avec un caractère plus diffus du sens activé dans ces contextes. Si l'on prend d'autres contextes d'apparition du *CT* « aspect » :

> *L'***aspect***, catégorie par laquelle l' énonciateur conçoit le déroulement interne d' un procès, est marqué en créole haïtien au moyen de particules marqueurs prédicatifs MP préposées au verbe.*

Ici, nous avons une *TO* de `aspect` car son contexte partage plus de mots spécifiques des *TO* du corpus d'entraînement que des *NTO* : *présent, déroulement, exprimer, passé, durée, langue, parler ...*.

Un autre exemple :

> L'**aspect** différentiel cède la place à une vision positive substantielle du lexique.

Ici nous avons une *NTO* car l'intersection avec les voisins spécifiques des *NTO* est plus grande. Cette méthode peut elle aussi générer de l'indécision pour des exemples non vus dans le jeu d'entraînement.

3.2.4 Approche fondée sur la saillance (*Salience Approach*, SA)

Cette approche conserve les observables utilisés dans les autres approches, à savoir les mots et les étiquettes syntaxiques (*POS tags*) Nous exploitons ces caractéristiques dans un cadre d'apprentissage supervisé, cadre naturel pour la désambiguïsation terminologique depuis [Quinlan, 1993] qui avait déjà fait usage d'un arbre de type C4. Dans cette approche, on utilise comme caractéristiques l'étiquette POS, le lemme et des caractéristiques discursives intrinsèques à l'occurrence considérée. L'hypothèse est que les occurrences terminologiques d'un candidat apparaissent plus probablement à des positions saillantes dans le document. Nous utilisons la saillance au sens où nous l'avons défini dans [Brixtel et al., 2013, Lejeune et al., 2015a] c'est-à-dire le phénomène par lequel les informations importantes du document sont placés à des positions particulièrement importantes du document, et généralement répétées. Pour les articles de presse il s'agit principalement du titre et du premier paragraphe (le chapeau).

Les textes scientifiques ne contiennent que quelques termes importants. Ces termes apparaissent prioritairement dans des positions saillantes afin de faciliter la compréhension du lecteur. De plus, lorsqu'un terme important apparaît, il s'accompagne d'autres termes importants, les *TO*

ont une forte tendance à se regrouper, un comportement que nous avons qualifié de grégaire [Lejeune and Daille, 2015]. ăAu contraire, les *NTO* sont distribués de manière plus uniforme dans le document. Dans mesure où le nombre de positions saillantes est limité, il est peu probable que des *NTO* occupent ces positions saillantes. Le corpus que nous exploitons ici est assez structuré au niveau sur-phrastique. Nous avons donc pris le parti de calculer les positions saillantes grâce à la structure Xml. ăLes principaux tags présents dans ce corpus sont :

text : le texte intégral, intégrant le titre du document (`title`) et le corps (`body`)

div : les sections, avec `head` leurs titres et `p` les paragraphes qui les composent

list : les listes à puces et leurs `item`

keywords : les mots-clés donnés par les auteurs

ref : les références bibliographiques

Le calcul des caractéristiques positionnelles est effectué de la façon suivante :
Pour chaque occurrence d'un *CT* :
— Pour chaque type de balise Xml (*tb*) du document :
  — On calcule la distance en caractères entre l'occurrence considérée et la balise de type *tb* la plus proche
  — On normalise la distance en fonction de la taille du document

Ces caractéristiques forment un vecteur que l'on combine avec l'information sur le lemme et l'information sur le *tag* pour entraîner un classifieur.

Dans le cas d'une unité multi-mots, le *POS-tag* est la concaténation des *POS-tags* des mots qui la composent, c'est en quelque sorte un motif syntaxique au sens de [Legallois et al., 2016].

Pour le choix des algorithmes de classification, nous nous sommes appuyés sur les travaux de [Yarowsky and Florian, 2002] qui montrent que les algorithmes de type *discriminant* (*discriminative*) tels que les arbres de décision sont plus efficaces lorsque l'on a un jeu de caractéristiques discriminantes mais en petit nombre. Pour l'expérience originale, nous avions utilisé la configuration par défaut de l'arbre de décision $C4.5$ de Weka [Witten and Fanck, 2005] à l'exception de la profondeur que nous avons fixé à 15. La figure 3.1 montre un exemple d'arbre généré de profondeur 5 généré sur un échantillon.

3.3 Jeu de données et résultats

3.3.1 Description du jeu de données

Le jeu de données utilisé comporte 55 articles scientifiques, 13 articles de journaux (144.000 tokens) et 42 articles de conférence (197.000 tokens) issus du corpus Scientext [Tutin and Grossmann, 2015]. Ce corpus a été enrichi au travers de plusieurs étapes. L'outil TermSuite [Daille et al., 2011] a été utilisé pour lemmatiser le corpus, l'étiqueter morpho-syntaxiquement puis en extraire les candidats termes (CTs).

Chaque occurrence d'un *CT* a été annoté manuellement par des experts au travers d'un processus comportant 3 étapes successives [Gaiffe et al., 2015]. À l'issue des 3 étapes de validation, l'occurrence est considérée comme terminologique. Si une étape n'est pas validé, l'occurrence est non-terminologique, à l'issue de la dernière étape in a donc les *TO* :

1. *CT* validé au niveau syntaxique. La validation repose sur des critères propres à chaque discipline.
2. *CT* validé au niveau disciplinaire. La validation repose sur l'appartenance effective du terme au champ scientifique dont relève les textes.

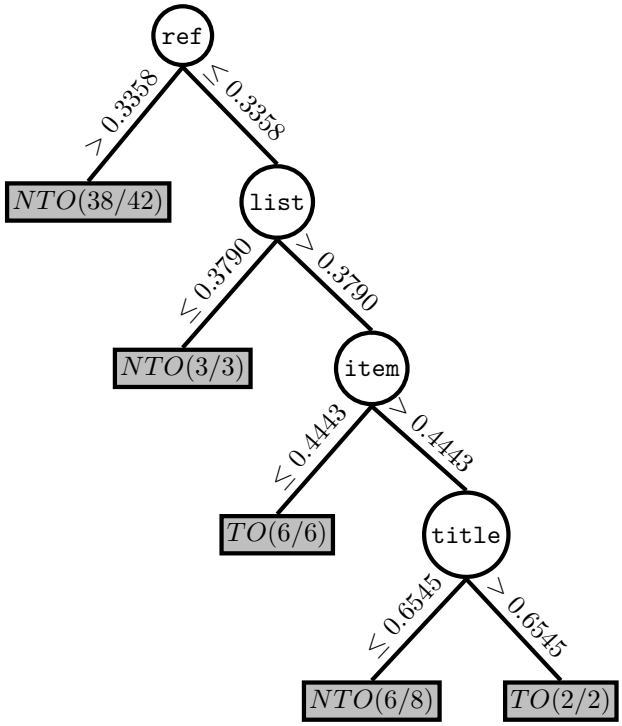


FIGURE 3.1 – Exemple de règle générée par l'arbre de décision (méthode SA)

3. CT validé au niveau terminologique. La validation repose sur un emploi véritablement terminologique dans le contexte du document où l'on retrouve le candidat.

Décrit de manière plus formelle, le processus de validation aboutit à une étiquette (du niveau 1 pour le candidat non validé syntaxiquement jusqu'au niveau 4 pour le candidat respectant toutes les contraintes) où les experts doivent valider l'étape $i + 1$ si et seulement si le candidat a été validé au niveau i. Le premier niveau de validation est préalable à l'intervention des experts puisqu'il s'agit de la sélection des CT par TERMSUITE. Ce découpage en étapes était destiné à faciliter le travail des annotateurs en morcelant la décision en plusieurs étapes objectivées.

4.204 CTs avaient été extraits par TERMSUITE avec au total 52.168 occurrences. Seules 33.10% de ces occurrences ont atteint le niveau 4 de validation et étaient donc considérées comme véritablement terminologiques.

À des fins d'évaluation, le jeu de données a été découpé en 8 strates. De manière à préserver le découpage des documents, un document apparaît en totalité dans le train ou dans le test. Chaque sous-corpus contient 48 documents pour l'entraînement et 8 documents pour le test. Dans les sous-corpus, l'équilibre entre les sous-domaines du corpus (Apprentissage de la Langue, Lexique, Linguistique Descriptive, Linguistique et Langue, TAL et Socio-Linguistique) a été autant que possible respecté.

3.3.2 Résultats

Les résultats obtenus sur le corpus de test sont présentés dans le tableau 3.3.

Les modalités d'évaluation sont les suivantes :

— Les Vrais Positifs (VP) sont les TO correctement classées

— Les Faux Positifs (FP) sont les NTO classées à tort comme TO.

— Les Faux négatifs (FN) sont les TO classées à tort comme NTO.

Certaines des méthodes (POL, LS et HB) ne donnent pas de réponse pour chaque candidat, cette indécision conduit à des instances non classées (annotées TO comme NTO). Cela peut affecter le calcul et l'analyse du rappel puisque on aura des VP, des FN et des *undecided*. Par conséquent, nous proposons deux définitions des Faux Négatifs (FN) :

— Les Faux Négatifs de type A (FN_A) sont les TOs qui ont été mal classés. Ce chiffre est utilisé pour calculer le rappel de type A.

— Si on ajoute aux FN_A, les TO pour lesquelles aucune décision n'a été prise, on obtient les FN_B qui sont utilisés pour calculer le rappel de type B

Le rappel de type A ne tient donc compte que des décisions prises, les *undecided* ne sont pas pris en compte. Le rappel de type B tient compte de toutes les décisions à prendre, les *undecided* sont considérés comme des NTO.

Le rappel de type A favorise les approches orientées précision, c'est-à-dire ces approches qui ne prennent pas systématiquement de décision de manière à prendre des décisions plus sûres. Pour différentes raisons, les approches POL, LS et HB peuvent tomber dans le cas où il n'y a pas assez d'informations pour trancher. Ce qui n'arrive pas avec la méthode ROL qui dans le doute classifie en NTO. C'est une sorte de *Fallback*. Cela n'est pas non plus possible avec la méthode SA qui se base sur des algorithmes de classification qui prennent des décisions quoi qu'il arrive. Au contraire, le rappel de type B met en avant les méthodes qui favorisent le rappel, dans cette configuration l'indécision est considérée comme une erreur. Nous ajoutons donc dans nos tableaux le taux de décision (DEC) qui permet de mesurer la proportion de décisions prises. Les différentes métriques d'évaluation sont ainsi les suivantes :

— Taux de Décision : $DEC = (VP + FN_A)/(VP + FN_B)$
— Précision : $P = VP/(VP + FP)$
— Rappel de type A : $R_A = VP/(VP + FN_A)$
— Rappel de type B : $R_B = VP/(VP + FN_B)$
— F_A-mesure : $F_{\beta_A} = (1 + \beta^2) * \frac{P_A * R_A}{(\beta^2 * P_A) + R_A}$
— F_B-mesure : $F_{\beta_B} = (1 + \beta^2) * \frac{P_B * R_B}{(\beta^2 * P_B) + R_B}$

La F-mesure est calculée avec $\beta = 1$ ainsi qu'avec $\beta = 0.5$ afin de mettre en valeur la précision.

Sur l'ensemble des strates utilisées pour la classification, nous avons 59 168 occurrences, parmi lesquelles 24 964 n'ont pas d'annotation par la méthode HB et 13 615 n'en ont pas par LS. 10 230 des instances ne sont classées par aucune de ces deux méthodes. On peut observer que ces occurrences sont particulièrement difficiles puisque la méthode SA obtient sur ce sous jeu de données les résultats suivants : 1 988 VP, 5 549 VN, 2 287 FP et 406 FN soit un rappel de 83.04 % mais une précision de seulement 46.5% soit 59% de F-mesure soit pour cette dernière mesure une perte de près de 11 points de pourcentage.

Nous pouvons voir une différence importante dans les résultats entre les deux configurations d'évaluation. Dans la configuration A, c'est l'approche HB qui donne les meilleurs résultats du fat d'une précision significativement meilleure. De la même manière, la faible performance des deux *baseline* s'explique par leur moindre précision. La configuration B, comme l'on pouvait s'y attendre, favorise les systèmes ayant un taux de décision (DEC) plus élevé.

Résultats de la méthode par Hypothèses

	POL	ROL	HB	LS	SA
VP	8034	11148	8864	10075	14355
FP	3478	5679	1577	4511	5314
DEC	78.8%	**100%**	53.5%	71.8%	**100%**
P	69.8%	66.2%	**84.9%**	69.1%	73.0%
FN_A	5398	9841	**2374**	4996	6634
R_A	59.8%	53.1%	**78.9%**	66.9%	68.4%
F_{1_A}	64.4%	59.0%	**81.8%**	67.9%	70.6%
$F_{0.5_A}$	65.4%	60.3%	**82.48%**	68.2%	71.1%
FN_B	12955	9841	12125	10914	**6634**
R_B	38.3%	53.1%	42.2%	48.0%	**68.4%**
F_{1_B}	49.4%	59.0%	56.4%	56.6%	**70.6%**
$F_{0.5_B}$	52.5%	60.3%	60.56%	58.8%	**71.1%**

TABLE 3.3 – Résultats pour les deux baselines, *Precision Oriented Lesk* (POL) et *Recall Oriented Lesk* (ROL) ainsi que des trois approches développées dans ce travail : fondée sur les Hypothèses (HB), Spécificité de Lafon (LS) et Saillance (SA). Dans chaque cas les deux configurations d'évaluation A et B sont utilisées

Analyse de la méthode SA La caractéristique POS est souvent très efficace pour la classification,en particulier du fait que si les occurrences de noms sont plus souvent des TO, les occurrences d'adjectifs sont plus souvent des TO Si l'on prend l'exemple de `radical` : en tant qu'adjectif on a 90% de NTO mais en tant que nom la proportion s'inverse puisque l'on a 80% de TO. Un autre exemple intéressant est le CT `linguistique`, on observe que ses occurrences dans la première phrase d'un paragraphe sont systématiquement des TO.

Afin de faciliter la comparaison avec les exemples pris sur la méthode HB (Tableau 3.4), nous présentons dans le Tableau 3.5 des résultats sur le terme candidat `sémantique`. C'est un exemple intéressant car seules 20% de ces occurrences sont des TOs, mais leur distribution est égale entre les POS Nom et Adjectif. Par contre, si 77% des occurrences nominales sont des TOs, seules 15% des occurrences adjectivales en sont. Ce n'est pas une surprise que l'usage en tant que nom, plus rare, soit plus terminologique. Dans le tableau nous pouvons voir des erreurs de classification typiques de la méthode SA. Les deux faux négatifs sont extraits de deux documents moins richement ment structurés que les autres, ce qui limite l'applicabilité des caractéristiques de saillance. Dans le cas du premier faux négatif, nous avons un document avec des paragraphes plutôt longs (voir la valeur de distance à p) et peu de références (voir la valeur de distance à ref) de sorte que l'occurrence adjectivale en question n'est pas saillante. Le Faux Négatif en tant que nom (deuxième ligne du tableau 3.5) est difficile à analyser tel quel mais quand on le compare au Vrai Positif on se rend compte que ce dernier était proche d'une référence (balise ref) ce qui est un indice fort pour la méthode SA. La saillance telle que nous la calculons ici ne permet donc pas de trancher dans le bon sens.

Règles d'association Afin d'obtenir une évaluation plus fine que les scores d'agrégation précédemment utilisés (Rappel, Précision et F-mesure), nous avons extrait les règles d'association entre les résultats des différentes annotations (manuelle ou automatique). Dans ce cadre, chaque occurrence d'un CT est décrit par un quadruplet (MA, HB, LS, SA) qui décrit les 4 annotations effectuées : MA pour l'annotation manuelle de référence et HB, LS et SA pour les annotations

CT	Effectif	Occurrences Positives	% Occurrences Terminologiques (*TO*)	Mots spécifiques à T_+	Hypothèses issues de T_+	Proportion d'Hypothèses Positives	Mots en Commun	Occurrences Négatives	Mots spécifiques à T_-	Hypothèses issues de T_-
adjectif	216	207	95.83%	966	301	97.41%	64	9	59	8
corpus	688	510	74.12%	1035	1347	81.93%	713	178	535	297
texte	568	266	46.83%	735	913	52.32%	772	302	792	832
relation	676	171	25.29%	159	183	11.48%	629	505	1427	1410
sémantique	413	80	19.37%	272	108	8.88%	560	333	1258	1107

TABLE 3.4 – Exemples de termes candidats traités par la méthode HB

Résultat	POS	title	head	p	item	ref
FN	ADJ	0.8131	0.5064	0.1914	0.8079	0.4875
FN	NOM	0.8324	0.5439	0.1116	0.8279	0.1516
VN	ADJ	0.8368	0.5483	0.1160	0.8323	0.1472
VP	NOM	0.8406	0.5717	0.1636	0.8318	0.0342
VN	ADJ	0.8523	0.5638	0.1315	0.8478	0.1317

TABLE 3.5 – Exemples de résultats de classification pour le *CT* `sémantique`

automatiques. Dans les règles d'association présentées dans le tableau 3.6, les annotations sont codées de la façon suivante *ID-RES* où *ID* est l'identifiant d'une méthode d'annotation (MA, HB, LS et SA donc) et *RES* est le résultat donné par cette méthode : TO, NTO ou UN (*unknown*) quand la méthode ne donne pas de résultat (ce qui n'est jamais le cas pour MA et SA). Il faut noter deux choses, la première est que ces règles n'ont pas de valeur prédictive, mais plutôt descriptive et d'autre part que chaque occurrence peut être décrite par plusieurs règles. Les règles sont décrites par deux scores : la confiance qui mesure la proportion d'instances pour lesquelles la véracité de la prémisse est associée à la véracité de la conclusion, le support (ou couverture) qui mesure quelle proportion des instances cette règle couvre. La première règle du tableau ([HB-NTO, LS-NTO] $\longrightarrow$ [SA-NTO]) se lit donc de la façon suivante : lorsque HB et LS donnent tous deux l'étiquette NTO alors SA donne le même résultat dans 93% des cas, couvrant 31% des instances du jeu de données. Dans ce tableau les règles sont présentées par score de confiance décroissant. Ce tableau montre que globalement les systèmes sont d'accord entre eux et que si la méthode SA a des meilleures performances de manière globale, il existe tout de même un plafond de verre au niveau des résultats avec des instances qu'aucun système n'arrive à classifier correctement.

3.4 Conclusion intermédiaire : quels observables pour quels usages ?

Il était possible sur le corpus TERMITH utilisé dans cette expérience d'exploiter la structure puisque celle-ci était marquée. Si nous n'avions disposé que de texte brut, où les seules vestiges

Confiance (in %)	Support (in %)	Règle d'association
0.93	0.31	[HB-NTO, LS-NTO] ⟶ [SA-NTO]
0.93	0.28	[MA-NTO, HB-NTO, LS-NTO] ⟶ [SA-NTO]
0.93	0.1	**[HB-TO, LS-TO] ⟶ [SA-TO]**
0.93	0.09	[MA-TO, HB-TO, LS-TO] ⟶ [SA-TO]
0.92	0.33	[MA-NTO, HB-NTO] ⟶ [SA-NTO]
0.92	0.13	[MA-TO, HB-TO] ⟶ [SA-TO]
0.91	0.4	**[MA-NTO, LS-NTO] ⟶ [SA-NTO]**
0.91	0.36	**[HB-NTO] ⟶ [SA-NTO]**
0.91	0.28	**[SA-NTO, HB-NTO, LS-NTO] ⟶ [MA-NTO]**
0.91	0.16	**[HB-TO] ⟶ [SA-TO]**
0.91	0.09	[MA-NTO, HB-UN, LS-UN] ⟶ [SA-NTO]
0.9	0.36	[HB-NTO] ⟶ [MA-NTO]
0.9	0.33	[SA-NTO, HB-NTO] ⟶ [MA-NTO]
0.9	0.3	[HB-NTO, LS-NTO] ⟶ [MA-NTO]
0.9	0.11	[MA-NTO, HB-UN, LS-NTO] ⟶ [SA-NTO]
0.89	0.15	[MA-TO, LS-TO] ⟶ [SA-TO]
0.89	0.11	[MA-NTO, LS-UN] ⟶ [SA-NTO]
0.88	0.4	[SA-NTO, LS-NTO] ⟶ [MA-NTO]
0.88	0.05	[MA-TO, HB-UN, LS-TO] ⟶ [SA-TO]
...	...	...
0.85	0.11	**[SA-NTO, HB-UN, LS-NTO] ⟶ [MA-NTO]**
0.85	0.09	**[SA-TO, HB-TO, LS-TO] ⟶ [MA-TO]**
...	...	...
0.7	0.09	**[SA-NTO, HB-UN, LS-UN] ⟶ [MA-NTO]**
...	...	...

TABLE 3.6 – Règles d'association entre les annotations (manuelles et automatiques), MA (manuelle) et HB, LS et SA triées par ordre décroissant de confiance. Le « - » sépare la méthode et le résultat de la méthode, de sorte que $HB - TO$ signifie la méthode HB a donné comme résultat TO

de la structuration opérée par l'émetteur auraient été des lignes vides (probablement traces de la séparation entre deux paragraphes) ou des lignes de texte sans ponctuation finale (probablement traces de titre ou sous-titres ...). Il aurait fallu également identifier les références, ce qui sans être un problème résolu est une tâche relativement simple à gros grain [Tkaczyk et al., 2018]. Bien sûr, si toutes les phrases du texte sont concaténées, cela devient un peu plus difficile.

Conserver la structure, si elle est disponible, ou au moins ne pas en écraser les traces permet de travailler avec autre chose que des mots dans des phrases. Cela me semble être une exigence importante pour permettre d'aller plus loin sur l'analyse automatique des textes. L'utilisation de la structure permet d'exploiter plus finement la notion de genre textuel, que ce soit dans des articles de presse en exploitant les phénomènes de répétition à des positions remarquables (voir [Lejeune et al., 2015a,Mutuvi et al., 2020a] pour une application à la veille épidémiologique) ou comme marqueur du statut de différents segments textuels notamment pour l'extraction automatique de tables des matières (voir [Giguet and Lejeune, 2019, Giguet et al., 2020] pour nos participations aux compétitions FinTOC) ou encore pour améliorer la segmentation en phrases de documents disponibles sous forme d'images ([Giguet and Lejeune, 2021b]). Je suis pragmatique par nature, il ne s'agit pas ici d'imaginer des normes de qualité des corpus qui soient inapplicables ou gênent fortement le passage à l'échelle au niveau de la taille des corpus traités. Par contre, il me semble possible d'imaginer une ligne de crête consistant à ne pas « sur-traiter » (ou maltraiter) les données, à ne pas imposer des sélections de caractéristiques qui restreignent fortement les possibilités qu'il y a d'exploiter ces données. On n'imaginerait pas fournir un corpus de données en français où l'on aurait enlevé les diacritiques, des corpus de dialogues fournis sous la forme d'un sac de phrases ou encore des données d'entraînement pour la correction d'OCR simplement sous la forme de fichiers textes sans les fichiers ALTO [58] ? Bien sûr, préparer les données facilite leur accès et évite que les utilisateurs des corpus aient à refaire des étapes de traitement déjà accompli par d'autres (en particulier pour des campagnes d'évaluation telles que TREC, DEFT ou encore FINTOC). Mais, en imposant une façon unique de représenter les données, un standard dans le sens de la correction (non dans le sens du format), on restreint la variété des approches. Il me semble donc fondamental de fournir les traces des différentes étapes de traitement des données textuelles de manière à pouvoir concevoir des approches plus holistes, plus globales, plus robustes. Ceci pose des questions importantes, il me semble, sur une constitution de corpus qui soient adaptés à la variété des données en entrée, à la variété des usages en sortie.

La prochaine partie de la présente habilitation sera justement consacrée à la question de la construction des corpus, de leur structuration et des contraintes qualitatives à exercer. Comment contrôler la collecte ? Comment favoriser une collecte large à moyens réalistes sans trop exclure une collecte en profondeur qui conserve au maximum l'intégrité des textes ? Comment prévenir, détecter ou encore exploiter les sauts qualitatifs dans les corpus ?

58. Format XML utilisé dans les outils d'OCR et qui contient non seulement les caractères extraits, mais aussi leur position dans le texte et le taux de confiance de l'OCR.

Troisième partie

La qualité des observatoires : le bruit et la trace dans les données non standards

Chapitre 4

Comment évaluer la qualité de données issues du web ?

Sommaire

4.1 Problématiques du TAL et des données issues du Web

La possibilité d'extraire du contenu textuel à partir de données Web a incontestablement révolutionné la manière d'appréhender le TAL puisque la quantité de données textuelles disponibles augmentait graduellement, qu'il s'agisse de contenus générés par des utilisateurs (*User Generated Content*) répondant à peu de contraintes de forme (*tweets* par exemple) ou de contenu répondant encore à des contraintes éditoriales. Dès lors, le développement d'outils facilitant la collecte de corpus à partir du Web est devenu très important. On peut subdiviser, à grand traits, cette tâche en en quatre grandes étapes [Barbaresi, 2021] : (I) repérage des `url` pertinentes et téléchargement des pages en question, (II) nettoyage/scraping, (III) évaluation de la qualité et (IV) archivage . Ici nous nous intéresserons en particulier aux étapes 2 et 3 puisqu'elles sont plus étroitement liées aux problématiques pratiques quotidiennes du TAL sur corpus issu du web.

4.1.1 Les contraintes posées par les données issues du Web

Je m'intéresse en particulier aux techniques d'extraction du contenu textuel à partir de pages Web, laissant de côté d'autres tâches de *scraping* comme par exemple l'extraction de données tabulaires [Chasins et al., 2018] qui correspond à un type beaucoup plus particulier de de données et d'usages. Les questions d'évaluation sont au cur de la problématique de l'équilibre entre la qualité et la quantité. Nous pouvons la formuler de cette façon : qu'est-ce qu'un bon outil de *scraping* apporte en terme de quantités de données traitées et qu'est-ce qu'un bon corpus (c'est-à-dire quelle est la qualité des données en sortie). Si l'on doit juger la qualité de l'outil à sa capacité à produire des données de qualité, les modalités d'expression de cette qualité doivent pouvoir se mesurer. Conjointement, est-ce que l'on peut choisir, ou conseiller, un outil indépendamment d'un usage très précis ? Les questions sur le sujet sont nombreuses et notamment pour savoir dans quelle mesure les outils peuvent et doivent être génériques. On serait tenté de dire que l'on va répondre à cette question par une évaluation, aussi objective que possible, mais est-ce que l'évaluation peut capturer efficacement cette généricité ou encore en détecter les failles. En particulier quand les données en entrée vont varier sur différents plans :

Linguistique : la variation des langues des documents traités ;

Diachronique : la variation du langage du Web, au travers des recommandations et des standards mais aussi de l'évolution des pratiques ;

Typologique : la variation dans la manière de construire une page, et l'on observe une évolution de la pratique du HTML indépendamment même de l'évolution de ces standards [59]

Je souhaiterais rapidement apporter quelques développements sur ce dernier point. Par bien des aspects le langage HTML se rapproche d'une langue naturelle puisque l'on a non bi-univocité entre le code et le rendu. On pourrait ainsi arguer, en reprenant les mots de Bruno Bachimont [Bachimont, 2007], que l'on se situe dans le cas où il n'y a pas identité entre le support d'enregistrement et le support de restitution. En effet, différents navigateurs vont générer différentes interprétation du même contenu. Plus encore, les navigateurs vont s'efforcer de générer un rendu quelque soit la conformité du code aux standards, y compris si ce code est tout simplement mal formé (au contraire de ce qui arriverait pour un document XML par exemple). La liberté laissée à l'émetteur du code donne donc pour mission au récepteur, le navigateur pour ce qui est du rendu ou l'outil d'extraction pour ce qui nous intéressera ici, de réaliser un certain nombres d'inférences pour expliciter la structure et hiérarchiser les données limitant la possibilité d'avoir une approche à la fois générique et pérenne.

Un premier choix à effectuer sera malgré tout celui qui amène à trancher entre des outils plutôt génériques et des outils *ad hoc* à un site web. Les outils génériques, tout terrain, par opposition aux outils *ad hoc*, vont favoriser une construction opportuniste de corpus [Barbaresi and Lejeune, 2020b] où la liste des sources pourra être amendée régulièrement sans contrainte de répartition équilibrée de documents par source. Cette modalité de construction se veut plus pragmatique dans le sens où elle sacrifie, en partie, la qualité du corpus au profit de la quantité. J'entends ici qualité dans le sens de conserver les bonnes propriétés que l'on attend de textes extraits automatiquement (complétude, homogénéité, conservation d'une structure, fidèle à l'original) et que l'on va accepter de laisser de côté, sans nécessairement nier leur pertinence mais simplement pour obtenir une quantité qui puisse mettre en valeur des méthodes de TAL ou parfois tout simplement justifier leur utilisation. Il me semble en effet que le constat qu'il y a une « grande quantité de données » ne peut justifier à lui seul l'utilisation d'une approche

59. Ce qui n'est pas sans effet par exemple sur les pratiques d'archivage du Web : `https://tinyurl.com/archives-web-bibliotheconomie` consulté le 2 octobre 2023

automatique (de type TAL) ou tout au moins d'un *distant reading* (de type textométrie par exemple), même si c'est sans doute un poncif bien pratique pour introduire le contexte d'une recherche. On s'ingénie, parfois artificiellement, à maintenir une approche quantitative, alors même que des méthodes plus simples pourraient être appliquées si les données étaient moins grandes mais de meilleure qualité et sélectionnés sous forme de sous-corpus homogène. Cette expression est un parfait pléonasme puisqu'un corpus au sens linguistique devant par définition être homogène un sous corpus devrait l'être plus encore. Mais le foisonnement excessif, pour rester positif, dans les corpus issus du web interdit souvent un usage qualitatif et quantitatif (quali-quanti) de ces données [Gautier, 2016]. L'évaluation proposée dans ce chapitre s'appuie sur des données de références en cinq langues (chinois, anglais, grec, polonais et russe) pour lesquelles nous disposons des données d'origine (en Html donc) ainsi que des données textuelles à extraire, données obtenues à partir d'une annotation manuelle.

La question qui se pose est celle des métriques d'évaluation, de leur interprétabilité, de leur capacité à représenter de façon fiable les qualités et les défauts de chaque méthode évaluée. On souhaiterait aussi pouvoir mesurer leur complémentarité qui peut être difficile à mesurer lorsque l'on a recours à des agrégations (Précision, Rappel, F-mesure) sur des *tokens* puis sur des documents et le tout moyenné. D'un point de vue d'utilisateur, pendant ma thèse par exemple, j'aurais bien aimé mieux comprendre les implications des chiffres présentés au-delà du fait de voir un outil qui est présenté comme « meilleur que les autres » . D'un point de vue de « conseilleur », dans le cadre de l'enseignement, du tutorat de stage, de l'encadrement doctoral ou plus récemment auprès des collègues de la faculté des lettres dans le cadre des activités de CERES, il faut aussi être certain de conseiller les bons outils ou en tout cas de transmettre les bonnes précautions à prendre vis à vis des résultats présentés qui sont souvent des agrégations (moyennes sur un grand nombre de documents le plus souvent). Or, ces agrégations vont, à des fins compréhensibles de simplicité des comparaisons, laisser de côté un certain nombre de dimensions pertinentes puisque l'on met tout sur un même plan (tous les *tokens* se valent par exemple). Un simple écart type ne suffit pas à rendre compte de l'ampleur de l'hétérogénéité des données, de l'intensité des sauts qualitatifs, au sein des documents traités et plus encore au sein du corpus dans son ensemble. Ces déséquilibres qualitatifs, masqués donc par les scores agrégés, peuvent en partie réapparaître en ayant recours à des partitions des corpus, par exemple par langue ou par domaine ou au moins à des agrégations de type macro-moyennes lorsque l'on a des partitions déséquilibrées. Les expériences présentées ici mettent en évidence la diversité d'encodage de la mise en page du Web selon les langues ou les pays d'édition. Ces écarts se traduisent par des performances divergentes, de sorte que le bon outil doit être choisi en fonction de plusieurs critères liés à la tâche envisagée. Les expérimentations et réflexions que je présente dans ce chapitre sont notamment le fruit d'une collaboration avec Adrien Barbaresi (BBAW Berlin) et Emmanuel Giguet (GREYC Caen), deux chercheurs qui comme moi se sont intéressés en profondeur à cette question de l'extraction de contenu, que Nadine Lucas avait proposé de qualifier de *détourage* (dans le sens que ce terme recouvre en photographie), terme que j'avais utilisé dans mes premiers travaux sur le sujet [Lejeune et al., 2015b]. Les expériences menées s'appuient sur une comparaison d'outils *open source* d'extraction de contenu que nous avons testé sur des pages en cinq langues différentes (chinois, anglais, grec, polonais et russe). Nous avons repris ici le corpus Daniel-scraping utilisé notamment dans [Lejeune and Zhu, 2018] pour comparer l'évaluation intrinsèque et extrinsèque du détourage de pages web. .

4.1.2 Les données du Web pour le TAL : un terrain miné ?

L'utilisation de grands corpus Web « hors ligne » est désormais la norme dans toutes les disciplines des communautés scientifiques qui s'intéressent au langage. Ceci implique la recherche de sources, leur téléchargement, puis leur nettoyage et éventuelle dé-duplication avant de procéder à l'analyse des données proprement dite à l'aide d'un outil d'Analyse de Données Textuelles ou de TAL [Kilgarriff, 2007]. Le web est fréquemment perçu comme un « réservoir indifférencié de textes à analyser » pour le TAL [Tanguy, 2013] et l'on peut affirmer sans risque que web et TAL poursuivent leur « histoire commune » (*op. cit.*). Les données issues du web y sont en effet omniprésentes, à la fois en tant qu'instantané d'un état de la langue, de données à analyser pour elles-mêmes, mais aussi de références destinées à construire des modèles de langue ou des ressources langagières. Bien que le texte soit une donnée omniprésente sur le Web, l'extraction d'informations à partir de pages Web peut s'avérer difficile. Les données textuelles se présentent sous différentes formes, principalement en raison de la grande variété de plates-formes et de systèmes de gestion de contenu, et notamment en fonction du contexte, par exemple des objectifs divergents suivis lors de la publication. Ce processus implique un nombre important de décisions de conception et de tournants dans le traitement des données. Selon l'objectif suivi lors de la collecte de données, l'importance d'un filtrage substantiel et une évaluation de la qualité peut être plus ou moins grande. Malgré une certaine impression de facilité quant à la construction de corpus, les méthodes utilisant des corpus web nécessitent des dispositifs expérimentaux et des instruments *ad hoc* [Valette, 2008] afin d'estimer leur qualité et leur adéquation aux tâches proposées. D'un point de vue épistémologique, la simple accumulation de données textuelles ne rend pas pour autant ce terrain intelligible et un retour analytique sur ces données s'avère nécessaire [60].

4.1.3 Sacrifier la qualité à la quantité ?

Récemment des approches, qui rentrent dans la catégorie des approches opportunistes du point de vue de la constitution de corpus, telles que celle de CommonCrawl [61] ont prospéré car elles permettent de séparer très explicitement les questions, et responsabilités, de la collecte des données et de leur utilisation. De la collecte au corpus, non seulement opportuniste [McEnery and Hardie, 2011] mais également « prêt à utiliser », il n'y a souvent qu'un pas. Le Common Crawl notamment s'est imposé comme source majeure pour des tâches variées, de la traduction automatique neuronale [Smith et al., 2013] à la construction et (dans une situation optimale) à l'affinage de modèles de langue basés sur des techniques d'apprentissage profond nécessitant des données massives [Suárez et al., 2019]. Cette évolution a conduit à des problèmes récurrents d'ordre éthique, à l'image du robot conversationnel Tay lancé par Microsoft en 2016 sur Twitter et stoppé 16 heures après son entrée en fonction en raison de l'ampleur et de la gravité des messages racistes et sexistes « appris » et ensuite (re-)publiés par le robot [62]. De même, des modèles entraînés sur des données massives (d'origine contrôlée ou non) intègrent une série de biais sociétaux [Caliskan et al., 2017].

Énumérer toutes les contraintes qui vont être rencontrées, ces précautions qui devront être

60. « La collecte et la mise en circulation des données dans des dispositifs adéquats aboutissent à une mise en ordre du monde qui relève dune cosmétique : ces données sont triées, classées, archivées. Ces opérations textuelles permettent l'accumulation et donc l'archivage mais ne rendent pas pour autant le terrain intelligible. Cette intelligibilité du terrain est le résultat dune deuxième opération, celle de mise en ordre des données accumulées, de leur traitement, de leur analyse et de leur restitution. » [Calberac, 2010, p. 104]

61. `https://commoncrawl.org` consulté le 2 octobre 2023

62. `https://en.wikipedia.org/wiki/Tay_(chatbot)` consulté le 2 octobre 2023

prises invite me semble-t-il à chercher à caractériser, en particulier dans le contexte de conseil technique et d'encadrement scientifique, les situations où la valeur ajoutée des outils est la plus évidente Pour l'utilisateur final, le bénéfice se situe principalement au niveau de la toute première étape du processus de collecte décrit ci-dessus à savoir : le téléchargement et la phase d'exploration [Habernal et al., 2016, Schäfer, 2016]. Au-delà du fait qu'il serait peut être préférable épistémologiquement parlant de trouver son propre chemin sur le Web, il est clair que ces données ne doivent pas être utilisées sans filtrage. Outre la découverte de sites Web pertinents, un problème majeur consiste à sélectionner le contenu approprié après téléchargement [Schäfer et al., 2013]. Ceci n'est pas nécessairement simple en raison de défauts et de biais inattendus dans les données ou de biais générés par le traitement lui-même. On peut s'attendre à ce que certains algorithmes dédiés au traitement à grande échelle atténuent les irrégularités. Cependant, l'utilisation sans contrôle de ces algorithmes nécessite l'assurance qu'il n'existe qu'une faible marge d'erreur, information que des approches de lecture rapprochée[63], appliquées à des échantillons ou des carottages dans le corpus ; sont sans doute les celles à pouvoir apporter. Mais, on ne peut bien sûr en attendre une évaluation précise et à grande échelle des sauts qualitatifs. On peut penser à des procédés des raffinements successifs dans la constitution et le traitement des corpus, par exemple dans le contexte d'une plateforme d'informations lexicales agrégées [Geyken et al., 2017].

Même si la quantité a ses propres vertus, il n'en reste pas moins que les données brutes, non situées, n'ont pas le même intérêt que des données qui ne sont pas un sac mais un ensemble ordonné d'éléments sur lesquels on peut utiliser des opérateurs de comparaison temporelle ou auctoriale. Or, les méta-données dont on peut disposer sur des corpus construits à partir du web souffrent d'incomplétude. Le caractère lacunaire des méta-données est renforcé par le manque d'informations ayant trait au contenu dont les frontières exactes sont sujet à une évaluation *post hoc* [Baroni et al., 2009] car les besoins sont extrêmement divers. Par exemple dans le cas qui nous intéresse ici, veut-on garder trace des auteurs, de la date, les légendes des images ... ?. La réponse est assurément que oui mais enrichir l'annotation en ralentit le processus ce qui en multiplie le coût et d'un autre côté accroît les désaccords [Fort et al., 2012]. Dès lors, Un défi essentiel est de pouvoir trouver un compromis entre l'extraction à grande échelle de données sur le web et la conservation d'une qualité qui n'affecte que raisonnablement les attentes d'un point de vue scientifique [Barbaresi, 2015], notamment la préservation de certains caractères singuliers des données. Par exemple, dans le corpus Daniel-scraping la structure n'est pas hiérarchique : seules les marques des titres (de quel niveau qu'ils soient) et des paragraphes sont retranscrites. Ce qui a sans doute simplifié l'annotation de même que l'évaluation mais conduit conséquemment à une sous-estimation des difficultés de la tâche d'extraction. Le choix de ne pas annoter explicitement le titre, qui est en fait dans le corpus le premier segment textuel entre balises `p` (Figure 4.1), est à ce titre discutable même s'il n'a pas d'influence sur les évaluations que nous présentons par la suite.

Bien entendu, devant la grande variété de l'offre et de la demande en données textuelles, que ce soit au niveau des cas d'utilisation ou des types de textes impliqués, il devient de plus en plus difficile de se mettre d'accord sur l'utilité, la pertinence ou la granularité de tels ou tels corpus tirés du web pour un objectif scientifique donné. Une énumération plus raisonnée des sources, *focused web crawling* est sans doute un bon compromis entre quantité et qualité des corpus construits [Schäfer et al., 2014, Barbaresi, 2016, Barbaresi, 2019].

Quelle que soit la méthode de construction choisie, une opération essentielle consiste à conserver le contenu souhaité tout en mettant de côté le reste, processus qui a reçu différentes dénomi-

63. Par opposition au *distant reading* imposé par la quantité en jeu

<p> </p>

<p> 2012 ,
, REGNUM 17
.</p>
<p> , . , ,
</p>
<p> 17 . 693 , 7 .
2 . , ,
.</p>
<p> 23,6%. -
, .
.</p>

FIGURE 4.1 – Exemple d'article du corpus DANIEL-SCRAPING

nations déterminé bien souvent par l'angle d'attaque choisi : collecte de données web, extraction de squelette de page, segmentation de page, nettoyage de page web ou extraction de données textuelles [Lejeune and Zhu, 2018]. La variété des contextes et des types de texte conduit à d'importantes décisions de conception lors de la collecte des textes : l'outillage pourrait-il et devrait-il être adapté à des sources particulières, soient qu'elles soient particulièrement ciblées soient tout simplement plus fréquentes, ou bien l'extraction doit-elle être aussi générique que possible pour fournir des moyens efficaces d'obtenir des textes provenant d'un grand nombre de sources différentes.

La construction de corpus à partir du web est devenue un élément si commun des chaînes de traitement de TAL que les détails techniques sur sa mise en uvre sont souvent omis. Or, peut-on réellement s'abstenir de présenter explicitement quelles informations, quel observables, sont intégrés dans des corpus ? Afin d'illustrer ce problème de contenu, nous avons comparé différents extracteurs récents, disponibles ou tout simplement populaires afin d'observer ce que des métriques d'évaluation spécialement conçues révèlent concrètement sur leur efficacité et sur l'adéquation entre les *desiderata* des utilisateurs finaux et la réalité de l'exécution technique. Nous laisserons de côté le choix des sources en elles-mêmes pour nous concentrer sur les résultats de l'extraction, qui forment la base de décisions quant à l'inclusion dans le corpus final [Schäfer et al., 2013], et nous intéressons en particulier à la question multilingue. En effet, les outils mis à disposition de la communauté sont très souvent conçus pour la langue anglaise, ou *a minima* évalués principalement sur elle. L'applicabilité à d'autres langues est souvent considérée comme allant de soi : si les extracteurs fonctionnent sur l'anglais, alors les mêmes ordres de grandeur de résultats seront observés sur d'autres langues.

4.2 *Web scraping* : contexte expérimental et outils

Il me semble important de rappeler ici que l'extraction de contenu textuel n'est pas une tâche résolue [64] et que son intérêt dépasse largement les frontières de l'ingénierie. L'intérêt des

64. Bien que dans une relecture d'article, j'ai déjà pu lire que le problème était de peu d'intérêt puisqu'il suffisait de suivre le `xpath` dans les documents HTML traités pour résoudre la question.

utilisateurs, et des communautés de chercheurs en particulier se mesure bien en examinant la profusion d'outils développés. Des problématiques très concrètes se posent comme la vitesse de rendu, notamment pour la lisibilité de contenu injecté de l'extérieur, on parle parfois de distillation de contenu par les DOM-distiller [65] pour les navigateurs. Ceux-ci doivent en effet faire face à un Web boursouflé (mon adaptation du concept de *Web bloat* [Ghasemisharif et al., 2019]). La tâche est d'autant moins résolue que les données et les pratiques dans les technologies du Web évoluent [Weninger et al., 2016], rendant illusoires des solutions de long terme et garantissant une obsolescence rapide aux systèmes fondés sur une photographie trop restrictive du problème. Le web sémantique et les approches connexes fondées sur un encodage explicite par les émetteurs de contenu se sont heurtés au problème de la vitesse de publication recherché par « les locuteurs » du web.

Le standard HTML 5 (datant déjà de 2008) a pu apparaître comme une solution au problème en tentant de « forcer » une sémantisation du balisage HTML à travers un nouveau jeu d'éléments structurels (tout en maintenant la rétro-compatibilité) avec de nouvelles balises : `main`, `section`, `article`, `header`, `footer`, `aside`, ou encore `nav`. Si le standard en lui même semble largement adopté (89,5% des sites selon une étude récente [66]) force est de constater que la validité du code HTML (sans parler du CSS) est beaucoup moins répandue qu'il ne pourrait y paraître, y compris sur les sites web les plus populaires [67]. L'engouement qui a accompagné la mise en place du standard [68] n'a pas amené une véritable révolution dans l'extraction de contenu dans le sens où cette tâche n'est pas plus facile depuis HTML5 qu'avant, à la fois du fait de méthodes explicites de lutte *anti-scraping* [Haque and Singh, 2015] mais aussi tout simplement du fait de choix très pragmatiques d'adoption partielle des standards [Weninger et al., 2016]. Les « producteurs » de pages Web n'adoptent pas tous les aspects des standards loin de là. La réalité est que cet espoir de fonder des analyses sur des méta-données déclaratives, et donc de faire confiance à l'émetteur s'avère infondé, des balises censées résoudre ces questions, comme la balise `article` n'offrent que des garanties partielles de confiance et de couverture. À ce titre, les réflexions de Cory Doctorow sur les *7 facts about metacrap* [69] il y a deux décennies semblent toujours d'actualité bien que déprimantes. Cory Doctorow y prophétise qu'« un monde de méta-données exhaustives et fiables est une utopie » mais aussi « une chimère fondée sur un auto-aveuglement, un orgueil technophile et des opportunités commerciales artificiellement exagérées » [70]. Le propos est développé en insistant sur deux types de limites à la confiance que l'on peut accorder aux méta-données :

1. des limites liées aux individus fournissant les méta-données qui mentent, peuvent être fainéants ou stupides et en général ne connaissent pas les limites de leur savoir
2. des limites plus méthodologiques en ce sens que les schémas de descriptions ne sont ni neutres ni incontestables et que la manière de mesurer influence les résultats.

Pour ce qui m'intéresse ici, je pense que l'assertion numéro 2 de la liste (*people are lazy*) suffit à expliquer cette faible fiabilité d'une approche fondée uniquement sur les méta-données déclaratives, puisque nous sommes dans un processus où l'usage a favorisé l'émission des messages

65. `https://chromium.googlesource.com/chromium/dom-distiller` consulté le 2 octobre 2023
66. `https://w3techs.com/technologies/details/ml-html5` consulté le 2 octobre 2023
67. `https://theseosystem.com/html-css-validation-statistics-10-biggest-websites-world/` consulté le 2 octobre 2023
68. Voir par exemple un article d'*InfoWorld* de 2010 qui illustre ces espoirs initiaux : `https://www.infoworld.com/article/2627336/html5-how-html5-will-change-the-web.html` consulté le 2 octobre 2023
69. `https://people.well.com/user/doctorow/metacrap.htm` consulté le 2 octobre 2023
70. La citation originale *A world of exhaustive, reliable metadata would be a utopia. It's also a pipe-dream, founded on self-delusion, nerd hubris and hysterically inflated market opportunities* (traduction personnelle)

au détriment de la réception, motivant donc le développement de solutions adaptées au fait que dans le monde réel du web il y a bien plus d'une manière de décrire quelque chose (réflexion 7 de Cory Doctorow). Ces solutions seront décrites dans les deux prochaines sous-sections.

4.2.1 Bref état de l'art sur le *web scraping*

Nous pouvons définir la tâche visée ici de la façon suivante : étant donné le code source d'une page web, le processus d'extraction de contenu consiste à détourer le contenu textuel utile (c'est-à-dire notamment sans les éléments de structure, la publicité ou encore les commentaires) et à identifier les méta-données. Concrètement, cette tâche implique une conversion du format HTML vers un autre format (souvent plein texte ou XML). Sachant que l'usage des modèles de page (*templates*) est très commun sur le Web [Bar-Yossef and Rajagopalan, 2002], bon nombre d'approches de l'extraction de contenu réalisent plus ou moins directement une détection du *template* par exemple via l'identification de contenu dupliqué entre différentes pages d'un même site [Rae et al., 2018]. Assez classiquement, les approches vont se fonder soit sur des heuristiques prédéfinies (des règles si l'on peut dire) ou sur l'application de méthodes de *machine learning* qui permettront de classer des blocs HTML comme faisant partie ou non du contenu. On rencontrera aussi des hybridations entre ces deux approches. Dans la première catégorie on trouve notamment des approches fondées sur la recherche de propriétés explicites des segments textuels [Kohlschütter and Nejdl, 2008], sur l'utilisation de segmentation visuelle [Cai et al., 2003] ou encore l'exploitation du code HTML proprement dit via le DOM (Document Object Model). L'idée, assez intuitive, étant de considérer le document HTML comme un arbre et de définir la tâche d'extraction comme la détection des nuds qui représentent du texte.

Il s'avère que les heuristiques sur des propriétés attendues du texte lui même (mots, caractères ou ponctuation) et la densité des blocs en balises ou en liens sont très efficaces pour discriminer les blocs textuels et non textuels. Historiquement, de premières approches plus ou moins indépendantes des langues sont apparues par exemple en appliquant des mesures d'entropie sur les liens, les balises et sur le texte lui même [Kao et al., 2004]. On peut citer ensuite la méthode *Content Extraction via Tag Ratios* (CETR) [Weninger et al., 2010] ou aussi *Content Extraction via Text Density* (CETD) [Sun et al., 2011]. Exploiter ces heuristiques pour sélectionner plus efficacement les nuds « de contenu » est apparu comme une amélioration de ces approches [Qureshi and Memon, 2012]. Par exemple, exploiter la sémantique des balises HTML (les balises paragraphes en particulier) et appliquer en second filtre les mesures de densité en balises forme une hybridation efficace [Carey and Manic, 2016].

Dès lors que l'on définit cette tâche comme une tâche de classification binaire, blocs de texte ou de non-texte, l'exploitation de méthodes d'apprentissage s'avère assez naturelle. On trouve notamment des champs conditionnels aléatoires (CRF ou *conditional random fields*) exploitant les propriétés des blocs en terme de balisage et de contenu en mots [Spousta et al., 2008]. Des SVM (*Support vector machines*) entraînées sur des propriétés linguistiques, structurelles et visuelles ont été testées dans [Bauer et al., 2007]. Plus récemment des approches d'apprentissage profond, des CNN (*Convolutional Neural Networks*) en l'espèce ont été appliqués en exploitant des propriétés du DOM [Vogels et al., 2018].

4.2.2 Échantillon de test d'outils de *Web Scraping*

Afin de circonscrire le champ des outils exploités pour cette évaluation, nous nous sommes focalisés sur les outils librement disponibles et en langage Python. Les outils propriétaires répondent à d'autres problématiques que celles de nos recherches et ne facilitent généralement

pas la reproductibilité des résultats (le versionnage des outils étant souvent moins précis) ou la pérennité des usages (au contraire des codes libres de droit comme on peut le voir avec la grande longévité de Boilerpipe). Bien entendu certains outils seront disponibles en Javascript ou en Java mais en général s'ils ont du succès, ou une certain efficacité, ils bénéficient d'un portage en Python ou au moins d'un *wrapper*.

Les outils que nous évaluons font partie de différentes catégories. Les outils de la première série (Type I) ont un fonctionnement qui peut s'apparenter à une sorte de copier-coller effectué sur le rendu de la page, de sorte que le rappel est généralement très élevé (il s'agit d'extraire tout le texte disponible) et la structure assez plate. Voici la liste des outils de cette catégorie qui seront traités ici :

— Html2text (HTML.2T) est véritablement l'outil le plus simple et le plus ancien, il est utilisable en ligne de commande comme en Python
— Inscriptis (INSCRI) peut se concevoir comme une amélioration d'Html2text avec une recherche de précision améliorée et une prise en compte de certains éléments structurels tels que les tableaux.

Arrivent ensuite des outils véritablement optimisés pour l'extraction du texte et en particulier pour ce qui est des articles de presse publiés sur le Web, ils se divisent en deux catégories. D'une part, ceux pour lesquels l'extraction du texte vise avant tout la lisibilité (Type II) que ce soit pour l'adaptation aux contraintes d'un terminal (écran de téléphone par exemple) ou pour la génération de version imprimable :

— Newspaper (NPAPER) est explicitement dédié aux articles de presse mais sans focalisation sur la structure ;
— News-please (NPLEASE) est un outil d'extraction de texte adossé à un web crawler dédié aux articles de presse [Hamborg et al., 2017] ;
— Python-Readability (READ) est une adaptation en Python de la bibliothèque Readability utilisée originellement dans Firefox pour afficher des articles de presse sans publicité ni contenu périphérique ;

Enfin, nous avons les outils véritablement dédiés à l'extraction de contenu textuel pour la constitution ou l'analyse subséquente de corpus (Type III) :

— Boilerpy3 (BP3) est une version Python du célèbre algorithme `boilerpipe` [Kohlschütter et al., 2010] destiné à l'élimination du squelette de page (*boilerplate removal* et l'extraction du contenu principal (*main text*), il s'agit d'un système par règles. Il existe dans différentes configurations, BP3 désignera la version par défaut, BP3_Art la version pour les articles de presse, BP3_KeepE celle orientée rappel (*KeepEverything*) et enfin BP3_Larg (*Largest*) pour une version moins bruitée de BP3_KeepE ;
— Goose3 (GOO) s'oriente vers l'extraction du contenu inséré dans un squelette sans intérêt particulier pour la structure ou les balises ;
— JusText (JT) est un outil initialement conçu pour la dé-duplication de pages web et dont l'usage a ensuite été étendu à l'extraction de blocs de textes à des fins de développement de ressources linguistiques [Pomikálek, 2011]. En plus de la version indépendante de la langue (JT), nous examinons JT_en qui est la version par défaut dans les tutoriels et enfin JT_truelg pour la version où la langue du document est donnée en amont ;
— Trafilatura (TRAF) est un outil complet qui opère sur toute la chaîne de traitement allant de la récupération des url, des pages web puis l'extraction du texte sous forme structurée dans différents formats [Barbaresi, 2021]. En plus de la configuration de base nous ajoutons la configuration *fallback* (TRAF_FB) qui utilise les résultats de JT et READ comme recours quand l'extraction avec les heuristiques de base échoue.

Ces outils sont utilisés tels quels (*out-of-the-box*) ou avec un paramétrage minimal, généralement

Cat.	Outil	Version	Adresse Github	Référence
I	HTML2TEXT	2020.1.16	Alir3z4/html2text	
I	INSCRIPTIS	1.0	weblyzard/inscriptis	
II	NEWSPAPER3K	0.2.8	codelucas/newspaper	
II	NEWS-PLEASE	1.4.25	fhamborg/news-please	[Hamborg et al., 2017]
II	READABILITY	0.7.1	buriy/python-readability	
III	BOILERPY3	1.0.2	jmriebold/BoilerPy3	[Kohlschütter et al., 2010]
III	DRAGNET	2.0.4	dragnet-org/dragnet	[Peters and Lecocq, 2013]
III	GOOSE3	3.1.6	goose3/goose3	
III	JUSTEXT	2.2.0	miso-belica/jusText	[Pomikálek, 2011]
III	TRAFILATURA	0.4.1	adbar/trafilatura	[Barbaresi, 2019]

TABLE 4.1 – Versions des outils utilisés dans cette expérience, classification proposée : outils orientés rappel (I), orientés lisibilité(II), spécifiquement dédiés à la tâche (III)

mis en uvre si les résultats s'avèrent trop éloignés de l'attendu et don probablement biaisés. Par exemple, un outil tel que JUSTEXT est sensible au fait de connaître ou non en amont la langue du texte et c'est à l'usage que l'on se rend compte que le modèle par défaut (conçu pour l'anglais) peut donner des résultats très décevants. On peut voir à ce titre les résultats de *JT_en* dans le tableau 4.6 page 89. Les codes ayant présidés à ces expériences sont présentés dans le projet WADDLE[71]. Les versions utilisées sont présentées dans le tableau 4.1.

Dans le tableau 4.2 sont présentés les temps de traitement du corpus complet pour les outils les versions les plus communes des outils étudiés ici. La dernière colonne présente le ratio de temps de traitement par rapport par rapport à l'outil le plus rapide (en l'occurrence INSCRI) qui est capable de traiter environ 5 000 documents par minute soit 700 000 documents par jour[72]. On voit ensuite que nous avons toute une série d'outils (BP3 et ses variantes ainsi que JT_EN) dont le temps de calcul sera environ deux fois plus élevé (350 000 documents/jour). READ HTML2T, NPAPER et JT_TRUELG affichent eux des temps de traitement plutôt 3 à 6 fois plus longs que la référence (entre 100 000 et 200 000 documents/jour). Enfin, GOO et JT (la variante de JUSTEXT indépendante de la langue) tomberaient nettement sous les 80 000 documents/jour tandis que NPLEASE paye certainement le fait qu'il n'est pas du tout conçu pour traiter de grandes quantités de données (moins de 4 000 documents/jour). Cette dimension du temps de calcul montre qu'une partie de ces outils supporteraient mal le passage à l'échelle sur de très gros jeux de données. Probablement que ces approches sont de toutes façons inadaptées pour participer à la constitution des très larges corpus utilisés notamment pour les modèles de langue de type BERT qui nécessitent des architectures très fortement optimisées [Abadji et al., 2021].

4.2.3 Corpus de référence pour l'évaluation du *scraping*

Nous reprenons un corpus que nous avions constitué ([Lejeune and Zhu, 2018]) qui comprend près de 1.700 documents en 5 langues (475 en anglais, 405 en chinois, 273 en grec, 274

71. `https://github.com/rundimeco/waddle` consulté le 2 octobre 2023
72. Cette étude a été menée sans parallélisation sur un ordinateur portable avec 4 curs 2,5 Ghz et 32 Go de RAM

Outil	Temps (secondes)	Ratio/plus rapide
INSCRI	**19,7**	x1
BP3_KeepE	37,5	x1,9
BP3_Larg	37,7	x1,9
BP3	38,1	x1,9
BP3_Art	39,8	x2,0
JT_en	41,5	x2,1
READ	56,8	x2,9
HTML2T	71,0	x3,6
NPAPER	105,5	x5,5
TRAF	109,9	x5,6
JT_langid	112,6	x5,7
GOO	191,3	x9,7
JT	322,0	x16,3
NPLEASE	3755,6	x190

Table 4.2 – Temps de traitement du corpus complet (1700 documents) pour chacun des outils testés, rangés par ordre croissant (moyenne sur 10 runs)

Données	# Lignes	# Tokens	# Caractères
Html brut (Brut)	1385 (±1303)	4726 (±3921)	75015 (±51924)
Vérité de terrain (VT)	13 (±10)	321 (±323)	2296 (±1982)
Ratio (VT/Brut)	0,9%	6,8%	3,1%

Table 4.3 – Taille du corpus (Html originaux et vérité de terrain), les balises sont comptées comme des tokens de même que les variables et attributs CSS/Javascript, le ratio entre le Html et le texte à extraire est également indiqué

en polonais et 267 en russe) pour lesquels figure la version HTML d'une part et une version de référence nettoyée manuellement d'autre part. Ce corpus de référence a été constitué à partir d'une sous-partie du corpus Daniel-dataset [Mutuvi et al., 2020b] pour laquelle les documents Html étaient encore disponibles en ligne. Le tableau 4.3 présente quelques statistiques sur ce corpus. Nous pouvons voir que d'un point de vue de classification, on s'attendra à devoir écarter beaucoup de lignes et que celles à écarter comporteront probablement peu de tokens. Les documents de ce corpus ont été collectés en 2011 et 2012 pour l'évaluation du système de veille épidémiologique Daniel. Il est à noter que le standard Html 5 standard datant d'une recommandation W3C de 2014, ces documents sont pour la plupart basés sur Html 4. Néanmoins, ce corpus reste à notre connaissance la vérité de terrain la plus grande et la plus variée en langues.

4.2.4 Mesures d'évaluation de la qualité du *Web Scraping*

Les mesures d'évaluation de la campagne Cleaneval [Baroni et al., 2008] sont fondées sur la préservation des séquences de tokens. Bien qu'imparfaites, elles ont le mérite d'être globalement utilisées par la communauté scientifique [Weninger et al., 2016]. Nous ajoutons une mesure plus simple, fondée sur la préservation du vocabulaire, qui donne des résultats tout à fait comparables. Cette évaluation nécessite une vérité de terrain, que nous appellerons GT et GT_{tok} pour la séquence de tokens correspondante. Nous nommons RES le résultat de l'extraction automa-

tique et RES_{tok} la séquence de tokens correspondante, en reprenant le tokeniseur fourni par Cleaneval. La mesure Cleaneval vérifie à quel point la séquence de tokens extraite automatiquement (RES_{tok}) est similaire à la séquence de référence (GT_{tok}). L'algorithme de Ratcliff/Obershelp [Ratcliff and Metzener, 1988] est utilisé pour détecter les plus longues séquences de tokens communes et non-redondantes, sa complexité quadratique est peu efficace et ses résultats ne sont pas immédiatement interprétables. Notre mesure plus simple (`occ_eval`) vérifie si le nombre d'occurrences des tokens correspond aux nombre d'occurrences attendues tandis que `voc_eval` correspond à la proportion de vocabulaire qui est conservée. C'est en quelque sorte un taux de lexicalité en fonction de la référence

Ce comparatif a également fait l'objet d'une démonstration à la conférence TALN [Lejeune and Barbaresi, 2020], ces résultats peuvent être reproduits en utilisant les données et scripts mis à disposition sur Waddle. Nous optons ici pour une version abrégée : une seule des configurations de Boilerpipe (BP3_Article), configuration par défaut pour Justext et Trafilatura. Enfin, nous avons avec Adrien Barbaresi et Emmanuel Giguet organisé un tutoriel à TALN 2021 (X-COTE) dédié aux problèmes méthodologiques et pratiques posés par ces outils. Ce tutoriel a réuni une trentaine de chercheurs académiques et quelques industriels sur une demi-journée [73].

4.3 Évaluation intrinsèque de la qualité du *web scraping*

Nous détaillons ici des résultats sur l'analyse des outils d'extraction de contenu textuel complémentaires de celle menée dans [Barbaresi and Lejeune, 2020a], nous nous concentrons sur la variabilité dans les résultats des outils.

Afin d'entrer directement dans le vif du sujet, le tableau 4.4 présente la taille de l'*output* global pour chacun des outils. On peut voir ici que les outils orientés rappel renvoient un contenu beaucoup plus conséquent, INSCRI semblant, comme attendu, être plus sélectif que HTML2T. Les résultats de la variante orientée rappel de BP3 (BP3_KeepE) se situent dans les mêmes ordres de grandeur. Le reste des outils peut se répartir en deux parties. D'un côté nous avons les outils très sélectifs (NPAPER, NPLEASE, GOOSE et JT_en) qui renvoient en moyenne moins de 10 lignes (avec un écart-type faible), ce qui est inférieur à la moyenne attendue de 13 lignes (Tableau 4.3). La taille moyenne en tokens est plus proche de l'attendu mais reste notablement inférieure (respectivement 205, 262, 203 et 170 VS 321 pour la référence). Les autres outils renvoient plus de lignes que l'attendu (de 14 à 22 lignes en moyenne VS 13 pour la référence) et un nombre de tokens plus proche des 312 attendus en moyenne (entre 286 pour BP3_Larg et 381 pour TRAF_FB).

Afin de présenter une évaluation plus précise de ces variations, en particulier sur les aspects multilingues, le tableau 4.5 présente pour chaque langue du corpus et pour chaque outil la proportion de documents pour lesquels la sortie est vide ou dont la taille est inférieure à 10% de l'*output* attendu. Un premier enseignement est que les outils de la catégorie orientés rappel (I) ne renvoient jamais de fichiers vides et que les outils de la catégorie orientés lisibilité (II) sont plus sujets à la variabilité de qualité entre les différentes langues du corpus. Deux exceptions à cela : READ qui obtient des résultats proches des outils de la catégorie III avec une certaine constance selon les langues tandis que GOO est de tous les outil dédiés à la tâche celui qui est le moins multilingue (à l'exception de JT_en qui est explicitement un système monolingue). Sur cet aspect multilingue, un deuxième enseignement est à lire en colonne cette fois : l'anglais est de très loin la langue où l'on attendra le moins de silence avec, dans une moindre mesure, le

73. X-COTE : Extraction de Contenus Textuels du Web `https://talnrecital2021.inria.fr/X-COTE/` consulté le 2 octobre 2023

Cat.	Outil	NB lignes	NB tokens	NB caractères
(I)	HTML2T	336 (±200)	1586 (±1307)	21246 (±13728)
(I)	INSCRI	248 (±176)	1421 (±1198)	22540 (±35889)
(II)	NPAPER	8 (±12)	205 (±314)	1305 (±2017)
(II)	NPLEASE	8 (±10)	262 (±352)	1713 (±2265)
(II)	READ	37 (±78)	410 (±421)	3271 (±3042)
(III)	BP3	22 (±26)	381 (±492)	2654 (±3085)
(III)	BP3_Ar	16 (±20)	314 (±352)	2283 (±2319)
(III)	BP3_KeepE	188 (±133)	1192 (±1009)	8259 (±6331)
(III)	BP3_Larg	14 (±22)	286 (±345)	2050 (±2265)
(III)	GOO	6 (±10)	203 (±297)	1303 (±2093)
(III)	JT	14 (±17)	383 (±500)	2513 (±3092)
(III)	JT_en	6 (±14)	170 (±435)	1013 (±2552)
(III)	JT_trueLg	14 (±17)	378 (±496)	2479 (±3073)
(III)	TRAF	21 (±22)	348 (±383)	2462 (±2360)
(III)	TRAF_FB	23 (±25)	381 (±370)	2696 (±2247)

TABLE 4.4 – Variation de la taille globale de l'output des différents outils ordonnés par catégorie : orientés rappels (I), orientés lisibilité (II) et dédiés à l'extraction proprement dite (III)

polonais. On peut soupçonner ici une influence du système d'écriture.

Le tableau 4.6a présente les résultats globaux avec la métrique `clean_eval`. La micro-précision et le micro-rappel sont des moyennes des précisions et rappels par document. La macro-précision et le macro-rappel sont des moyennes des résultats par sous-corpus, par langue donc ici, ce qui diminue l'importance relative de l'anglais. Le tableau 4.6b présente les résultats obtenus avec `occ_eval`, ceux-ci diffèrent assez peu. L'ordre de grandeur des résultats et la hiérarchie entre les outils sont conservés, à ceci près que certains outils à fort rappel semblent pénalisés par la mesure `clean_eval`.

4.3.1 Variation selon les langues

D'un point de vue général, l'outil le plus fiable semble être BP3_ART, READABILITY, TRAFILATURA et JUSTEXT se situant juste derrière. Toutefois, les moyennes (micro ou macro) masquent des différences entre les langues, comme nous le montrons dans les tableaux 4.7a à 4.7c qui présentent les résultats sur les sous-corpus anglais, russe et chinois pour une sélection d'outils parmi les plus efficaces. Nous avons marqué en grisé les performances des 4 outils les plus efficaces sur le corpus multilingue, ce qui permet de voir qu'ils sont bien placés, sauf sur le sous-corpus anglais où des outils très spécialisés sont plus performants. L'anglais est évidemment la langue où l'on trouve les meilleurs résultats puisque pour 9 des 11 systèmes testés on a une F-mesure au dessus de 80% (contre seulement 2 en grec et 3 en polonais). En ce qui concerne les performances par outil, BP3_ART, le meilleur outil selon la micro-moyenne générale, est inégal selon la langue traitée : très efficace comparativement aux autres sur le chinois, il se situe un ton en dessous de ses concurrents pour ce qui est de la langue russe. JUSTEXT s'impose sur cette langue, ce qui semble valider la robustesse de son approche multilingue fondée sur les mots outils. Ses résultats sont compétitifs sur l'anglais et c'est sans doute sur le chinois qu'il perd la confrontation à distance avec BP3_ART. En effet les modèles langagiers de JUSTEXT sur les mots outils ne

Cat,	Outil	el	en	pl	ru	zh	Macro-moy	Micro-moy
(I)	HTML2T	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
(I)	INSCRI	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
(II)	NPAPER	**95,2%**	0,8%	**21,9%**	**95,4%**	**19,9%**	**46,7%**	**38,8%**
(II)	NPLEASE	**46,5%**	0,8%	5,1%	**65,0%**	**15,1%**	**26,5%**	**22,3%**
(II)	READ	0,0%	1,3%	0,4%	0,0%	0,2%	0,4%	0,5%
(III)	BP3	**31,9%**	6,9%	2,2%	5,7%	0,5%	9,4%	8,5%
(III)	BP3_Art	**30,8%**	7,1%	2,2%	5,7%	0,5%	9,3%	8,3%
(III)	BP3_KeepE	0,0%	3,6%	0,7%	1,9%	0,0%	1,2%	1,4%
(III)	BP3_Larg	**30,8%**	6,9%	2,2%	5,7%	0,5%	9,2%	8,3%
(III)	GOO	**99,3%**	1,3%	**11,3%**	**65,4%**	**18,1%**	**39,1%**	**32,7%**
(III)	JT	1,8%	3,8%	0,0%	0,4%	**18,9%**	5,0%	5,9%
(III)	JT_en	**98,2%**	3,8%	**99,3%**	**99,6%**	**19,9%**	**64,1%**	**53,3%**
(III)	JT_trueLg	1,8%	3,8%	0,0%	0,4%	**18,9%**	5,0%	5,9%
(III)	TRAF	**12,5%**	2,3%	3,6%	2,7%	0,5%	4,3%	3,8%
(III)	TRAF_FB	0,0%	0,0%	0,0%	0,0%	0,2%	0,0%	0,1%

TABLE 4.5 – Proportion de documents vides ou quasi-vides (taille<10% de la taille attendue) pour chacun des outils analysés, ordonnés par catégorie. Les proportions supérieures à 10% sont indiquées en gras.

sont pas applicables à la langue chinoise. Si l'on partait des résultats sur l'anglais pour choisir un outil d'extraction de contenu, nous pourrions être tentés de choisir NEWSPAPER ou encore GOOSE. Mais leurs performances sont très variables, en particulier pour le grec (moins de 6% de F-mesure avec un rappel très faible). NEWSPAPER et NEWSPLEASE apparaissent véritablement spécialisés sur l'anglais.

4.3.2 Visualisation de la variation à l'échelle des documents

Afin de mieux visualiser ces variations nous présentons dans la Figure 4.2 les résultats par document pour le corpus global pour les 6 outils les plus performants du point de vue monolingue et multilingue. Les écarts types sont plutôt élevés en général ($\pm$24 sur la précision pour JUSTEXT par exemple, ou sur des sous-corpus particuliers, $\pm$17 sur la précision de GOOSE sur l'anglais). Les graphiques permettent de saisir d'un coup d'il l'importance de cette variabilité. On peut ainsi observer que les documents en anglais sont mieux traités (en bleu). Les points correspondant au grec (en gris) sont peu nombreux, ce qui correspond à un plus faible nombre de sources (les points groupés sur la même abscisse). La dispersion des points et les codes couleur permettent ainsi de saisir des informations sur la composition du corpus lui-même, et d'en déduire ici qu'il peut être intéressant de tenir compte du nombre de documents par source ou de retenir la macro-moyenne sur les sources.

En regard de ces graphiques, le comportement des outils peut être classé en trois catégories distinctes. GOOSE est plutôt efficace en termes de précision, d'où le grand nombre de points situés à droite du graphique mais aussi les problèmes afférents bien visibles par des pics de faible rappel. Au contraire, JUSTEXT offre un bon rappel pour plus de documents, d'où le grand nombre de points en haut de la courbe. Enfin, READABILITY laisse apparaître une diagonale, ce qui suggère plus de résultats équilibrés entre la précision et le rappel et explique pour partie les bons résultats

Outil	Macro-F	Macro-Préc.	Macro-Rappel	Micro-F	Micro-Préc.	Micro-Rappel
TRAFFallback	**82,77** (±4, 1)	77,87 (±5, 5)	88,31 (±3, 5)	**83,33**	78,32 (±25)	89,02 (±19)
BP3Article	79,12 (±7, 4)	81,29 (±7, 0)	77,07 (±9, 1)	80,25	82,29 (±26)	78,31 (±35)
BP3Largest	77,48 (±5, 0)	**83,62** (±4, 4)	72,18 (±6, 9)	78,12	84,37 (±26)	72,73 (±35)
READABILITY	76,54 (±5, 8)	68,27 (±7, 7)	87,10 (±4, 8)	76,64	68,15 (±22)	87,56 (±19)
READ_py	76,05 (±8, 4)	66,39 (±10, 9)	**89,00** (±3, 6)	74,77	64,61 (±35)	88,72 (±17)
JT	75,80 (±36, 7)	81,35 (±6, 9)	70,95 (±38, 9)	73,98	81,70 (±25)	67,59 (±42)
BP3	73,72 (±7, 3)	73,93 (±8, 5)	73,50 (±14, 6)	74,32	75,04 (±24)	73,61 (±33)
TRAF	72,69 (±11, 4)	68,87 (±10, 0)	76,95 (±13, 6)	74,73	70,71 (±31)	79,24 (±33)
NEWSPLEASE	63,80 (±32, 2)	83,16 (±7, 4)	51,76 (±35, 7)	65,38	**84,39** (±21)	53,36 (±46)
GOO	51,07 (±40, 5)	80,37 (±10, 7)	37,42 (±39, 7)	54,80	82,25 (±22)	41,09 (±44)
HTML-text	47,94 (±6, 2)	32,45 (±5, 4)	91,68 (±1, 5)	48,58	33,04 (±19)	91,73 (±9)
NEWSPAPER	46,07 (±43, 5)	77,06 (±11, 8)	32,86 (±43, 4)	51,16	79,38 (±21)	37,74 (±45)
INSCRIPTIS	42,36 (±6, 6)	27,09 (±5, 2)	97,08 (±2, 1)	42,68	27,32 (±17)	**97,43** (±9)
HTML2TEXT	34,08 (±10, 2)	20,82 (±7, 3)	93,78 (±2, 8)	34,35	21,01 (±16)	94,13 (±13)
JT_english	28,74 (±36, 6)	73,11 (±9, 4)	17,88 (±37, 7)	37,24	75,10 (±19)	24,76 (±40)

(a) Mesure `clean_eval`

Outil (différence p.p)	Macro-F	Macro-Préc.	Macro-Rappel	Micro-F	Micro-Préc.	Micro-Rappel
TRAFFallback (−1, 65)	**81,12** (±8, 8)	75,47 (±11, 6)	**87,69** (±4, 5)	**81,12**	75,25 (±28)	87,99 (±16)
BP3Article (−1, 72)	77,40 (±8, 4)	78,72 (±8, 2)	76,12 (±9, 9)	77,93	78,95 (±25)	76,94 (±32)
BP3Largest (−1, 27)	76,21 (±8, 3)	**81,48** (±7, 0)	71,57 (±9, 6)	76,39	**81,54** (±24)	71,85 (±33)
JT (+0, 29)	76,09 (±29, 4)	78,54 (±9, 2)	73,79 (±34, 7)	74,34	78,27 (±23)	70,80 (±38)
BP3 (−0, 08)	73,64 (±8, 4)	72,22 (±7, 3)	75,12 (±14, 1)	73,93	72,66 (±23)	75,25 (±31)
READ_py (−3, 39)	72,66 (±17, 6)	64,45 (±16, 5)	83,27 (±19, 3)	70,59	62,16 (±33)	81,67 (±24)
TRAF (−0, 75)	71,94 (±11, 0)	67,79 (±11, 3)	76,63 (±12, 3)	73,30	68,74 (±31)	78,50 (±31)
READABILITY (−5, 74)	70,80 (±20, 8)	62,46 (±22, 1)	81,71 (±13, 7)	69,64	61,08 (±29)	80,99 (±21)
NEWSPLEASE (+0, 93)	64,73 (±28, 1)	80,29 (±10, 3)	54,22 (±33, 2)	66,26	80,84 (±19)	56,14 (±43)
GOO (+1, 63)	52,70 (±37, 5)	77,44 (±12, 2)	39,94 (±38, 0)	56,43	78,67 (±19)	43,99 (±43)
NEWSPAPER (−1, 75)	47,82 (±41, 3)	74,09 (±12, 0)	35,31 (±42, 1)	52,87	75,77 (±18)	40,60 (±44)
HTML-text (−0, 06)	39,88 (±19, 8)	26,73 (±14, 1)	78,52 (±30, 9)	38,96	26,20 (±21)	75,98 (±31)
INSCRIPTIS (−5, 02)	37,34 (±15, 7)	23,27 (±11, 0)	94,47 (±6, 6)	36,64	22,74 (±18)	**94,15** (±12)
HTML2TEXT (−0, 92)	33,16 (±13, 2)	20,25 (±9, 0)	91,49 (±9, 1)	33,12	20,24 (±17)	91,01 (±14)
JT_english (+2, 62)	31,36 (±36, 1)	69,94 (±7, 7)	20,21 (±37, 4)	39,78	71,37 (±15)	27,58 (±39)

(b) Mesure `occ_eval` (avec entre parenthèses la différence en points de pourcentage par rapport à `clean_eval` en macro F-mesure). Les lignes sur fond gris indiquent les différences d'ordre dans le classement.

Table 4.6 – Extrait de l'évaluation sur le corpus multilingue (tableau complet page 138) en Précision (Préc.), Rappel et F-mesure (F) avec les macro-moyennes (sur les langues) et micro-moyennes. NB : la micro F-mesure est calculée à partir des micro-précision et micro-rappel de sorte qu'il n'y a pas d'écart-type

en F-mesure, et notamment en chinois (en rouge dans la Figure 4.2). Trafilatura se manifeste par une dispersion des points plus homogène que les autres outils, signe que l'outil n'a pas de réel point faible mais pas non plus de point fort. Cette performance est mesurée par le meilleur résultat en macro-moyenne, tandis que les problèmes de précision en anglais notamment visibles ici permettent d'expliquer la performance plus faible capturée par la micro-moyenne.

4.4 Conclusion intermédiaire : exploitabilité de données textuelles issues du web

La collecte et l'usage de données web sont sujets à une série de problèmes éthiques, méthodologiques et épistémologiques qui méritent l'attention de la communauté scientifique. Il appert que les approches opportunistes présidant à l'établissement de grands corpus tirés du web ne sont pas sans poser un certain nombre de difficultés. Nous avons apporté des preuves empiriques de leur impact, tout d'abord en étudiant la forme des documents obtenus à travers la compa-

Outil	F-mes.	Préc.	Rap.
NPAPER	**91,32**	91,34	91,31
GOO	90,69	**92,94**	88,54
NPLEASE	88,91	87,89	89,96
READ	87,16	84,31	90,21
BP3_Art	87,00	87,50	86,51
JT	84,86	83,16	86,62
TRAF	82.58	74.28	92.97
INSCRI	45,84	29,88	**98,46**
HTML2T	44,61	28,98	96,84

(a) `occ_eval` (Anglais)

Outil	F-mes.	Préc.	Rap.
JT	**76.29**	71.64	81.59
READ	74.27	72.29	76.36
TRAF	71.20	64.80	79.02
BP3_Art	69.31	70.11	68.53
NPLEASE	42.64	93.16	27.64
GOO	40.24	90.96	25.83
INSCRI	32.53	19.77	**91.75**
HTML2T	29.55	17.63	91.35
NPAPER	5.14	**92.34**	2.64

(b) `occ_eval` (Russe)

Outil	F-mes.	Préc.	Rap.
BP3_Art	**63.30**	71.28	56.93
TRAF	55.48	46.81	68.09
READ	42.36	48.00	37.91
GOO	20.60	82.54	11.77
JT	19.19	82.32	10.86
NPAPER	19.17	82.72	10.84
HTML2T	13.83	7.62	74.87
NPLEASE	13.31	**97.52**	7.14
INSCRI	12.97	7.06	**79.52**

(c) `occ_eval` (Chinois)

TABLE 4.7 – `occ_eval` par langue, sur fond gris les systèmes les plus performants sur le corpus complet avec les 5 langues

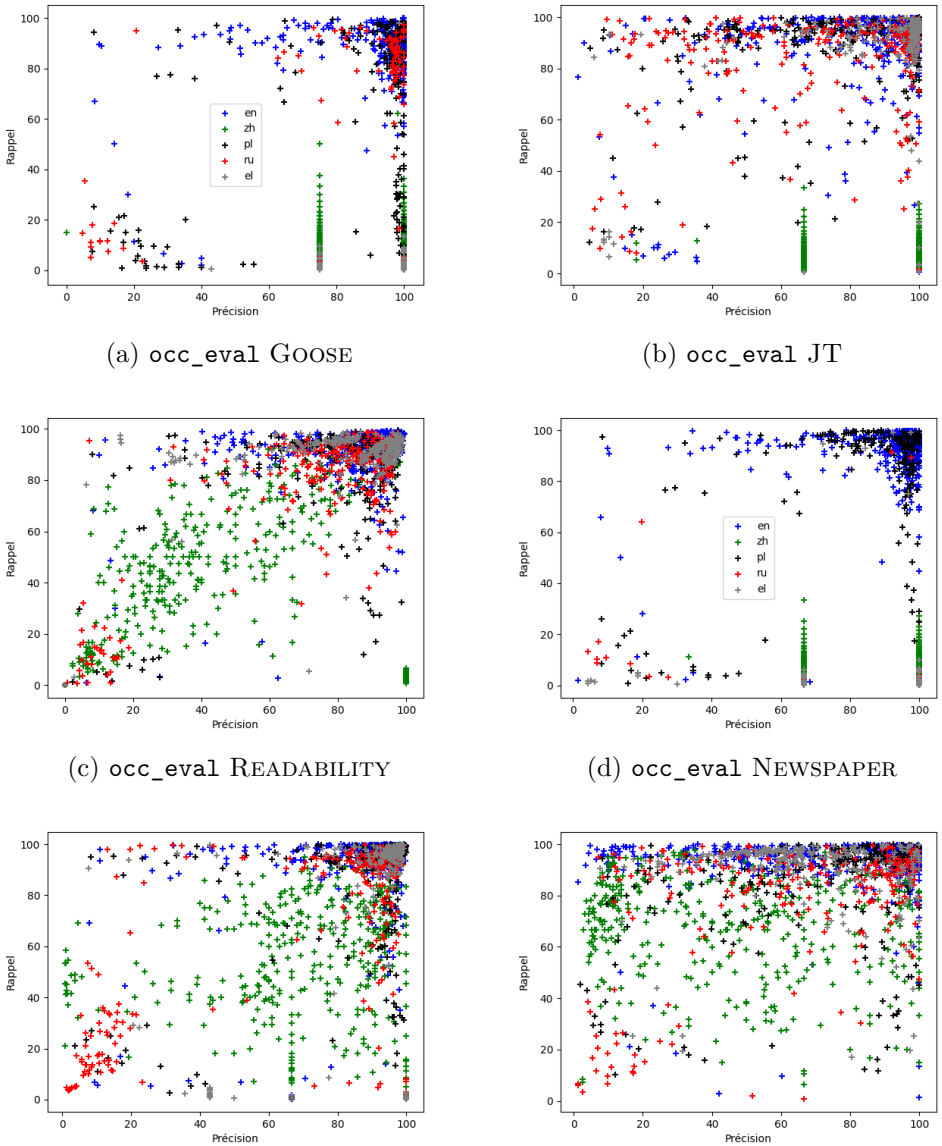
raison de méthodes d'extraction des données et ensuite en recensant des problèmes centrés sur le contenu des corpus et liés aux méthodes d'acquisition opportuniste des données. La faible supervision conduit à un *far west*, *« Wild West Web Crawling »* selon [Jo and Gebru, 2020], tandis qu'une approche plus supervisée et maîtrisée ne suffit pas à résoudre des problèmes posés par l'extraction de texte.

Au phénomène de dispersion des segments textuels visible sur les graphiques d'évaluation répond une probabilité élevée de cerner certaines communautés (hobby précis ou frange politique) et genres textuels (petites annonces et annuaires). Nous pouvons ainsi voir dans la Figure 4.3 un exemple d'articles en russe [74] présentant des blocs de type Faux Positif. Le premier est un bloc date + source (second bloc : 26.11.11), ces blocs avaient été considérés comme ne faisant pas partie du texte dans l'annotation. Ceci est discutable d'un certain point de vue mais permettait d'écarter les cas où la date ne serait pas présente dans le document où présente à un autre nud du HTML. Le second cas, plus problématique bien sûr est celui des commentaires ( ) qui posent des questions de stabilité du texte d'une part, dépendant donc du moment de récupération de la donnée, et de la sur représentation de certains éléments de langage. Enfin, nous avons dans ces vrais positifs toute une série de blocs sur les articles liés (: et suivants) ou sur les modalités de partage sur les réseaux sociaux. On voit donc que des textes ou éléments indésirables se trouvent dans des corpus destinés à la recherche en linguistique et en TAL, d'une part à cause de l'impossible contrôle des sources et adaptation à certains types de pages dès que la taille du corpus atteint un certain ordre de grandeur, et d'autre part en raison de l'application d'outils génériques et supposés adéquats sans vérification de leur efficacité pour des textes, langues ou sujets divergents, problématique connue en apprentissage artificiel par la notion d'adaptation de domaine.

Sur la forme, les corpus web peuvent receler des documents incomplets et tronqués ainsi que des doublons et des segments génériques, dans une proportion variable qui pourrait bien être inconnue ou mal estimée par la communauté scientifique. Par ailleurs, des problèmes de fond substantiels peuvent surgir. Si l'on prend l'exemple du document en grec de la Figure 4.4 [75], on observe qu'un certain nombre de blocs ont été écartés à tort (Faux Négatifs). Nous pouvons faire l'hypothèse qu'il s'agit d'une mauvaise interprétation du bloc A qui annonce en fait une liste d'items et qui a été interprété par JT comme l'annonce d'une séquence d'éléments trop réguliers pour être interprétés comme du contenu textuel. En effet, nous avons plusieurs

74. Document `20111128_krasnoturinsk.info_5a3bfe4c4a2968237c0aaee00b904f9d78a311ee826c1913c7b15bc0` du corpus DANIEL-SCRAPING, indisponible en ligne

75. Document `20111121_www.iatronet.gr_daefc5bd65a1d9515057b7a1afbf8d982af1d43e9005e12c065047f8` du Corpus DANIEL-SCRAPING, indisponible en ligne



(a) `occ_eval` Goose (b) `occ_eval` JT

(c) `occ_eval` Readability (d) `occ_eval` Newspaper

(e) `occ_eval` BP3_Article (f) `occ_eval` Trafilatura

Figure 4.2 – Visualisation mettant en rapport la précision (abscisse) et le rappel (ordonnée) pour chaque document du corpus. Les couleurs correspondent à la langue : *el* = grec (gris), *en* = anglais (bleu), *pl* = polonais (noir), *ru* = russe (rouge), *zh* = chinois (vert)

segments consécutifs, encodés dans le Html original comme des listes, débutant par ... (l'actionnaire et) et se terminant par des sommes d'argent ( ). Comme ces segments ne sont pas considérés comme des segments textuels, l'agorithme de Justext « disqualifie » aussi le segment annonciateur A : (analyse). Ici une bonne partie de l'information est donc manquante.

Pour conclure, il faudrait pouvoir non seulement décrire ces problèmes mais également les circonscrire, ce qui implique de trouver des façons de mesurer ainsi que des heuristiques de limitation. Alors que le détourage peut être évalué et résolu par des approches quantitatives (comprenant étalons et métriques), les difficultés d'ordre qualitatif sont plus difficiles à cerner et à

Vérité de Terrain	**Document détouré** (Boilerpipe)
, , 26 , […] , „ „ . […] , .[…] . , .[…] " !" . - - 1000 .[…]	**26.11.11** , , 26 , […] , „ „ . […] , .[…] . , .[…] " !" . - - 1000 .[…] " , . , ". [. . .] *. . . 6 autres blocs de commentaires* **:** . ... [. . .] *. . . 16 blocs divers*

Figure 4.3 – Exemples de document présentant des blocs de type Faux Positifs en russe (tiré de krasnoturinsk.info (.) du 28 novembre 2011

Vérité de Terrain	**Document détouré** (Justext)
Ι : Π Π 26.500 Ι , 5 Δ.Σ. Δ.Σ. , 30.000 . Α : Δ.Σ. […] **1.059,60** . Δ.Σ. . Γ Σ […] **7.200** . Δ Σ . Ε Π […] **7.200** . Δ Σ . Π Π […]**3.600** . Σ 30.000 ΙΑΣΩ Α.Ε 21.600 , Ε Μ , Δ Σ Γ Μ .[. . .]	Ι : Π Π 26.500 Ι , 5 Δ.Σ. Δ.Σ. , 30.000 .

Figure 4.4 – Exemples de document présentant des blocs de type Faux Négatifs en grec tiré de Iatronet (iatronet.gr) du 21 novembre 2011

étalonner, sachant que leurs potentielles conséquences éthiques voire pénales sont plus graves (des développements sur ce sujet sont proposés dans [Barbaresi and Lejeune, 2020b]). L'utilisation de méthodes d'évaluation extrinsèque du *scraping* en observant les résultats de classification [Lejeune and Zhu, 2018] ou les *word embeddings* [Chiu et al., 2016] tirés des corpus peuvent sans doute aider à améliorer les réponses à ces questions. La difficulté étant qu'il est difficile de distinguer ce qui relève à proprement parler de la donnée en *input* ou de la méthode, les effets de bord étant toujours possibles. En effet, si un document faux positif, c'est-à-dire considéré comme pertinent dans la version vérité de terrain du corpus, est très dégradé (par exemple le document devient vide ou presque) alors il peut devenir un vrai négatif. Dans [Lejeune and Zhu, 2018, Nguyen et al., 2020], nous avons ainsi montré qu'une détérioration du texte en entrée (mauvais détourage ou mauvaise océrisation) pouvait améliorer la précision. Inversement, il est possible de voir des documents faux négatifs devenir des vrais positifs lorsque des segments textuels erronés sont intégrés au document en entrée. Dans la veille, il arrive sur certains documents que l'ajout par erreur des commentaires de lecteurs ou des titres des articles liés amène des mots déclencheurs (par exemple des noms de maladie dans la veille épidémiologique). Ceci impose que lorsque l'on est « hors données de laboratoire » il faille raffiner l'opposition traditionnelle Vrais/Faux Positifs VS Vrais/Faux Négatifs. Ce sera un des sujets du prochain chapitre où les variations étudiées sont cette fois dues à l'OCR.

Chapitre 5

Comment exploiter des données non standards issues d'OCR ?

Sommaire

Au chapitre précédent, je me suis intéressé à la question de la qualité des corpus dans un cas de documents nativement numériques (ou nés numériques). J'ai cherché à montrer que bien que la tâche puisse semble aisée par certains aspects, extraire de documents Html le texte et rien que le texte est une tâche difficile. Dans ce chapitre, il est question d'un cas différent : la question de l'extraction de texte dans des documents numérisés, donc des images. Il s'agira d'une part de s'intéresser à la mesure de la qualité des documents. Qualité qui, rappelons le, peut être mesurée de manière intrinsèque, c'est-à-dire que l'on cherche à savoir à quel point le texte que l'on va analyser est fidèle à l'original, ou bien de manière extrinsèque, c'est-à-dire en fonction de la dégradation occasionnée sur les traitements automatiques réalisés.

Il sera question en premier lieu de corpus avec vérité de terrain, un cas « plus facile » dans lequel on peut effectivement mesurer l'impact du bruitage sur les résultats en comparant les résultats obtenus sur les données idéales aux résultats obtenus sur les données imparfaites ou bruitées. Nous pourrons donc comparer comme au chapitre précédent les résultats d'uné évaluation intrinsèque de l'extraction de contenu textuel, un cas supervisé grâce à la vérité de

terrain, aux résultats d'une évaluation extrinsèque, un cas semi-supervisé puisque la référence sera la sortie obtenue sur les données dites « propres ». Dans l'autre cas, en l'absence de vérité de terrain, il sera question de vérifier l'applicabilité de méthodes de TAL et des possibilités d'évaluation non-supervisée de la qualité des données.

Les données bruitées dont il est question ici sont de mon point de vue un cas particulier de données non standards. Ces données sont « non standards » dans la mesure où elles dévient par rapport à un attendu de langue : il s'agit d'une altération par rapport à un état de texte initial. On peut voir trois situations :

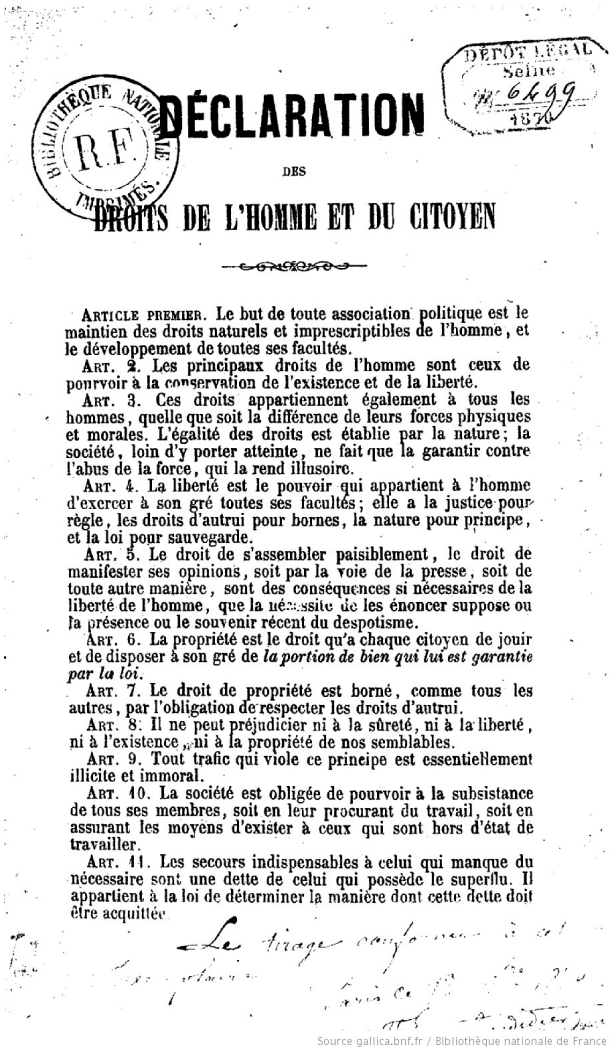
1. Cet attendu peut être connu (cas avec vérité de terrain),
2. L'attendu peut-être identifiable (cas où la qualité de la numérisation ne pose pas de problème à l'analyse à l'il nu)
3. L'attendu peut aussi être justifiable *a posteriori* (par analyse manuelle des résultats).

Bien entendu, dans le cas de données produites dans d'autres périodes temporelles, de nouvelles questions de non-standardisation se posent : variation orthographique, évolution du lexique (néologismes et archaïsmes) ou instablité interne du corpus. C'est ce dernier cas qui introduit l'hétérogénéité dans le sens où l'on aura différents états de langue au sein du même corpus, des variations aussi de la qualité, par exemple variation du taux de lexicalité. Ces questions seront traitées ici mais de manière accessoire, la problématique de chapitre restant donc celle du caractère bruité des données. Ce terme de bruité s'entend pour moi dans un sens d'altération des données, c'est-à-dire que la donnée émise est différente de la donnée reçue. Je ne traiterai pas particulièrement des problèmes de réception liés aux différences de réprésentation intervenant avec le temps mais plutôt des problèmes liés aux formes brutes.

Il ne s'agit donc pas de bruit au sens le plus commun où l'on aurait simplement une donnée ou une instance en trop, mais bien du bruit dans ses deux aspects : si quelque part il y a quelque chose en trop, il y a très probablement aussi quelque chose qui manque, qui a été remplacé. Une variation au grain caractère impliquera ainsi certainement une variation des mots. Le bruit dans la classe positive implique ainsi du silence dans la classe négative, le faux positif est aussi un faux négatif. D'un point de vue de recherche documentaire, cela semble assez évident. D'un point de vue d'extraction de textes cela signifie que l'on va avoir des séquences textuelles, des mots par exemple, qui disparaissent tandis que d'autres apparaissent. La figure 5.1 (page 95) présente côté à côte la version image de la Déclaration des Droits de l'Homme et du Citoyen disponible sur Gallica et une version océrisée par Kraken (avec le modèle du 17$^{\text{ème}}$ siècle développé par Simon Gabay dont nous parlerons ultérieurement).

On voit que l'entête pose problème, mêlant différentes graisses et tailles de police avec en sus un tampon. Mais si l'on s'intéresse simplement au contenu textuel on voit dès l'article premier que l'on va avoir du bruit et du silence puisque « ARTICLE PREMIER. » devient « ARIcs nsun, » et « Le but de toute association politique » est reconnu comme « Le but de teute asseciation politique ». Nous avons donc des tokens disparus, c'est-à-dire du silence (Article, premier, toute et association), de même que des tokens nouveaux, donc du bruit (ARIcs, nsun, teute et asseciation). Se pose donc ici la question de la possibilité d'utiliser ces documents en l'état, avec une correction automatique, semi-automatique ...

Au-delà de mes travaux personnels, cette problématique était l'objet principal de deux thèses que j'encadre et a été un sujet annexe d'une autre thèse où j'étais encadrant scientifique. La thèse de Caroline Parfait sur le sujet « *Literary space analysis : Machine learning and evaluation of recognition systems of named entities* » (Analyse de l'espace littéraire : apprentissage et évaluation de systèmes de reconnaissance d'entités nommées) co-encadrée avec Glenn Roe et Motasem Alrahabi de l'Obtic qui fera l'objet des premiers développements. D'autre part,



ARTICLE PREMIER. Le but de toute association politique est le maintien des droits naturels et imprescriptibles de l'homme, et le développement de toutes ses facultés.

ART. 2. Les principaux droits de l'homme sont ceux de pourvoir à la conservation de l'existence et de la liberté.

ART. 3. Ces droits appartiennent également à tous les hommes, quelle que soit la différence de leurs forces physiques et morales. L'égalité des droits est établie par la nature; la société, loin d'y porter atteinte, ne fait que la garantir contre l'abus de la force, qui la rend illusoire.

ART. 4. La liberté est le pouvoir qui appartient à l'homme d'exercer à son gré toutes ses facultés; elle a la justice pour règle, les droits d'autrui pour bornes, la nature pour principe, et la loi pour sauvegarde.

ART. 5. Le droit de s'assembler paisiblement, le droit de manifester ses opinions, soit par la voie de la presse, soit de toute autre manière, sont des conséquences si nécessaires de la liberté de l'homme, que la nécessité de les énoncer suppose ou la présence ou le souvenir récent du despotisme.

ART. 6. La propriété est le droit qu'a chaque citoyen de jouir et de disposer à son gré de *la portion de bien qui lui est garantie par la loi.*

ART. 7. Le droit de propriété est borné, comme tous les autres, par l'obligation de respecter les droits d'autrui.

ART. 8. Il ne peut préjudicier ni à la sûreté, ni à la liberté, ni à l'existence, ni à la propriété de nos semblables.

ART. 9. Tout trafic qui viole ce prineipe est essentiellement illicite et immoral.

ART. 10. La société est obligée de pourvoir à la subsistance de tous ses membres, soit en leur procurant du travail, soit en assurant les moyens d'exister à ceux qui sont hors d'état de travailler.

ART. 11. Les secours indispensables à celui qui manque du nécessaire sont une dette de celui qui possède le superilu. Il appartient à la loi de déterminer la manière dont cette dette doit être acquittée

Le tirage conforme à l'

(a) Vue par l'humain (Source : Gallica)

```
'teI 1- ( sel l :. 1

ELIBITION C.
1.n
3
ni
r
n 1 T1n
- ::
MI
5u3 DNS S II IIIII II I0 0III

  ARICS NSUN. Le but de teute asseciation politique est le
maintien des droits natarels et imprescriptibles de l'homme, et
le developpement de toutes ses facult6s.
  ia. i priiiii ai d ros et eas 4
ponrvoir a la cnrseralian de l'existerce et de la lierte.
  ARR. 3. Ces droits appertiennent egaslement teus les
hommes, quelle qe soit la difference le leurs forces physiques
et morales. Lgilit des droits est etblie par la iature ; la
societe, loin d'y porter atteinte, ne fait que la garantir contre
l'abus de la fore, qui la rend illusoiro.
  ANT. I. L liberle est le pouoir qui appartient lhomme
d'eercer son gre toutes ses facultes ; elle a la justice pour
regle, les droits d'autrui pour hornes, la nature peur prineipe,
et la loi ponr sauvegarde.
  . . ILs ani de assemnbler paisiblemnent, le droit de
manifester ses opinions, sit par la oie de la presse, soit de
oute autre maniere, sont des conseuences si necessaires de la
liberte de l'homme, ue la ue :::ssile ie les enoncer suppose ou
la presence ou le souenir recent du despotisme.
  AR,. 6, La propriete est le droit qu'a chaque citoyen de jouir
et de disposer a son gre de la portion de ie qui lui est garantie
par l loi.
  ARR. T. ILe dreit de propriete est borne, comme tous les
autres, par l'obligation derespecter les droits d'autrui,
    ARI, 8 : II ne peut prejudicier ni a la srele, ni a la liherte,
ni b l'existece„i a la propriete de nos semblables
  ANR. 9. Tout trafic uii viole ce prineipe est essentiellement
illicite et immoral.
  ANR. 10. La societe est ohligee de pourvoir a la subsistanee
de tous ses membres, soit en leur procurant du travail, soil en
t les moyens d'exisuer a ceux qui sont hors d'etat de
travailler.
  AAR, 11. Les secours indispensables a celui qi manqe d
necessaire sont une dette de celui qui possede le superilu. II
appartient a la loi de determiner la maniere dant cette lette doil
elre acquitlee

-eL-za4-
. : a-. .c . -'
.-
```

(b) Vue par la machine (océrisation de Kraken-17 sur les données Gallica)

FIGURE 5.1 – La déclaration des droits de l'Homme et du Citoyen vue par l'il humain et par l'il numérique

la thèse de Jean-Baptiste Tanguy sur « Laccessibilité et lexploitation des documents textuels numérisés » co-encadrée par Glenn Roe et Karine Abiven (STIH). Dans la thèse de Nhu Khoa Nguyen sur l'analyse de la presse, qui ne sera pas évoquée en détails ici, une partie du travail avait consisté à mesurer l'impact du bruit sur la REN elle même ainsi que sur la classification et la détection d'évènements (épidémiologiques en l'espèce), voir [Boros et al., 2022] pour une synthèse de ce qui figure à ce sujet dans la thèse [Nguyen, 2023].

5.1 Influence du bruitage des données sur les méthodes de TAL : le cas de la reconnaissance d'Entités Nommées

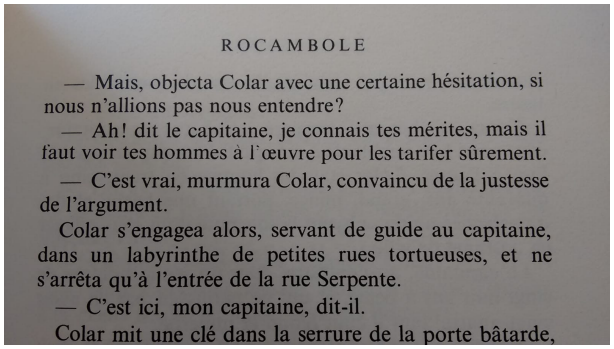
Parmi les problèmes qui se posent au chercheur en TAL qui s'intéresse à la collecte de corpus figure l'applicabilité, ou disons les conditions d'applicabilité, des méthodes et approches connues aux données construites. Parmi les tâches qui vont intéresser les chercheurs, en particulier dans un contexte comme celui qui est le mien à la faculté des lettres, figure la reconnaissance d'entités nommées (REN). On peut ainsi se demander dans quelle mesure les méthodes et systèmes de REN vont pouvoir s'appliquer quand l'entrant varie, en particulier sous l'effet de la reconnaissance optique de caractères appliquée à des données qui ne sont pas nées numériques. Le cas auquel nous nous sommes intéressés est le cas du français mais nous avons déjà dans les données collectées une variation importante :

— Qualité de la numérisation : allant d'une mauvaise orientation des pages à des problèmes de résolution en passant par des étrangetés matérielles [76] ;
— Complexité de la structure de page : multicolonnage, images, lettrines ... ;
— Distribution inattendue des caractères et de leur forme matérielle par rapport à ce sur quoi les systèmes de TAL, notamment de REN, ont été entraînés.

Le choix de s'intéresser à la REN réside dans une collaboration avec Glenn Roe et Motasem Alrahabi de l'OBTIC (Observatoire des Textes et des Connaissances, équipe projet SCAI/Sorbonne Université [77]) qui relayaient des besoins sur des possibilités de navigation dans les corpus (littéraires majoritairement dans leur cas) au moyen de la REN mais aussi une certaine déception vis-à-vis des résultats obtenus. Nous avons donc conçu un projet de thèse pour lequel nous avons obtenu un financement pour 3 ans de l'initiative SCAI de Sorbonne Université.

Le constat que nous avions fait en préambule de ce projet était que l'amélioration de l'accès à l'information et aux textes [Linhares Pontes et al., 2019] supposé intervenir grâce à la REN n'était pas tout à fait au rendez-vous. Étant donné que l'intérêt principal côté OBTIC concernait les lieux (*location* dans la terminologie anglophone du domaine) nous avons concentré nos efforts sur ce cas particulier. Un des objectifs et de comparer l'application de la REN sur différents « états » d'un même texte : son état idéal, la retranscription parfaite de la version canonique du texte ainsi que ses états plus ou moins dégradés résultant de la numérisation et de la reconnaisance optique de caractères (ci-après OCR pour garder le sigle anglais, plus fréquemment utilisé). Par simplicité, nous parlerons ici de version : version de référence et différentes versions OCR. Les variations que l'on observera en comparant la version de référence et les différentes versions OCR n'ont pas toutes la même influence sur les tâches de TAL réalisées en aval de l'OCR. Certaines variations sont complètement anodines, imaginons par exemple la présence d'un espace double au lieu d'un espace simple, tandis que d'autres vont notablement affecter certaines tâches. Par exemple, si la casse est modifiée c'est souvent problématique pour la REN. Pour nommer les cas

76. Dans sa thèse Jean-Baptiste tanguy montre ainsi l'exemple de la numérisation d'une main intégrée à un livre sur Google Livres [Tanguy, 2022]
77. `https://obtic.sorbonne-universite.fr/` consulté le 2 octobre 2023



ROCAMBOLE

— Mais, objecta Colar avec une certaine hésitation, si nous n'allions pas nous entendre?

— Ah! dit le capitaine, je connais tes mérites, mais il faut voir tes hommes à l'œuvre pour les tarifer sûrement.

— C'est vrai, murmura Colar, convaincu de la justesse de l'argument.

Colar s'engagea alors, servant de guide au capitaine, dans un labyrinthe de petites rues tortueuses, et ne s'arrêta qu'à l'entrée de la rue Serpente.

— C'est ici, mon capitaine, dit-il.

Colar mit une clé dans la serrure de la porte bâtarde,

FIGURE 5.2 – Extrait de « Rocambole, l'héritage mystérieux » (Pierre Alexis de Ponson du Terrail, 1857), l'antepénultième ligne mentionne le lieu « rue Serpente »

où ces variations dans le texte ont une influence, négative ou positive, sur la qualité du traitement informatique des textes nous avions proposé le terme d'*interférences* [Koudoro-Parfait et al., 2021]. Nous définissons les *interférences* comme l'ensemble des perturbations pouvant affecter la transmission de l'information au sein de la chaîne de traitement et *in fine* la donnée extraite en sortie.

Les cas d'utilisation de la REN spatiales dans des données littéraiires sont nombreux mais l'on peut citer le suivi des déplacements d'un personnage dans une uvre, la mise en parallèle de parcours similaires dans des uvres différentes ou encore la mise en relation de lieux emblématiques ou non avec les récits réels ou fictifs qui s'y sont produits. On peut aussi penser au point de vue de l'historien, ou du curieux, qui chercherait des informations sur un lieu particulier. Qu'il me soit permis, en tant que curieux, de donner dans la Figure 5.2 un exemple de repérage d'entité nommée spatiale dans une uvre littéraire. C'est un exemple qui nous intéresse tout particulièrement à la faculté des Lettres de Sorbonne Université puisque c'est le lieu où se trouve la maison de la Recherche où j'officie. Nous pouvons déjà imaginer plusieurs emplois très simples de la REN ici : (I) naviguer dans des corpus au gré des mentions de noms de lieu, (II) chercher spécifiquement un ou plusieurs lieux dans des corpus ou (III) observer l'évolution de la représentation des lieux au cours du temps. En l'espèce il n'était pas étonnant pour Ponson du Terrail que la rue Serpente héberge les activités illicites du brigand Colar, alors qu'aujourd'hui sans doute l'image de la rue est plus lisse.

Bien évidemment, il y a un décalage entre les entités visibles à l'il nu et celles qui seront visibles à l'il électronique et donc extraites par REN. C'est l'objet des évaluations que de chercher à mesurer cet écart entre la sortie désirée et la sortie obtenue. Le cas est particulièrement intéressant quand nous sommes confrontés à des données réelles, hors conditions de laboratoire, qui vont présenter différentes variations (dont le bruit produit par l'OCR) par rapport aux données sur lesquels les systèmes de REN sont usuellement entrainés et évalués.

Le premier aspect variationnel auquel on songe est évidemment le bruitage de la donnée provoqué par la phase d'OCR. Ceci inclut l'insertion, l'effacement ou la substitution de caractères ; il est ainsi relativement courant qu'il y ait une confusion entre le « a » et le « o » ou entre le « i » et le « l » : « Paris » étant ici retranscrit « Poris » ou encore « Parls » [78]. Nous prenons dans le tableau 5.1 (page 98) des exemple tiré de la version numérisée de l'Education Sentimentale de Flaubert où les erreurs de retranscription sont surlignées et les caractères disparus

78. Exemple obtenus avec Kraken sur une version numérisée de Savarus de Balzac et de l'éducation sentimentalee de Flaubert

Original	[...]Mystè	res	de Paris		tira de sa poche	un brûle-gueule
Tesseract-fr	[...]_ystè	res	de Paris	,	tira de sa poche	un brûle-gueule,
Tesseract-Base	[...]dystè	res	de Paris	,	tira de sa poche	un brile-gueule,
Kraken-17	[...]_ sfe	res	de Poris	,	tira de sa poche	un brûle-gueule,
Kraken-Base	[...]Iyste	res	de Paris	,	tira de sa poche	un br_le-gueule

TABLE 5.1 – Variations apportées par différents OCR sur un même segment de « L'Éducation Sentimentale ».

remplacés par "_".

Au-delà de ces cas disons simples, figurent des situations plus complexes où la séquence de caractères issue de l'OCR ressemble assez peu à du langage naturel. Ceci est particulièrement le cas quand le modèle de page est complexe ou quand il y a des illustrations et autres lettrines.

Les interférences vont générer des faux positifs, des entités extraites par erreur, de même que des faux négatifs, des entités qui seront manquantes. Si dans notre exemple on extrait l'entité « Poris » on a un faux Positif de même qu'un faux négatif puisque, en l'état l'entité « Paris » est manquante. Les variations sont aussi présentent dans le contexte des entités ce qui peut provoquer des erreurs d'étiquetage, un nom de personne pouvant être étiqueté nom de lieu. Bien évidemment, il faut pouvoir différencier les cas qui viennent véritablement d'interférences dans l'entrant (l'*input*) de celles qui ce seraient de toutes façons produites y compris avec « l'entrée idéale », la version de référence. On observe également, même s'ils sont plus rares, des cas où le système de REN reconnaît correctement une entité sur la version bruitée du texte mais ne le reconnaît pas, ou l'étiquete incorrectement, sur la version de référence (cf. Figure 5.5 page 105).

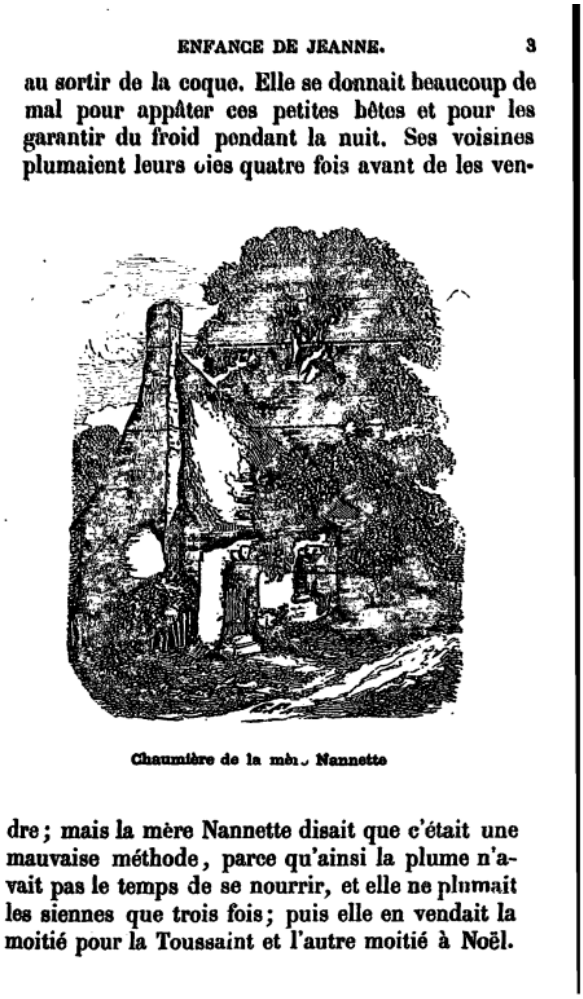
Les expériences présentées dans les pages qui suivent ont été réalisées dans le cadre de la thèse de Caroline Parfait et une partie de ce travail a fait l'objet d'une publication dans un atelier spécialisé dans les humanités géolocalisées [Koudoro-Parfait et al., 2021]. L'idée a été d'examiner dans un premier temps les intersections et les différences entre les résultats de différents modèles de REN en fonction de différentes entrées : texte de référence et différents résultats d'OCR. Ensuite, nous avons proposé une typologie des erreurs rencontrées. Enfin, nous avons cherché à mesurer la corrélation entre le bruit mesuré en entrée et le bruit résultant en sortie. Avec cette question : est-ce que le meilleur système d'OCR donne forcément le meilleur résultat en REN ? Et son corollaire : est-ce que l'on mesure bien le bruitage si l'on ne trouve pas de corrélation entre bruitage de l'entrée et bruitage de la sortie ?

5.2 La Reconnaissance d'Entités Nommées dans des textes océrisés

Dans la section 5.2.1 je rappelle certaines solutions proposées dans la littérature pour la REN en situation bruitée. Ensuite, je présente (Section 5.2.2) les données textuelles exploitées et comment elles ont été constituées.

5.2.1 Quelles approches pour la REN sur des données bruitées ?

L'importance de la reconnaissance des Entités Nommées dans des données OCR prend sa source dans des besoins exprimés par les chercheurs en humanités numériques [Hamdi et al., 2020], en particulier par ce que de nombreuses requêtes formulées sur des corpus vont concerner des entités nommées [van Strien et al., 2020]. Le problème de la reconnaissance d'entités nommées



ENFANCE DE JEANNE. 3

au sortir de la coque. Elle se donnait beaucoup de mal pour appâter ces petites bêtes et pour les garantir du froid pendant la nuit. Ses voisines plumaient leurs oies quatre fois avant de les ven-

Chaumière de la mère Nannette

dre ; mais la mère Nannette disait que c'était une mauvaise méthode, parce qu'ainsi la plume n'avait pas le temps de se nourrir, et elle ne plumait les siennes que trois fois ; puis elle en vendait la moitié pour la Toussaint et l'autre moitié à Noël.

FIGURE 5.3 – Numérisation d'une page avec illustration et légende (Zulma Carraud, La petite Jeanne, source GALLICA)

dans des sorties OCR a fait l'objet d'un certain nombre d'études même si la REN et l'OCR sont généralement étudiées isolément [Huynh et al., 2020]. Il est intéressant de noter que les études conjointes de ces deux problèmes ont souvent pour toile de fond les Humanités Numériques et la question de l'usage que l'utilisateur final peut avoir de grandes quantités de données numérisées qui ne peuvent matériellement être corrigées en détails [Huynh et al., 2020]. Bien qu'il soit rare d'obtenir des sorties OCR sans défaut, de nombreux travaux montrent que, fort heureusement, ces données restent utilisables pour des tâches de TAL comme la REN [Gupta et al., 2015]. Dans la compétition CLEF-HIPE (*Identifying Historical People, Places and other Entities*) en 2020, les données d'entraînement étaient des données issues d'OCR [Ehrmann et al., 2020] afin de se rapprocher des conditions réelles d'application de la REN à des données en humanités. Avant cela, [Traub and Van Ossenbruggen, 2015] s'étaient intéressés à l'expression des besoins des chercheurs en humanités qui cherchent à traiter des données issues d'OCR.

Étant donné que la qualité du texte est considérée comme un facteur limitant pour l'extraction d'entités nommées, une préparation des données [79] est parfois appliquée : correction de scories orthographiques, standardisation ou encore révision des erreurs de segmentation. À l'opposé, d'autres approches se sont intéressées à l'utilisation de la sortie OCR en l'état sans chercher spécifiquement à évaluer l'impact des erreurs d'OCR sur les résultats, comme nous l'avions proposé dans la compétition multilingue (Allemand, Anglais et Français) CLEF-HIPE 2020 *dentifying Historical People, Places and other Entities* avec Pedro Ortiz Suarez, Yoann Dupont et Tian Tian [Suárez et al., 2020]. [Huynh et al., 2020] a proposé d'examiner les sorties de la REN pour chercher à savoir quelles sont les erreurs d'OCR qui posent particulièrement problème, quelles sont donc les interférences les plus régulières. Les auteurs observent que si la correction de l'entrée a un impact sur la qualité de la sortie, elle va amener aussi de nouvelles erreurs notamment du fait de la sur-correction. [Petkovic et al., 2022] ont ainsi montré que les vrais positifs « gagnés » grâce à la correction étaient largement moins nombreux que les faux positifs engendrés.

D'un point de vue général, il y a un consensus pour dire que le bruit dans les retranscriptions OCR influe négativement sur les résultats des systèmes de TAL placés en aval [Lopresti, 2009]. [van Strien et al., 2020] montre que toutefois ceci est beaucoup plus prégnant pour des tâches telles que la segmentation en phrases, voir par exemple [Giguet and Lejeune, 2021b] pour un travail sur des documents financiers, ou [Hwang et al., 2020] pour l'analyse en syntaxe de dépendance, que pour la REN. Cette affirmation est concordante avec ce que nous avions remarqué dans CLEF-HIPE, à savoir que retrouver une bonne segmentation en phrase était difficile bien que tout à fait souhaitable d'un point de vue de qualité des résultats [Suárez et al., 2020]. [Hamdi et al., 2020] montrent ainsi que s'il y a une relation entre la dégradation de l'OCR et la dégradation de la REN, un taux d'erreur au mot (WER pour *Word Error rate* de 8% peut permettre néanmoins d'obtenir une exactitude supérieure à 90%. [Alex et al., 2012] ont montré que des corrections ciblées telles que la reconstitution des mots découpés par la césure et la standardisation des « s longs » (voir Annexes Figure A.1 page 135) amélioraient significativement la reconnaissance des entités.

Le problème de l'évaluation étant aussi de disposer d'une vérité de terrain sur différentes versions d'OCR, des tests ont été pratiqués à partir d'une dégradation artificielle (dans le sens de simulée) des documents images avant océrisation. [Linhares Pontes et al., 2019] ont évalué ceci directement sur la REN tandis que dans [Nguyen et al., 2020] nous l'avons fait sur des tâches de classification impliquant la reconnaissance d'entités.

Parmi les erreurs d'importance produites par l'OCR figure la modification de la casse (en

79. Pour ne pas dire des pré-traitements.

règle générale cela se manifeste par le passage de majuscule en minuscule plutôt que l'inverse). Ceci a un impact important sur la REN [Kettunen et al., 2020] en particulier évidemment pour les langues, comme le français ou l'anglais par exemple, où la présence de la casse est importante, voire capitale, pour détecter les noms propres [Powalski and Stanislawek, 2020].Retrouver la casse correcte (*true-casing*) devient ainsi un élément autonome dans la correction des textes et retrouver la vraie casse améliore souvent les résultats de certaines tâches de TAL, voir [Niu and Carpuat, 2020] pour un exemple sur la traduction automatique.

Du point de vue de l'évaluation, un problème important est de pouvoir aligner les entités trouvées dans le texte bruité avec celles du texte de référence. Ce problème ne se posait pas dans CLEF-HIPE 2020 car seule une version du texte était disponible (bruitée en l'occurrence) et les données fournies étaient déjà tokenisées (format CONLL). Ceci facilite une évaluation précise de la performance que ce soit dans un scénario strict (où tous les tokens de l'entité doivent être détectés correctement) ou flou (où un token correct suffit) [Ehrmann et al., 2020]. Mais si l'on veut obtenir ceci sur différentes versions du texte (la version de référence et une voire plusieurs retranscriptions OCR) alors il faut réaliser un alignement précis (*a minima* au grain phrase) [Boros et al., 2020], se posent alors la question de la granularité de cet alignement. Cet alignement est coûteux à réaliser manuellement, et n'est pas du tout trivial à réaliser automatiquement [Chiron et al., 2017, Hamdi et al., 2020] en particulier sur des documents longs (des livres par exemple) par opposition à des documents de moindre taille tels que des articles de presse [Kettunen et al., 2016]. Dans les expériences présentées ici, nous ne pratiquons pas d'altération artificielle mais nous comparons différents OCR et différents formats et qualités d'images en entrée. L'évaluation est ici réalisé de manière globale sans alignement préalable, pour une évaluation plus fine après alignement voir [Koudoro-Parfait et al., 2022].

5.2.2 Contexte expérimental : différentes versions en entrée et différents approches pour la REN en sortie

Je donnerai ici les éléments ayant trait à la constitution d'un jeu de données de référence pour l'étude de la robustesse de la REN en contexte bruité puis dans la section 5.2.3 je présente les outils utilisés pour cette étude et les résultats obtenus.

Un corpus de romans en français

Le corpus est composé de dix romans appartenant au corpus ELTeC [80]. L'intérêt de ce corpus est de regrouper des versions de référence pour un nombre conséquents de romans publiés entre 1840 et 1920 dans une dizaine de langues d'Europe. Nous nous sommes limité ci au cas du français mais cette étude mériterait bien évidemment d'être réalisée dans un cadre multilingue. Les versions de référence forment un corpus de plus de 3 000 pages (tableau 5.2). Pour générer les versions OCR, nous avons collecté les versions PDF disponibles sur la plateforme Gallica [81].

Les versions OCR ont été obtenues à partir de deux outils librement disponibles. En premier lieu, nous avons utilisé Kraken [82], qui est régulièrement utilisé dans la communauté des Humanités Numériques [Kiessling et al., 2019] et qui offre une certaine facilité d'utilisation dès lors que l'on a besoin de ré-entraîner pour une langue [Weichselbaumer et al., 2020] ou une période particulière [Gabay et al., 2020]. Nous avons utilisé le modèle par défaut pour le français. En-

80. European Literary Text Collection : `https://www.distant-reading.net/eltec/` consulté le 2 octobre 2023
81. `https://gallica.bnf.fr/` consulté le 2 octobre 2023
82. `https://github.com/mittagessen/kraken` consulté le 2 octobre 2023

Roman	Auteur	Année	# pages	# mots
« Le château de Pinon, vol. I »	Comtesse Dash [84]	1844	332	56 $*10^3$
« Albert Savarus. Une fille d'Ève. »	Honoré de Balzac	1853	60	49 $*10^3$
« Les trappeurs de l'Arkansas »	Gustave Aimard	1858	450	112 $*10^3$
« Mon village »	Juliette Adam (Lambert)	1860	200	26 $*10^3$
« Le petit chose »	Alphonse Daudet	1868	292	107 $*10^3$
« L'Éducation sentimentale » [85]	Gustave Flaubert	1880	520	188 $*10^3$
« Une vie »	Guy de Maupassant	1883	337	93 $*10^3$
« La petite Jeanne »	Zulma Carraud	1884	220	64 $*10^3$
« La belle rivière »	Gustave Aimard	1894	339	177 $*10^3$
« La nouvelle espérance »	Anna de Noailles	1903	325	67 $*10^3$
« Marie-Claire »	Marguerite Audoux	1925	120	44 $*10^3$

TABLE 5.2 – Statistiques sur les versions de référence des textes du corpus

suite, nous avons utilisé TESSERACT [83] [Smith, 2007] qui en plus d'être lui aussi ré-entraînable, a la réputation de bien se comporter dans sa version par défaut [Clausner et al., 2020]. Nous avons utilisé deux modèles : le modèle par défaut et le modèle optimisé pour le français (TESS-FR).

Pour chaque roman nous avons donc quatre versions : la version de référence issue d'ELTEC est les versions océrisées par KRAKEN, TESSERACT et TESS-FR.

Outils de reconnaissance d'entités nommées

Nous avons fait le choix d'utiliser des outils de REN de grande diffusion afin de nous placer dans la situation des collègues en humanités numériques qui vont chercher à exploiter des outils faciles d'utilisation et « disponibles sur l'étagère » (*off-the-shelf*). Le choix est forcément contestable, puisque l'on ne s'intéresse qu'à une partie des outils disponibles, mais reflète ce que l'on observe à ce jour dans les tutoriels par exemple. Et c'est là que de nouveaux utilisateurs d'outils de TAL iront piocher ce dont ils ont besoin. Cette comparaison me semble donc pertinente du point de vue d'un utilisateur qui cherche simplement à voir si ses données sont utilisables en l'état ou si la REN peut être pertinente pour son problème. L'outil SPACY [86] y est omniprésent, ce qui en fait il me semble un bon candidat pour savoir ce qu'un utilisateur lambda peut attendre d'un système de REN. SPACY propose différents modèles (*small, medium and large*) qui, au prix du temps de calcul, permettent *a priori* d'obtenir des résultats de meilleure qualité. Bien que le développement de SPACY ne soit pas fait dans un contexte académique, il reste un outil très utilisé dans différentes communautés [van Strien et al., 2020]. STANZA [87] [Qi et al., 2020] est aussi un outil populaire et sans doute le plus populaire parmi les outils développés en contexte académique.

5.2.3 La question de l'évaluation non supervisée de la REN sur des données OCR

Faute de vérité de terrain côté Entités Nommées, puisque nous n'avons pas de corpus de référence de grande taille sur la période, nous procédons à une évaluation extrinsèque dans

83. `https://github.com/tesseract-ocr/tesseract` consulté le 2 octobre 2023
86. `https://spacy.io/` consulté le 2 octobre 2023
87. `https://stanfordnlp.github.io/stanza/ner.html` consulté le 2 octobre 2023

l'esprit des travaux de [van Strien et al., 2020]. Nous comparons ainsi les sorties obtenues sur les différentes versions par plusieurs modèles de REN [88].

Nous nous focalisons sur les lieux puisque c'est le cas d'utilisation le plus fréquent chez les chercheurs en humanités [89] mais cette étude pourrait là encore être étendue à tous les types d'entités. L'étude proposée s'est déroulé en trois étapes :

1. Évaluation manuelle sur un échantillon des 160 sorties obtenues : 10 romans, avec chacun 4 versions (nous éliminons le modèle Kraken-17 qui est inadapté à nos données) et 4 modèles de REN pour les traiter ;
2. Questionnement sur l'évaluation via une première proposition de redéfinition des concepts de Vrai/Faux positifs et de Vrai/faux négatifs ;
3. Évaluation automatique et recherche de corrélation entre le bruit de l'entrée et le bruit de la sortie.

Évaluation manuelle : quelques exemples d'erreurs d'extraction

Nous cherchons ici à observer et à caractériser l'interaction entre l'OCR et les modèles de REN. Nous utiliserons par la suite le terme de *configuration* pour désigner le fait d'utiliser un système de REN X sur la version Y d'un texte. Nous pourrons ainsi comparer les résultats sur deux axes principaux : (I) assez classiquement comparer différents systèmes de REN sur la même version textuelle et (II) observer le comportement d'un même système de REN sur différentes versions d'un même texte.

Dans le tableau 5.3 figure une série d'exemples de variation des entités détectées par le modèle `small` de spaCy en fonction des versions qui lui sont soumises. Il s'agit d'une sélection parmi les cas où nous avons observé une variation de l'entrée et de la sortie. Il est intéressant de noter que les variations subies par l'entité elle-même (exemple 2 avec Kraken, Châlons → Ch_lons) ne perturbent pas nécessairement sa reconnaissance , même si se posera la question d'associer l'entité trouvée avec l'entité réelle. Les deux cas où l'entité n'est pas trouvée (sur la version Kraken dans l'exemple 1 et sur la version Tesseract dans l'exemple 3) sont intéressants : le contexte gauche a été bruité et l'entité n'a pas été reconnue alors même que dans le deuxième cas l'entité est intacte, l'interférence est donc dans le contexte.

Dans le tableau 5.4 sont exposées les sorties d'un modèle plus évolué de spaCy, en l'espèce `spacy_lg` qui est comparé à Stanza. L'idée est de regarder les différentes variantes d'un même terme, variantes dans chaque version particulière, et de voir si elles sont effectivement détectées par les deux modèles. Nous nous intéressons ici au lieu « Morlincourt » qui est fréquent dans le roman d'Adam mais dont les occurrences vont être retranscrites de manière variable dans les différentes versions OCR.

Nous pouvons déjà remarquer que la différence entre les deux outils s'exprime sur la version de référence, Morlincourt est identifié 18 fois par `spacy_lg` et 16 fois par Stanza. Certaines des occurrences ne sont pas détectées dans la version de référence mais le sont par contre dans la version OCR, sous une forme graphique altérée par rapport à l'attendu. Ceci est quelque peu contre intuitif puisque la version dégradée fait disparaître certains faux négatifs. Or on s'attend plutôt à voir apparaître des faux positifs, du fait du bruit OCR, voir en disparaître certains si la retranscription est tellement mauvaise que le système n'extrait rien (voir [Lejeune and Zhu, 2018]

88. Les résultats présentés ici sont rendus disponibles sur le Github de Caroline Koudoro-Parfait : `https://github.com/These-SCAI2023/NER_GEO_COMPAR` consulté le 2 octobre 2023

89. Ceci a notamment été observé par Caroline Parfait et Jean-Baptiste Tanguy dans le cadre d'une consultation lancé via les listes DH et ATALA en 2021 et prolongé par des entretiens avec différents chercheurs, [Tanguy, 2022]

Version	Contexte		Sortie	
Ref.	en faïence	de Hollande.		Hollande
Kraken	en faience	_e kollande		()
Tess	en faience	___ Hollande.. 7 _		Hollande
Tess fr	en faïence	___ Hollande.		Hollande
Ref.	prendre la	diligence de Châlons		Châlons
Kraken	prendre la	diligence de Ch_lons		Ch_lons
Tess	prendre la	diligence de Chalons		Chalons
Tess fr	prendre la	diligence de Châlons		Châlons
Ref.	on se voit	forcé d'aller	à Morlincourt.	Morlincourt
Kraken	on se voit	force d'aller	_ Mlorlincourt.	Mlorlincourt
Tess	on se v01t	force (1 al- '- :3 ler	£1 Morlincourt.	()
Tess fr	on se v01t	force (1al- *-S.' . %ler	à Morlincourt.	Morlincourt

TABLE 5.3 – Exemples de variations graphiques dans les différentes versions et de leur influence dans la sortie de `spacy_sm` (les caractères erronés figurent en rouge, les tirets bas (« _ ») marquent les caractères manquants)

Version	Modèle REN	Entité : Effectif	# Faux Négatifs
Ref.	`spacy_lg`	Morlincourt : 18	N/A
	STANZA	Morlincourt : 16	N/A
Kraken	`spacy_lg`	Morlincourt : 8 Mlorlincourt : 1 Mlorlincourtl : 1	8
	STANZA	Morlincourt : 6 Mlorlincourt : 3 Mlorlincourtl : 1	6
Tess fr	`spacy_lg`	Morlincourt : 11 Morlincourt : 1	6
	STANZA	Morlincourt : 9 Morlincourt : 1	6
Tess	`spacy_lg`	Morlincourt : 9 Morlin : 1	8
	STANZA	Morlincourt : 7 Morlinco'urt : 1	8

TABLE 5.4 – Variantes graphiques de l'entité « Morlincourt » trouvées dans les différentes versions de « Mon Village » (Juliette Adam)par STANZA et `spacy_lg`

Version	spacy_sm	spacy_md	spacy_lg	Stanza
Ref.	Grèce	Grèce **bleue (M)**	Grèce **bleue (M)**	Grèce **bleue (M)**
Kraken	Grece	Grece	Grece bleue	Grece **bleue (M)**
Tess	Grèce **(P)**	Grèce	Grèce	Grèce **bleue**
Tess fr	Grèce	Grèce	Grèce	Grèce **bleue (M)**

TABLE 5.5 – Traitement des variations graphiques de l'entité « Grèce » dans différentes configurations (l'étiquette P correspond à Personne et M à Misc.)

pour de tels exemples d'amélioration de la précision dans des versions bruitées de documents). Ce phénomène n'aura pas la même importance pour une entité fréquente, pour la quelle la variation de quelques occurrences par rapport à l'attendu ne modifie pas la dynamique globale, que pour une entité avec peu d'occurrences qui peut disparaître complètement de la sortie REN. Afin d'illustrer plus avant cette problématique, nous nous intéressons à la reconnaissance de l'entité « Grèce » dans « La nouvelle espérance (Anna de Noailles), le contexte d'apparition en est le suivant :

Oh ! oui, tous les pays du monde... la ***Grèce*** *bleue et la belle Anatolie !*

Le tableau 5.5 montre que le modèle `spacy_lg` a une acception plus large de l'entité (en incluant « bleue » dans deux versions) que le modèle `spacy_sm` supposé moins efficace. Par contre nous observons une erreur d'étiquette dans la version Tesseract (Personne au lieu de Lieu).

Nous pouvons observer que la disparition de l'accent dans la version KRAKEN n'a pas perturbé les modèles de REN. Par contre, il est intéressant de noter que dans 3 cas sur 4 l'étiquette appliquée à l'entité dans la version de référence est erronée (Misc au lieu de Location), erreur qui se produit moins dans les version océrisées. Ceci est sans doute dû au fait que le déterminant *la* n'apparaît pas dans le contexte gauche immédiat de *Grèce bleue* dans les versions OCR du fait de la segmentation. On peut donc faire l'hypothèse qu'ici *Grèce bleue* est envisagée comme un groupe nominal par trois des quatre outils de REN. Par contre, nous n'avons pas trouvée d'explication spécifique à l'étiquette Personne attribué à *Grèce* par `spacy_sm`.

En guise de conclusion partielle de cette partie, nous pouvons dire que :

— Les modèles de REN utilisés sont en mesure de reconnaître des variantes graphiques incorrectes des entités (« Chlons » au lieu de « Châlons » par exemple)

— Le contexte syntaxique, ou à tout le moins la qualité de la segmentation en phrase semblent importants

— Il n'y a pas de garantie que les entités trouvées dans des parties exemptes d'erreurs à première vue ne soient pas en fait des interférences.

Typologie des erreurs et problèmes d'évaluation

Les quelques exemples présentés dans la section précédente amènent quelques questions sur la définition des différents cas qui vont se présenter au moment de l'évaluation de la sortie REN sur des entrées présentant différentes versions en terme de bruit :

— Comment définir précisément les Vrais Positifs (VP), Faux Positifs (FP), Faux Négatifs (FN) et Vrais Négatifs (VN) ?

— Comment comparer des listes d'entités ou d'occurrences d'entités sans alignement fin (au grain token) ou plus lâche (au grain phrase ou page par exemple) ?

— Quelle souplesse s'autorise t-on dans la comparaison des formes graphiques ? Comment cela se combine avec une évaluation stricte ou floue ?

Est une entité dans		**Interprétation**		
Référence	Version OCR	Verdict naïf	Commentaires	Verdict révisé
Oui	Oui	VP	Entité *a priori* correcte	Vrai VP
Oui	Oui	VP	Faux positif, erreur de REN	Faux VP
Non	Non	VN	N'est pas une entité et n'est annotée comme telle dans aucune version	Vrai VN
Non	Non	VN	Entité manquante dans les deux versions	Faux VN (FN ?)
Oui	Non	FN	Entité manquante dans la version OCR	Vrai FN
Oui	Non	FN	Entité surnuméraire dans la version de référence	Faux FN
Non	Oui	FP	Entité surnuméraire dans la version OCR	Vrai FP
Non	Oui	FP	Entité manquante dans la version de référence	Faux FP (VP)
Non	Oui	FP	Problème d'*entity linking* (interférence dans le nom de l'entité)	Faux FP (VP ?)

TABLE 5.6 – Proposition de typologie des Vrais Positifs, Vrais Négatifs, Faux Positifs et Faux Négatifs dans un contexte d'évaluation non supervisée

Ceci amène à s'interroger sur la pertinence, ou au moins sur l'étanchéité, des notions classiques de VP, FP, FN et FN. On peut bien sûr partir d'une définition très intuitive :

— VP : occurrence présente à la fois dans la version de référence et dans la version OCR ;
— VN : occurrence qui n'est identifiée ni dans la version de référence ni dans la version OCR ;
— FP : occurrence identifiée dans la sortie OCR mais pas dans la version de référence ;
— FN : occurrence identifiée dans la version de référence mais pas dans la version OCR.

Mais ceci ne correspond pas forcément à la réalité puisque cela part du présupposé que : ce qui est trouvé dans le texte de référence est nécessairement correct, et inversement que ce qui n'y est pas trouvé est incorrect. Le tableau 5.6 propose une typologie un peu plus détaillée des cas qui se présentent. Dans le tableau 5.7 nous présentons quelques exemples, tirés de notre corpus, pour les différentes catégories de cette typologie à l'exception des Faux VN dont n'avons pas rencontré d'occurrence à ce stade et des Vrais VN qui n'ont eux aucun intérêt.

Un cas qui semble particulièrement intéressant est celui des Faux Faux Négatifs : une entité trouvée dans la référence mais pas dans l'OCR (ce qui constituerait en fait un Faux Négatif) mais qui est en fait une entité erronée. Il s'agira en règle générale d'un effet de bord, une amélioration des résultats pour de mauvaises raisons (cf. [Lejeune and Zhu, 2018] pour une discussion sur ces effets de bord en classification).

Un second cas intéressant, dans la série des faux faux, est celui des Faux FP. Il s'agit en l'espèce d'une interférence attendue : le bruit dans l'entrée amène du bruit dans la sortie. Dans les faits il semble que ce soient moins des entités erronées que des entités détectées sous une forme légèrement fautive. C'est par exemple le cas de la forme « Mlorlincourtl » (Table 5.4) trouvée dans la version OCR en lieu et place de la chaîne attendue dans la référence « Morlincourt ». Il semble raisonnable de penser qu'une légère correction de la sortie permettrait de transformer ce Faux FP en VP.

Effectuer une comparaison plus systématique nécessiterait nous l'avons dit un alignement

Verdict naïf	Contexte Réf	EN Lieu Réf.	Contexte OCR	EN Lieu OCR	Verdict révisé
VP	les Watteville sont de Suisse	Suisse	Watteville sont de Suisse,[Tesseract-Fr]	Suisse	**Vrai VP**
FP	...madame de Vandesse, lui parler de Laure et de Beatrix.		...madame de Vandesse, lui parler de Laure et de Béatrix. [Tesseract-Fr]	Beatrix	**Vrai FP**
FN	ou sur les collines d'Italie	d'Italie	N/A *segment non retranscrit par* [Kraken-17]		**Vrai FN**
VP	On vendit alors la maison de monsieur de Watteville pour s'établir rue de la Préfecture...	Watteville	On vendit alors la maison de monsieur de Watteville pour s'etablir rue de la Prefecture... [Kraken-base]	Watteville	**Faux VP**
FP	Un mot sur cette dame, le personnage féminin le plus considérable peut-être de Besançon	Besançon	Pour peul-etre de Besangcon.	Besangcon	**Faux FP**

Table 5.7 – Exemples illustrant notre typologie, dans la partie haute de ce tableau figurent les cas où le verdict 1 (naïf) est correct dans la seconde partie les cas où ce verdict est incorrect

un peu fin des différentes versions afin de pouvoir vérifier la correction des sorties occurrence par occurrence. Un tel alignement sur des longs textes devient un problème en soit et plusieurs facteurs le rendent ardu :

— le grain page n'est pas marqué dans la version de référence
— il peut y avoir du texte en trop, par exemple du fait des illustrations, ou des portions qui n'apparaissent pas du tout, dans le cas de pages où l'encre est peu marquée
— Le nombre de lignes va significativement varier entre la version de référence (où cela correspondra plus ou moins à la notion de paragraphe) et la version OCR (où cela correspond avec la notion de segment visuel), il faudrait donc assurer un alignement entre m segments et n segments avec $m \neq n$
— dans certains cas, l'alignement ne se fait pas sur un seul segment mais sur plusieurs (par exemple deux segments de la version de référence correspondent à trois segments dans la version OCR)
— il peut y avoir des changements d'ordre dans les segments dans le cas de multicolonnage

5.2.4 Comparaison automatique des sorties REN

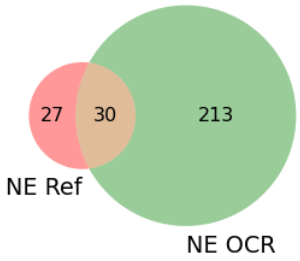
Comparaison globale avec des diagrammes de Venn Dans une première approximation, nous utilisons des diagrammes de Venn pour observer la variation dans les entités découvertes dans différentes configurations . Ceci permet d'observer les intersections (ce qui est en commun) et les différences (ce qui apparaît ou disparaît d'une configuration à l'autre). Ici le travail n'est pas fait par occurrence mais par entité, les entités sont traitées comme des ensembles. On ne tient donc pas compte de la fréquence, du nombre d'occurrences de chaque entité, mais de sa présence ou non dans la sortie. De ce fait, c'est une mesure plutôt stricte puisque les versions bruitées vont générer des hapax (formes contaminées) et des nullax (entités hapax dans la référence et qui sont les plus susceptibles de disparaître complètement dans la version bruitée).

La Figure 5.4 (page 109)permet de comparer les sorties REN de la version de référence et des versions Kraken et Tess-fr. En lisant la Figure de haut en bas, nous pouvons voir l'influence de l'OCR et de gauche à droite l'influence du modèle de REN. Avec Tess-fr on améliore l'intersection (les VP) pour `spacy_lg` mais le phénomène est inverse pour Stanza. Par contre, les potentiels Faux Positifs sont nettement diminués dans les deux cas dans la version Tess-fr par rapport à la version Kraken. En lisant de gauche à droite, on peut voir l'impact du modèle REN. On voir que `spacy_lg` détecte 57 entités différentes là où Stanza n'en trouve que 51. Globalement Stanza semble plus sélectif que `spacy_lg`.

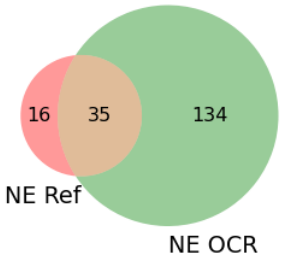
Mesurer la corrélation entre le bruitage OCR et la sortie REN avec des mesures de distance

Même si l'on ne dispose pas d'alignement fin, on peut toutefois de mesurer de manière globale la corrélation entre la variation de l'entrée OCR et la variation de la sortie REN.

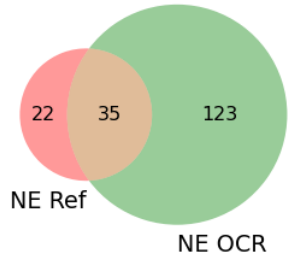
Étant donné la longueur des textes et l'absence d'alignement à la page par exemple, il est malaisé computationnellement parlant de chercher à calculer un taux d'erreur au mot (WER, *Word Error Rate*) ou pire un taux d'erreur au caractère (CER *Character Error Rate*). Notons que c'est le problème qui se pose aussi dans le nettoyage de pages web avec la mesure Cleaneval (voir section 4.2.4 page 85). Le choix s'est porté sur des mesures de distances sur des vecteurs, avec des valeurs binaires ou non, présentées dans la Figure 5.5. Ces mesures sont l'indice de Jaccard (en bleu), la distance de Bray-Curtis (en orange), le coefficient de Dice (vert) et la distance cosinus. Pour mesurer l'impact de la binarisation (effectuée nativement dans Kraken)



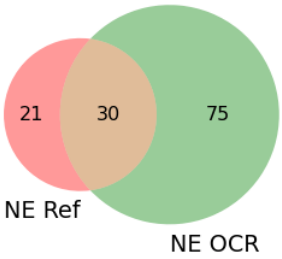
(a) Kraken – `spacy_lg`



(b) Kraken – Stanza



(c) Tess-fr – `spacy_lg`

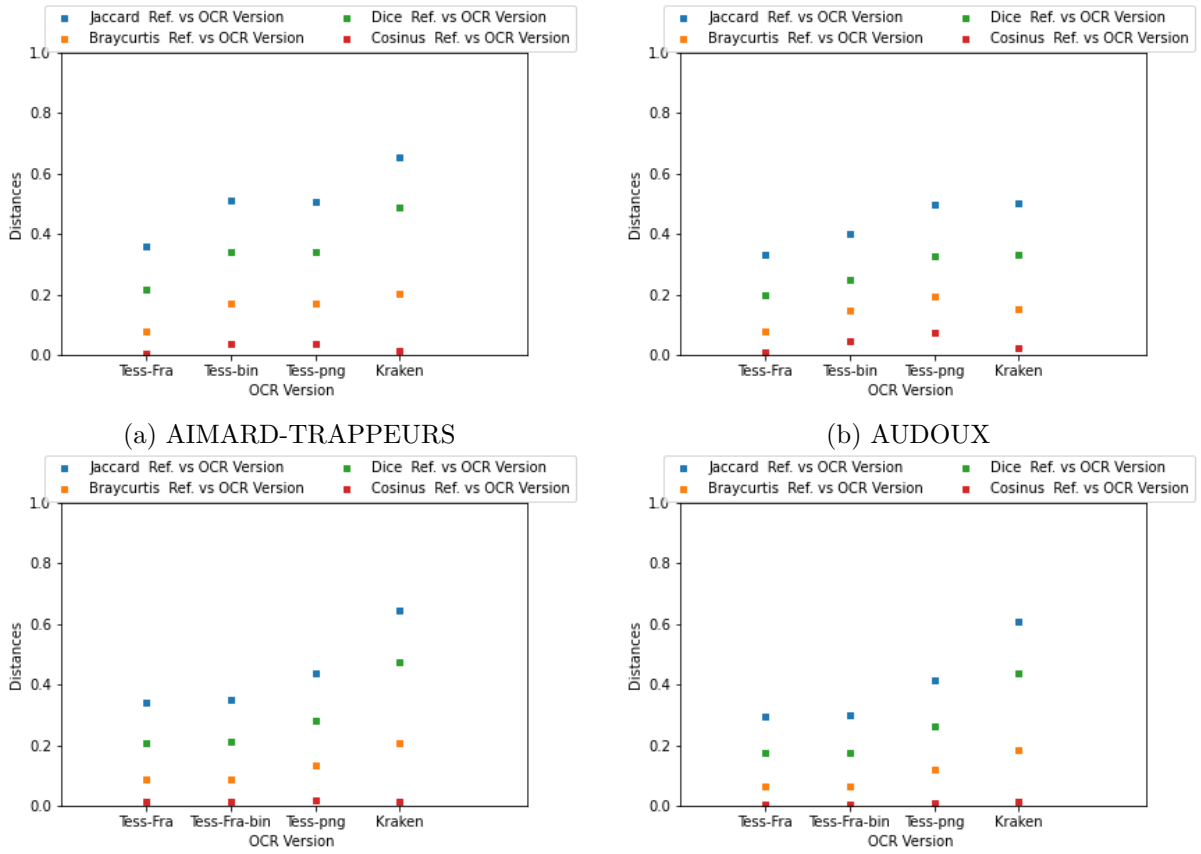


(d) Tess-fr – Stanza

Figure 5.4 – Intersections et différences entre les entités trouvées par `spacy_lg` et Stanza dans la version de référence (NE Ref) et les versions Kraken et Tess-fr pour le roman « Mon village » de Juliette Adam

nous avons testé Tesseract sur ces fichiers binarisés (`Tess-png`). On observe globalement que les résultats de Tesseract sont meilleurs en particulier lorsque l'on utilise le modèle approprié Tess-Fra. D'autre part, la binarisation des images détériore les résultats de Tess-Fra. Ce que l'on voit aussi c'est que les quatre mesures donnent des résultats assez différents. Les valeurs obtenues par Dice et Jaccard sont plutôt élevées et reflètent un fait déjà évoqué : les éléments les plus marqués par le bruitage des données sont les éléments rares. Or, comme ces mesures se basent sur une représentation binaire (présence ou absence d'un token), les hapax disparus et les nullax apparus ont une grande influence sur les résultats. On observe que la distance obtenue avec l'indice de Dice reste moins élevée que celle obtenue avec Jaccard. Si l'on s'intéresse à l'autre famille de métriques, on va observer que la distance cosinus est significativement plus « tolérante » par rapport aux écarts. Il est difficile de juger à ce stade si c'est la mesure cosinus qui sous-estime la distance ou au contraire qu'avec Bray-Curtis on la surestime. Ce qui est certain c'est que le choix de la mesure de distance n'est pas innocent et que cela questionne un des paradigmes du TAL qui est le fait que la distance cosinus est utilisée très régulièrement sans être forcément comparée à d'autres mesures, pour une étude de ces mesures sur une tâche de similarité entre phrases voir [Buscaldi et al., 2020] ou nous sommes intéressés à cet autre paradigme du TAL.

La Figure 5.6 représente cette fois les distances entre les sorties REN vues comme des séquences d'entités. Au contraire de l'évaluation à partir des diagrammes de Venn, ici l'intensité de la présence d'une entité (le nombre d'occurrences) pourra être mesurée (avec cosinus et Bray-



(a) AIMARD-TRAPPEURS

(b) AUDOUX

(c) DASH

(d) NOAILLES

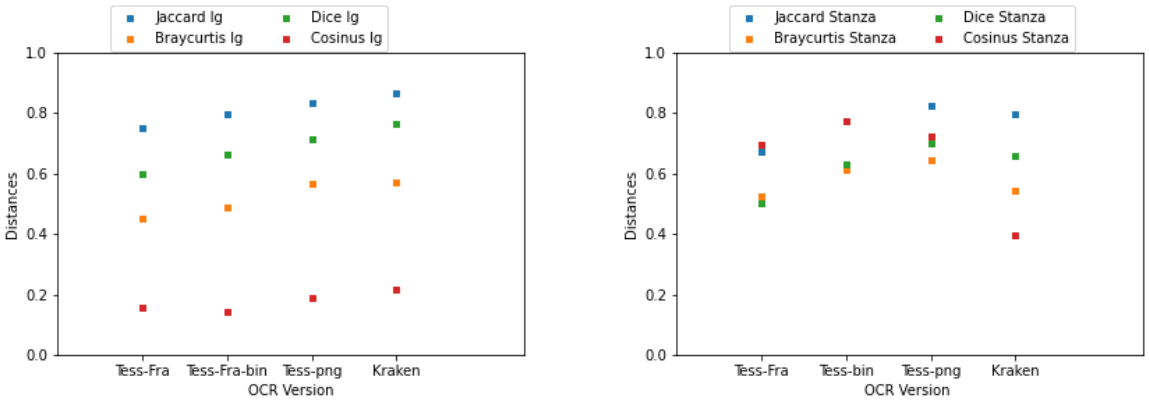
FIGURE 5.5 – Mesure de la qualité des retranscriptions OCR de Audoux, Aimard, Dash et Noailles

curtis). On peut voir que les distances mesurées sont plus grandes sur la sortie que sur l'entrée. Ceci s'explique en grande partie par les phénomènes rares. En effet, l'écart observé est moins grand sur le couple Jaccard/Dice que sur le couple Cosinus/Bray-Curtis. On peut soupçonner que la distance cosinus minimise les écarts en donnant une importance (trop) grande aux entités assez fréquentes. Ce qui expliquerait pourquoi la distance cosinus se rapproche fortement de la distance de Bray-curtis dans les résultats de STANZA, modèle dont on a montré qu'il était plus sélectif.

Quand la distance cosinus est supérieure aux autres comme dans le cas de la figure 5.7d (entités détectées par STANZA dans la version TESS-PNG), cela signifie que les résultats sont effectivement très mauvais. Et en effet, c'est corrélé au fait que l'intersection au niveau des entités est très petite (17 éléments).

Comme montré plus haut, la sortie STANZA est globalement moins bruitée, en particulier par ce que ce modèle extrait peu de variantes graphiques en comparaison des différents modèles SPACY, cette observation est cohérente avec les résultats présentés par [Qi et al., 2020]. Un élément important qui est ressorti des tests effectués ici est le temps de calcul. Traiter l'intégralité du corpus avec SPACY_SM nécessite quelques minutes (et moins d'une heure pour faire tourner l'ensemble des trois modèles) tandis qu'avec STANZA il faut compter pas moins de 15h [90].

90. Test mené sur un ordinateur portable avec 4 curs de 1,8 Ghz et 16 Go de RAM



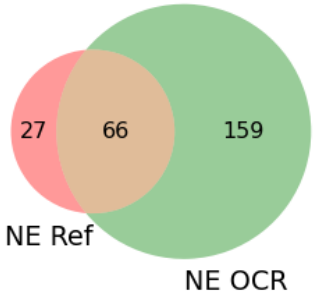
(a) `spacy_lg` (b) Stanza

Figure 5.6 – Mesure de la distance entre les sorties REN obtenues par `spacy_lg` et Stanza sur « Mon village » de Juliette Adam

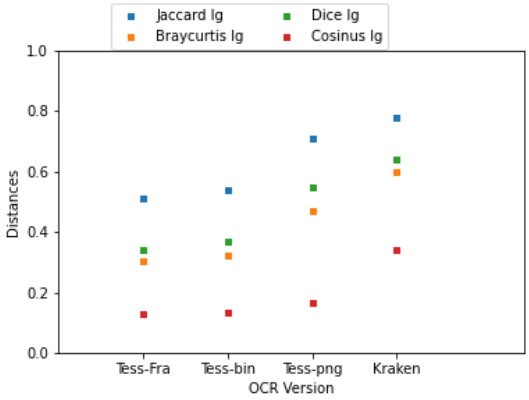
Version :: Modèle	Jaccard	Dice	Bray-curtis	Cosinus
Tesseract ::`spacy_sm`	0.900 **(p=0.000)**	0.900 **(p=0.000)**	0.600 (p=0.051)	0.682 **(p=0.021)**
Tesseract ::`spacy_md`	0.900 **(p=0.000)**	0.900 **(p=0.000)**	0.873 **(p=0.000)**	0.627 **(p=0.039)**
Tesseract ::`spacy_lg`	0.900 **(p=0.000)**	0.900 **(p=0.000)**	0.891 **(p=0.000)**	0.645 **(p=0.032)**
Tess-fr ::`spacy_sm`	0.827 **(p=0.002)**	0.827 **(p=0.002)**	0.600 (p=0.051)	0.500 (p=0.117)
Tess-fr ::`spacy_md`	0.836 **(p=0.001)**	0.836 **(p=0.001)**	0.627 **(p=0.039)**	0.709 **(p=0.015)**
Tess-fr ::`spacy_lg`	0.809 **(p=0.003)**	0.809 **(p=0.003)**	0.873 **(p=0.000)**	0.582 (p=0.060)
Kraken ::`spacy_sm`	0.682 **(p=0.021)**	0.682 **(p=0.021)**	0.700 **(p=0.016)**	-0.109 (p=0.750)
Kraken ::`spacy_md`	0.400 (p=0.223)	0.400 (p=0.223)	0.664 **(p=0.026)**	-0.191 (p=0.574)
Kraken ::`spacy_lg`	0.618 **(p=0.043)**	0.618 **(p=0.043)**	0.664 **(p=0.026)**	-0.264 **(p=0.433)**

Table 5.8 – Corrélation de Spearman (*p-value* entre parenthèses, en gras si significative) entre la distance mesurée sur le texte et la distance mesurée sur les entités pour les différentes versions et modèles étudiés

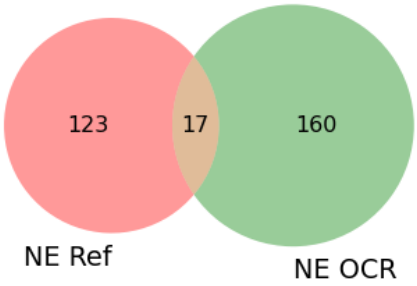
Dans le tableau 5.8 figure un étude de la corrélation entre la distance entre les versions en entrée (référence VS OCR) et la distance entre les entités en sortie. Ce calcul est effectué sur l'ensemble du corpus, il s'agit donc de la corrélation entre deux séries de 11 valeurs. En premier lieu, on observe que la corrélation est très forte avec Dice et Jaccard ce qui paraît confirmer l'observation déjà faite que la qualité des données a avant tout un impact sur les entités rares. Pour ces deux distances, il n'y a que dans le cas de `spacy_md` appliqué à Kraken qu'il n'y a pas de corrélation, ni d'ailleurs de p-value significative. Du point de vue des versions, c'est avec les versions Tesseract que les corrélations les plus fortes sont observées. Du point de vue de Kraken, cela pourrait signifier qu'une partie des erreurs OCR ne causent pas d'interférences pour la REN, ce qui amène sans doute à reconsidérer la manière dont on calcule la qualité : toutes les erreurs ne se valent pas. Encore une fois, l'amélioration mesurée de la qualité des données ne correspond pas toujours à une amélioration de la qualité des sorties.



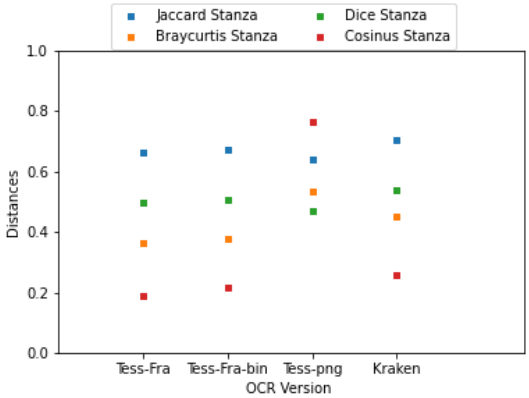
(a) Diagramme de Venn des entités extraites par `spacy_lg`



(b) Distances au niveau des entités pour `spacy_lg`

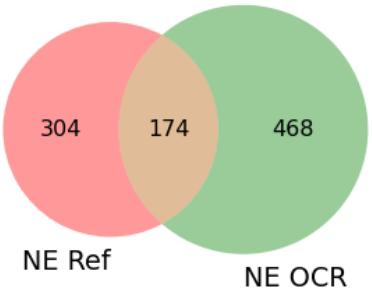


(c) Diagramme de Venn des entités extraites par Stanza

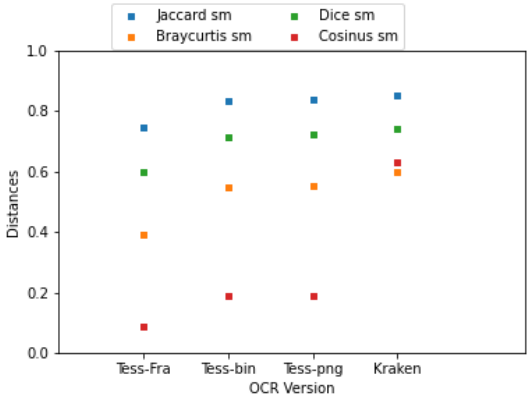


(d) Distances au niveau des entités pour Stanza

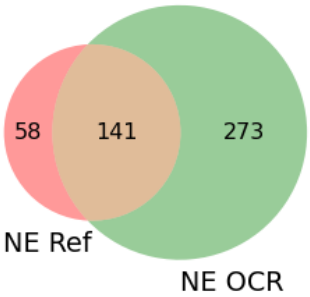
Figure 5.7 – Résultats de `spacy_lg` et Stanza sur la version Tess-fr de *La nouvelle espérance* d'Anna de Noailles



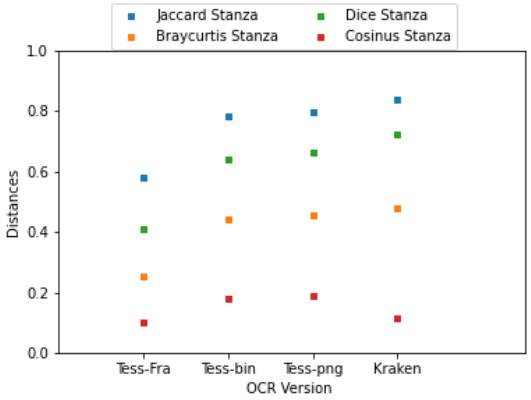
(a) Diagramme de Venn des entités extraites par `spacy_sm`



(b) Distances au niveau des entités pour `spacy_sm`



(c) Diagramme de Venn des entités extraites par Stanza



(d) Distances au niveau des entités pour Stanza

Figure 5.8 – Résultats obtenus par `spacy_sm` et Stanza sur la version Tess-fr de « La belle rivière » de Gustave Aimard

5.2.5 Conclusion sur les interférences OCR – REN

L'étude sur l'impact des interférences de l'OCR sur les sorties REN que nous avons présenté ici, a débouché sur plusieurs conclusions et au moins autant de questions. En comparant 16 situations (4 OCR et 4 modèles de REN combinés) sur 11 romans différents nous avons pu observer que la robustesse des modèles de REN, bien que ce ne soit plus des modèles état de l'art, est correcte. En effet, les faux positifs que l'on peut détecter sont régulièrement des variantes orthographiques générées par l'OCR, faux positifs qui pourraient pour une bonne partie d'entre eux se résoudre par de l'*entity linking*. Nous avons parlé ainsi de faux faux positifs (ou faux FP). Le problème du silence, des faux négatifs, semble plus important sans pour autant amener à la conclusion que les interférences soient trop grandes pour espérer obtenir un rappel raisonnable. D'une part, bon nombre d'entités ne sont de toutes façons pas détectées par les modèles REN sur les données de référence. En l'espèce du point de vue du résultat de l'OCR on pourrait parler en étendant la typologie de faux vrai négatif puisque ce sont des entités absentes de la sortie REN de référence mais pour de mauvaises raisons. Nous avons aussi le cas des faux faux négatifs (faux FN) puisque le bruitage amenant des décisions différentes de la part des modèles de REN, des entités erronées sont parfois détectées sur la version de référence mais pas sur la version OCR. Finalement, la corrélation entre la qualité de l'OCR et la qualité de la REN n'est pas systématique et n'est pas aussi grande que l'on pourrait l'imaginer. Cela concerne avant tout les entités rares, en particulier hapax, qui sont plus susceptibles de disparaître des sorties dans une version détériorée. Ceci pose la question du choix de la métrique. Cherche-t-on a trouver les entités rares, autrement dit l'ensemble du vocabulaire géographique utilisé ? Ou bien cherche-t-on plutôt les occurrences, donnant ainsi plus d'importance aux phénomènes fréquents (voire évidents) ? La question n'est pas évidente à trancher. Au-delà de ces choix d'évaluation, bien évidemment l'amélioration de la robustesse des modèles REN aux micro-variations permettra de régler un certain nombre d'erreurs. Du point de vue de l'entrée, on observe que toutes les erreurs ne se valent pas et que des corrections ciblées permettraient sans doute d'améliorer significativement les résultats à moindre coût. Trois grands types de bruits semblent particulièrement à l'uvre :

— La présence de mots concaténés par erreur du fait d'une mauvaise détection de l'espace typographique
— Le changement de la première lettre d'un mot, en particulier quand il s'agissait d'une capitale dans la référence
— La disparition de la notion de phrase lorsque la ponctuation devient erratique

Une étude plus fine des phénomènes illustrés nécessiterait sans doute une évaluation manuelle *a posteriori*, ou en amont l'alignement des sorties sur des échantillons du corpus. Il serait peut être intéressant de chercher les zones où la différence entre les résultats des OCR et des modèles REN sont particulièrement grandes et qui seraient sans doute instructives sur les verrous à lever. Il serait sans doute aussi utile d'examiner plus avant l'impact de ces entités rares et à quel point l'*entity linking* serait une solution pour limiter le problème. Ensuite, comparer ces résultats avec des résultats dans d'autres langues serait sans doute très fécond. Si l'on avait par exemple d'un côté l'anglais comme langue très bien doté en outils de qualité et de l'autre côté une langue comme l'espagnol, on pourrait sans doute mesurer mieux à quel point la robustesse des systèmes OCR et REN en anglais est plus élevée qu'en français ou en espagnol. Tester d'autres alphabets, le cyrillique ou le chinois, serait sans doute également intéressant. Enfin, on pourrait tester des langues où la casse est une caractéristique moins importante pour la détection des entités : est-ce que cela rend la tâche plus difficile ou est-ce qu'au contraire les modèles sont plus robustes du fait que ce critère de capitalisation est moins opérant ? Dans la section suivante, nous présenterons

un autre aspect de la problématique des données bruitées en s'intéressant à un corpus plus ancien (17ème siècle), à la fois hétérogène et bruité, le corpus des Mazarinades.

5.3 La standardisation des données est-elle indispensable à la (bonne) application de méthodes de TAL ?

Si l'on souhaite exploiter des données textuelles qui ne soient pas hors-sol, on doit bien évidemment disposer de corpus qui conservent les diverses caractéristiques des textes originels. Mais on sait aussi que le choix d'avoir des corpus peu structurés (voire déstructurés) correspond aussi à une vision pragmatique : il n'est pas évident de conserver tous les observables. Les observables obtenus résultent d'un arbitrage entre la donnée d'entrée (ce que l'on a réellement), la méthode (ce que l'on veut pratiquer comme traitements) et la tâche (ce que l'on s'est fixé comme objectif).

Nous avons évoqué l'impact du bruitage des données sur les chaînes de traitement de TAL ou plus généralement les tâches que l'on cherche à remplir. Nous nous plaçons dans un cas d'usage où la recherche de corpus trop épurés pourrait nuire au traitement automatique puisque l'on s'interdirait d'utiliser les données avant qu'elles ne satisfassent un degré très élevé de qualité. Si le coût d'obtention d'une telle qualité des données s'avère trop important, le passage à l'échelle dans l'analyse de corpus devient impossible.

5.3.1 Des données en ligne à un corpus manipulable

Les expériences et réfléxions présentées ici sont issues du projet Antonomaz [91] mené avec Karine Abiven et qui a servi de cadre à la thèse de Jean-Baptiste Tanguy [Tanguy, 2022]. Je présente ici un des aspects de cette collaboration avec l'analyse stylistique d'un corpus bruité, les questionnements méthodologiques et les résultats expérimentaux afférents (voir [Abiven et al., 2021] pour une description plus complète des enjeux). Il s'agissait d'explorer un ensemble d'écrits dits « burlesques » [92]. On estime que sur les 5 500 brefs imprimés pendant la Fronde (1648-1653), traditionnellement appelés « mazarinades », plusieurs centaines relèveraient de cette pratique d'écriture, caractérisée par le mélange. Le chiffre de 1 300 écrits burlesques parus pendant la Fronde, soit un quart des écrits du corpus, a été évoqué par certains auteurs [Carrier, 1996]. Malgré ce nombre relativement modeste, il s'avère en pratique difficile d'exploiter ce sous-corpus via une lecture de près, plusieurs auteurs choisissant en conséquence de n'en traiter qu'une sous-partie [Briot, 1993, Leclerc, 2013].

On peut dès lors trouver tentant d'éloigner quelque peu la focale et de chercher une autre échelle pour faire émerger des caractéristiques, notamment métriques et lexicales. On pourrait ainsi exposer la spécificité de cette pratique d'écriture poétique. C'est en constituant une liste de titres exploitable en format numérique et en regroupant les données librement disponibles à ce jour pour les constituer en corpus de Mazarinades burlesques que nous avons posé les jalons d'une telle étude. En effet, il n'existait pas de corpus des Mazarinades représentatif et librement disponible. Nous avons donc dû en constituer un.

Le site du projet « Recherches Internationales sur les Mazarinades »(RIM) [93], propose bien un corpus entièrement revu à main, mais il est plus petit (2 000 pièces VS 3 000) que celui

91. `https://antonomaz.huma-num.fr/` consulté le 2 octobre 2023

92. *Burlesque* qualifiant tantôt style, registre, ou « genre d'écrire ». En l'espèce, le qualificatif genre ou registre reste soumis à discussion [Nédelec, 2004]

93. `http://mazarinades.org/` consulté le 2 octobre 2023

que nous avons pu rendre disponibles sur Antonomaz. De plus, l'interface du site RIM ne permet pas le téléchargement du corpus complet et n'autorise que des requêtes lexicales et non la confrontation à des corpus contrastifs. Ce qui a justifié la constitution par nos soins d'un corpus plus bruité mais plus représentatif [Abiven et al., 2022]. L'intérêt d'aller au-delà d'un repérage purement lexical, notamment en constituant des sous-corpus, doit permettre de mieux saisir les spécificités internes de ces données. L'analyse purement quantitative du lexique des Mazarinades a d'ailleurs été critiquée par [Jouhaud, 2009]. Dans ce contexte, et afin de justifier le passage à l'échelle et l'intérêt d'employer des méthodes automatiques, nous avons utilisé une méthodologie plus exigeante impliquant : (I) l' analyse de la tâche et la production des hypothèses, (II) la constitution du corpus de référence et du corpus de travail, (III) le traitement instrumenté de ces corpus en exploitant le contraste et (IV) l'interprétation des résultats et le retour aux sources pour validation [Rastier, 2011]. Bien sûr se pose le problème de l'acquisition des données, l'intégralité du corpus des mazarinades, de même que sa sous partie burlesque, n'étant pas accessible numériquement. Jugeant impossible l'atteinte de la complétude du corpus, nous en avons construit un échantillon, en choisissant d'exploiter des données ni traitées ni enrichies, mais directement sorties du processus d'océrisation des données textuelles contenues dans une image numérique, sachant bien sûr que cela génère nécessairement du bruit mais aussi du silence puisque des formes peuvent disparaître.

Ce choix doublement pragmatique, incomplétude du corpus constitué et caractère bruité du corpus obtenu, pose une double question méthodologique : à quel point peut-on se fier (I) à la transcription automatique et (II) aux résultats statistiques obtenus sur celle-ci. Cette réflexion se situe donc à l'articulation entre l'acquisition automatisée des données textuelles et la pensée de leur usage en corpus. Entre une approche orientée *data paper* et des exemples applicatifs, il s'est agi à la fois d'exposer des données et de prouver par quelques conclusions stylistiques que leur qualité inégale n'est pas un obstacle à leur exploration et qu'elle n'interdit pas de penser concomitamment leur possible amélioration. Cette démarche rejoint des questionnements qui agitent une partie de la communauté linguistique de corpus [94].

5.3.2 À la recherche des traits d'écriture burlesque en contexte bruité

La production imprimée « en vers burlesques »est un véritable phénomène éditorial à l'époque étudiée, de sorte que les mazarinades sont parfois réduites à ce « style ». L'écriture burlesque est notamment caractérisée par une visée ludique, voire railleuse. Ce « genre d'écrire »est le plus souvent décrit en extension, c'est-à-dire par ses manifestations concrètes plutôt qu'en intension, par des traits définitoires, sauf pour ce qui est de l'innovation lexicale. Il est alors tentant de se focaliser sur ce point. Ceci est intéressant pour le TAL puisque l'on sait (I) que le repérage des innovations lexicales et néologismes au sens large pose un certain nombre de problèmes méthodologiques de la représentativité des dictionnaires de référence (ou dictionnaires d'exclusion [Cartier and Sablayrolles, 2009]) au positionnement temporel de ceux-ci [Yapomo and Lejeune, 2022] en passant par le bruit inhérent au processus d'isolation et de comparaison des lexies [Lejeune and Cartier, 2017]. Les travaux sur le sujet s'accordent sur le fait que le style burlesque est principalement repérable par la présence de néologismes et l'usage différentes variations du français : diachronique, diatopique, et diastratique à travers du jargon technique, des tours populaires et de l'argot principalement. En somme, il s'agit pour nous ici de repérer la déviation par rapport à un standard. Les acquis stylistiques et littéraires ont été

94. Voir par exemple le numéro à venir de la revue Corpus « Bruit de fond ou valeur ajoutée ? Gérer le bruit lors des traitements informatiques des corpus linguistiques » `https://journals.openedition.org/corpus/2577` consulté le 2 octobre 2023

jusqu'ici obtenus par des intuitions appuyées sur l'analyse d'échantillons, par induction à partir de « quelques coups de sonde »(voir par exemple [Nédelec, 2004]), ce qui est sans doute incontournable quand le terrain d'étude est problématique d'accès. La méthode statistique ici adoptée permet, grâce à un corpus relativement important et sélectionné, la vérification de ces hypothèses confrontées à la réalité des données.

5.3.3 Constitution d'un corpus d'écrits burlesques de la Fronde

Lister des titres : des bibliographies papier à leurs numérisations

Si aucune liste exhaustive recensant l'ensemble des « mazarinades »n'existe, des bibliographies de référence [95] permettent d'atteindre précisément 5 059 références, sans compter les « textes fantômes »(dont on sait qu'ils ont existé mais dont on n'a pas conservé d'exemplaires). Entre documents introuvables physiquement et documents introuvables sous forme numérique, on voit que le bruitage au niveau local n'est qu'un tout petit aspect des problématiques méthodologiques traitées ici. Nous avons constitué par océrisation [96] et mis en ligne sous forme structurée une bibliographie complète des Mazarinades avec les méta-données associées [97]. Au-delà de l'intérêt pratique, aucune ressource complète et interrogeable n'existant préalablement [98] , ce travail a été un banc d'essai très utile sur la fiabilité des données bruitées. Par exemple, sur la recherche automatique en ligne de documents dont le titre est imparfaitement retranscrit dans la ressource et dont les manifestations en ligne peuvent comporter des variations (non-dissimilation du *u* et du *v* par exemple).

Le corpus burlesque est constitué par ceux des 5 059 titres qui contiennent « burlesque » - ou une variante graphique de ce terme (« bvrlesque »par exemple). Supposer qu'une pièce est burlesque si son titre l'indique donne sans doute un peu trop de confiance aux étiquettes endogènes mais la dimension commerciale de ces écrits implique souvent un étiquetage explicite pour un argument de vente de choix. Le sous-titre *en vers burlesques* est si fréquent qu'il a toutefois fallu ôter les faux positifs liés au fait que le titre ne renvoie pas au contenu, mais seulement à l'octosyllabe (mètre presque systématique dans l'écriture burlesque). À l'opposé, pour réduire le silence, un repérage manuel a permis d'ajouter quelques pièces dépourvues de titraille explicite [99].

Cette liste s'élève, en l'état, à 250 titres [100] 179 de ces pièces sont librement disponibles au téléchargement en ligne [101]. Nous présentons deux exemples de problèmes de numérisation observés dans un même document *La nappe renversée, chez Renard, en vers burlesques*, téléchargées sur Google Livres [102]. La page 6 (Figure 5.9) est très difficilement lisible car on y devine par transparence la page suivante dont l'encre a déteint ; le phénomène est fréquent pour des

95. *Bibliographie des Mazarinades* de Célestin Moreau et ses suppléments successifs ([Moreau, 1850, Moreau, 1862, Moreau, 1869] auxquels il faut ajouter des ajouts ultérieurs ([Socard, 1876, Labadie, 1904])

96. Avec Kraken 2.0

97. `https://antonomaz.huma-num.fr/tools/Biblio_Moreau.html` consulté le 2 octobre 2023

98. Un travail plus minutieux de refonte et de mise en ligne de la bibliographie des mazarinades est entrepris par la Bibliothèque Mazarine depuis 2019 avec 1200 notices atteintes en janvier 2020, soit près de 20% de l'ensemble des données attendues à terme, aucun chiffre plus récent n'étant disponible (`https://tinyurl.com/mazarine-pages-presentation`, consulté le 2 octobre 2023).

99. En prose (*Lettre de remerciment envoyee au Cardinal Mazarin... avec la harangue de dame Denise*) comme en vers (*Le Médecin politique, qui donne un souverain remède, pour guérir la France malade à l'extrémité*)

100. Ce qui amène à relativiser l'estimation à 1 300 libelles en vers burlesques évoquée plus haut.

101. Entre Gallica (`https://gallica.bnf.fr/`) et Google Livres (`https://books.google.fr/`) consultés le 2 octobre 2023

102. `https://tinyurl.com/nappe-renversee-renard` consulté le 2 octobre 2023

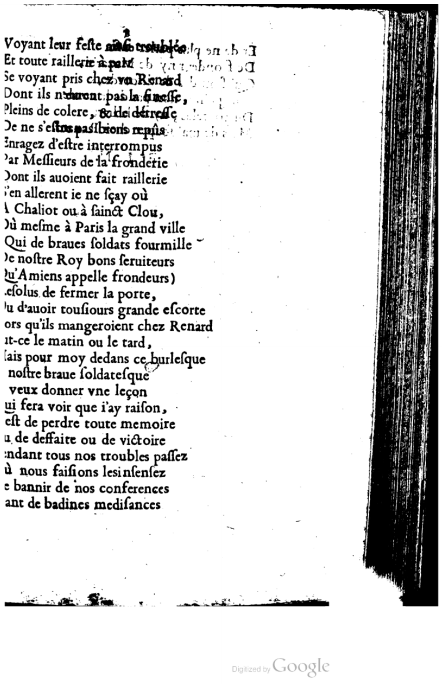


FIGURE 5.9 – Page 6 de La nappe renversée, chez Renard, en vers burlesques, téléchargé sur Google Livres

imprimés furtifs comme les mazarinades avec du papier fin et de mauvaise qualité. La page 12 (Figure 5.10) documente elle la présence humaine nécessaire pour numériser, et le le bruitage induit.

Océrisation et évaluation

Les océrisations ont été réalisées avec KRAKEN [103], outil choisi car on dispose d'un modèle spécialement entraîné pour le français du XVII[e]siècle [104] [Gabay et al., 2020].

Classiquement, le taux d'erreur au caractère (CER) est utilisé pour évaluer la qualité d'une sortie d'OCR. Cette mesure supervisée nécessite des transcriptions, manuelles par exemple, de référence, dont le présent projet ne dispose pas encore. Aussi, nous privilégions ici des méthodes d'évaluation non supervisées, il en existe un certain nombre :

- **Taux de confiance** : [Springmann et al., 2016] exploitent les taux de confiance, au caractère ou au mot, fournis par logiciels d'OCR ;
- **Ressources lexicales** : les mêmes auteurs explorent une autre piste en cherchant, pour chaque mot de la sortie d'OCR, la distance de Levenshtein le séparant de sa forme moderne la plus proche ;
- **Bounding boxes** [Gupta et al., 2015] utilisent les rectangles résultants de la segmentation de la page pour calculer la proportion de rectangles représentant du bruit ;
- **Modèles de langue** : [Tanguy, 2020] entraîne des modèles de langue au grain caractère sur un corpus de référence puis récupère les probabilités d'apparition des caractères de la sortie d'OCR.

103. `https://pypi.org/project/kraken/`, (consulté le 2 octobre 2023).
104. `https://github.com/e-ditiones/OCR17` (consulté le 2 octobre 2023)

FIGURE 5.10 – Page 12 de La nappe renversée, chez Renard, en vers burlesques, téléchargé sur Google Livres

Nous avons utilisé deux adaptations des méthodes proposées par [Springmann et al., 2016] : la moyenne des taux de confiance et le taux de lexicalité. La première, désignée T_{con}, calcule la moyenne, pour une page, des taux de confiance donnés par le logiciel d'OCR pour chaque caractère. Une page affichant un taux de confiance au caractère supérieur à la moyenne devrait donc être bien océrisée puisque Springmann *et al.* notent que « [...] tous les caractères ayant un taux de confiance supérieur à la moyenne (0,93) sont corrects » [105]. La seconde mesure, le T_{lex}, calcule, pour une page, la proportion de mots de la sortie d'OCR qui appartiennent à un lexique de référence. Le lexique utilisé ici est le LGERM [Souvay and Pierrel, 2009] qui compte plus de trois millions de formes fléchies. Cette mesure est ici invoquée pour savoir si, globalement, les mots présents dans les sorties d'OCR correspondent à une forme attestée.

Les figures 5.11 et 5.12 montrent que les taux de confiance sont globalement supérieurs au seuil de 0,93 cité plus haut. En moyenne, T_{con} est égal à 0,98, ce qui suggère une qualité au moins correcte des données océrisées, avec les précautions de rigueur puisque l'expérience montre que ce taux est très souvent supérieur à 0,90. D'ailleurs, l'analyse des taux de lexicalité nuance ces chiffres. En effet la lexicalité gravite autour de 0,4 pour l'ensemble des pages soit un peu moins de la moitié des mots qui appartiennent au LGERM. On peut constater trois causes : (I) l'incomplétude du LGERM d'un point de vue purement lexical (beaucoup de tokens existant réellement sont absents du lexique) , (II) le manque de variété des formes recensées pour un même token (le LGERM référençant une langue plutôt standardisée) et bien sûr (III) les erreurs d'OCR proprement dites.

105. *all characters with a confidence above average (0.93) are correct*

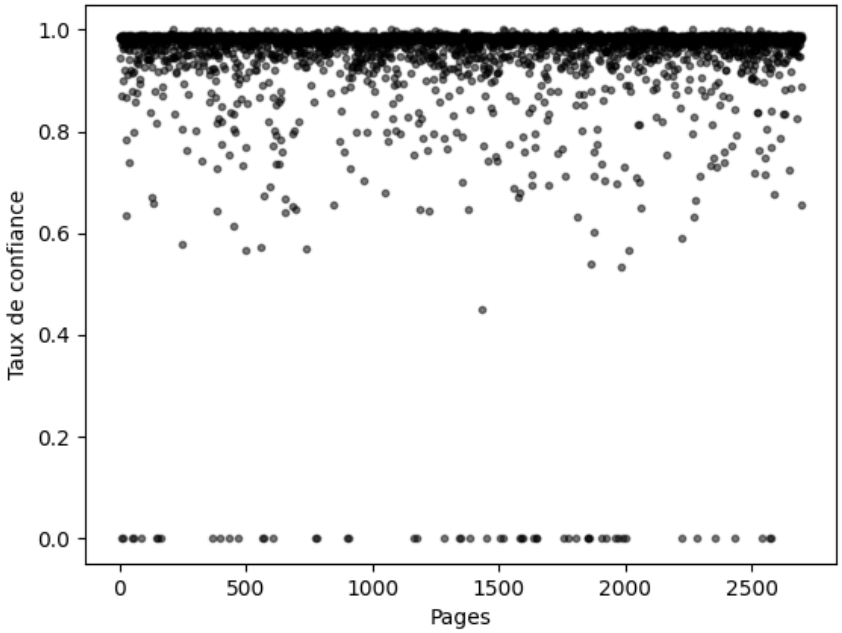


FIGURE 5.11 – Taux de confiance (T_{con}) calculé pour toutes les pages du corpus

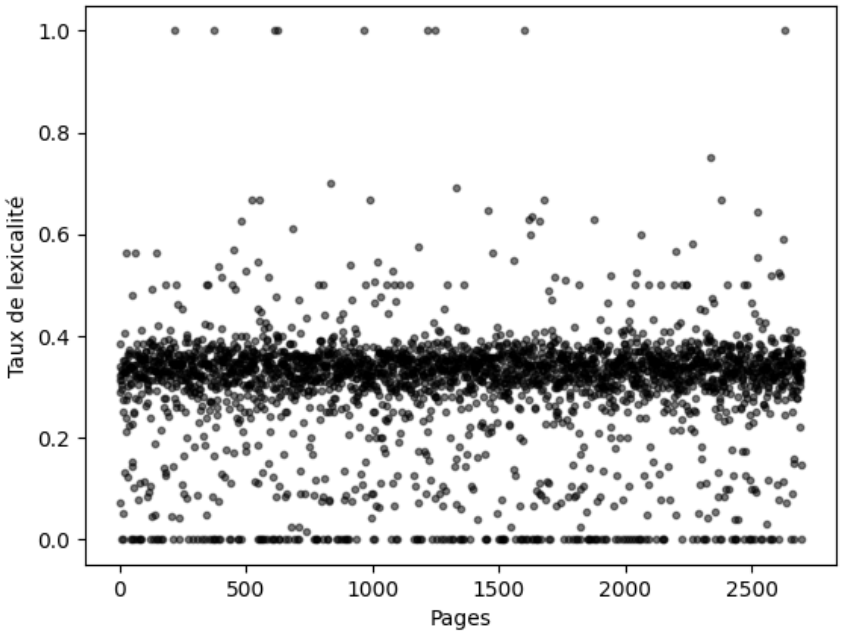


FIGURE 5.12 – Taux de lexicalité (T_{lex}) calculé pour toutes les pages du corpus

Influence du bruit et du silence sur les tables de fréquences

L'acquisition des données textuelles par OCR, dans une démarche d'analyse des faits de langue et de style sur corpus, pose on l'a dit les deux problèmes du bruit et du silence. Le premier, pluriel en tant que résultat d'une océrisation d'une portion de texte inexistante [106] mais aussi en tant que confusion entre caractères « proches » [107], a pour double effet d'ajouter de fausses entrées dans les tables de fréquences et de diminuer la fréquence des entrées mal reconnues. Le second, en tant qu'absence de reconnaissance de portions de texte pourtant présentes dans le document numérisé, implique une diminution des fréquences des formes non reconnues - et certains hapax deviennent des nullax.

L'analyse des tables de fréquences sera donc surtout utile pour l'observation de phénomènes saillants, peu susceptibles de disparaître complètement ou d'apparaître fortuitement. C'est une des conclusions positives de ce travail, qui rejoint des observations faites par exemple sur la langue allemande par [Lorenzen, 2021]. Une des conclusions (négative) de cette étude est que la quête des néologismes ne peut pas être menée avec une telle qualité de données sans un minimum de supervision.

Exploration contrastive des écrits burlesques de la Fronde

La caractérisation des spécificités d'une langue ou d'un genre peut se faire au moyen de corpus contrastifs, ou corpus de référence dans un outil comme Antconc. L'établissement de tels ensembles doit répondre à des critères explicites, mais, lorsque tout est à construire, il a parfois fallu trouver des compromis entre ce qui est possible et ce qui est désirable [Hunston, 2008].

Description des corpus contrastifs

La délimitation de ce que sont les textes burlesques pose en soi un problème comme vu plus haut, nous nous sommes fondés sur les 45 titres cités par Francis Bar [Deloffre, 1963]. Concernant l'obtention des données elles-mêmes, elle a été contrainte par la disponibilité numérique et l'obtention d'une taille suffisante pour être statistiquement représentatif. Le tableau 5.9 propose une représentation statistiques des corpus constitués. 56 œuvres dites « burlesques »ont été rassemblées pour constituer un corpus burlesque contrastif (hors Fronde, donc). Parallèlement, 1 092 œuvres non burlesques ont été rassemblées pour constituer un second corpus contrastif.

Exploitation du Corpus

Nous avons montré dans l'article précité que ces données bruitées étaient tout à fait utilisables avec un outil « Grand Public »tel que Antconc sans nécessiter d'outillage complexe ni de préparation intensive des données en amont. Á Titre d'exemple, la mesure de *keyness* (fondée au moyen de la *log-likelihood* et du χ^2 se sont avérées très robustes au bruitage des données, montrant ainsi que si l'absence de régularité dans le bruitage gène la correction , elle présente l'avantage de ne pas perturber les grandes tendances statistiques.

Ainsi, on peut vérifier l'existence de lexies référant à l'actualité en exploitant le contraste entre les écrits burlesques produits pendant la Fronde et ceux produits avant ou après (Burlesque Fronde *vs.* Burlesque hors Fronde). On observe sans surprise une fréquence supérieure de

106. Par exemple, la fausse reconnaissance de texte dans une illustration.
107. Par exemple, le *G* et le *O*.

Corpus	**# œuvres**	**# occurrences**	**# formes**
Non Burlesque	1 092	7 873 298	231 283
Burlesque	235	3 525 879	256 909
Burlesque Fronde	56	3 123 098	230 889
Burlesque hors Fronde	179	402 781	50 642
Burlesque Vers	202	1 733 044	170 243
Burlesque Prose	33	1 792 835	134 855

TABLE 5.9 – Description synthétique des corpus : nombre d'œuvres, nombre d'occurrences et nombre de formes (calculées avec Antconc)

noms propres dans le premier corpus Apparaissent ainsi dans les premières positions de la table de fréquence les principaux anthroponymes attendus : (I) ceux des personnalités impliquées dans les événements (*Mazarin*, *Condé*, *Beaufort*) ou (II) des partis en présence (*Parlement*) et de leurs appellatifs (*Messieurs* pour les parlementaires et *partisans* pour les financiers enrichis et associés à Mazarin). Notons bien sûr que la spécificité des noms propres aurait certainement été moins saillante avec un corpus contrastif d'imprimés dits « occasionnels », imprimés eux aussi en lien avec un fait d'actualité) mais l'objectif ici est de montrer que si le bruit (dans les données) produit du silence (dans les résultats), cela ne gêne que marginalement le repérage de phénomènes saillants, y compris au niveau lexical.

Un autre cas de spécificité détecté est la figure de la *muse*. Il apparaît dans les hautes fréquences du corpus BURLESQUE FRONDE, au rang 520 sur les 50 642 formes du corpus (63 « muse »et 14 « muses »). Elle apparaît également dans les spécificités positives du corpus BURLESQUE face au corpus NON BURLESQUE. Si ce résultat n'est pas en soit original, puisque le motif de la muse est très fréquent en poésie (donc dans le corpus BURLESQUE, majoritairement poétique), il montre là aussi que le caractère bruité des données n'implique pas l'impossibilité de vérifier par la lecture distante des hypothèses fondées sur la lecture rapprochée. On a ainsi pu observer différentes marques de l'interlocution : *ma muze*, *ma muse*, *notre muse*, *toy muse* ainsi que l'invocation à une muse « burlesque » ou « grotesque » déjà repérée comme typique du burlesque dès avant la Fronde [Nédelec, 2020]. L'étude a aussi montré que ce motif n'impliquait pas l'utilisation de termes liés à la mythologie, absents des traits saillants identifiés par contraste, sauf à penser que ces traces avaient, bien que fréquentes dans le document, été particulièrement atteintes par le bruit.

Un second exemple de trait spécifique repéré (ou vérifié) malgré le bruitage des données est l'utilisation du « burlesque On ». Les spécificités lexicales obtenues en opposant le corpus BURLESQUE FRONDE au corpus BURLESQUE HORS FRONDE montrent une sur-représentation des indéfinis, et en particulier donc du pronom personnel indéfini *on* (cf. Tableau 5.10) utilisé avec marquage stylistique dans deux types d'emploi : la personnification et l'anaphore.

Entre autres constats, nous avons pu voir qu'il apparaît largement dans des énoncés invoquant ou propageant une rumeur. L'observation simple du concordancier couplée à un tri des occurrences (selon leur contexte droit ou gauche) a permis de repérer un nombre important d'occurrences associées au verbe de discours rapporté *dire* : *on dit que*, *qu'on dit, comme on dit*, *on m'a dit*, *comme l'on dit, ...*

Bien que cette étude soit parcellaire, et nous n'entendons pas prétendre ici que tout corpus bruité est utilisable pour une analyse lexicométrique, elle donne je pense à voir l'utilisabilité de corpus bruités pour vérifier des hypothèses ou découvrir des phénomènes inconnus. Si, à petite

Corpus d'étude	Corpus de référence	Spécificité
Burlesque	Non Burlesque	822
Burlesque hors Fronde	Non Burlesque	370
Burlesque Fronde	Burlesque hors Fronde	333
Burlesque Fronde	Burlesque + Non Burlesque	571

TABLE 5.10 – Spécificité du « on »(*Keyness*) dans différentes situations de contraste

échelle, les données (raisonnablement) bruitées peuvent faire peur visuellement, à grande échelle elles sont statisiquement fiables à l'exception encore une fois des tokens de faible fréquence. Donner ainsi accès aux données permet aussi de questionner les pratiques du TAL et une certaine obsession que notre communauté scientifique peut avoir pour la standardisation des données.

5.4 Conclusion intermédiaire : de l'intérêt des données imparfaites et des moyens raisonnés de leur standardisation

Nous avons montré dans ce chapitre deux exploitations de corpus bruités pour des usages tout à faits différents : la reconnaissance d'entités nommées d'une part et la textométrie d'autre part. Ces premières explorations permettent d'assurer l'intérêt de considérer la fécondité des données bruitées utilisées en corpus et leur exploitabilité. Si cette exploitabilité est prouvée concernant les phénomènes fréquents, nous avons tout de même dessiné certaines limites au niveau de l'étude de la rareté.

Les données océrisées, imparfaites au sens où elles sont « bruitées » soit par la segmentation en mots ou par la transcription des caractères, n'interdisent pas l'utilisation de méthodes de TAL ou d'ADT. Ainsi, la transcription manuelle, si coûteuse en temps et en ressources (ou en problématiques éthiques si elle est sous traitée), n'est pas un passage nécessaire. Ceci autorise donc une approche pragmatique, opportuniste exploitant des données telles quelles sans considérer comme indispensable la conformité avec un standard quelconque. La quantité des données mise en jeu compense positivement, encore une fois pour une partie de ce que l'on cherche à observer, la qualité intrinsèque de celles-ci.

Des questions restent bien sûr en suspens : le « bruit » des données océrisées relève-t-il plus particulièrement du *bruit* ou du *silence* ? Autrement dit, dans les altérations observées quelles sont la proportion et l'influence des faux positifs (mots fantômes par exemple) ou des faux négatifs (phénomènes rares devenus absents).

Au demeurant, nous avons démarré plusieurs expérimentations pour améliorer tant la qualité intrinsèque des données textuelles que celle des méta-données qui les accompagnent. Nous avons constitué une vérité de terrain comprenant une retranscription manuelle de 120 pièces complètes ainsi que des premières pages de chacune des pièces collectées (3 059 à ce jour). Ceci nous permettra de ré-entraîner un modèle d'OCR sur nos données mais aussi, et ce qui nous semble plus prometteur scientifiquement, de chercher à combiner différentes versions OCR d'un même texte pour en fabriquer une nouvelle de meilleure qualité. Les approches en ce sens reviennent quelque peu sur le devant de la scène dans la communauté *Digital Libraries* reprenant des idées très utilisées en traitement de la parole telles que le ROVER [Fiscus, 1997], les systèmes à base de vote [Petrescu et al., 2019] ou encore les chaînes de Markov [Riddell, 2022]. Trouver des manières astucieuses de combiner des versions permettrait aussi de chercher la résolution optimale utile pour l'OCR, avec un impact non négligeable sur la soutenabilité à travers le stockage des images

(est-il nécessaire d'avoir pour chaque page un TIFF de 3 mégaoctets ?) et le temps de calcul. Conjointement, plus du côté de la linguistique de corpus, nous avons réalisé une structuration en XML-TEI nonobstant le bruitage du corps du texte (voir [Abiven and Lejeune, 2022]). Nous obtenons ainsi un corpus certes bruité mais avec des méta-données de grande qualité, permettant notamment de confronter des sous-corpus [108].

108. Voir `https://github.com/Antonomaz/Corpus/tree/main/Mazarinades` consulté le 2 octobre 2023

Conclusion générale

J'ai cherché dans ce travail à poser, à défaut de résoudre, la question de la prise en compte de la variation dans la conception et l'évaluation de méthodes de TAL. J'ai défini la variation comme l'ensemble des phénomènes qui contribuent à établir une différence entre un cas d'usage particulier et des conditions expérimentales déjà connues et bien maîtrisées. La variation pose la question de l'adaptabilité de méthodes état de l'art à des données dont les propriétés vont varier ou à des tâches qui vont s'éloigner des canons. J'ai cherché à montrer en quoi la variation gêne l'adaptabilité car elle pointe directement les limites de procédés de normalisation/standardisation des données que l'on peut observer régulièrement dans la littérature : suppression de la structure, du balisage, correction des fautes dans les tweets, post-correction de l'OCR ou encore modernisation d'états de langue anciens. J'ai identifié différents types de variations des conditions expérimentales, à travers les données, qui affectent l'adaptabilité des méthodes de TAL et qui décrivent les grands axes de mes recherches futures.

Il y a tout d'abord la variation en genre, chaque genre textuel apportant son lot de spécificités, faut-il chercher à représenter tout texte comme une séquence de phrases, tout corpus comme une agrégation de séquences de phrases ? Ou bien doit-on chercher à tirer le maximum du matériau d'origine en exploitant aussi, entre autres choses, la structure des documents ? Un exemple d'utilisation de certaines des propriétés du genre textuel a été présenté dans le chapitre 3 pour ce qui est du genre des articles scientifiques. Ces travaux avaient mené à d'autres développements que je n'ai pas abordé ici, notamment sur l'extraction/génération de tables des matières dans des documents PDF (voir [Giguet and Lejeune, 2019,Giguet et al., 2020]) dans une approche descendante plutôt qu'ascendante. Cet intérêt pour l'extraction de structure se manifeste également dans mon travail au sein du projet ANR Ecole [109] pour la période 2023-2026. Dans ce cadre va se poser la question de la comparaison en diachronie longue de la structure de différentes versions d'un même document. Différents genres textuels sont repérables dans les corpus du projet et nécessiteront certainement un traitement différencié qui tienne compte de leurs spécificités.

Ce projet me permettra également de continuer à m'intéresser à la question de la variation en langue : multilinguisme, frontière entre différentes langues ou états de langues ou encore instabilité des formes. En l'espèce, le passage du moyen français au français classique est un processus étalé dans le temps avec des dynamiques diatopiques propres. Il s'agit d'une extension de mes travaux précédents sur le multilinguisme, qui ont notamment connu des avancées dans le cadre du co-encadrement de la thèse de Steve Mutuvi, soutenue en novembre 2022, sur l'extraction d'évènements épidémiologiques multilingues [Mutuvi, 2022]. Dans ces travaux ont été successivement proposés une analyse comparative des approches de classification multilingue [Mutuvi et al., 2020b, Mutuvi et al., 2020a] puis une mise en lumière de l'importance du fine-tuning de modèles de langues [Mutuvi et al., 2021a] et de l'utilisation de *sub-words* du domaine de

109. `http://www.univ-paris3.fr/anr-ecole-747549.kjsp` consulté le 2 octobre 2023

spécialité pour améliorer la « stabilité multilingue » des résultats [Mutuvi et al., 2023]. Le multilinguisme fera l'objet d'un projet de thèse que j'aimerais mener à bien sur l'analyse diachronique des imprimés de large diffusion à l'époque moderne (écrits de colportage, notamment), dans une continuation scientifique du projet Antonomaz. Différents aspects liés à la variation linguistique et matérielle sont en jeu dans ce sujet, par exemple l'étude de la phraséologie et de son aspect diastratique, entre une phraséologie considérée comme populaire et une phraséologie plus noble (ou littéraire).

Un troisième axe que je prévois d'explorer est l'exploration des corpus de chansons car ce sont des données qui vont présenter des types particuliers de variation. Il s'agit en effet d'un genre textuel tout à fait particulier qui présente un certain nombre d'intérêts pour le traitement automatique des langues. En premier lieu, pour la socio-linguistique via l'étude de néologismes formels ou sémantiques. En second lieu, pour la linguistique des corpus [110] puisque se posent des questions, (en reprenant les critères d'acceptabilité de Bénedicte Pincemin [Bommier-Pincemin, 2000]), sur la représentativité (avec une sur-représentation de textes d'origine récente, mais aussi de certains genres musicaux), sur la régularité (des textes issus de *scraping* automatique, d'OCR, de transcription automatique ou « à l'oreille » de la parole) et la complétude (quid des versions successives d'un texte, de ceux dont les méta-données sont incomplètes ?).

Je co-encadre depuis octobre 2022 la thèse de Julien Bezançon qui porte sur la détection et la production de défigements linguistiques. Le défigement est l'opération par laquelle un énoncé idiomatique (proverbe ou citation par exemple), va être modifié par un locuteur pour produire un effet particulier (des jeux de mots notamment). Par exemple, à partir de l'expression devenue figée « Travailler plus pour gagner plus » on pourra défiger en « travailler plus pour gagner moins » [Bezançon and Lejeune, 2023]. Pour effectuer cette association, que l'on a choisi de décrire comme une tâche de similarité, un travail sur différents niveaux linguistiques (phonétique, morphologique, syntaxique ...) avec une réflexion sur comment calculer la similarité, sur quel grain d'analyse (phonème, morphème ...) et enfin comment agréger les résultats. Se pose aussi la question de savoir comment définir la tâche : le but est-il simplement de produire un système de détection que l'on va évaluer sur un corpus de référence annoté manuellement ? Ou bien est-ce une occasion, en liaison avec les linguistes du projet de poser des questions plus complexes sur la valeur ajoutée de l'automatisation ? Sur la capacité d'une même approche à fonctionner sur des tweets, des articles de presse, ou encore des données issues de retranscription de l'oral ou de la détection de slogans originaux dans des images ?

Ces cas de variation, parmi d'autres, illustrent les limites de la standardisation de la langue et de l'adaptabilité (pour ne pas dire de la robustesse) des représentations classiques en TAL. Cela, on l'a vu, est particulièrement prégnant dans les contextes bruités. Il serait intéressant de pouvoir diagnostiquer automatiquement si, au sein d'un même corpus, tous les documents (selon leur état de conservation pour des imprimés, selon leur « audibilité » pour des documents sonores ...) pourront être utilisés pour les mêmes tâches ou si l'on doit diféfrencier ces cas d'usage. Dans le domaine des humanités numériques (mais aussi dans d'autres) ceci permettrait de redessiner la place de l'expert métier en proposant un traitement différencié des données. Cela peut amener une intervention sur certaines instances particulières à des fins de validation ou de correction de résultats douteux, intervention qui dépasse la simple annotation destinée à l'apprentissage ou au ré-apprentissage. On peut y voir de nombreuses applications pratiques, et d'un point de vue général, j'entrevois de nombreux intérêts : (I) questionner les méthodes de TAL en les confrontant régulièrement à des situations nouvelles, (II) pouvoir questionner au fur et à mesure de l'exploration des corpus, la pertinence des méthodes et des objectifs en liaison

110. Le choix de l'article « des » est tout à fait intentionnel.

avec des besoins variés en humanités numériques et (III) contribuer au questionnement de la communauté TAL sur la gourmandise en ressources des dernières approches état de l'art.

Table des figures

Liste des tableaux

Annexe A

Ressources et tables de correspondance

FIGURE A.1 – Graphies du S long, la seconde est la plus fréquente dans les corpus anciens tels que les Mazarinades, Source : Wikipédia

Étiquette	Partie du discours
ADJ	Adjectif
ADV	Adverbe
ADVWH	Adverbe interrogatif
CC	Conjonction de coordination
CLO	Pronom clitique object
CLR	Pronom clitique réflechi
CLS	Pronom clitique sujet
CS	Conjonction de subordination
DET	Déterminent
DETWH	Déterminant interrogatif
ET	Mot étranger ou inconnu
I	Interjection
NC	Nom commun
NPP	Nom propre
P	Préposition
P+D	Préposition et déterminant combiné
P+PRO	Préposition and pronom combiné
PONCT	Ponctuation
PRO	Pronom
PROREL	Pronom relatif
PROWH	Pronom interrogatif
V	Forme verbale à l'indicatif
VIMP	Forme verbale à l'impératif
VINF	Forme verbale à l'infinitif
VPP	Forme verbale au participe passé
VPR	Forme verbale au participe présent
VS	Forme verbale au subjonctif

TABLE A.1 – Jeu d'étiquette du PennTreeBank utilisé par le Stanford POS tagger

Annexe B

Tableaux de résultats complets

Données supplémentaires pour le chapitre 4

Outil	Macro-F-mesure	Macro-Précision	Macro-Rappel	Micro-F-mesure	Micro-Précision	Micro-Rappel
TRAFFallback	**82,77** (±4, 1)	77,87 (±5, 5)	88,31 (±3, 5)	**83,33**	78,32 (±25)	89,02 (±19)
BP3Article	79,12 (±7, 4)	81,29 (±7, 0)	77,07 (±9, 1)	80,25	82,29 (±26)	78,31 (±35)
BP3Largest	77,48 (±5, 0)	**83,62** (±4, 4)	72,18 (±6, 9)	78,12	84,37 (±26)	72,73 (±35)
READABILITY	76,54 (±5, 8)	68,27 (±7, 7)	87,10 (±4, 8)	76,64	68,15 (±22)	87,56 (±19)
READ_py	76,05 (±8, 4)	66,39 (±10, 9)	**89,00** (±3, 6)	74,77	64,61 (±35)	88,72 (±17)
JT	75,80 (±36, 7)	81,35 (±6, 9)	70,95 (±38, 9)	73,98	81,70 (±25)	67,59 (±42)
JT_langid	75,41 (±36, 5)	81,46 (±7, 1)	70,21 (±38, 5)	73,51	81,79 (±25)	66,75 (±42)
JT_trueLg	75,41 (±36, 5)	81,46 (±7, 1)	70,21 (±38, 5)	73,51	81,79 (±25)	66,75 (±42)
TRAFFallbackComments	74,23 (±7, 4)	71,86 (±8, 3)	76,75 (±6, 4)	74,79	72,49 (±33)	77,25 (±31)
BP3	73,72 (±7, 3)	73,93 (±8, 5)	73,50 (±14, 6)	74,32	75,04 (±24)	73,61 (±33)
TRAF	72,69 (±11, 4)	68,87 (±10, 0)	76,95 (±13, 6)	74,73	70,71 (±31)	79,24 (±33)
TRAFComments	70,61 (±10, 8)	67,68 (±9, 5)	73,79 (±13, 1)	72,66	69,52 (±31)	76,09 (±35)
TRAF_BL	64,78 (±9, 0)	59,59 (±7, 0)	70,96 (±11, 9)	66,18	60,80 (±30)	72,62 (±32)
NEWSPLEASE	63,80 (±32, 2)	83,16 (±7, 4)	51,76 (±35, 7)	65,38	**84,39** (±21)	53,36 (±46)
GOO	51,07 (±40, 5)	80,37 (±10, 7)	37,42 (±39, 7)	54,80	82,25 (±22)	41,09 (±44)
HTML-text	47,94 (±6, 2)	32,45 (±5, 4)	91,68 (±1, 5)	48,58	33,04 (±19)	91,73 (±9)
BP3KeepEverything	47,00 (±6, 3)	31,21 (±5, 3)	95,11 (±2, 0)	47,50	31,67 (±20)	94,99 (±18)
NEWSPAPER	46,07 (±43, 5)	77,06 (±11, 8)	32,86 (±43, 4)	51,16	79,38 (±21)	37,74 (±45)
INSCRIPTIS	42,36 (±6, 6)	27,09 (±5, 2)	97,08 (±2, 1)	42,68	27,32 (±17)	**97,43** (±9)
HTML2TEXT	34,08 (±10, 2)	20,82 (±7, 3)	93,78 (±2, 8)	34,35	21,01 (±16)	94,13 (±13)
JT_english	28,74 (±36, 6)	73,11 (±9, 4)	17,88 (±37, 7)	37,24	75,10 (±19)	24,76 (±40)

(a) Mesure `clean_eval`

Tool	Macro-F-mes.	Macro-Précision	Macro-Rappel	Micro-F-mes.	Micro-Précision	Micro-Rappel
TRAFFallback	**81,12** (±8, 8)	75,47 (±11, 6)	**87,69** (±4, 5)	**81,12**	75,25 (±28)	87,99 (±16)
BP3Article	77,40 (±8, 4)	78,72 (±8, 2)	76,12 (±9, 9)	77,93	78,95 (±25)	76,94 (±32)
BP3Largest	76,21 (±8, 3)	**81,48** (±7, 0)	71,57 (±9, 6)	76,39	**81,54** (±24)	71,85 (±33)
JT	76,09 (±29, 4)	78,54 (±9, 2)	73,79 (±34, 7)	74,34	78,27 (±23)	70,80 (±38)
JT_langid	75,86 (±29, 3)	78,78 (±9, 4)	73,15 (±34, 4)	74,03	78,48 (±22)	70,06 (±38)
JT_trueLg	75,86 (±29, 3)	78,78 (±9, 4)	73,15 (±34, 4)	74,03	78,48 (±22)	70,06 (±38)
TRAFFallbackComments	75,12 (±8, 5)	71,65 (±10, 7)	78,94 (±6, 0)	75,43	71,84 (±32)	79,38 (±25)
BP3	73,64 (±8, 4)	72,22 (±7, 3)	75,12 (±14, 1)	73,93	72,66 (±23)	75,25 (±31)
READ_py	72,66 (±17, 6)	64,45 (±16, 5)	83,27 (±19, 3)	70,59	62,16 (±33)	81,67 (±24)
TRAF	71,94 (±11, 0)	67,79 (±11, 3)	76,63 (±12, 3)	73,30	68,74 (±31)	78,50 (±31)
READABILITY	70,80 (±20, 8)	62,46 (±22, 1)	81,71 (±13, 7)	69,64	61,08 (±29)	80,99 (±21)
TRAFComments	70,20 (±9, 8)	66,97 (±10, 5)	73,75 (±11, 4)	71,59	67,92 (±31)	75,68 (±32)
NEWSPLEASE	64,73 (±28, 1)	80,29 (±10, 3)	54,22 (±33, 2)	66,26	80,84 (±19)	56,14 (±43)
TRAF_BL	56,05 (±21, 7)	52,11 (±18, 6)	60,64 (±25, 9)	55,53	51,63 (±32)	60,07 (±37)
GOO	52,70 (±37, 5)	77,44 (±12, 2)	39,94 (±38, 0)	56,43	78,67 (±19)	43,99 (±43)
NEWSPAPER	47,82 (±41, 3)	74,09 (±12, 0)	35,31 (±42, 1)	52,87	75,77 (±18)	40,60 (±44)
BP3KeepEverything	42,62 (±14, 6)	27,53 (±11, 0)	94,35 (±4, 2)	42,25	27,25 (±20)	93,91 (±16)
HTML-text	39,88 (±19, 8)	26,73 (±14, 1)	78,52 (±30, 9)	38,96	26,20 (±21)	75,98 (±31)
INSCRIPTIS	37,34 (±15, 7)	23,27 (±11, 0)	94,47 (±6, 6)	36,64	22,74 (±18)	**94,15** (±12)
HTML2TEXT	33,16 (±13, 2)	20,25 (±9, 0)	91,49 (±9, 1)	33,12	20,24 (±17)	91,01 (±14)
JT_english	31,36 (±36, 1)	69,94 (±7, 7)	20,21 (±37, 4)	39,78	71,37 (±15)	27,58 (±39)

(b) Mesure `occ_eval`

TABLE B.1 – Évaluation sur le corpus multilingue, F-mesure calculée à partir des micro-moyennes de la Précision et du Rappel

Bibliographie

[Abadji et al., 2021] Abadji, J., Ortiz Suárez, P. J., Romary, L., and Sagot, B. (2021). Ungoliant : An Optimized Pipeline for the Generation of a Very Large-Scale Multilingual Web Corpus. In *CMLC 2021 - 9th Workshop on Challenges in the Management of Large Corpora*, Limerick / Virtual, Ireland.

[Abbasi and Chen, 2008] Abbasi, A. and Chen, H. (2008). Writeprints : A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2) :7.

[Abiven et al., 2022] Abiven, K., Bartz, A., Lejeune, G., and Tanguy, J.-B. (2022). Vers une collection numérique des libelles parus pendant la fronde, ou comment relier des mazarinades. *Le Verger*, 1(23).

[Abiven and Lejeune, 2019] Abiven, K. and Lejeune, G. (2019). Analyse automatique de documents anciens : tirer parti dun corpus incomplet, hétérogène et bruité. *Recherche d'information, document et web sémantique*, 2(Numéro 1).

[Abiven and Lejeune, 2022] Abiven, K. and Lejeune, G. (2022). Des données au corpus : l'exploitation numérique des mazarinades. pages 181–192.

[Abiven et al., 2021] Abiven, K., Tanguy, J.-B., and Lejeune, G. (2021). Exploiter un corpus de données textuelles sans post-traitement : lécriture burlesque de la Fronde. *Humanités Numériques*, 1(4).

[Agrawal and Srikant, 1994] Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, page 487499, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Alex et al., 2012] Alex, B., Grover, C., Klein, E., and Tobin, R. (2012). Digitised historical text : Does it have to be mediocre ? In Jancsary, J., editor, *11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing, Vienna, Austria, September 19-21, 2012*, volume 5 of *Scientific series of the ÖGAI*, pages 401–409. ÖGAI, Wien, Österreich.

[Bachimont, 2007] Bachimont, B. (2007). *Indexation et archivage de contenus multimédias*. Ed. Techniques Ingénieur.

[Baledent et al., 2020] Baledent, A., Hiebel, N., and Lejeune, G. (2020). Dating Ancient texts : an Approach for Noisy French Documents. In *Language Technologies for Historical and Ancient Langues (LT4HLA) @LREC2020*.

[Baledent and Lejeune, 2020] Baledent, A. and Lejeune, G. (2020). Automatic Stylistic Analysis : a search for efficient and interpretable descriptors to characterize individual writing style. In Fesenmeier, L. and Novakova, I., editors, *Phraséologie et stylistique de la langue littéraire / Phraseology and Stylistics of the Literary Language*, pages 329–342. Peter Lang.

[Bar-Hillel, 1960] Bar-Hillel, Y. (1960). The present status of automatic translation of languages. *Advances in computers*, 1 :91–163.

[Bar-Yossef and Rajagopalan, 2002] Bar-Yossef, Z. and Rajagopalan, S. (2002). Template Detection via Data Mining and its Applications. In *Proceedings of the 11th International Conference on World Wide Web*, pages 580–591.

[Barbaresi, 2015] Barbaresi, A. (2015). *Ad hoc and general-purpose corpus construction from web sources*. PhD thesis, École Normale Supérieure de Lyon.

[Barbaresi, 2016] Barbaresi, A. (2016). Efficient construction of metadata-enhanced web corpora. In Cook, P., Evert, S., Schäfer, R., and Stemle, E., editors, *Proceedings of the 10th Web as Corpus Workshop*, pages 7–16. Association for Computational Linguistics.

[Barbaresi, 2019] Barbaresi, A. (2019). Generic Web Content Extraction with Open-Source Software. In *Proceedings of KONVENS 2019, Kaleidoscope Abstracts*, pages 267–268. GSCL.

[Barbaresi, 2021] Barbaresi, A. (2021). Trafilatura : A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing : System Demonstrations*, pages 122–131.

[Barbaresi and Lejeune, 2020a] Barbaresi, A. and Lejeune, G. (2020a). Out-of-the-Box and Into the Ditch ? Multilingual Evaluation of Generic Text Extraction Tools. In *Proceedings of the 12th Web as Corpus workshop (WAC-XII)*. ELRA. à paraître.

[Barbaresi and Lejeune, 2020b] Barbaresi, A. and Lejeune, G. (2020b). Que recèlent les données textuelles issues du web ? In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)*, pages 19–28. ATALA ; AFCP.

[Baroni et al., 2009] Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web : a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3) :209–226.

[Baroni et al., 2008] Baroni, M., Chantree, F., Kilgarriff, A., and Sharoff, S. (2008). Cleaneval : a Competition for Cleaning Web Pages. In *Proceedings of LREC*, pages 638–643. ELRA.

[Bauer et al., 2007] Bauer, D., Degen, J., Deng, X., Herger, P., Gasthaus, J., Giesbrecht, E., Jansen, L., Kalina, C., Kräger, T., Märtin, R., Schmidt, M., Scholler, S., Steger, J., Stemle, E., and Evert, S. (2007). FIASCO : Filtering the internet by automatic subtree classification. In *Building and Exploring Web Corpora : Proceedings of the 3rd Web as Corpus Workshop (WAC-3)*, pages 111–121.

[Béchet et al., 2012] Béchet, N., Cellier, P., Charnois, T., and Crémilleux, B. (2012). Fouille de motifs séquentiels pour la découverte de relations entre gènes et maladies rares. In *Ingénierie des Connaissances*, pages 149–164.

[Ben Ltaifa et al., 2022] Ben Ltaifa, I., Boubehziz, T., Briglia, A., Chutaux, C., Dupont, Y., González-Gallardo, C.-E., Koudoro-Parfait, C., and Lejeune, G. (2022). Stylo@DEFT2022 : Notation automatique de copies d'étudiant×e×s par combinaisons de méthodes de similarité. In Grouin, C. and Illouz, G., editors, *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*, pages 11–22, Avignon, France. ATALA.

[Bender, 2019] Bender, E. (2019). The #BenderRule : On naming the languages we study and why it matters. *The Gradient*.

[Bender, 2009] Bender, E. M. (2009). Linguistically naïve != language independent : Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction Between Linguistics and Computational Linguistics : Virtuous, Vicious or Vacuous ?*, ILCL '09, pages 26–32. ACL.

[Bengio et al., 2003] Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null) :11371155.

[Bertin-mahieux et al., 2011] Bertin-mahieux, T., Ellis, D. P. W., Whitman, B., and Lamere, P. (2011). The million song dataset. In *In Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*.

[Bezançon and Lejeune, 2023] Bezançon, J. and Lejeune, G. (2023). Reconnaissance de défigements dans des tweets en français par des mesures de similarité sur des alignements textuels. In Servan, C. and Vilnat, A., editors, *18e Conférence en Recherche d'Information et Applications – 16e Rencontres Jeunes Chercheurs en RI – 30e Conférence sur le Traitement Automatique des Langues Naturelles – 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 56–67, Paris, France. ATALA.

[Bojanowski et al., 2017] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5 :135–146.

[Bommier-Pincemin, 2000] Bommier-Pincemin, B. (2000). *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*. PhD thesis, Paris 4.

[Boros et al., 2020] Boros, E., Hamdi, A., Linhares Pontes, E., Cabrera-Diego, L. A., Moreno, J. G., Sidere, N., and Doucet, A. (2020). Alleviating digitization errors in named entity recognition for historical documents. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 431–441, Online. Association for Computational Linguistics.

[Boros et al., 2021] Boros, E., Moreno, J. G., and Doucet, A. (2021). Event detection as question answering with entity information.

[Boros et al., 2022] Boros, E., Nguyen, K., Lejeune, G., and Doucet, A. (2022). Assessing the Impact of OCR Noise on Multilingual Event Detection over Digitised Documents. *International Journal on Digital Libraries*, 1(23) :241–266.

[Boré and Bosredon, 2016] Boré, C. and Bosredon, C. (2016). La phrase dans le texte. lexemple de phrases de dialogue dans un corpus décole élémentaire. *LIDIL*, 54 :1115–133.

[Brennan et al., 2012] Brennan, M., Afroz, S., and Greenstadt, R. (2012). Adversarial stylometry : Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)*, 15(3) :12.

[Briot, 1993] Briot, F. (1993). La rime et la déraison : l'épître en vers dans le temps de la fronde. *Littératures classiques*, 18(1) :159–171.

[Brixtel, 2015] Brixtel, R. (2015). Maximal repeats enhance substring-based authorship attribution. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 63–71, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

[Brixtel et al., 2015] Brixtel, R., Lecluze, C., and Lejeune, G. (2015). Attribution d'Auteur : approche multilingue fondée sur les répétitions maximales. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2015)*, pages 208–219.

[Brixtel et al., 2013] Brixtel, R., Lejeune, G., Doucet, A., and Lucas, N. (2013). Any Language Early Detection of Epidemic Diseases from Web News Streams. In *International Conference on Healthcare Informatics (ICHI)*, pages 159–168.

[Bunt, 2009] Bunt, H. (2009). The dit++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24.

[Burrows et al., 2014] Burrows, S., Uitdenbogerd, A., and Turpin, A. (2014). Comparing techniques for authorship attribution of source code. *Software – Practice and Experience*, 44(1) :1–32.

[Buscaldi et al., 2020] Buscaldi, D., Felhi, G., Ghoul, D., Le Roux, J., Lejeune, G., and Zhang, X. (2020). Calcul de similarité entre phrases : quelles mesures et quels descripteurs ? In Cardon, R., Grabar, N., Grouin, C., and Hamon, T., editors, *Traitement Automatique des Langues Naturelles (TALN, 27e édition). Atelier DÉfi Fouille de Textes*, pages 14–25, Nancy, France. ATALA.

[Buscaldi et al., 2017] Buscaldi, D., Grezka, A., and Lejeune, G. (2017). Tweetaneuse : Fouille de motifs en caractères et plongement lexical à l'assaut du deft 2017. In *Actes du 13e Défi Fouille de Texte*, pages 65–76, Orléans, France. Association pour le Traitement Automatique des Langues.

[Buscaldi et al., 2018] Buscaldi, D., Le Roux, J., and Lejeune, G. (2018). Modèles en caractères pour la détection de polarité dans les tweets. In *Actes du 14e Défi Fouille de Texte*, pages 249–258, Rennes, France. Association pour le Traitement Automatique des Langues.

[Cai et al., 2003] Cai, D., Yu, S., Wen, J.-R., and Ma, W.-Y. (2003). VIPS : a Vision-based Page Segmentation Algorithm. Technical report, Microsoft Research.

[Calberac, 2010] Calberac, Y. (2010). *Terrains de géographes, géographes de terrain. Communauté et imaginaire disciplinaires au miroir des pratiques de terrain des géographes français du XXe siècle*. PhD thesis, Université Lumière Lyon 2.

[Caliskan et al., 2017] Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334) :183–186.

[Camacho-Collados et al., 2014] Camacho-Collados, J., Billami, M. B., Jacquey, E., and Kister, L. (2014). Approche statistique pour le filtrage terminologique des occurrences de candidats termes en texte intégral. In *Jadt'14*, Paris, France.

[Carassus, 1970] Carassus, É. (1970). Maurice barrès feuilletoniste. *Revue d'Histoire littéraire de la France*, 70(1) :90–97.

[Carey and Manic, 2016] Carey, H. J. and Manic, M. (2016). HTML web content extraction using paragraph tags. In *25th International Symposium on Industrial Electronics (ISIE)*, pages 1099–1105. IEEE.

[Carrier, 1996] Carrier, H. (1996). Les muses guerrières. *Paris : Klincksieck*.

[Cartier and Sablayrolles, 2009] Cartier, E. and Sablayrolles, J.-F. (2009). Néologismes, dictionnaires et informatique. *Cahiers de lexicologie*, 2008(93) :175–192.

[Cellier et al., 2015] Cellier, P., Charnois, T., Plantevit, M., Rigotti, C., Crémilleux, B., Gandrillon, O., Klema, J., and Manguin, J.-L. (2015). Sequential pattern mining for discovering

gene interactions and their contextual information from biomedical texts. *Journal of Biomedical Semantics*, 6 :1–27.

[Chambers and Jurafsky, 2011] Chambers, N. and Jurafsky, D. (2011). Template-based information extraction without the templates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 976–986, Portland, Oregon, USA. Association for Computational Linguistics.

[Chasins et al., 2018] Chasins, S. E., Mueller, M., and Bodik, R. (2018). Rousillon : Scraping distributed hierarchical web data. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pages 963–975.

[Chaski, 2001] Chaski, C. E. (2001). Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 8 :1–65.

[Chen et al., 2006] Chen, Y., Zhou, M., and Wang, S. (2006). Reranking answers for definitional qa using language modeling. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1081–1088.

[Chiron et al., 2017] Chiron, G., Doucet, A., Coustaty, M., Visani, M., and Moreux, J.-P. (2017). Impact of OCR errors on the use of digital libraries Towards a better access to information. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Toronto, Canada. IEEE.

[Chiu et al., 2016] Chiu, B., Korhonen, A., and Pyysalo, S. (2016). Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 1–6, Berlin, Germany. Association for Computational Linguistics.

[Clausner et al., 2020] Clausner, C., Antonacopoulos, A., and Pletschacher, S. (2020). Efficient and effective ocr engine training. *International Journal on Document Analysis and Recognition (IJDAR)*, 23(1) :73–88.

[Collier, 2011] Collier, N. (2011). What's unusual in online disease outbreak news ? *Journal of Biomedical Semantics*, 1(2).

[Cothière-Robert, 2007] Cothière-Robert, D. (2007). Stratégies des restitutions des constructions verbales sérielles du créole haïtien en français l2. In *Autour des langues et du langage : perspective pluridisciplinaire*. Presses Universitaires de Grenoble.

[Crabbé and Candito, 2008] Crabbé, B. and Candito, M. (2008). Expériences d'analyse syntaxique statistique du français. In *TALN 2008*, pages pp. 44–54.

[Daille et al., 2016] Daille, B., Jacquey, E., Lejeune, G., Melo, L. F., and Toussaint, Y. (2016). Ambiguity Diagnosis for Terms in Digital Humanities. In *Language Resources and Evaluation Conference*, Portorož, Slovenia.

[Daille et al., 2011] Daille, B., Jacquin, C., Monceaux, L., Morin, E., and Rocheteau, J. (2011). TTC TermSuite : une chaîne de traitement pour la fouille terminologique multilingue. In *18ème Conférence francophone sur le Traitement Automatique des Langues Naturelles Conference (TALN 2011).*, Montpellier, France. Démonstration.

[de Loupy and El-Bèze, 2000] de Loupy, C. and El-Bèze, M. (2000). Using few clues can compensate the small amount of resources available for word sense disambiguation. *behaviour*, 3(1004) :279.

[Deloffre, 1963] Deloffre, F. (1963). Le genre burlesque en france au xviie siècle. étude de style.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[Domingo et al., 2018] Domingo, M., Garca-Martnez, M., Helle, A., Casacuberta, F., and Herranz, M. (2018). How much does tokenization affect neural machine translation ?

[Dong and Li, 1999] Dong, G. and Li, J. (1999). Efficient mining of emerging patterns : Discovering trends and differences. In *ACM SIGKDD*, pages 43–52, New York, NY, USA. ACM.

[Doualan et al., 2012] Doualan, G., Boucher, M., Brixtel, R., Lejeune, G., and Dias, G. (2012). Détection de mots-clés par approches au grain caractère et au grain mot. In *JEP-TALN-RECITAL 2012, Atelier DEFT 2012 : DÉfi Fouille de Textes*, pages 41–48, Grenoble, France. ATALA/AFCP.

[Drouin, 2007] Drouin, P. (2007). Identification automatique du lexique scientifique transdisciplinaire. *Revue française de linguistique appliquée*, 12(2) :45–64.

[Du and Cardie, 2020] Du, X. and Cardie, C. (2020). Event extraction by answering (almost) natural questions. *CoRR*, abs/2004.13625.

[Ducel et al., 2022a] Ducel, F., Fort, K., Lejeune, G., and Lepage, Y. (2022a). Do we name the languages we study ? the #BenderRule in LREC and ACL articles. In *LREC 2022*, Marseille, France. ELRA.

[Ducel et al., 2022b] Ducel, F., Fort, K., Lejeune, G., and Lepage, Y. (2022b). Langues par défaut ? Analyse contrastive et diachronique des langues non citées dans les articles de TALN/ACL. In *RECITAL 2022 - Conférence sur le traitement automatique des langues naturelles (TALN)*, Avignon, France.

[Dupont et al., 2021] Dupont, Y., González-Gallardo, C.-E., Lejeune, G., Millour, A., and Tanguy, J.-B. (2021). QUEER@DEFT2021 : Identification du profil clinique de patients et notation automatique de copies d'étudiants (QUEER@DEFT2021 : Patients clinical profile identification and automatic student grading). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*, pages 95–107, Lille, France. ATALA.

[Eco, 1985] Eco, U. (1985). *Lector in Fabula.* Grasset.

[Ehrmann et al., 2020] Ehrmann, M., Romanello, M., Flückiger, A., and Clematide, S. (2020). Overview of clef hipe 2020 : Named entity recognition and linking on historical newspapers. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 288–310. Springer.

[El Bouanani and Kassou, 2014] El Bouanani, S. E. M. and Kassou, I. (2014). Authorship analysis studies : A survey. *International Journal of Computer Applications*, 86 :22–29.

[El-Khoury, 2007] El-Khoury, T. (2007). Les procédés de métaphorisation dans le discours médical arabe : étude de cas. In *Autour des langues et du langage : perspective pluridisciplinaire.* Presses Universitaires de Grenoble.

[Falk et al., 2014] Falk, I., Bernhard, D., Gérard, C., and Potier-Ferry, R. (2014). Étiquetage morpho-syntaxique pour des mots nouveaux. In *21ème conférence sur le Traitement Automatique des Langues Naturelles*, pages 431–438. Poster.

[Fiscus, 1997] Fiscus, J. (1997). A post-processing system to yield reduced word error rates : Recognizer output voting error reduction (rover). In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 347–354.

[Forsyth and Holmes, 1996] Forsyth, R. S. and Holmes, D. I. (1996). Feature-finding for text classification. *Literary and Linguistic Computing*, 11(4) :163–174.

[Fort et al., 2012] Fort, K., Nazarenko, A., and Rosset, S. (2012). Modeling the complexity of manual annotation tasks : a grid of analysis. In *International Conference on Computational Linguistics*, pages 895–910.

[Gabay et al., 2020] Gabay, S., Clérice, T., and Reul, C. (2020). OCR17 : Ground Truth and Models for 17th c. French Prints (and hopefully more). working paper or preprint.

[Gaiffe et al., 2015] Gaiffe, B., Husson, B., Jacquey, E., and Kister, L. (2015). Smarties : Consultation des fichiers annotés manuellement, domain scientext 2014, available at http ://apps.atilf.fr/smarties/index.php ?r=text/listtext. Technical report.

[Gale et al., 1992] Gale, W. A., Church, K., and Yarowsky, D. (1992). Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*.

[Gautier, 2016] Gautier, H. (2016). Méthodes digitales.. approches quali/quanti des données numériques. *Terminal. Technologie de l'information, culture & société*, 1(118).

[Geyken et al., 2017] Geyken, A., Barbaresi, A., Didakowski, J., Jurish, B., Wiegand, F., and Lemnitzer, L. (2017). Die Korpusplattform des "Digitalen Wörterbuchs der deutschen Sprache" (DWDS). *Zeitschrift für germanistische Linguistik*, 45(2) :327–344.

[Ghasemisharif et al., 2019] Ghasemisharif, M., Snyder, P., Aucinas, A., and Livshits, B. (2019). SpeedReader : Reader Mode Made Fast and Private. In *Proceedings of the World Wide Web Conference*, pages 526–537.

[Ghoul and Lejeune, 2019] Ghoul, D. and Lejeune, G. (2019). MICHAEL : Mining Character-level Patterns for Arabic Dialect Identification (MADAR Challenge). In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 229–233, Florence, France. Association for Computational Linguistics.

[Ghoul and Lejeune, 2020] Ghoul, D. and Lejeune, G. (2020). Comparison between Voting Classifier and Deep Learning methods for Arabic Dialect Identification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020 @COLING2020)*, Barcelona, Spain.

[Ghoul and Lejeune, 2021] Ghoul, D. and Lejeune, G. (2021). Sarcasm and Sentiment Detection in Arabic : investigating the interest of character-level features. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP 2021@EACL2021)*, pages 329–333, Kiev, Ukrain.

[Giannetti et al., 2019] Giannetti, C., Lucini, B., and Vadacchino, D. (2019). Machine learning as a universal tool for quantitative investigations of phase transitions. *Nuclear Physics B*, 944 :114639.

[Giguet, 2011] Giguet, E. (2011). *De l'analyse syntaxique automatique à l'analyse automatique du discours dans les collections multilingues de documents numériques composites*. Mémoire d'habilitation à diriger des recherches, Université de Caen Basse-Normandie.

[Giguet and Lejeune, 2019] Giguet, E. and Lejeune, G. (2019). Daniel@FinTOC-2019 shared task : TOC extraction and title detection. In *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)*, pages 63–68, Turku, Finland. Linköping University Electronic Press.

[Giguet and Lejeune, 2021a] Giguet, E. and Lejeune, G. (2021a). Daniel at the FinSBD-2 task : Extracting list and sentence boundaries from PDF documents, a model-driven approach to PDF document analysis. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 67–74, Kyoto, Japan. -.

[Giguet and Lejeune, 2021b] Giguet, E. and Lejeune, G. (2021b). Daniel at the FinSBD-2 Task : Extracting List and Sentence Boundaries from PDF Documents, a model-driven approach to PDF document analysis. *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 67–74.

[Giguet et al., 2020] Giguet, E., Lejeune, G., and Tanguy, J.-B. (2020). Daniel@fintoc2 shared task : Title detection and structure extraction. In *1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation @COLING2020*, pages 174–180.

[Grieve, 2007] Grieve, J. (2007). Quantitative authorship attribution : An evaluation of techniques. *Literary and linguistic computing*, 22(3) :251–270.

[Grishman et al., 2002] Grishman, R., Huttunen, S., and Yangarber, R. (2002). Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics*, 35(4) :236–246. Sublanguage - Zellig Harris Memorial.

[Gupta et al., 2015] Gupta, A., Gutierrez-Osuna, R., Christy, M., Capitanu, B., Auvil, L., Grumbach, L., Furuta, R., and Mandell, L. (2015). Automatic assessment of ocr quality in historical documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

[Habernal et al., 2016] Habernal, I., Zayed, O., and Gurevych, I. (2016). C4Corpus : Multilingual Web-size corpus with free license. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 914–922.

[Hamborg et al., 2017] Hamborg, F., Meuschke, N., Breitinger, C., and Gipp, B. (2017). news-please : A generic news crawler and extractor. In Gaede, M., Trkulja, V., and Petra, V., editors, *Proceedings of the 15th International Symposium of Information Science*, pages 218–223.

[Hamdi et al., 2020] Hamdi, A., Jean-Caurant, A., Sidère, N., Coustaty, M., and Doucet, A. (2020). Assessing and Minimizing the Impact of OCR Quality on Named Entity Recognition. In *Digital Libraries for Open Knowledge 24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020, Lyon, France, August 25–27, 2020, Proceedings*, pages 87–101.

[Haque and Singh, 2015] Haque, A. and Singh, S. (2015). Anti-scraping application development. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 869–874.

[Heiden et al., 2010] Heiden, S., Magué, J.-P., and Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In *JADT 2010*, pages 1021–1032, Rome, Italy.

[Hessel and Schofield, 2021] Hessel, J. and Schofield, A. (2021). How effective is BERT without word ordering ? implications for language understanding and data privacy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2 : Short Papers)*, pages 204–211, Online. Association for Computational Linguistics.

[Hovy, 2015] Hovy, D. (2015). Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

[Hunston, 2008] Hunston, S. (2008). Collection strategies and design decisions. In *Corpus linguistics : An international handbook*, pages 154–168. de Gruyter.

[Huttunen et al., 2011] Huttunen, S., Vihavainen, A., von Etter, P., and Yangarber, R. (2011). Relevance prediction in information extraction using discourse and lexical features. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 114–121.

[Huttunen et al., 2002] Huttunen, S., Yangarber, R., and Grishman, R. (2002). Diversity of scenarios in information extraction. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

[Huynh et al., 2020] Huynh, V.-N., Hamdi, A., and Doucet, A. (2020). When to Use OCR Post-correction for Named Entity Recognition ? In *22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020*, pages 33–42.

[Hwang et al., 2020] Hwang, W., Yim, J., Park, S., Yang, S., and Seo, M. (2020). Spatial dependency parsing for semi-structured document information extraction. *arXiv preprint arXiv :2005.00642*.

[Ilie and Smyth, 2011] Ilie, L. and Smyth, W. F. (2011). Minimum unique substrings and maximum repeats. *Fundamenta Informaticae*, 110(1) :183–195.

[Jacques, 2003] Jacques, M.-P. (2003). *Approche en discours de la réduction des termes complexes dans les textes spécialisés*. PhD thesis, Toulouse 2.

[Jijkoun et al., 2004] Jijkoun, V., Mur, J., and de Rijke, M. (2004). Information extraction for question answering : Improving recall through syntactic patterns. In *COLING 2004 : Proceedings of the 20th International Conference on Computational Linguistics*, pages 1284–1290, Geneva, Switzerland. COLING.

[Jin et al., 2009] Jin, P., McCarthy, D., Koeling, R., and Carroll, J. A. (2009). Estimating and exploiting the entropy of sense distributions. In *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume : Short Papers*, pages 233–236.

[Jo and Gebru, 2020] Jo, E. S. and Gebru, T. (2020). Lessons from Archives : Strategies for Collecting Sociocultural Data in Machine Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 306–316.

[Jouhaud, 2009] Jouhaud, C. (2009). *Mazarinades. La Fronde des mots*. Aubier, Paris.

[Joulin et al., 2016] Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016). Fasttext.zip : Compressing text classification models. *arXiv preprint arXiv :1612.03651*.

[Kacmarcik and Gamon, 2006] Kacmarcik, G. and Gamon, M. (2006). Obfuscating document stylometry to preserve author anonymity. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 444–451. Association for Computational Linguistics.

[Kagan, 2004] Kagan, R. (2004). *Of Paradise and Power : America and Europe in the New World Order*. Vintage Books.

[Kageura, 2012] Kageura, K. (2012). *The quantitative analysis of the dynamics and structure of terminologies*, volume 15. John Benjamins Publishing.

[Kao et al., 2004] Kao, H.-Y., Lin, S.-H., Ho, J.-M., and Chen, M.-S. (2004). Mining web informative structures and contents based on entropy analysis. *IEEE Transactions on Knowledge and Data Engineering*, 16(1) :41–55.

[Kärkkäinen et al., 2006] Kärkkäinen, J., Sanders, P., and Burkhardt, S. (2006). Linear work suffix array construction. *Journal of the ACM*, 53(6) :918–936.

[Kettunen et al., 2020] Kettunen, K., Koistinen, M., and Kervinen, J. (2020). Ground truth ocr sample data of finnish historical newspapers and journals in data improvement validation of a re-ocring process. *Liber Quarterly*, 30(1).

[Kettunen et al., 2016] Kettunen, K., Mäkelä, E., Ruokolainen, T., Kuokkala, J., and Löfberg, L. (2016). Old content and modern tools - searching named entities in a finnish ocred historical newspaper collection 1771-1910. *CoRR*, abs/1611.02839.

[Kiessling et al., 2019] Kiessling, B., Tissot, R., Stokes, P., and Ezra, D. S. B. (2019). escriptorium : An open source platform for historical document analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19. IEEE.

[Kilgarriff, 2007] Kilgarriff, A. (2007). Googleology is bad science. *Computational Linguistics*, 33(1) :147–151.

[Kohlschütter et al., 2010] Kohlschütter, C., Fankhauser, P., and Nejdl, W. (2010). Boilerplate detection using shallow text features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM 10*, pages 441–450.

[Kohlschütter and Nejdl, 2008] Kohlschütter, C. and Nejdl, W. (2008). A Densitometric Approach to Web Page Segmentation. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 1173–1182.

[Koppel et al., 2009] Koppel, M., Schler, J., and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1) :9–26.

[Koppel et al., 2011] Koppel, M., Schler, J., and Argamon, S. (2011). Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1) :83–94.

[Koudoro-Parfait et al., 2022] Koudoro-Parfait, C., Lejeune, G., and Buth, R. (2022). Reconnaissance d'entités nommées sur des sorties OCR bruitées : des pistes pour la désambiguïsation morphologique automatique. In Estève, Y., Jiménez, T., Parcollet, T., and Zanon Boito, M., editors, *Actes de la 29e Conférence TALN, Atelier Humanités Numériques*, pages 45–55, Avignon, France. ATALA.

[Koudoro-Parfait et al., 2021] Koudoro-Parfait, C., Lejeune, G., and Roe, G. (2021). Spatial named entity recognition in literary texts : What is the influence of ocr noise ? In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities*, pages 13–21.

[Kuhn, 1962] Kuhn, T. S. (1962). *The Structure of Scientific Revolutions (fourth ed.)*. University of Chicago Press.

[Kuznetsov, 2001] Kuznetsov, S. O. (2001). Machine learning on the basis of formal concept analysis. *Autom. Remote Control*, 62(10) :1543–1564.

[Kuznetsov, 2004] Kuznetsov, S. O. (2004). Complexity of learning in concept lattices from positive and negative examples. *Discrete Applied Mathematics*, 142(13) :111 – 125. Boolean and Pseudo-Boolean Functions.

[Labadie, 1904] Labadie, E. (1904). *Nouveaux Suppléments à la bibliographie des Mazarinades*. Henri Leclerc, Paris.

[Labbé and Labbé, 2013] Labbé, C. and Labbé, D. (2013). Lexicométrie : quels outils pour les sciences humaines et sociales ? In *Usages de la lexicométrie en sociologie*.

[Lafon, 1980] Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1 :127–165.

[Lardilleux, 2010] Lardilleux, A. (2010). *Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle*. PhD thesis, Université de Caen.

[Le and Mikolov, 2014] Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents.

[Leclerc, 2013] Leclerc, J. (2013). a faim des parisiens en vers burlesques : sur quelques mazarinades du blocus (hiver 1649). *Les Histoires de Paris (XVIe-XVIIIe siècle)*, 2(1) :199–218.

[Lecluze and Lejeune, 2014] Lecluze, C. and Lejeune, G. (2014). DEFT 2014, analyse automatique de textes littéraires et scientifiques en langue française. In *Actes de DEFT 2014 : 10ème DÉfi Fouille de Textes*, pages 11–19, Marseille, France.

[Legallois, 2016] Legallois, D. (2016). Caractériser le style d'un auteur par des patrons lexico-grammaticaux : une nouvelle approche en stylistique. In *Méthodes stylistiques : Unités et paliers de pertinences*. Presses Univ. de Lyon.

[Legallois et al., 2016] Legallois, D., Charnois, T., and Poibeau, T. (2016). Repérer les clichés dans les romans sentimentaux grâce à la mérthode des ń motifs ż. *Lidil. Revue de linguistique et de didactique des langues*, 1(53) :95–117.

[Lejeune, 2009] Lejeune, G. (2009). Structure patterns in information extraction : a multilingual solution ? In *Advances in Methods of Information and Communication Technology*, pages 105–111.

[Lejeune, 2013] Lejeune, G. (2013). *Veille épidémiologique multilingue : une approche parcimonieuse au grain caractère fondée sur le genre textuel*. PhD thesis, Université de Caen.

[Lejeune and Barbaresi, 2020] Lejeune, G. and Barbaresi, A. (2020). Bien choisir son outil dextraction de contenu à partir du Web. In *Actes de la conférence JEP-TALN-RECITAL 2020, Démonstrations*. ATALA. à paraître.

[Lejeune et al., 2015a] Lejeune, G., Brixtel, R., Doucet, A., and Lucas, N. (2015a). Multilingual event extraction for epidemic detection. *Artifcial Intelligence in Medicine*, 65(2) :131–143.

[Lejeune et al., 2011] Lejeune, G., Brixtel, R., Giguet, E., and Lucas, N. (2011). Deft 2011 : appariements de résumés et d'articles scientfiques fondés sur des distributions de chaînes de caractères. In *Proceedings of DEfi Fouille de Texte (DEFT'11)*, pages 53–64.

[Lejeune et al., 2015b] Lejeune, G., Brixtel, R., and Lecluze, C. (2015b). Évaluation intrinsèque et extrinsèque du nettoyage de pages web. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 95–101.

[Lejeune et al., 2013] Lejeune, G., Brixtel, R., Lecluze, C., Doucet, A., and Lucas, N. (2013). Added-value of automatic multilingual text analysis for epidemic surveillance. In *Artificial Intelligence in Medicine (AIME)*, pages 284–294.

[Lejeune and Cartier, 2017] Lejeune, G. and Cartier, E. (2017). Character based pattern mining for neology detection. In *Proceedings of Subword & Character Level Models in NLP (SCLeM), EMNLP 2017 Copenhagen*, pages 25–30, Copenhagen, Danemark. Association for Computational Linguistics.

[Lejeune and Daille, 2015] Lejeune, G. and Daille, B. (2015). Vers un diagnostic d'ambiguïté des termes candidats d'un texte. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2015)*, pages 446–452.

[Lejeune et al., 2010] Lejeune, G., Doucet, A., and Lucas, N. (2010). Tentative d'approche multilingue en Extraction d'Information. In *JADT 2010*, pages 1259–1268. JADT.

[Lejeune and Dumonceaux, 2015] Lejeune, G. and Dumonceaux, F. (2015). Une approche stylométrique pour la fouille d'opinion. In *Actes de la 11e Défi Fouille de Texte*, pages 12–15, Caen, France. Association pour le Traitement Automatique des Langues.

[Lejeune et al., 2016] Lejeune, G., Rioult, F., and Crémilleux, B. (2016). Highlighting psychological features for predicting child interventions during story telling. In *INTERSPEECH 2016*.

[Lejeune and Zhu, 2018] Lejeune, G. and Zhu, L. (2018). A new proposal for evaluating web page cleaning tools. *Computación y Sistemas*, 22(4).

[Lesk, 1986] Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries : How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, pages 24–26, New York, NY, USA. ACM.

[L'Homme, 2004] L'Homme, M.-C. (2004). *La terminologie : principes et techniques*. Presses de l'Université de Montréal.

[Li et al., 2019] Li, X., Yin, F., Sun, Z., Li, X., Yuan, A., Chai, D., Zhou, M., and Li, J. (2019). Entity-relation extraction as multi-turn question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350, Florence, Italy. Association for Computational Linguistics.

[Linhares Pontes et al., 2019] Linhares Pontes, E., Hamdi, A., Sidère, N., and Doucet, A. (2019). Impact of OCR Quality on Named Entity Linking. In *International Conference on Asia-Pacific Digital Libraries 2019*, Kuala Lumpur, Malaysia.

[Longrée and Mellet, 2013] Longrée, D. and Mellet, S. (2013). Le motif : une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours*. *Langages*, 1(189) :65–79.

[Lopresti, 2009] Lopresti, D. P. (2009). Optical character recognition errors and their effects on natural language processing. *Int. J. Document Anal. Recognit.*, 12(3) :141–151.

[Lorenzen, 2021] Lorenzen, M. (2021). Testing hypotheses with dirty ocr and web-based tools in periodical studies1. *Digital Humanities Research/ Volume*, page 131.

[Lucas, 2009] Lucas, N. (2009). *Modélisation différentielle du texte, de la linguistique aux algorithmes*. PhD thesis, Université de Caen.

[Luong et al., 2013] Luong, T., Socher, R., and Manning, C. (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, Sofia, Bulgaria. Association for Computational Linguistics.

[Lyon et al., 2011] Lyon, A., Nunn, M., Grossel, G., and Burgman, M. (2011). Comparison of Web-Based Biosecurity Intelligence Systems : BioCaster, EpiSPIDER and HealthMap. *Transboundary and Emerging Diseases*.

[LHomme, 2005] LHomme, M.-C. (2005). Sur la notion de ńătermeăż. *Meta*, 50(4) :1112–1132.

[Maeda et al., 2000] Maeda, A., Sadat, F., Yoshikawa, M., and Uemura, S. (2000). Query term disambiguation for web cross-language information retrieval using a search engine. In *Proceedings of the Fifth International Workshop on on Information Retrieval with Asian Languages*, IRAL '00, page 2532, New York, NY, USA. Association for Computing Machinery.

[Manning, 2011] Manning, C. D. (2011). Part-of-Speech Tagging from 97% to 100% : Is It Time for Some Linguistics ? *SpringerLink*, pages 171–189.

[Martin et al., 2020] Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., and Sagot, B. (2020). Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[Maslow, 1966] Maslow, A. (1966). *The Psychology of Science : A Reconnaissance*. New York, Harper and Row.

[Mathet et al., 2008] Mathet, Y., Charnois, T., Doucet, A., and Rioult, F. (2008). Trois approches du greyc pour la classification de textes. *Actes du quatrième DÉfi Fouille de Textes*, page 37.

[Mayaffre, 2005] Mayaffre, D. (2005). Rôle et place du corpus en linguistique. réflexions introductives. In *Actes du colloque JETOU'2005*, pages 5–17. Université de Toulouse-Le Mirail.

[McEnery and Hardie, 2011] McEnery, T. and Hardie, A. (2011). *Corpus linguistics : Method, theory and practice*. Cambridge University Press.

[Melo-Mora and Toussaint, 2015] Melo-Mora, L.-F. and Toussaint, Y. (2015). Automatic validation of terminology by means of formal concept analysis. In *International Conference on Formal Concept Analysis (ICFCA)*.

[Miikkulainen and Dyer, 1991] Miikkulainen, R. and Dyer, M. G. (1991). Natural language processing with modular pdp networks and distributed lexicon. *Cognitive Science*, 15(3) :343–399.

[Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Millour, 2020] Millour, A. (2020). *Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées.* PhD thesis, Sorbonne Université, France.

[Molinié and Viala, 1993] Molinié, G. and Viala, A. (1993). *Stylème*, pages 31–42. Perspectives littéraires. Presses Universitaires de France, Paris cedex 14.

[Montague, 1970] Montague, R. (1970). English as a formal language. *Formal philosophy, Yale university press*, page 188.

[Moreau, 1850] Moreau, C. (1850). *Bibliographie des Mazarinades*. Jules Renouard, Paris.

[Moreau, 1862] Moreau, C. (1862). *Supplément à la Bibliographie des mazarinades*, pages 786–829. Techener, Paris.

[Moreau, 1869] Moreau, C. (1869). *Supplément à la Bibliographie des mazarinades*, pages 61–81. Techener, Paris.

[Mutuvi, 2022] Mutuvi, S. (2022). *Epidemic Event Extraction in Multilingual and Low-resource Settings*. Theses, La Rochelle Université.

[Mutuvi et al., 2020a] Mutuvi, S., Boros, E., Doucet, A., Jatowt, A., Lejeune, G., and Odeo, M. (2020a). Multilingual epidemiological text classification : A comparative study. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6172–6183, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[Mutuvi et al., 2023] Mutuvi, S., Boros, E., Doucet, A., Jatowt, A., Lejeune, G., and Odeo, M. (2023). Analyzing the impact of tokenization on multilingual epidemic surveillance in low-resource languages. In *Document Analysis and Recognition - ICDAR 2023*, pages 17–32, Cham. Springer Nature Switzerland.

[Mutuvi et al., 2021a] Mutuvi, S., Boros, E., Doucet, A., Lejeune, G., Jatowt, A., and Odeo, M. (2021a). Multilingual epidemic event extraction. In *23rd International Conference on Asia-Pacific Digital Libraries ICADL 2021, Online*, volume of , pages 139–156. Springer.

[Mutuvi et al., 2021b] Mutuvi, S., Boros, E., Jatowt, A., Lejeune, G., Odeo, M., and Doucet, A. (2021b). Token-level multilingual epidemic dataset for event extraction. In Berget, G., Hall, M. M., Brenn, D., and Kumpulainen, S., editors, *Linking Theory and Practice of Digital Libraries*, Lecture Notes in Computer Science, pages 55–59. Springer.

[Mutuvi et al., 2020b] Mutuvi, S., Doucet, A., Lejeune, G., and Odeo, M. (2020b). A dataset for multi-lingual epidemiological event extraction. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4139–4144, Marseille, France. European Language Resources Association.

[Navigli, 2009] Navigli, R. (2009). Word sens disambiguation : A survey. *ACM Computing Surveys*, 41(2).

[Navigli and Velardi, 2005] Navigli, R. and Velardi, P. (2005). Structural semantic interconnections : a knowledge-based approach to word sense disambiguation. *IEEE transactions on pattern analysis and machine intelligence*, 27(7) :1075–1086.

[Nédelec, 2004] Nédelec, C. (2004). *Les États et empires du burlesque*. Lumière classique. Honoré Champion.

[Nédelec, 2020] Nédelec, C. (2020). Les muses camuses des burlesques. *Littératures classiques*, 1(2) :81–92.

[Neveu, 2008] Neveu, F. (2008). Pour une description terminographique des sciences du langage. In Publication de l'Université, P. V. E., editor, *Cahiers du CIEL "Langues de spécialité"*, page à paraître. C. Cortès.

[Nguyen, 2023] Nguyen, N. K. (2023). *Emerging Trend Detection in News Articles*. PhD thesis, La Rochelle Universite.

[Nguyen et al., 2020] Nguyen, N. K., Boros, E., Lejeune, G., and Doucet, A. (2020). Impact analysis of document digitization on event extraction. In *Proceedings of the 4th Workshop on Natural Language for Artificial Intelligence (NL4AI), 19th International Conference of the Italian Association for Artificial Intelligence*, pages 17–28, Roma, Italy. -.

[Niu and Carpuat, 2020] Niu, X. and Carpuat, M. (2020). Controlling neural machine translation formality with synthetic supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8568–8575.

[Ogundepo et al., 2022] Ogundepo, O., Zhang, X., and Lin, J. (2022). Better than whitespace : Information retrieval for languages without custom tokenizers. *arXiv preprint arXiv :2210.05481*.

[Patel et al., 2018] Patel, R., Yang, Y., Marshall, I., Nenkova, A., and Wallace, B. C. (2018). Syntactic patterns improve information extraction for medical search. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, page 371. NIH Public Access.

[Peng et al., 2003] Peng, F., Schuurmans, D., Wang, S., and Keselj, V. (2003). Language independent authorship attribution using character level language models. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 267–274. Association for Computational Linguistics.

[Perlis, 1982] Perlis, A. J. (1982). Special feature : Epigrams on programming. *ACM Sigplan Notices*, 17(9) :7–13.

[Peters and Lecocq, 2013] Peters, M. E. and Lecocq, D. (2013). Content extraction using diverse feature sets. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 89–90.

[Petkovic et al., 2022] Petkovic, L., Alrahabi, M., and Roe, G. (2022). Impact de la correction automatique de l'OCR/HTR sur la reconnaissance d'entités nommées dans un corpus bruité. *JIS - Journal of Information Sciences*, 21(2) :42–57.

[Petrescu et al., 2019] Petrescu, R., Manolache, S., Boiangiu, C., Vlasceanu, G., Avatavului, C., Prodan, M., and Bucur, I. (2019). Combining tesseract and asprise results to improve ocr text detection accuracy. *Journal of Information Systems & Operations Management*, 13(1) :57–64.

[Piskorski et al., 2011] Piskorski, J., Belayeva, J., and Atkinson, M. (2011). Exploring the usefulness of cross-lingual information fusion for refining real-time news event extraction : A preliminary study. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 210–217.

[Pomikálek, 2011] Pomikálek, J. (2011). *Removing boilerplate and duplicate content from web corpora*. PhD thesis, Masaryk University.

[Pöttker, 2003] Pöttker, H. (2003). News and its communicative quality : the inverted pyramidwhen and why did it appear ? *Journalism Studies*, 4(4) :501–511.

[Pöttker, 2003] Pöttker, H. (2003). News and its communicative quality : the inverted pyramidwhen and why did it appear ? *Journalism Studies*, 4(4) :501–511.

[Powalski and Stanislawek, 2020] Powalski, R. and Stanislawek, T. (2020). Unicase–rethinking casing in language models. *arXiv preprint arXiv :2010.11936*.

[Qi et al., 2020] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza : A python natural language processing toolkit for many human languages.

[Quine, 1960] Quine, W. V. (1960). *Word and Object*. MIT Press.

[Quiniou et al., 2012] Quiniou, S., Cellier, P., Charnois, T., and Legallois, D. (2012). What about sequential data mining techniques to identify linguistic patterns for stylistics ? In *CICLing 2012*, pages 166–177.

[Quinlan, 1993] Quinlan, J. R. (1993). *C4.5 : programms for machine learning*. Morgan Kaufmann Publishers, San Francisco, CA, USA.

[Qureshi and Memon, 2012] Qureshi, P. A. R. and Memon, N. (2012). Hybrid model of content extraction. *Journal of Computer and System Sciences*, 78(4) :1248–1257.

[Rae et al., 2018] Rae, A. R., Kim, J., Le, D., and Thoma, G. R. (2018). Main Content Detection in HTML Journal Articles. In *Proceedings of the ACM Symposium on Document Engineering 2018*, pages 1–4, New York, NY, USA. ACM.

[Ramshaw and Marcus, 1999] Ramshaw, L. A. and Marcus, M. P. (1999). Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

[Rastier, 2002] Rastier, F. (2002). Enjeux épistémologiques de la linguistique de corpus. In G., W., editor, *Deuxièmes journées de la linguistique de corpus*, pages 31–46, Lorient, France. Presses Universitaires de Rennes.

[Rastier, 2005] Rastier, F. (2005). Mésosémantique et syntaxe. *Texto !*

[Rastier, 2011] Rastier, F. (2011). *La mesure et le grain. Sémantique de corpus*. Collection Lettres numériques. Paris : Champion.

[Ratcliff and Metzener, 1988] Ratcliff, J. W. and Metzener, D. E. (1988). Pattern Matching : The Gestalt Approach. *Dr. Dobb's Journal*, 13(7) :46.

[Ravishankar et al., 2019] Ravishankar, V., Gökırmak, M., Øvrelid, L., and Velldal, E. (2019). Multilingual probing of deep pre-trained contextual encoders. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 37–47, Turku, Finland. Linköping University Electronic Press.

[Riddell, 2022] Riddell, A. B. (2022). Reliable editions from unreliable components : estimating ebooks from print editions using profile hidden markov models. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pages 1–5.

[Rioult, 2004] Rioult, F. (2004). Mining strong emerging patterns in wide SAGE data. In *ECML/PKDD'04 Discovery Challenge, Pisa, Italy*, pages 127–138.

[Rioult, 2005] Rioult, F. (2005). *Extraction de connaissances dans les bases de données comportant des valeurs manquantes ou un grand nombre d'attributs*. Theses, Université de Caen.

[Rioult, 2017] Rioult, F. (2017). *Fouille de données : motifs minimaux, redescription d'espace et analyse du (e-)sport*. Habilitation à diriger des recherches, Université de Caen Normandie.

[Rust et al., 2020] Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., and Gurevych, I. (2020). How good is your tokenizer ? on the monolingual performance of multilingual language models. *arXiv preprint arXiv :2012.15613*.

[Saenger, 2001] Saenger, P. (2001). *Histoire de la lecture dans le monde occidental*, chapter Lire aux derniers siècles du moyen âge. Cavallo, Guglielmo and Chartier, Roger.

[Schäfer et al., 2013] Schäfer, R., Barbaresi, A., and Bildhauer, F. (2013). The Good, the Bad, and the Hazy : Design Decisions in Web Corpus Construction. In *Proceedings of the 8th Web as Corpus Workshop*, pages 7–15.

[Schäfer et al., 2014] Schäfer, R., Barbaresi, A., and Bildhauer, F. (2014). Focused Web Corpus Crawling. In *Proceedings of the 9th Web as Corpus workshop (WAC-9)*, pages 9–15.

[Schäfer, 2016] Schäfer, R. (2016). CommonCOW : Massively Huge Web Corpora from CommonCrawl Dataand a Method to Distribute them Freely under Restrictive EU Copyright Laws. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, pages 4500–4504.

[Slodzian, 2014] Slodzian, M. (2014). Pourquoi réunifier la traductologie. In *Actes des Journées d'Analyse Statistique des Données Textuelles (JADT)*, pages 1–12, Paris, france.

[Small and Rieger, 1982] Small, S. and Rieger, C. (1982). Parsing and comprehending with word experts (a theory and its realization). *Strategies for natural language processing*, pages 89–147.

[Smith et al., 2013] Smith, J., Saint-Amand, H., Plamadă, M., Koehn, P., Callison-Burch, C., and Lopez, A. (2013). Dirt cheap web-scale parallel text from the Common Crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1374–1383.

[Smith, 2007] Smith, R. (2007). An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

[Socard, 1876] Socard, E. (1876). *Supplément à la Bibliographie des Mazarinades*. H.menu, Paris.

[Socher et al., 2012] Socher, R., Huval, B., Manning, C. D., and Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211, Jeju Island, Korea. Association for Computational Linguistics.

[Sollaci and Pereira, 2004] Sollaci, L. B. and Pereira, M. G. (2004). The introduction, methods, results, and discussion (imrad) structure : a fifty-year survey. *Journal of the medical library association*, 92(3) :364.

[Souvay and Pierrel, 2009] Souvay, G. and Pierrel, J.-M. (2009). Lgerm lemmatisation des mots en moyen français. *Revue TAL*, 50(2) :21.

[Sparks and Ganschow, 1993] Sparks, R. and Ganschow, L. (1993). Searching for the cognitive locus of foreign language learning difficulties : Linking first and second language learning. *The modern language journal*, 77(3) :289–302.

[Sperber and Wilson, 1998] Sperber, D. and Wilson, D. (1998). *Relevance Theory : Applications and Implications*, pages 283–293. Carston R., Uchida S., Amsterdam : John Benjamins.

[Spousta et al., 2008] Spousta, M., Marek, M., and Pecina, P. (2008). Victor : the Web-Page Cleaning Tool. In *4th Web as Corpus Workshop (WAC-4)*, pages 12–17.

[Springmann et al., 2016] Springmann, U., Fink, F., and Schulz, K. U. (2016). Automatic quality evaluation and (semi-) automatic improvement of ocr models for historical printings. *arXiv preprint arXiv :1606.05157*.

[Stamatatos, 2006] Stamatatos, E. (2006). Ensemble-based author identification using character n-grams. In *Proceedings of the 3rd International Workshop on Text-based Information Retrieval*, pages 41–46.

[Stamatatos, 2009] Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3) :538–556.

[Stamatatos et al., 2016] Stamatatos, E., Tschnuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., and Potthast, M. (2016). Clustering by authorship within and across documents. In *CLEF 2016*, pages 691–715.

[Suárez et al., 2020] Suárez, P. J. O., Dupont, Y., Lejeune, G., and Tian, T. (2020). Sinner@ clef-hipe2020 : Sinful adaptation of sota models for named entity recognition in french and german. In *CLEF 2020 Working Notes. Working Notes of CLEF 2020-Conference and Labs of the Evaluation Forum*.

[Suárez et al., 2019] Suárez, P. J. O., Sagot, B., and Romary, L. (2019). Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *Challenges in the Management of Large Corpora (CMLC-7) 2019*, pages 9–16.

[Sun et al., 2011] Sun, F., Song, D., and Liao, L. (2011). DOM-based content extraction via text density. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 245–254.

[Sun et al., 2012] Sun, J., Yang, Z., Liu, S., and Wang, P. (2012). Applying stylometric analysis techniques to counter anonymity in cyberspace. *Journal of Networks*, 7(2).

[Tanguy, 2020] Tanguy, J.-B. (2020). Exploiter des modèles de langue pour évaluer des sorties de logiciels d'OCR pour des documents français du XVIIe siècle. In Benzitoun, C., Braud, C., Huber, L., Langlois, D., Ouni, S., Pogodalla, S., and Schneider, S., editors, *Traitement Automatique des Langues Naturelles (TALN, 27e édition), Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL*, pages 205–217, Nancy, France. ATALA.

[Tanguy, 2022] Tanguy, J.-B. (2022). *Océriser pour accéder aux données ? Vers une évaluation non supervisée du bruit dans les données textuelles issues dOCR de documents du XVIIème siècle*. PhD thesis, Sorbonne Université, France.

[Tanguy, 2013] Tanguy, L. (2013). La ruée linguistique vers le Web. *Texto ! Textes et Cultures*, 18(4).

[Tkaczyk et al., 2018] Tkaczyk, D., Collins, A., Sheridan, P., and Beel, J. (2018). Machine learning vs. rules and out-of-the-box vs. retrained : An evaluation of open-source bibliographic reference and citation parsers. In *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries*, pages 99–108.

[Tomkins, 1963] Tomkins, S. (1963). *Computer Simulation of Personality : Frontier of Psychological Theory*. New York, Wiley.

[Traub and Van Ossenbruggen, 2015] Traub, M. and Van Ossenbruggen, J. H. L. (2015). Impact analysis of OCR quality on research tasks in digital archives. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages pp. 252–263.

[Tutin and Grossmann, 2015] Tutin, A. and Grossmann, F. (2015). Scientext : Un corpus et des outils pour étudier le positionnement et le raisonnement dans les écrits scientifiques, available at `http://scientext.msh-alpes.fr/scientext-site/spip.php?article8`.

[Tweedie et al., 1996] Tweedie, F. J., Singh, S., and Holmes, D. I. (1996). Neural network applications in stylometry : The Federalist papers. *Computers and the Humanities*, 30(1) :1–10.

[Ukkonen, 2009] Ukkonen, E. (2009). Maximal and minimal representations of gapped and non-gapped motifs of a string. *Theoretical Computer Science*, 410(43) :4341–4349.

[Umemura and Church, 2009] Umemura, K. and Church, K. (2009). Substring statistics. In *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing 09, page 5371, Berlin, Heidelberg. Springer-Verlag.

[Valette, 2008] Valette, M. (2008). Pour une science des textes instrumentée. *Syntaxe et sémantique*, 9 :9–14.

[Valette, 2016] Valette, M. (2016). Analyse statistique des données textuelles et traitement automatique des langues. une étude comparée. In *JADT 2016*, volume 2, pages 697–706.

[van Strien et al., 2020] van Strien, D., Beelen, K., Ardanuy, M., Hosseini, K., McGillivray, B., and Colavizza, G. (2020). Assessing the Impact of OCR Quality on Downstream NLP Tasks. In *In Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1 : ARTIDIGH*, pages 484 – 496.

[Vogels et al., 2018] Vogels, T., Ganea, O.-E., and Eickhoff, C. (2018). Web2text : Deep structured boilerplate removal. In *European Conference on Information Retrieval*, pages 167–179. Springer.

[Weichselbaumer et al., 2020] Weichselbaumer, N., Seuret, M., Limbach, S., Dong, R., Burghardt, M., and Christlein, V. (2020). New approaches to ocr for early printed books. *DigItalia*, 2 :74–87.

[Weninger et al., 2010] Weninger, T., Hsu, W. H., and Han, J. (2010). CETR : content extraction via tag ratios. In *Proceedings of the 19th international conference on World Wide Web*, pages 971–980.

[Weninger et al., 2016] Weninger, T., Palacios, R., Crescenzi, V., Gottron, T., and Merialdo, P. (2016). Web content extraction : A metaanalysis of its past and thoughts on its future. *SIGKDD Explor. Newsl.*, 17(2) :1723.

[Wilson, 2006] Wilson, D. (2006). The pragmatics of verbal irony : Echo or pretence ? *Lingua*, 116 :1722–1743.

[Witten and Fanck, 2005] Witten, I. H. and Fanck, E. (2005). *Data Mining : Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, San Francisco.

[Wu and He, 2019] Wu, S. and He, Y. (2019). Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 23612364, New York, NY, USA. Association for Computing Machinery.

[Yapomo and Lejeune, 2022] Yapomo, M. and Lejeune, G. (2022). Les innovations lexicales dans le domaine des énergies renouvelables : exploitation du contraste de corpus. *Neologica*, 16(16).

[Yarowsky, 1992] Yarowsky, D. (1992). Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, page 454460, USA. Association for Computational Linguistics.

[Yarowsky, 1995] Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL '95, page 189196, USA. Association for Computational Linguistics.

[Yarowsky and Florian, 2002] Yarowsky, D. and Florian, R. (2002). Evaluating sense disambiguation across diverse parameter spaces. *Natural Language Engineering*, 8(4) :293–310.

[Zeng et al., 2015] Zeng, D., Liu, K., Chen, Y., and Zhao, J. (2015). Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal. Association for Computational Linguistics.

[Zouaq et al., 2012] Zouaq, A., Gasevic, D., and Hatala, M. (2012). Linguistic patterns for information extraction in ontocmaps. In *Proceedings of the 3rd International Conference on Ontology Patterns - Volume 929*, WOP'12, page 6172, Aachen, DEU. CEUR-WS.org.

Résumé

Cette habilitation à diriger les recherches traite de la variation des données textuelles et de son influence sur l'application de méthodes de Traitement Automatique des Langues (TAL). Différents types de variation sont examinés : variation de la langue, variation de la qualité des données, variation de l'homogénéité des corpus et variation du genre textuel. Nous posons, d'une part, la question des observables du TAL. Il s'agit d'interroger la pertinence du paradigme, majoritaire dans le domaine, consistant à envisager les documents avant tout à travers des représentations en mots, très sensibles aux variations de toutes sortes, au détriment par exemple d'approches en chaînes de caractères plus robustes. D'autre part, nous interrogeons les observatoires du TAL en proposant des pistes pour exploiter les genres textuels des documents et tirer des corpus desquels ils sont tirés des propriétés utiles au traitement automatique à rebours d'une approche où les documents sont simplement des séquences de mots et/ou de sous-mots. Nous montrons notamment comment la structure des documents et le genre textuel peuvent être exploités pour concevoir des modèles de TAL.

Mots-clés : Tokenisation, n-grammes de caractères, sous-mots, genre textuel, collecte de corpus, nettoyage de pages Web, reconnaissance optique de caractères, reconnaissance d'entités nommées, données bruitées, variation linguistique

Abstract

This habilitation thesis deals with variation in textual data and its influence on the application of Natural Language Processing (NLP) methods. Different types of variation are examined : language variation, quality variation, homogeneity variation and textual genre variation. On the one hand, we raise the question of NLP observables. This involves questioning the relevance of the paradigm, majority in the field, consisting in considering documents primarily through word-based representations, highly sensitive to variations of all kinds, to the detriment, for example, of more robust character n-gram based representations. On the other hand, we question the observatories of NLP by proposing ways of exploiting the textual genres of documents and deriving useful properties for automatic processing from the corpora from which they are drawn. We show that there is a great interest in considering that documents are more than mere sequences of words and/or subwords.

Keywords : Tokenization, character n-grams, subwords, text genre, corpus collection, Web Scraping, Optical Character Recognition, Named Entity Recogntition, noisy data, linguistic variation