Océriser pour accéder aux données ? Vers une évaluation non supervisée du bruit dans les données textuelles issues d'OCR de documents du XVIIème siècle

Jean-Baptiste Tanguy

SORBONNE UNIVERSITÉ

ÉCOLE DOCTORALE III

Laboratoire de recherche CELLF – UMR 8599

T H È S E

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ SORBONNE UNIVERSITÉ

Discipline : Littérature

Présentée et soutenue par :

Jean-Baptiste TANGUY

le : 16 septembre 2022

Océriser pour accéder aux données ?

Vers une évaluation non supervisée du *bruit* dans les données textuelles issues d'OCR de documents du XVIIème siècle

Sous la direction de :

M. Glenn ROE – Professeur, Université Sorbonne Université, OBTIC
Mme Karine ABIVEN – Maîtresse de Conférence, Université Sorbonne Université, STIH
M. Gaël Lejeune – Maître de Conférence, Université Sorbonne Université, STIH

Membres du jury :

M. Glenn ROE – Professeur, Sorbonne Université, OBTIC – Directeur
Mme Karine ABIVEN – Maîtresse de Conférence, Sorbonne Université, STIH – Co-directrice
M. Gaël Lejeune – Maître de Conférence, Sorbonne Université, STIH – Co-directeur
Mme Ioana GALLERON – Professeure, Université Sorbonne Nouvelle, LATTICE – Rapporteuse
M. Antoine DOUCET – Professeur, Université de La Rochelle, L3I – Rapporteur
M. Olivier KRAIF – Professeur, Université de Grenoble Alpes, LIDILEM – Examinateur

Remerciements

Mes remerciements les plus vifs vont à :
— Karine Abiven et Gaël Lejeune, pour m'avoir supporté tant de temps, avec tous les hauts et bas, en même temps que de m'avoir guidé avec tant de finesse ;
— Glenn Roe, pour avoir éclairé cette thèse ;
— Alice Millour, ma marraine ;
— tous les résidents de la salle D206 ;
— Karën Fort ;
— Simon Gabay ;
— chaque membres du jury : Ioana Galleron, Antoine Doucet et Olivier Kraif ;
— Yann Sordet, Christophe Vellet et Anne Weber de la bibliothèque Mazarine ;
— la région *Île de France*.

Résumé

Cette thèse propose un questionnement sur l'exploitabilité des données textuelles océrisées en contexte non supervisé. Si le travail se concentre sur un « corpus » qu'on appelle les mazarinades, il ne s'y limite pas pour rendre compte plus généralement des phénomènes sur des documents du XVII^{e} siècle. Il s'agit de poser fermement la question : les données textuelles issues d'OCR peuvent-elles être utilisées avec intérêt, à défaut de disposer d'une transcription de référence (vérité de terrain) ? La réponse n'est évidemment pas univoque. D'abord, nous menons une étude sur l'impact des erreurs d'OCR pour certaines tâches de TAL [1] pour montrer le caractère erratique de cet impact (fonction des tâches donc, mais aussi des corpus utilisés). Ensuite, nous montrons que nous pouvons rassembler un faisceau d'indices qui ne nécessite pas de vérité de terrain pour apprendre un modèle de prédiction du taux d'erreurs. L'enjeu est de faire l'économie des transcriptions de référence pour juger de la qualité des modèles que l'on souhaite utiliser. Enfin, nous prenons l'exemple de deux tâches de TAL (la textométrie et la similarité textuelle) pour admettre qu'il existe certaines tâches où les données n'ont pas besoin d'être spécialement corrigées pour offrir des performances satisfaisantes mais que d'autres sont impossibles à résoudre dans cet état. La question de la non supervision des évaluations en TAL est posée en conclusion.

1. Traitement Automatique des Langues (Naturelles).

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Sommaire

Cette thèse s'intéresse aux *données textuelles océrisées* ainsi qu'à leur *exploitabilité*. L'acronyme OCR, d'où l'on dérive en néologisme le qualificatif *océrisé*, signifie *Optical Character Recognition*, ou « Reconnaissance Optique de Caractères » en français. Il s'agit de la technologie mettant en œuvres plusieurs traitements distincts pour reconnaître, dans un *document numérique*, le *texte* qu'il contient éventuellement. En clair, il s'agit de l'outil informatique qui, quand on lui donne en entrée un document PDF, JPEG, TIFF, etc., nous donne en sortie un fichier texte brut (de type TXT) ou éventuellement un document plus structuré (de type XML-ALTO). Cette technologie est avancée pour les documents *nés-numériques* mais son utilisation sur des *numérisations* de documents anciens pose encore beaucoup de questions [Nagy, 2016]. Questions soulevées par les erreurs qu'elle génère. Par exemple, l'encre qui disparaît les siècles passant rend plus difficile la reconnaissance du texte.

Si les motivations de l'utilisation de l'OCR en général sont variées [1], celles de l'utilisation de l'OCR pour des documents anciens sont plus unifiées. En effet, il s'agit le plus souvent de pouvoir faire de la recherche *plein texte*, c'est-à-dire que

1. De la reconnaissance d'adresses postales, sur des courriers pour les adresser plus rapidement (voir les travaux de Yann Le Cun, en particulier [Le Cun and Fogelman-Soulié, 1987]), à la recherche textométrique.

les chercheurs ont pour but de dépasser la nécessaire lecture occulaire de tous les documents qu'ils questionnent pour trouver un passage pertinent pour leur cheminement intellectuel. Concrètement, il s'agit de pouvoir faire un *Control+F* pour trouver un mot ou encore de sélectionner un passage pour le copier sans nécessairement le *re*-copier.

Tel qu'il a été conçu lors de sa candidature au financement *Paris Region PhD2 2019*[2], ce projet de thèse se concentrait sur deux axes principaux : i) l'amélioration automatique de ces données textuelles océrisées et ii) leur exploitation avec comme postulat qu'un texte illisible par l'œil humain est exploitable par la machine. En substance, il s'agissait de chercher d'un côté à améliorer les données textuelles océrisées de documents anciens, et de les partager aux chercheurs et au grand public, en même temps que de viser une utilisation performante de ces données par la machine (non conforme à la lecture « humaine » car contenant des erreurs[3]). Utilisation *performante* car des données contenant des erreurs peuvent être plus rentables algorithmiquement que leur version corrigée. Par exemple, un même caractère d'impression conduisant à la même erreur peut mettre sur la piste d'un imprimeur en particulier.

Les enjeux d'une telle affirmation sont clairs. D'abord, serait faite l'économie d'une correction, manuelle ou automatisée, pour l'utilisation de ces données par des algorithmes. Ensuite, cela amène à penser que les erreurs sont porteuses d'un sens, peut-être caché qu'il faudrait découvrir. Enfin, qu'un modèle d'OCR n'affichant pas des performances acceptables pour la lecture occulaire serait en mesure de satisfaire une machine. Mais cette dernière hypothèse reviendrait à remettre en question totalement l'intérêt de toute étude cherchant à améliorer les performances des modèles d'OCR. On pourrait aussi croire en disant ceci que plus il y a d'erreurs de reconnaissance, plus la machine sera en mesure de faire des découvertes ; ce qui est un non-sens. Ou encore, que la correction des sorties d'OCR est néfaste au bon fonctionnement des algorithmes de TAL.

Telles hypothèses sont peut-être trop anticipées, ou hâtives. D'abord, les hypothèses révélées dans le paragraphe précédent dérivées du second axe de recherche sont en contradiction évidente avec le premier axe. Pourquoi chercher à améliorer la reconnaissance quand les erreurs peuvent être tolérables (voire bénéfiques ?) pour la machine ? Ensuite, s'il semble vrai que les erreurs faites par l'OCR ne sont pas de la même nature que celles faites par les humains[4], peut-être ne sont-elles pas *bénéfiques* aux algorithmes, tout au plus *transparentes* à ceux-ci. Enfin, car c'est bien de ce type de document dont il s'agit ici, le caractère ancien de certains imprimés rend plus difficile la lecture, parfois même impossible. Il va donc de soi que l'OCR introduise des erreurs. Quel sens donner à celles-ci, à part que la lecture occulaire est hardue ?

Si les deux axes d'étude nous semblent pertinents, les questions qui en découlent méritent d'être précisées. Aussi peut-on se poser les questions suivantes :

2. Partenariat avec la Bibliothèque Mazarine – programme Phd2 2019, DIM STCN (sciences des textes et connaissances nouvelles).

3. C'est dans ce cas que nous parlerons de données *bruitées*.

4. Les erreurs d'OCR étant considérées itératives donc repérables, alors que les erreurs humaines de transcription ne le sont pas (voir sous-section 2.2.3).

1. Comment fonctionne l'OCR et peut-on en déduire plusieurs types d'erreurs ?
2. Comment mesurer la qualité d'une sortie d'OCR, c'est-à-dire la quantité d'erreurs ?
3. Que peut-on faire avec les modèles et les données textuelles océrisées qui existent déjà, lesquels contiennent possiblement beaucoup d'erreurs ?

Ces trois questions sont celles qui mèneront notre argumentaire, qui est le suivant. Il nous apparaît essentiel de poser la question de la *mesure* de la qualité de l'OCR pour pouvoir juger, en conséquence, les modèles existants ainsi que leur pertinence pour des documents historiques. Ensuite, si l'on peut s'engager sur la voie de l'amélioration des modèles existants, il apparaît aussi pertinent de chercher à comprendre ce qu'il est possible de réaliser avec cet existant. Car, apprendre de nouveaux modèles d'OCR est très couteux : constitution d'une vérité de terrain (données transcrites manuellement) et temps d'apprentissage, entre autres. Pour cela, nous travaillerons avec un ensemble d'imprimés connus sous le nom de « mazarinades ». Il s'agit de documents imprimés au XVII^{e} siècle, parfois reliés au XIX^{e}, extrêmement politisés dans un contexte de guerre civile en France : la Fronde (1648-1653).

1.1 Les mazarinades

Courts imprimés, en moyenne six feuillets [5], parus en France pendant la Fronde (1648-1653), les « mazarinades » correspondent à un phénomène radicalement nouveau dans l'histoire de l'imprimé [6]. Il s'agit de sa politisation radicale, nouveau à cette échelle [Jouhaud, 2009] [7]. Sans qu'il soit question de colporter une opinion publique fantasmée, les mazarinades sont de véritables *armes* politiques. En tant que telle, elles prennent des formes variées, si bien que le terme flou de « mazarinades » n'est qu'une commodité. En effet, dans leur forme comme dans leur fond, ces imprimés n'ont pas véritablement d'unité qui les érigerait en corpus cohérent.

Le bibliographe Moreau en dresse une liste d'environ 5 000 entrées [Moreau, 1851]. Néanmoins, les événements historiques que constituent la Fronde, au moins, en circonscrivent la temporalité. Il s'agit, pour une grande partie, de document officiels, déclarations ou actes royaux, ou émanant des grandes cours souveraines [8], mais aussi de pamphlets contre le Cardinal Mazarin (et ce sont les plus connus, d'où même le nom de « mazarinade » qu'on attribue à la plume de Scarron), de lettres plus ou moins fictionnelles, de chansons, etc. Il s'agit donc d'une masse profuse.

D'un point de vue du contenu, ces imprimés sont écrits dans un état du français qu'on appelle souvent « pré-classique » (en référence au français dit

5. La plupart du temps, ils sont constitués de 2 ou 4 feuillets, ce qui représente 4 ou 8 pages.
6. Au moins en France, car il y a eu des phénomènes encore plus virulents en Angleterre.
7. L'imprimé avait déjà été politisé, dans une moindre mesure, au siècle précédent durant la Ligue (année 1580).
8. C'est-à-dire englobant le Parlement et les autres instances décisionnelles du royaume.

« classique » qui se constitue après les années 1660, avec l'arrivée du premier dictionnaire de l'Académie à la fin du siècle). Pendant la Fronde donc, le français est en pleine mutation. Il s'agit néanmoins d'un état de langue proche du français contemporain dans la mesure où, après enseignement de quelques règles limitées, tout lecteur du français contemporain peut faire l'expérience de la lecture de cet état de langue. Au plan orthotypographique, quelques caractéristiques fondamentales :

— assimilation du *i* et du *j* [9] ;
— assimilation du *v* et du *u* [10] ;
— présence du *s* long (*ſ*) [11].

1.2 Questionner le rendement de la recherche assistée par ordinateur

En tant qu'objet inédit dans l'histoire française de l'imprimé, les mazarinades intéressent la recherche en littérature et en histoire. Toutefois, une étude systématique de ces imprimés n'apparaît pas – de prime abord – envisageable, tant on a dit qu'aucune étude sérielle n'était possible [Jouhaud, 2009, p.39] et qu'il faut les étudier par petits « faisceaux » cohérents (par exemple, tel fait divers, ou tel événement en particulier : les barricades d'août 1648, le blocus de Paris de l'hiver 1649, l'emprisonnement des Princes en 1650, etc.). Néanmoins, la recherche outillée sur ces textes permettrait leur confrontation d'une part mais aussi leur pleine préhension en tant qu'ensemble d'autre part, en faisant l'économie d'une lecture « de près » de chaque texte de l'ensemble. C'est dans ce sens que ces imprimés intéressent aussi les Humanités Numériques [12] (HN), en battant en brèche l'hypothèse précitée de l'impossibilité de leur étude par la machine. Si [Jouhaud, 2009, p.39] propose de justes résistances à l'approche computationnelle, il s'agit bien ici de poser le défi de leur exploitabilité outillée.

Aussi cet ensemble de textes est-il fécond pour la recherche en HN : la pertinence du *distant reading*, c'est-à-dire de la lecture assistée par ordinateur, pourra être éprouvée en même temps qu'elle sera mise à l'épreuve. L'enjeu est donc plus général que la seule étude outillée des mazarinades ; il s'agit bien de questionner la numérisation des documents anciens et leur océrisation, mais aussi le *rendement* d'une telle étude par le numérique. Et les mazarinades sont ici exemplaires de cette problématique générale. L'absence de rareté physique implique une disponibilité numérique *a priori*. Le resserremenet dans le temps présente l'avantage d'une neutralisation du problème de l'évolution diachronique (de l'état de langue par exemple – voir [Ayres-Bennett, 2004, p. 83]). Enfin, il existe des bibliographies classiques disponibles permettant de disposer de bonnes métadonnées pour un ensemble massif.

9. Par exemple, *iufte* pour le *juste* contemporain.
10. Par exemple, *morueux* pour le *morveux* contemporain.
11. Par exemple, *ceſſe* pour le *cesse* contemporain.
12. Nous appelons *Humanités Numériques* la transdiscipline regroupant les Sciences Humaines et Sociales qui utilise l'informatique comme moyen de recherche [Mounier, 2010].

1.3 Numériser, océriser et patrimonialiser : une textualité changée

Les campagnes de numérisation des fonds documentaires des bibliothèques s'inscrivent au carrefour de deux enjeux majeurs. D'une part, la pérennisation des documents, en leur garantissant un avenir (numérique donc) ; d'autre part, leur accessibilité, en offrant à leurs publics un accès à ces numérisations. La Bibliothèque Nationale de France (BNF), qui a commencé la numérisation de ses fonds au début des années 1990 (avec l'arrivée de Gallica en 1997 [Bermes, 2020]), et la Bibliothèque Mazarine, qui a engagé en 2014 la numérisation de sa collection d'incunables et en 2015 celle de ses mazarinades, sont ici exemplaires.

La *numérisation* apparaît donc comme une solution technique à la *pérennisation* : numériser les fonds documentaires pour en garantir non seulement un accès facilité par indexation mais surtout pour constituer un moyen d'échapper aux ravages matériels du temps (dégradation de l'objet physique *livre*). En plus de l'archivage physique des œuvres telles qu'on les connaît depuis les *codex* jusqu'à l'arrivée du numérique, l'archivage revêt désormais une dimension nouvelle : celle de l'organisation digitale de ces mêmes œuvres.

La *patrimonialisation* prétend dès lors constituer une « mémoire numérique » [Habert, 2012, section 4.4], non pas seulement comme la *réplication* d'une véritable mémoire physique qui lui pré-existerait, mais bien comme mémoire désormais, et consubstantiellement, physique et numérique. Cela signifie qu'avec l'archivage numérique est né un ensemble d'attributs et de propriétés nouvelles, non présentes déjà dans les bibliothèques, qu'il convient de repérer pour (re)définir la notion de *textualité*. [Mayaffre, 2006, section 1.1] aborde d'ailleurs cette nouvelle conception de la textualité, ou plutôt la remise en cause de son caractère « intangible », dans sa première note à propos de la saisie des textes :

> Que saisit-on exactement ? Le corps du texte seulement ? La couverture et les en-têtes ? Et quelle édition choisir ? Quel format de restitution demander au logiciel de reconnaissance de caractères ? Même lorsqu'ils sont tirés de documents papier, les documents électroniques ne peuvent être la reproduction exacte d'originaux, à moins de seulement photographier les textes. Mais précisément, nous aurions alors affaire à des images et non plus à des textes. Les derniers développements du format PDF sont intéressants à ce sujet. Pendant longtemps le PDF était la reproduction fidèle et intangible du format papier. Seulement, la manipulation de ces fichiers images a très vite paru rigide pour l'utilisateur. Aussi, il est désormais possible de transformer avec *PDF Converter* l'image en texte,... et le caractère intangible du contenu se trouve remis en cause.

La numérisation apparaît donc comme une patrimonialisation par le numérique.

En définissant la *patrimonialisation* comme « processus de valorisation culturelle et identitaire d'un objet donné », [Noille, 2020, section 1] observe trois conséquences, « car la patrimonialisation est loin d'être une opération neutre » :

— les textes engagés dans un tel processus sont d'emblée jugés comme

« exemplification » (témoignage d'une culture) ou comme « exemplarité » (référence pour une culture) ;
— la « pensée de la perte » est présente dans notre rapport au texte ;
— il s'agit d'une « opération d'extraction et de réification de la textualité, en l'engageant en dehors de son territoire d'usages et de pratiques, dans des protocoles de pérennisation et d'accessibilité. »

Plus fort encore, la numérisation des textes que leur patrimonialisation engendre est vue comme « le protocole par lequel la textualité est transformée en cellules d'information » [Noille, 2020, section 5]. Cela signifie que lorsque nous lisons un énoncé long mis en ligne en mode texte, nous n'avons plus à faire à un *textus*, « un *tissu* tramé de mots et de sens », mais bien à une suite d'information mises en scène, à « une *data* ».

La vision enthousiasmée des nouvelles textualités décrites par [Mayaffre, 2006] est ici contrecarrée par [Noille, 2020] qui, pour reprendre son titre, cherche à « sortir les textes du musée computationnel » car si « [l]'imprimé me donne un texte à lire, le numérique me donne une base de données à consulter ». Ici, le *distant reading* n'est même pas une *machine reading* mais une *not reading*, une non-lecture, car « elle exclut tous les usages de la lecture ».

Et il est vrai que le résultat d'une OCR est parfois si incompréhensible qu'il est plus proche d'une base de données que d'un texte imprimé. Il est peut-être néanmoins trop attendu de parler d'une « *not reading* » quand la lecture est impossible, alors que nous pourrions parler plutôt du *résultat* de la « *machine reading* ». Aussi peut-être est-il pertinent de proposer au *distant reading* cette *machine reading*.

1.4 Le paradoxe de l'océrisation

Bien plus que de constituer des éditions web donc [13], la numérisation des fonds documentaires permet leur exploitation automatique à grande échelle. L'acquisition des données textuelles permet d'indexer les documents numérisés, ce qui constitue un réel intérêt, tant pour la communauté scientifique que pour le grand public. Mais cette acquisition des données textuelles ne peut être faite manuellement pour des campagnes de numérisation massive. À titre d'exemple, Gallica franchissait en février 2020 le « cap symbolique des 6 millions de documents numérisés » [14] (et proposait même le mot-dièse *#Gallica6Millions*). Dans ce contexte, les logiciels de reconnaissance optique de caractère sont de bons alliés car ils permettent de rendre automatique cette acquisition des données textuelles.

Néanmoins un paradoxe se pose d'emblée lorsqu'on pense cette acquisition automatique des données textuelles, cette transcription automatique du texte contenu dans des numérisations. D'un côté, l'accès à ce contenu textuel est facilité car est faite l'économie de la transcription manuelle ; de l'autre, cet accès

13. Bien *moins* aussi, car une édition numérique suppose quantité d'enrichissements scientifiques que ne permet pas la pure numérisation de masse.
14. https://tinyurl.com/yyr2j4f4 consulté le 28 juin 2022.

est d'autant plus limité qu'on ne *connait pas* les données textuelles ainsi acquises au sens où l'on ne peut statuer sur la fiabilité de celles-ci. La transcription automatique par OCR est donc à la fois levier et frein à l'acquisition des données textuelles.

Les considérations de Benoît Habert dans [Habert, 2012] inciteraient à penser la nature de ces données textuelles océrisées et ce qu'on veut en garder, ou en transmettre aux communautés futures. Entre « répétition » et « momification » d'un côté [15] et « remémoration » et « (re)création » de l'autre [16], la distinction entre la *sauvegarde* et l'*archive* s'impose.

> Sauvegarder n'est pas archiver, pas plus que répliquer (copier sur un autre site) : c'est effectuer une « copie » (éventuellement multiple et sur des sites différents) de la donnée telle quelle, sans vérifier sa bonne conformité, sans lui associer obligatoirement des métadonnées (autres que celles du nom du fichier et de son type éventuel) et sans la « monitorer », c'est-à-dire maintenir les conditions de sa restitution la plus fidèle possible. La sauvegarde relève plutôt de la répétition, tandis que l'archivage ressortit plutôt à la remémoration, par les choix qu'il suppose, par la sollicitude continuée qu'il implique. [Habert, 2012, section 3.3]

L'océrisation de collections numériques ne donne ni accès à l'*information* contenue dans les œuvres de ces collections pas plus qu'elle ne garantit la fidèle copie de ces *données*.

Certains travaux assurent cependant que l'OCR est un vecteur privilégié de l'accessibilité pérenne des documents. Citons par exemple [Hamdi et al., 2019, introduction] :

> Avec les bibliothèques numériques, une grande quantité de documents imprimés sont scannés et archivés comme images. L'extraction de texte utilisant la reconnaissance optique de caractères [...] est dès lors nécessaire pour l'indexation des documents ce qui est essentiel à l'accessibilité de ceux-ci. [17]

Mais c'est sans prendre en compte la précaution fondamentale pointée par [Habert, 2012, section 4] : « Le monde numérique qui est le nôtre entraîne des modifications importantes des opérations intellectuelles et sensibles de base : lire, écrire, mais aussi se souvenir. » En l'occurrence, en proposant des sorties d'OCR comme *texte*, « lire » et « se souvenir » (opposé par Benoît Habert à « avoir des souvenirs ») sont modifiés, sinon altérés, et l'on ne doit plus parler d'« accès aux données » mais bien d'« accès *médié* aux données ». De manière particulière pour la lecture, [Mayaffre, 2006, section 1.1] rappelle qu'avec l'arrivée de la « philologie numérique »,

15. L'auteur donne ici l'exemple de *Total Recall*, où Gordon Bell propose de « numériser sa vie entière » (*lifelogging*) ce qui peut aboutir à une « incapacité à catégoriser ».

16. L'exemple d'*Un Spécialiste* est ici donné car le film reconstruction éclairée d'un événement historique à partir d'images miticuleusement choisies.

17. *« In digital libraries, large quantities of printed documents are scanned and archived as images. Text extraction using Optical Character Recognition [...] is therefore necessary for indexing documents which is an essential feature for the accessibility to these documents. »*

> [o]nt été relevées [...] trois lectures électroniques complémentaires à la lecture oculaire linéaire traditionnelle : lecture quantitative (complémentaire de la lecture qualitative), lecture paradigmatique (complémentaire de la lecture syntagmatique), lecture hypertextuelle (complémentaire de la lecture textuelle).

Bien que ces différents niveaux de lecture offrent de nouveaux horizons très féconds (et c'est bien le point de vue de [Mayaffre, 2006]), l'indexation de données textuelles océrisées, leur exploitation par des logiciels de TAL, mais aussi simplement leur lecture, sont autant d'exercices qui restent modifiés, sinon altérés, par l'OCR.

1.5 Des mazarinades à l'étude des données bruitées

Le processus d'océrisation, en tant que processus automatique, n'est pas complètement et *a priori* fiable. Cela signifie que le passage de l'image au texte se fait très probablement par l'introduction d'erreurs (gardons volontairement ce substantif flou). Cet *autre* texte, qui certes n'est pas un texte *différent*, est autre par son support (papier *vs.* image *vs.* format texte) mais aussi par son contenu. En clair, certains mots ou caractères peuvent être confondus, ajoutés ou omis. Et c'est de cette confusion, ajout ou omission dont il est ici précisément question.

La collection des mazarinades, paradoxalement profuse et circonscrite (voir section 1.1), constitue un laboratoire privilégié pour l'étude de ce processus d'océrisation. En tant que documents historiques, ces imprimés sont un enjeu pour l'océrisation. [Abiven and Lejeune, 2019] exposent un ensemble d'éléments rendant l'étude de ces documents historiques particulièrement complexe : variantes graphiques, abréviations, orthographe erratique mais aussi un « état inégal de conservation [des] imprimés souvent produits dans l'urgence et l'économie de moyens (papier et encre de mauvaise qualité, notamment) ». Ces données océrisées, qu'on caractérise communément de *bruitées*[18], contiendront probablement beaucoup d'erreurs, ou en tout cas relativement plus qu'une collection de textes plus modernes, imprimés dans des formes plus usuellement identifiées par les logiciels de reconnaissance automatique de caractères.

L'introduction d'erreurs, ou de *bruit*, dans les données océrisées perturbe autant de tâches allant de la simple lecture du texte aux chaînes de traitements du TAL comme la tokenisation, la classification thématique de textes ou encore l'attribution d'auteur. Introduire du bruit dans un texte, c'est donc impacter toutes les tâches en aval ; étudier le bruit de ces textes, c'est donc étudier son impact sur des tâches en aval ; connaître l'impact des erreurs d'OCR c'est donc un moyen de connaître le bruit dans les données océrisées. D'où une circularité entre mesure des erreurs d'OCR et étude de leur impact.

18. Particulièrement en référence au domaine du traitement du signal, le bruit étant ce qui vient pertuber la compréhension d'un signal.

1.6 Pour une (ré)évaluation positive des données bruitées

Le passage du format papier ou image au format texte, s'il n'est pas automatique par OCR, est manuel. Or cette transcription, qui, si elle est bien faite, doit être réalisée par plusieurs annotateurs permettant de calculer un accord entre ceux-ci, est très coûteuse en temps, en ressources humaines et cognitives et en moyens financiers. Il y a donc un intérêt évident à procéder automatiquement par OCR à cette transcription. Toutefois, et c'est le cas dans bien des campagnes de numérisation et d'océrisation, une partie limitée des œuvres océrisées est tout de même transcrite pour permettre le calcul d'un taux d'erreurs, lequel permet de juger de la qualité de l'OCR. Cette évaluation *supervisée* (par comparaison à la transcription de référence) est bien souvent considérée comme nécessaire pour, on l'a dit, juger de la qualité de l'OCR, mais aussi et par voie de conséquence pour comparer, par exemple, les performances relatives de différents logiciels et/ou modèles d'OCR.

Néanmoins, les données océrisées peuvent aussi et en elles-mêmes donner des indications sur la qualité de ces logiciels et/ou modèles. Juger de la qualité de ces données sans rendre nécessaire la transcription d'une partie de la collection à océriser constitue une évaluation *non supervisée*. Telle évaluation est un enjeu de taille dans les campagnes de numérisation car ne travailler qu'avec les données océrisées et bruitées permet l'économie de la constitution couteuse et fastidieuse d'une vérité de terrain. Si l'on peut montrer que de telles données bruitées peuvent suffir à l'évaluation de leur qualité, un autre enjeu est de montrer qu'elles peuvent aussi servir à leur étude. En clair, qu'il n'y ait pas besoin de post-correction [19] des erreurs d'OCR impliquerait une économie des moyens dans les processus d'ingénerie du TAL – puisque le résultat direct de l'OCR servirait à l'étude de ces textes (d'un genre nouveau car possiblement difficiles à lire).

En somme, il s'agit de considérer les données bruitées issues d'OCR comme *exploitables* sans aucun traitement, que ce soit dans l'évaluation non supervisée de leur qualité comme dans leur exploitation.

1.7 De l'usage des données bruitées

La réévaluation positive des données bruitées implique une étude de l'impact du bruit dans la réalisation des tâches considérées (dans notre cas, la textométrie et la similarité textuelle). Toutefois, montrer qu'il peut être possible, dans certaines conditions, d'exploiter des données textuelles océrisées ne signifie pas qu'elles le seront de fait. La question du *bruit* est en effet intimement liée à celle du *bruit ressenti*. Un texte océrisé, disons du XVII^{e} siècle, peut comporter beaucoup d'erreurs rendant parfois difficile l'exercice de lecture. Que dire alors de l'usage de ce texte par la communauté des chercheurs, même s'il a été montré par ailleurs que ce texte océrisé est peut-être exploitable par la machine ? Il est

19. Qu'elle soit manuelle ou automatisée.

bien évident qu'un texte difficilement lisible sera plus facilement écarté d'une étude sur corpus, même si, on le montrera, l'étude de certains phénomènes est largement possible. L'enjeu n'est pas seulement de montrer que ces données bruitées sont fécondes, mais bien en plus de rompre avec la défiance qu'elles inspirent.

Ceci nécessite donc de procéder à une interrogation de la communauté quant à son *rapport* au bruit. Les données textuelles océrisées sont-elles utilisées par les chercheurs, autres que celles et ceux qui les étudient pour ce qu'elles sont ? Si oui, à quel moment de la recherche cet usage intervient-il ? Des résultats issus de recherches sur données bruités sont-ils publiés ? En gros, l'OCR a-t-il un usage autre que l'indexation de documents numérisés dans des bibliothèques numériques ? Car tout ce qui fait baisser la qualité de l'OCR « réduit la précision de la reconnaissance ce qui affecte le traitement des documents et, surtout, l'usage des bibliothèques numériques » [20] [Mutuvi et al., 2018, introduction].

Après avoir étudié le *ressenti* de la communauté scientifique face à la notion de *bruit* (Préambule), d'un point de vue quantitatif avec un questionnaire mis en ligne et qualitatif avec une série d'entretiens, l'étude portera sur le lien circulaire entre mesure des erreurs d'OCR et étude de leur impact (Partie I). C'est après avoir rendu compte du processus d'océrisation pour des documents historiques (Chapitre 3) que l'étude du bruit (Chapitre 4) sera permise. Cette recherche sur le bruit dans les données textuelles océrisées issues de documents historiques permettra de réaliser une évaluation intrinsèque mais non supervisée des erreurs d'OCR (Partie II). Un modèle d'évaluation non supervisé sera conçu (Chapitre 5) puis testé sur un sous-corpus des mazarinades (Chapitre 6). Enfin, les données textuelles ainsi acquises et mesurées seront étudiées extrinsèquement à travers deux tâches usuelles en TAL (Partie III) : la textométrie (Chapitre 7) et la similarité textuelle (Chapitre 8). Radicalement différentes dans leur conception, ces deux tâches ont été choisies dans un but probatoire de la problématique initiale.

Ce chemin permettra de conclure sur deux affirmations majeures. D'une part, il est non seulement possible mais plus encore souhaitable de mesurer la qualité des données textuelles acquises par OCR en utilisant un modèle prédictif du taux d'erreurs car celui-ci permet très bien de définir quel modèle offre le moins d'erreurs. D'autre part, telles données textuelles sont exploitables sans post-correction et le *distant reading* ou plutôt le *mediated distant reading* [21] (*mediated* car il y a l'introduction d'erreurs) est réalisable en permettant la construction de connaissances nouvelles [22]. Cette thèse cherche à montrer que de telles données bruitées, parfois difficiles à appréhender par simple lecture occulaire, sont non seulement largement plus rapides à acquérir mais en plus qu'elles sont riches et

20. *« [...] reduce the accuracy of recognition which affects the processing of the documents and, overall, the use of digital libraries. »*

21. Cette expression est une tournure que nous apprécions mais qui n'est pas attestée par une référence.

22. Un enjeu est de savoir de quel type de connaissances il s'agit (étude possible des phénomènes fréquents en textométrie ?) et à quelle échelle (le caractère, le mot, la phrase, l'œuvre ou même l'ensemble des textes ?).

que leur exploitation ne doit pas être évitée. Nous nuancerons ce propos par la question du caractère supervisé des tâches considérées...

Chapitre 2

Enquêter auprès de la communauté : quel rapport au « bruit » ?

[Traub et al., 2015] mènent des entretiens, auprès de quatre chercheurs en histoire, pour savoir quels sont leurs usages et leur perception des données contenant des erreurs dans leurs recherches. Il s'agit de comprendre i) à quel moment et pourquoi de telles données « bruitées » peuvent être utilisées et ii) dans quelle mesure ces données sont considérées comme « bruitées ». Ces entretiens permettent aux auteurs de conclure que ce genre de données, océrisées donc, sont perçues plus bruitées qu'elles ne le sont, où en tout cas leur caractère bruité est perçu comme freinant leur usage dans la recherche. Et pour cause, un des quatre chercheurs interrogés a complètement arrêté de les utiliser quand les autres ne travaillent avec uniquement dans la phase inaugurale et prospective d'un projet de recherche, pour construire des idées qui seront, ensuite, étayées par la lecture de véritables attestations. En outre, il y a une défiance à publier les données contenant des erreurs de transcription.

Il apparaît donc nécessaire d'interroger plus largement la communauté sur cette technologie ayant connu ses premiers fantasmes il y a plus d'un siècle (voir 3.1.1) : l'OCR. L'utilisent-ils, et à quel moment de la recherche ? La corrigent-ils ? La mesurent-ils ? La définissent-ils ? Pour répondre à ces questions sur les usages et perceptions de l'OCR par la communauté des HN, nous proposons deux types d'études : un questionnaire en ligne (section 2.1) et une série d'entretiens (section 2.2).

Ce travail a été réalisé en étroite collaboration avec Caroline Parfait. Une lecture de ses travaux (notamment [Koudoro-Parfait et al., 2021]) donnera au lecteur un angle nouveau à propos de la reconnaissance d'entités nommées spatiales en contexte océrisé. Elle aussi travail dans le laboratoire OBTIC de Sorbonne Université ; sa thèse est aussi dirigée par Glenn Roe.

2.1 Questionnaire en ligne

Dans le but d'interroger un large éventail de ce que la communauté des HN recouvre, nous avons choisi de concevoir un questionnaire en ligne s'adressant à un large public – des plus experts aux plus novices. Il s'agit là d'entrevoir ce qui unit et différencie les usages et les perceptions de l'OCR.

2.1.1 Conception

Afin de constituer le questionnaire, plusieurs thématiques ont été identifiées, en plus des questions orientées sur le profil du répondant : Corpus, Transcription, OCR et Logiciel. De là, un ensemble de questions ont été formulées.

1. **Profil du répondant** :
   (a) Dans quel(s) domaine(s) travaillez-vous ?
   (b) Quel est le nom de votre poste ? Si autre, précisez.
   (c) Quel est votre ressenti sur votre capacité à naviguer dans le monde numérique ?
   (d) Dans votre travail, utilisez-vous des outils numériques ?
   (e) À quelle fréquence utilisez-vous ces outils numériques ?
   (f) Pouvez-vous lister les outils numériques que vous utilisez le plus ?
   (g) Dans quel but utilisez-vous des outils d'OCR ?
2. **Corpus** :
   (a) Utilisez-vous des corpus de textes dans votre travail ?
   (b) Si vous utilisez des corpus de textes, sont-ils numériques ?
   (c) Utilisez-vous des outils d'OCR pour construire ces corpus ?
   (d) Quel type de document traitez-vous ? (Imprimés, manuscrits, etc.)
3. **OCR** :
   (a) Êtes-vous familier avec le concept d'OCR et son usage ?
   (b) Utilisez-vous des outils d'OCR dans votre travail ?
   (c) Avez-vous déjà utilisé un logiciel d'OCR ? Lequel ?
   (d) Quelle(s) bonne(s) surprise(s) avez-vous eu en utilisant les logiciels d'OCR ?
   (e) Quelle(s) déception(s) avez-vous eu en utilisant les logiciels d'OCR ?
   (f) Quel est votre sentiment sur les performances de ces outils d'OCR ?
   (g) Procédez-vous à des évaluations de ces outils (*benchmark*) ?
   (h) Si vous évaluez ces outils, comment le faites-vous ? (CER, WER, précision, lecture...)
   (i) Est-ce que la qualité de l'OCR vous empêche d'exploiter directement votre corpus ?
   (j) Procédez-vous à des corrections de ces sorties d'OCR ? (Manuelles, automatiques, semi-automatiques)
   (k) Lorsque les sorties d'OCR ne sont pas parfaites, quel(s) terme(s) utilisez-vous ?

2.1.2 Déploiement et diffusion

L'ensemble de ces questions a été déployé sur la plate-forme framaforms.org (consulté le 28 juin 2022) pour les raisons suivantes :

- c'est une plate-forme libre qui n'utilise pas les données dans un but commercial ;
- les répondants n'ont pas à s'identifier et ne sont pas tracés ;
- elle est gratuite ;
- elle dispose des fonctionnalités nécessaires au bon fonctionnement d'un tel questionnaire (saisie plein-texte, choix unique, choix multiples, etc.) ;
- elle dispose enfin d'un module permettant, et automatiquement, la représentation graphique de certaines questions.

La communication de ce questionnaire a été faire par les listes d'e-mail *atala-LN* et *atala-DH* lors d'une première vague lancée le 3 mai 2021, et sur le site de l'ObTIC [1] et la liste de mails *CERES*[2] lors d'une seconde vague le 27 mai 2021.

2.1.3 Résultats

Public

La distribution des statuts est relativement égale entre les doctorants, ingénieurs et MCF (maîtres de conférences), quoique légèrement inférieure pour les étudiants et les bibliothécaires, ce qui constitue un bon panel du point de vue de la représentativité (voir Figure 2.1). Toutefois, un biais de l'étude se situe dans les listes de diffusion utilisées pour partager ce questionnaire. En effet, elles sont toutes orientées *lettre* (linguistes, philologues, etc.) ; ainsi déplorons-nous le manque de variété dans les disciplines dans lesquelles travaillent les répondants. *Toutes*, non : la liste *atala-DH* aurait pu permettre de glaner d'autres disciplines, ce qui n'a pas été le cas.

La majorité des participants se sentent très à l'aise dans le monde numérique quand les autres le sont aussi sans en être autant assurés. Tous les participants travaillent avec des outils numériques quotidiennement.

Usages de l'OCR

La majorité des participants utilisent les textes obtenus en sortie du processus d'OCR pour les publier (voir figure 2.3). Il apparaît aussi que ces données textuelles sont utiles pour réaliser de la prospection et valider ou réfuter des hypothèses pré-existantes, quand une minorité de participants utilisent ces données pour formuler des hypothèses. Les données textuelles océrisées interviennent donc en *début* (construction des données et prospection) et en *fin* de recherche (validation/réfutation d'hypothèses et publication des données) ; elles ne font pas partie de l'élaboration même de la recherche.

1. obtic.sorbonne-universite.fr (consulté le 28 juin 2022)
2. http://www.ceres-sorbonne-universite.fr/ (consulté le 28 juin 2022)

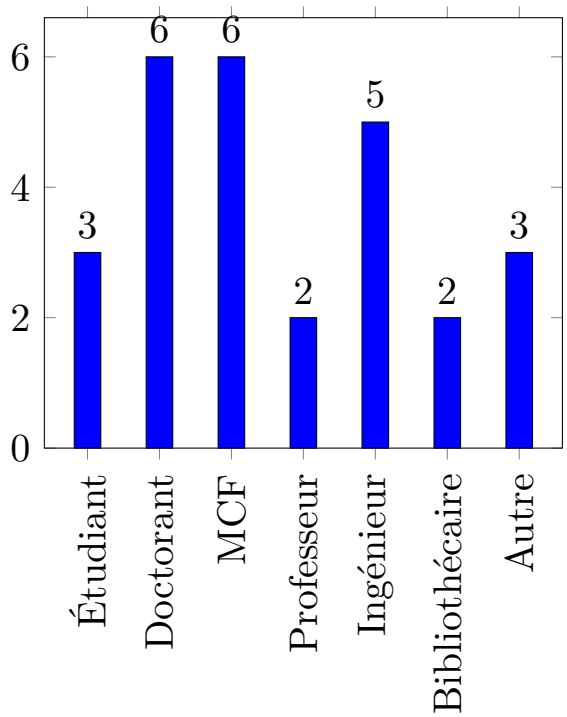


FIGURE 2.1 – Distribution des status des participants

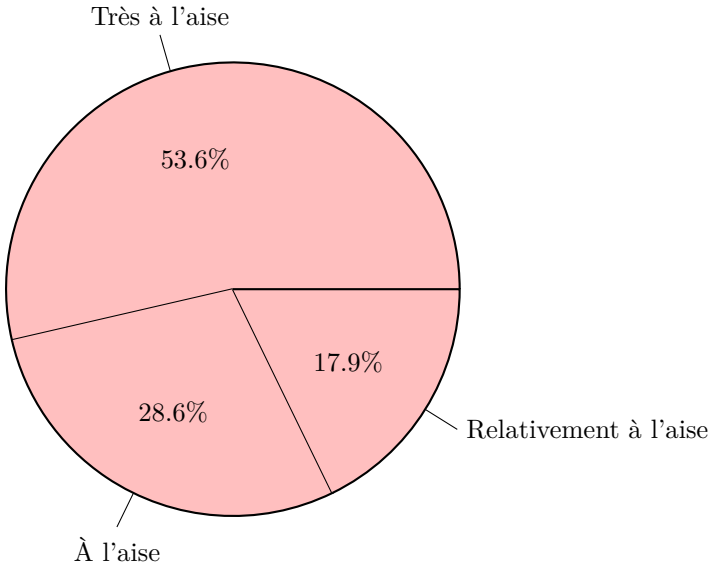


FIGURE 2.2 – Ressenti des participants sur leur capacité à naviguer dans le monde numérique

Plus de 96% des participants utilisent dans leurs travaux des corpus, ceux-ci étant numériques pour 82% d'entre eux. 75% des corpus numériques ont été construits par OCR.

Perception de l'OCR

Les participants semblent donc avoir une vision positive de ces outils d'OCR (figure 2.5), alors que 48% d'entre eux ne procèdent pas à une évaluation des performances de ces outils (en utilisant des métriques et des références) [3] (figure 2.6).

Plusieurs bonnes surprises concernant les performances de l'OCR ont été relevées par les participants (voir la figure 2.4 pour la répartition des logiciels d'OCR utilisés) :

— résultats concluant pour la textométrie, la qualité est suffisante (ABBY et KRAKEN) ;
— efficacité de la segmentation des éléments d'une page (ABBY) ;
— efficacité à gérer les caractères italiques (TRANSKRIBUS) ;
— reconnaissance de caractères difficilement lisibles par l'humain (TRANSKRIBUS).

Plusieurs mauvaises surprises cependant :

— les erreurs découragent la lecture (ABBY et KRAKEN) ;

3. La distribution gloable des 50-50 pour le *benchmark* se vérifie aussi pour chaque réponse *Très bonnes*, *Bonnes* et *Moyennes*.

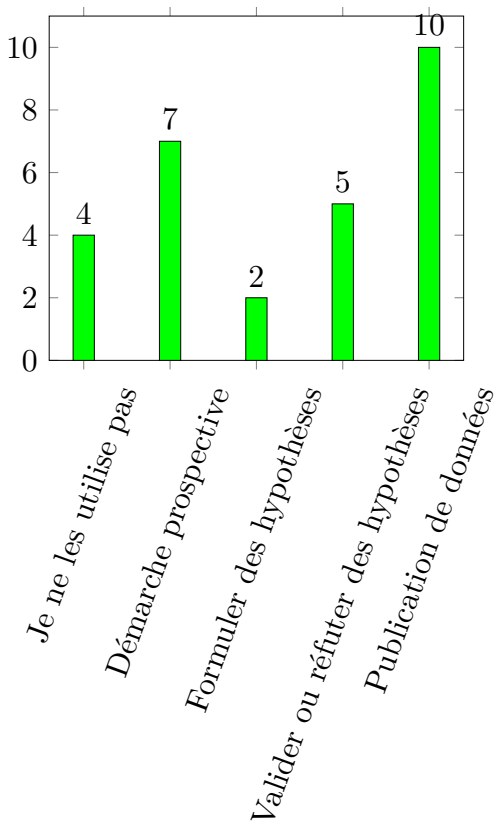


FIGURE 2.3 – Dans quel but utilisez-vous des outils d'OCR ?

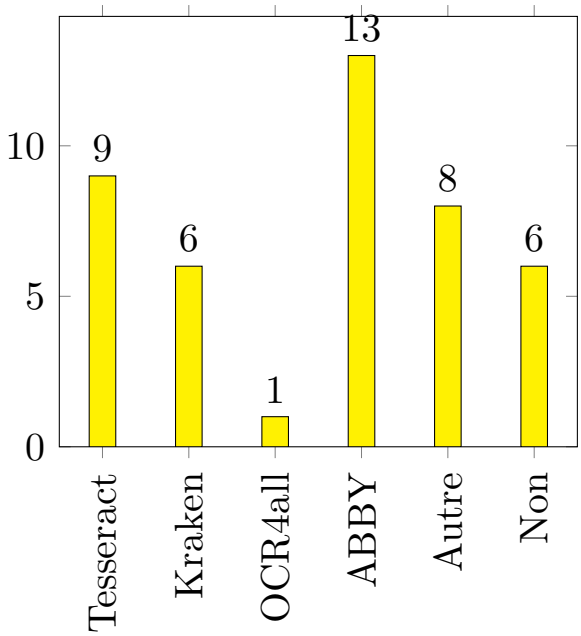


FIGURE 2.4 – Avez-vous déjà utilisé un logiciel d'OCR ?

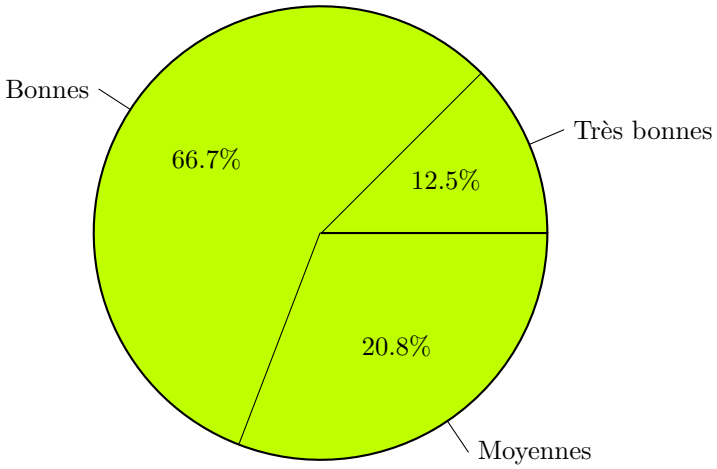


FIGURE 2.5 – Quel est votre sentiment sur les performances de ces outils d'OCR ?

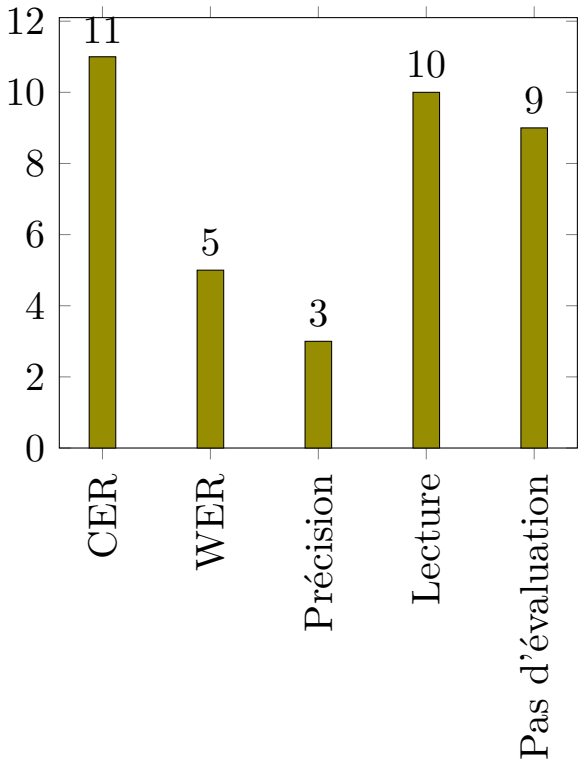


FIGURE 2.6 – Si vous évaluez ces outils, comment le faites-vous ?

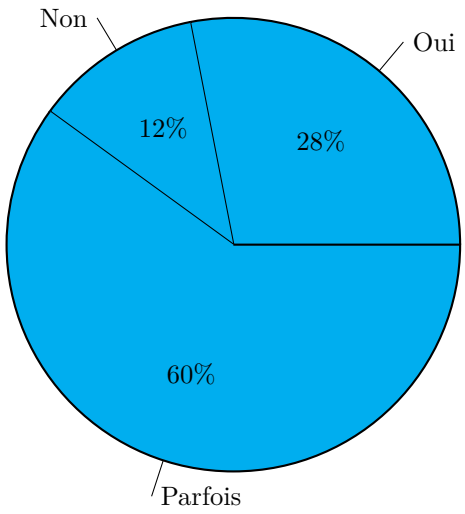


FIGURE 2.7 – Est-ce que la qualité de l'OCR vous empêche d'exploiter directement votre corpus ?

— parfois, mélange des éléments d'une page (ABBY) ;
— mauvaise gestion des italiques et autres marqueurs textuels (ABBY) ;
— plus globalement, trop de corrections nécessaires.

Parmi ceux qui qui déclarent chercher à évaluer ces outils, environ la moitié le fait par simple lecture de la sortie d'OCR, c'est-à-dire sans métrique d'évaluation. Quand il y a métrique en revanche, c'est le CER (taux d'erreur au caractère) qui est le plus utilisé (Figure 2.6).

Si l'on a vu que la qualité des sorties d'OCR était jugé comme « bonne » ou « très bonne » par la majorité des participants, paradoxalement, l'utilisation directe de ces corpus océrisés est gênée sinon empêchée par les erreurs qu'ils contiennent (Figure 2.7).

Ces erreurs qui empêchent l'exploitation directe des corpus ont pour conséquence une post-correction, manuelle en majorité (Figure 2.8).

Ces corpus océrisés, qui contiennent des « erreurs » (terme employé majoritairement), sont aussi jugés « bruités » (deuxième terme majoritaire) ou même plus spécifiquement « sales » ou « imparfaits » (Figure 2.9).

2.1.4 Conclusion

L'hypothèse de [Traub et al., 2015] selon laquelle les utilisateurs d'OCR ressentent un bruit plus fort qu'il ne l'est n'est ici pas vérifiée car les participants jugent plutôt bonne la qualité de l'OCR même s'ils sont bloqués en majorité par le bruit et que la correction manuelle est très souvent nécessaire. On vérifie le contraire : le bruit dans les données semble « sous-ressenti ».

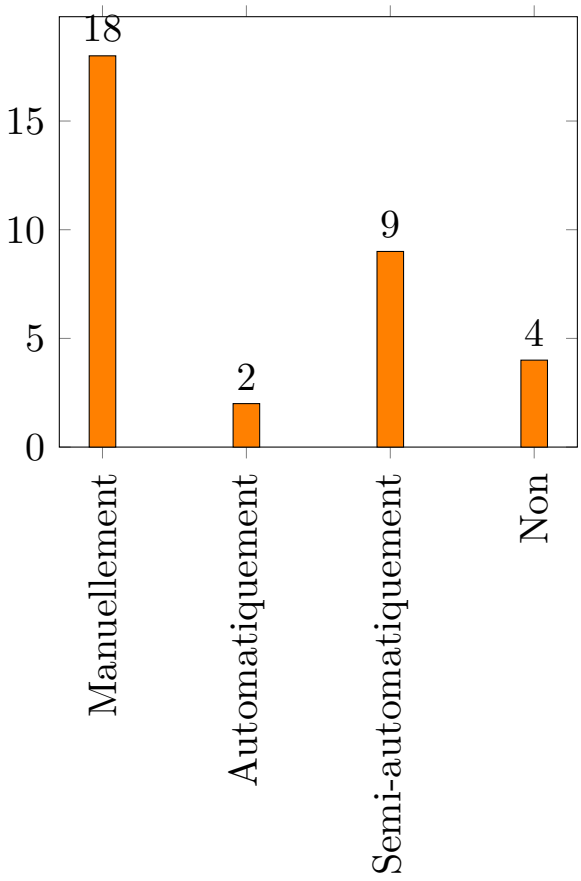


FIGURE 2.8 – Procédez-vous à des corrections de ces sorties d'OCR ?

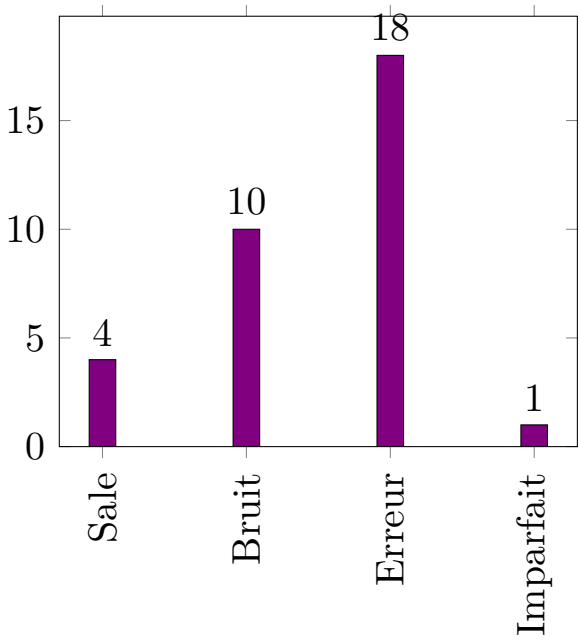


FIGURE 2.9 – Lorsque les sorties d'OCR ne sont pas parfaites, quel(s) terme(s) utilisez-vous ?

Sous-ressenti dans l'usage, mais pas dans la lecture. Cette expérience de la lecture oculaire de données océrisées est utilisée en majorité pour juger de la qualité d'une telle sortie de logiciel. Cette dernière est souvent complexe et même « décourageante ». Ce qui amène à penser que les participants distinguent bien usage direct de ces données comme avec la lecture et usage médié de ces données comme avec la textométrie.

Un paradoxe apparaît donc. La lecture est utilisée pour juger de la qualité des sorties d'OCR, globalement bonne selon les résultats du questionnaire, et en même temps cette lecture est jugée « décourageante »...

2.2 Entretiens

Les conclusions de la section précédente – l'évaluation des données textuelles issues d'OCR par simple lecture occulaire, le « sous-ressenti » du bruit lorsque le *distant reading* se manifeste ainsi que l'éviction de telles données océrisées au cœur même de la recherche – nécessitent plus de précisions. C'est pourquoi plusieurs entretiens ont été menés, sept au total, avec des acteurs de l'OCR d'horizons différents.

2.2.1 L'OCR : quels objectifs ?

G. travaille sur un corpus d'incunables. La difficulté éprouvée pour l'OCR est la segmentation de la page (ou, en anglais, le *layout analysis*). Il utilise l'OCR (avec *pdf2txt*[4]) pour la rapidité de l'acquisition des données textuelle mais « il y a une perte au passage ». En sous-texte on comprend qu'il y a une perte d'information linguistique. Cela lui pose problème car l'objectif de l'utilisation de l'OCR est de réaliser de la linguistique sur corpus – il s'intéresse aux chaînes graphiques. L'OCR est en réalité pour lui une solution à court terme, « l'idéal [étant] d'avoir un corpus parfait d'un point de vue philologique et avec une annotation linguistique en *part of speech*[5] ».

Dans son travaille avec l'OCR, G. a l'idée d'avancer même si la correction de l'OCR est lente. La rapidité de l'OCR permet de « voir ce qu'on peut faire de plus qu'avec un repérage occulaire [...] ; l'idée c'est de changer d'échelle ». L'OCR est donc utilisée pour sa rapidité d'une part mais aussi et par voie de conséquence pour la quantité de données qu'elle permet d'acquérir, car G. travaille sur les hautes fréquences des phénomènes pour éviter de les comparer avec des textes trop différents. Il procède ainsi à une étude intrinsèque de ses textes à travers leurs hautes fréquences. Au fond, l'OCR est « une façon d'aller à la pêche ».

Expert de cette technologie, J. abstrait et identifie « deux objectifs distincts » à l'OCR qui impliquent « deux niveaux d'exigences ». D'un côté, il y a l'édition de textes qui implique une correction après coup des erreurs d'OCR pour atteindre un taux d'erreur de 0%. À l'inverse, il y a le TAL (tokenisation, lemmatisation,

4. Voir http://manpages.ubuntu.com/manpages/trusty/man1/pdf2txt.1.html (consulté le 31 janvier 2022).

5. Parties du discours.

stylométrie, etc.) qui exige une masse mais pas de correction et donc un taux d'erreurs au caractère non nul est acceptable.

On peut alors identifier un travail deux vitesses quand il s'agit d'OCR. D'abord, un travail d'atelier, sur des données en masse, non préparées ou corrigées, où il est possible d'observer des phénomènes sur des fréquences fortes. Il s'agit d'un travail quantitatif et non qualitatif dont les résultats sont exploités comme piste de recherche. Les interrogés parlent volontiers de leur « soupe interne ». D'autre part, des recherches précises et singulières en vue de publications scientifiques des résultats. Il y a préparation du corpus, nettoyage et correction. Ici, c'est la rareté d'un phénomène qui est cherchée.

2.2.2 Comment et pourquoi mesurer la qualité des sorties d'OCR ?

Pour G., le jugement de la qualité de l'OCR se fait principalement « à l'œil » même s'il utilise parfois un taux d'erreurs. Une autre démarche qu'il tente d'entreprendre est l'évaluation en termes de gain de recherche : est-ce que si un collègue qui n'avait pas trouvé tel phénomène aurait au contraire été capable de le trouver sans l'utilisation de l'OCR ? En se posant cette question, il met bien l'accent sur le fait qu'il s'agit d'un monde nouveau pour lui (les disciplines comme les statistiques et l'informatique) et que dans une certaine mesure il va « contre » sa communauté en travaillant ainsi.

Pour R., « [i]l faut avoir des exigences qui se modulent en fonction des besoins. » Selon lui par exemple, l'évaluation dépend des usages : une bonne qualité sera exigée pour la recherche stylométrique ou encore l'édition de textes alors que, pour lui encore, une qualité inférieure n'est pas dérangeante pour la lecture occulaire.

Comme R., selon J. « la qualité de l'OCR dépend de la tâche et de l'objectif ».

2.2.3 Des erreurs de natures différentes

R. utilise aussi l'OCR dans ses travaux. Ce qu'il nous permet d'apprendre est que l'OCR implique des erreurs, certes, mais qui ne sont « pas de la même nature » que les erreurs faites par les humains lorsqu'ils transcrivent. Effectivement, ces derniers peuvent prendre un mot pour un autre alors que les erreurs d'OCR sont itératives donc plus reconnaissables *a posteriori*. Plus fort encore, il continue en remarquant que « la correction [des erreurs d'OCR] peut ajouter des erreurs ».

2.2.4 Un changement de paradigme : du mot au caractère

J. note qu'une représentation des données textuelles océrisées en bigrammes et en trigrammes est très robuste au bruit car ce sont des n-grammes très fréquents.

2.2.5 Une défiance de la communauté

L'entretien de G. se conclut par une remarque essentielle : « le bruit dans les données [acquises par OCR] implique une défiance des chercheurs et les corpus ne sont pas utilisés ».

2.3 Conclusion

Finalement, on retient de ces enquêtes que la communauté prise dans son ensemble ne croit pas à l'OCR car les erreurs semblent sur-estimées lors de la lecture occulaire. À l'inverse, les chercheurs qui travaillent avec ce genre de données textuelles bruitées semblent sous-estimer les erreurs car le jugement de la qualité, souvent jugée « bonne », ne se fait en général pas avec une métrique objective. La qualité de l'OCR est, en fin de compte, jugée selon le propre niveau d'idéalisme de chacun. Dans tous les cas, la correction *a posteriori* apparaît trop couteuse pour être menée au-delà d'un échantillon qui permettrait de calculer un taux d'erreurs.

Toutefois, les données océrisées ont plusieurs avantages majeurs en comparaison à l'acquisition manuelle des données textuelles :

— cette acquisition est beaucoup plus rapide ;
— les erreurs de l'OCR sont itératives, régulières, et donc susceptibles d'être corrigées automatiquement ;
— si l'étude des corpus textuels sans erreur de transcription se fait au grain mot, l'étude des corpus océrisés doit probablement se faire au grain caractère car les n-grammes de caractères sont très robustes au bruit.

L'OCR, qui cherche à acquérir une *masse*, est par essence faite pour travailler en *distant reading* ; le *close reading* nécessite les œuvres elles-mêmes, l'OCR est inutile à ce niveau. En outre, force est de constater que la qualité de l'OCR dépend bien de la tâche et de l'objectif.

On remarque finalement qu'il manque à la communuté un outil *user friendly* où l'océrisation ainsi que le taux d'erreurs puissent être donnés « clefs en main ».

Première partie

OCR de documents anciens et impact du bruit dans les données textuelles océrisées

Chapitre 3

Océriser des documents historiques

Sommaire

L'océrisation de documents historiques appartenant au patrimoine est une tâche complexe en appelant d'autres en amont. Effectivement, les fonds profus des bibliothèques contiennent souvent pour un même titre plusieurs *éditions* et *états d'une même édition*, sans même évoquer que plusieurs *items* d'un même état d'édition peuvent être présents dans les rayons de la bibliothèque. Un long travail bibliographique est donc nécessaire d'abord pour renseigner un ensemble de métadonnées pour chaque item (comme le titre, l'auteur et la date d'édition s'ils sont connus, mais aussi les signatures des feuillets, etc.) mais aussi pour sélectionner l'item à numériser [1].

Il est important de noter d'entrée de jeu que ce temps long de la recherche bibliographique se distingue de la temporalité de la recherche universitaire. En effet, ce travail bibliographique, nécessaire mais minutieux et chronophage, peut faire ralentir des travaux dépendant de cette étape préalable ; une recherche en linguistique sur de tels corpus anciens et requérant une approche philologique ne

1. Celui-ci sera sélectionné selon plusieurs critères au premier rang desquels se trouve sa *rareté* ou son caractère unique ; il s'agit de valoriser les *items* rares ou uniques pour proposer, notamment aux chercheurs, des numérisations qui ne sont pas des doublons des mêmes *items* numérisés dans d'autres campagnes.

peut se faire que sur un corpus, lequel ne peut être disponible si le travail des bibliothécaires n'est pas achevé.

Ensuite, le processus de numérisation est lui aussi complexe. D'abord, en tant que documents du patrimoine, les ouvrages ne peuvent être manipulés comme bon nous semble, « [p]arce qu'ils ont été valorisés comme éléments identitaires en danger » [Noille, 2020, section 1] [2]. Aussi est-il nécessaire pour une lecture d'utiliser un coussin de lecture pour limiter le degré d'ouverture de l'ouvrage et protéger sa reliure. *A fortiori*, lors de la numérisation, cette précaution est de mise. Avec de tels documents, fragiles et précieux, le prix de la numérisation augmente car le travail des prestataires est plus complexe et plus lent.

Notons aussi, pour ajouter un autre élément rendant les étapes précédant l'océrisation plus longues encore, que, s'agissant de fonds publics, un appel d'offre est souvent lancé pour établir un contrat avec un prestataire donné. En ce qui concerne la numérisation des mazarinades, le groupe *Puce & Plume* a été sélectionné par la Bibliothèque Mazarine. Ceux-ci proposent en livrable, pour chaque document numérisé, différents formats de numérisation : PDF, JPEG et TIFF. Sont disponibles aussi autant de fichiers ALTO [3] que de pages numérisées comportant des transcriptions automatiques (OCR) réalisées par Tesseract 4.0.

Une fois les numérisations rendues disponibles, l'océrisation peut commencer. Toutefois, le chercheur qui souhaite océriser une collection documentaire numérisée sera confronté à une multitude de logiciels – *open source* ou propriétaires – et de modèles d'OCR, ce qui peut rendre confus ce processus. En outre, les numérisations peuvent aussi être pré-traitées en amont de l'OCR (la binarisation, qui consiste à transformler tous les pixels d'une image en noir ou en blanc, est ici exemplaire). Logiciel, modèle, prétraitements, mais aussi format de la numérisation sont autant de variables pouvant influencer la qualité de l'OCR. Aussi proposons-nous ici un état de l'art sur l'OCR en général et sur l'OCR de documents historiques en particulier pour montrer toute la spécificité de ce type de documents.

3.1 L'OCR : une technologie à deux vitesses

De ses premiers balbutiements remontant au début du XX^{e} siècle jusqu'aux années 2020, l'OCR a connu un intérêt toujours croissant – quoique changeant – ce qui a permis une amélioration constante de ses performances. Cependant, si l'OCR de documents nés-numériques offre des performances très convaincantes, l'OCR de documents historiques est encore au stade de recherche [Nagy, 2016]. Le caractère « bruité » de tels documents (au sens où ils contiennent possiblement

2. « Mais s'il importe de muséaliser l'objet patrimonial, c'est qu'il est crucial pour une société donnée de le sauver, d'en garder la mémoire. » [Noille, 2020, section 1] Au concept de « patrimonialisation », Christine Noille associe en aval celui de la « muséalisation » : certaines œuvres sont mises à part, considérées comme échantillon ou monument d'un phénomène culturel. Car, « [l]'idée de patrimoine est liée à celle de perte avérée ou de disparition potentielle. » Christine Noille est, en outre, opposée à ce figement muséal et prône le contraire.

3. Standard XML : *Analysed Layout and Text Object*.

des tâches, que l'encre a pu déteindre, etc.) met cette technologie au défit en même temps qu'on observe beaucoup de campagnes de numérisation des fonds documentaires – tant au niveau national (*Gallica*), européen (*IMPACT*) qu'international (*Google Books*). Aux début des années 2020, l'enjeu est donc moins l'OCR de chèques bancaires ou d'avis d'imposition que l'OCR de collections documentaires historiques.

On comprend donc bien comment le contexte culturel et les préoccupations d'une société façonnent une technologie.

3.1.1 Histoire de l'OCR

Au commencement, les États-Unis et les années 1950

La première moitié du XX^{e} siècle inaugure l'intérêt pour l'OCR avec le fantasme d'une machine capable de *lire* des documents – en particulier pour l'assistance aux personnes mal voyantes [Burgy et al., 2020, Nagy, 2016]. Mais c'est à partir des années 1950 que l'OCR devient une entreprise commerciale compétitive [Stevens, 1961]. David Shepard fonde l'*Intelligent Machine Corporation*. Dans les années 60, Jacob Rabinow crée le premier lecteur d'adresses postales étasuniennes [Burgy et al., 2020, section 2.1.1]. Quelques dix années plus tard, la machine de Kurweil permet de lire des textes aux personnes mal voyantes [Burgy et al., 2020, section 2.1.1].

Dans les années 1960, plus de cinquante entreprises d'OCR existent aux États-Unis seulement. « Avec l'avènement des microprocesseurs et des scanners optiques bon marché dérivés des télécopieurs, le prix de l'OCR est passé de dizaines et de centaines de milliers de dollars à celui d'une bouteille de vin de table. » [4] [Nagy, 2016, section 1.2]

À partir des années 1960, lorsque les méthodes de codage des caractères se développent (comme l'ASCII pour les caractères latins), l'OCR prend un tournant multilingue [Nagy, 2016, section 1.2]. On a donc un berceau étasunien rapidement étendu à d'autres pays.

Toutefois, on a globalement assez peu d'informations détaillées sur la manière dont fonctionne ces technologies à cette époque [Nagy, 2016, section 1.2]. Effectivement, entremêlées dans le jeu des brevets et l'arsenal juridique, les entreprises privées ont assez peu d'intérêt à dévoiler leurs méthodes. Néanmoins, une exception demeure : *History of OCR, optical character recognition* (1982) d'Herbert F. Schantz [Schantz, 1982] dans lequel « est retracé la croissance de la REI (à l'origine *Recognition Equipment Inc.*), qui a été l'une des grandes réussites de l'OCR des années 60 et 70 » [5] [Nagy, 2016, section 1.2].

L'OCR connaît ensuite une amélioration progressive et continue de sa qualité jusqu'à atteindre des taux d'erreur au caractère atteignant 0,5 %, quoiqu'il reste important de garder en tête qu'il s'agit toujours de « dix erreurs par page

4. *« With the advent of microprocessors and inexpensive optical scanners derived from facsimile machines, the price of OCR dropped from tens and hundreds of thousands of dollars to that of a bottle of table wine. »*

5. *« it traced the growth of REI (originally Recognition Equipment Inc.), which was one of the major OCR success stories of the 1960's and 1970's. »*

imprimée » [Nagy, 2016, section 1.2]. Si l'auteur ne précise pas ce chiffre, la remarque reste pertinente dans la mesure où un faible taux d'erreur peut cacher bien des disparités et que le taux d'erreur est à entendre au prisme de la tâche que ces données satisferont.

Le paysage contemporain

Les années 2000 connaissent trois événements marquant un renouveau. Né en 2004, le projet conquérant de *Google Books*, selon lequel tout document aspirerait à être numérisé par le géant américain, a permis « une grande reconnaissance de la technologie OCR » [Burgy et al., 2020]. En 2005 est mis à disposition le premier logiciel libre d'OCR : Tesseract [Smith, 2007]. Ceci permet une large diffusion de la technologie OCR car elle est désormais disponible pour tous [Burgy et al., 2020, section 2.1.1]. En 2008 les premiers financements du projet de recherche européen *IMPACT* commencent. Ce projet vise à proposer « des outils et des méthodes de travail permettant l'océrisation de documents historiques numérisés avec un très haut niveau de précision » [Burgy et al., 2020, section 2.1.1] citant [Balk and Ploeger, 2009].

[Nagy, 2016, section 1.3] dresse comme suit le paysage industriel de l'OCR : « *FineReader* (*Abbyy*, originellement de Moscou), *OmniPage* (ascendance du produit : *Palantir* puis *Calera* puis *CAERE*, désormais *Nuance*), *Readiris* (*Iris*) et les produits *open source Tesseract* et *OCRopus* (à l'origine *ReadRight* de Hewlett Packard, désormais sponsorisé et promu par *Google*) » [6] [Nagy, 2016, section 1.3]. Il existe beaucoup d'entreprises proposant des produits fondés sur l'OCR mais la grande majorité utilise un des produits sus-cités.

3.2 L'océrisation : un processus

Pour commencer, avant l'OCR, une numérisation de qualité est souvent jugée indispensable [Burgy et al., 2020, section 2.1.2] (sans que cela ait été, à notre connaissance, démontré). Le TIFF non compressé apparaît dès lors comme privilégié, quoique très lourd ce qui pose des problème de stockage lorsqu'il y a numérisation de larges collections.

Le processus d'OCR est globalement fondé sur quatre étapes qui se suivent [Burgy et al., 2020] (voir aussi [Smith, 2007] pour un panorama de ce processus, centré sur Tesseract). D'abord, le pré-traitement des images (ou *pre-processing*). On trouvera : le rognage des bords vides [Zhou, 2010], la binarisation (par seuil, lequel peut être local ou global), le redressement des lignes (ou *deskewing*), etc. [Blanke et al., 2012, section 2.2.1]. Ensuite, suit l'étape de segmentation (ou *layout analysis*). La page est découpée en autant de rectangles que de colonnes, paragraphes, lignes, mots et caractères. Suit ensuite l'étape de reconnaissance des caractères à proprement parler (ou *recognition*). Après segmentation, chaque

6. *« FineReader (Abbyy, originally from Moscow), OmniPage (product ancestry : Palantir then Calera then CAERE, now Nuance), Readiris (Iris) and Open Source Tesseract and OCRopus (originally Hewlett Packard's ReadRight, now sponsored and promoted by Google) »*

caractère est décrit selon un ensemble de propriétés (ou attributs, ou *features* en anglais) : projection de tous les pixels sur un unique axe horizontal, vertical, etc. et une classification est réalisée pour attribuer à la forme un caractère [Anugrah and Bintoro, 2017, section C et D]. Finalement, une étape de post-traitement ou post-correction (ou *post-processing*) peut être réalisée. Elle vise à corriger la sortie de l'OCR après coup, pour en diminuer le taux d'erreur [Blanke et al., 2012, section 4].

3.2.1 Post-correction de sortie d'OCR

Cette dernière étape peut être réalisée manuellement, semi-automatiquement ou automatiquement – notamment grâce à l'apprentissage automatique, parfois couplé à l'utilisation de modèle de langue et de dictionnaire [Kissos and Dershowitz, 2016, section III]. Mais la post-correction automatique nécessite une vérité de terrain pour procéder à l'apprentissage, ce qui est très coûteux. Certaines recherches tentent de dépasser cette nécessaire transcription : [Ghosh et al., 2016].

PoCoTo, du projet *IMPACT*, permet quant à lui d'accélérer le processus de correction automatique en rendant la correction sérielle [Vobl et al., 2014].

Le *crowdsourcing* est également utilisé [Traub et al., 2018, Clematide et al., 2016] en faisant appel à des volontaires pour corriger des erreurs d'OCR. Une *gamification*[7] peut être envisagée.

3.3 Perfectionnements contemporains – réseaux de neurones artificiels

Si l'apprentissage automatique *supervisé* est déjà « une pratique bien établie dans le domaine » [Burgy et al., 2020, section 2.2.1], l'apprentissage *non supervisé* est en pleine expansion par l'utilisation des réseaux de neurones artificiels. [Rehman and Saba, 2014] posent en 2014 les premiers jalons d'une étude de ce genre d'apprentissage automatique non supervisé en établissant un premier état de l'art sur l'utilisation des réseaux de neurones pour le pré-traitement des images avant océrisation. Ils justifient leur utilisation croissante dans la communauté grâce à leur « capacité à généraliser »[8] [Rehman and Saba, 2014]. « Cette technologie permet en effet de faciliter le repérage des lignes de textes, la segmentation et l'extraction de *features* » [Burgy et al., 2020, section 2.2.1]

Deux années plus tard, [Afroge et al., 2016] proposent l'utilisation de tels réseaux de neurones pour la reconnaissance (et la classification) à proprement parler des caractères à partir de *features* dérivés des pixels. Les résultats sont encourageants pour les caractères alphanumériques anglais quoique ces résultats sont nuancés par les auteurs en rappelant que, dans le cas d'autres systèmes

7. C'est-à-dire rendre la tâche plaisante en lui faisant prendre la forme d'un jeu (*game*, en anglais).

8. *« generalization abilities »*

d'écriture, notamment ceux présentant des ligatures, les résultats sont peu satisfaisants.

Toutefois le temps de calcul et la quantité de données d'entraînement nécessaires au bon apprentissage peut constituer un frein à l'utilisation de ce genre d'apprentissage non supervisé. [Anugrah and Bintoro, 2017] préconisent de procéder à un débruitage des images avant l'océrisation de celles-ci pour réduire le temps de calcul, trop conséquent sinon.

3.4 Océriser des documents historiques

Encore à l'état de recherche [Nagy, 2016], l'océrisation de documents historiques en constitue un domaine phare [Burgy et al., 2020, section 2.2.3.1]. Cet engouement (qui atteint son paroxysme avec le projet européen IMPACT) s'explique par les nombreuses problématiques que pose ce type d'OCR :

- la technique d'imprimerie n'étant qu'à ses débuts, les typographies ne sont pas harmonisées et elles varient d'un imprimeur à l'autre, lesquels créent souvent leurs propres fontes [Berg-Kirkpatrick and Klein, 2014] ;
- il en va de même pour l'édition, où la mise en page des textes peut être particulièrement « riche » avec plusieurs tailles de caractères, plusieurs fontes différentes, des lignes en capitales, en minuscules, en italiques, etc. ;
- le système écrit comprend des ligatures, particulièrement complexes à détecter et segmenter [Hill and Hengchen, 2019, section 2.3] ;
- la qualité de l'encre et du papier peut varier d'une impression à une autre, rendant les ravages du temps plus ou moins visibles [9] [Abiven and Lejeune, 2019].

Par ailleurs, la post-correction peut être rendue elle aussi difficile en raison du nombre limité (voire inexistant pour certains état de langue) de ressources (lexicales et linguistiques) disponibles – quoique le projet IMPACT a tenté de répondre à ce manque. Il en va de même pour les langues contemporaines « peu dotées » comme, par exemple, l'alsacien [Millour and Fort, 2019]. Ce manque de ressources pose problème à plusieurs égards. D'une part, elles sont nécessaires à la post-correction, qu'elle fonctionne par interrogation de lexique ou par apprentissage de modèles de langues. D'autre part, ce manque de ressource corrobore un manque de vérités de terrain (transcriptions diplomatiques) nécessaire à l'apprentissage de modèle d'OCR spécialisés. Sur ce second point, [Springmann et al., 2014] montrent que l'apprentissage de modèle d'OCR spécialisé sur des vérités de terrain de l'état de langue étudié permet de dépasser le seuil symbolique de 90 % de précision.

> Les bons résultats d'OCR pour les imprimés historiques dépendent de la disponibilité de modèles entraînés sur des transcriptions diplomatiques comme la vérité de terrain, ce qui est à la fois rare et chronophage. » [10] [Springmann et al., 2014, résumé]

9. C'est aussi ce qu'on appelle le « bruit » en faisant référence aux taches, aux pages adjacentes qui ont déteint, etc.

10. *« Good OCR results for historical printings rely on the availability of recognition models*

Mais deux années plus tard, [Springmann et al., 2016] montrent qu'il est possible d'apprendre des « modèles mixtes » [11] sur des œuvres datant de siècles différents (en l'occurrence, dans [Springmann et al., 2016], 6 œuvres datant de 1471 à 1686) et d'atteindre des précisions de plus de 90 % sur des œuvres non présentes dans l'ensemble d'apprentissage. Cela démontre la capacité de cette technologie à dépasser la « barrière typographique » en généralisant lors d'un apprentissage mixte.

3.5 Conclusion

Le processus d'océrisation connaît donc des avancées inégales, selon le type de document à océriser. Pour les documents nés-numériques, les performances semblent satisfaisantes et les technologies qui les sous-tendent matures. Pour les documents historiques, l'enjeu est toujours de taille, en particulier face à une variation multimodale (des fontes, des états de langue, de la qualité des papiers et des encres, etc.). Il est donc encore nécessaire de procéder à des études évaluatives et comparatives pour créer de la connaissance dans ce domaine et ne pas être aveuglé par l'illusion séduisante selon laquelle l'OCR est une tâche résolue.

trained on diplomatic transcriptions as ground truth, which is both a scarce resource and time-consuming to generate.

11. *« mixed models »*

Chapitre 4

Le « bruit » : définition, impacts et représentations

Sommaire

4.1 Définir le bruit : entre singulier et pluriel

Parfois, le dimanche, j'entendais les cloches, la cloche de Lincoln, d'Acton, de Bedford ou de Concord, lorsque le vent se trouvait favorable, comme une faible, douce, et eût-on dit, naturelle mélodie, digne d'importation dans la solitude. À distance suffisante par-dessus les bois, ce bruit acquiert un certain bourdonnement vibratoire, comme si les aiguilles de pin à l'horizon étaient les cordes d'une harpe que ce vent effleurât. Tout bruit perçu à la plus grande distance

> possible ne produit qu'un seul et même effet, une vibration de la lyre universelle, tout comme l'atmosphère intermédiaire rend une lointaine arête de terre intéressante à nos yeux par la teinte d'azur qu'elle lui impartit. Il m'arrivait, en ce cas, une mélodie que l'air avait filtrée, et qui avait conversé avec chaque feuille, chaque aiguille du bois, telle part du bruit que les éléments avaient reprise, modulée, répétée en écho de vallée en vallée. L'écho, jusqu'à un certain point, est un bruit original, d'où sa magie et son charme. Ce n'est pas simplement une répétition de ce qui valait la peine d'être répété dans la cloche, mais en partie la voix du bois [...] [Thoreau, 1854, p. 69]

En procédant à cette citation, [Plas, 2015] rappelle l'importance de la « réévaluation radicale du bruit dans le contexte musical ». Ainsi abordée, la notion de bruit revêt une dimension tout à fait singulière et bien loin de la corruption à l'harmonie qu'on peut, de prime abord, lui associer. Car cette notion, tout aussi polymorphe que plurielle, s'inscrit dans des champs aussi variés que l'informatique, l'écologie, la physique ou encore la musique. C'est en embrassant cette variété que nous comptons dessiner un portrait, nouveau peut-être, du « bruit ».

4.1.1 Pourquoi la métaphore du bruit ?

Une rupture à l'harmonie

« Ensemble de sons, d'intensité variable, dépourvus d'harmonie, résultant de vibrations irrégulières. » C'est ainsi qu'est inaugurée la définition de bruit dans le dictionnaire du CNRTL [1]. Pour les musiciens, le bruit – comme le bruit des percussions – n'est pas une note de musique et se distingue en tant que rythme de la mélodie. En musique donc, le bruit apparaît comme un ensemble de sons complexes romptant avec l'harmonie et la mélodie. De la même manière, dans un environnement, le bruit – qui est souvent source de litiges – est une polution sonore, donc une *concentration* sonore, qui se distingue de l'ensemble des sons normalement tolérés. Il y a bien cette idée de *quelque chose en plus*, ou plutôt *quelque chose en trop*, qui vient détonner contre ce qui est habituellement et par consensus admis comme harmonieux. Toutefois, et c'est particulièrement le cas dans un paysage urbain, on ne peut dire que l'harmonie ait comme propriété le silence. Il suffit d'imaginer une capitale sans plus aucun son pour l'animer pour s'en rendre compte. Entre présence et absence donc, le *bruit* se définit dans un premier temps relativement et par opposition à une *harmonie*. Toujours est-il que le *bruit* est un phénomène saillant, parfois dérangeant ou désagréable, mais toujours *perçu*. « Faire grand bruit » révèle ainsi l'aspect inévitable de la réception de ce bruit, cette rumeur. Il s'agit d'un éclat, d'un retentissement, qu'on ne peut que percevoir.

1. https://cnrtl.fr/definition/bruit (consulté le 28 juin 2022)

Aux carrefours des paradoxes

Qu'en est-il du « bruit de fond », par exemple en radiophonie ? Il y a bien une rupture à l'harmonie que serait l'émission « pure » mais il ne s'agit pas d'un éclat momentané comme le serait le son d'une percussion, mais bien d'un phénomène continu gênant la bonne compréhension de l'émission diffusée. En cela, on peut apprécier une source de bruit, discrète ou continue, comme dans [Shannon, 1949, section 11, p. 49], qui s'ajouterait de manière *persistante* à la « vraie » source d'information – il s'agit alors d'un *flux*. Il y aurait donc une *superposition*, brouillante ou corruptrice, à un signal considéré comme utile. Il y a donc une *nuisance* à la compréhension. Mais le bruit peut aussi être grand, retentissant, résonnant comme petit, murmurant, doucement de manière persistante. En témoigne le foisonnement du vocabulaire associé, dont : *vacarme, boucan, brouhaha* ou *barouf* d'un côté et *glouglou, cliquetis* ou encore *bourdonnement* de l'autre. Cette intensité du bruit, si elle existe en soi en étant mesurable, est à la fois le résultat de *mon* interprétation. Les bruits du corps du partenaire, aussi, qui finissent par agacer et insupporter pourraient être un exemple satisfaisant. Ici se dessine à nouveau un paradoxe : pourrait-on être sourd au bruit ? N'ai-je jamais entendu auparavant ces bruits du corps, pourtant maintenant si insupportables ? N'avais-je jamais considéré l'ampleur de la nuisance que représente le petit bruit quotidien des pas de mon voisin du dessus ?

Ceci va dans le sens de la perception relative à l'idéalisme de chacun quant aux erreurs d'OCR vue dans la conclusion du Chapitre 2.

Des objets *derrière*

Que le bruit considéré soit grand, petit, réel, interprété, perçu ou non, il semble qu'on puisse trouver, au moins souvent, un lien (causal en ce qui concerne les sons) entre le bruit et sa source. Ainsi entendons-nous la voiture en approche derrière le son du klaxon, la porte derrière le grincement ou encore l'électroménager derrière les alarmes. Il n'est peut-être pas anodin qu'on dise entendre un piano plutôt que le son d'un piano. Être sourd au bruit, ce serait alors refuser de voir ce qui se cache *derrière* ces bruits, ce serait refuser d'étudier le lien (causal, de ressemblance, d'écho, etc.) qui existerait entre le bruit et sa source mais aussi son environnement.

Tirer parti des bruits

Si les bruits sont évités, contournés, gommés, il arrive aussi qu'on les « piège » et les « annexe ». Les bruits ambiants sont de bons exemples en musique. Alors apparaît un changement, qui consiste à substituer,

> grâce aux travaux de Von Förster, à l'idée de l'engendrement de l'ordre par l'ordre, celle de la production de l'ordre à partir du bruit, à l'*order from order principle*, un *order from noise principle*, alors il

> faut assimiler l'œuvre à une chréode de Waddington [2] : à un ensemble autocorrelé vivant, du genre système ouvert, qui ne rejette pas le bruit ambiant mais sait le piéger et l'annexer. Et le critère de la modernité résidera, comme l'a établi Cage, dans la capacité d'une musique à intégrer les perturbations. Ou à piéger les bruits. [Plas, 2015] (citant un article de Daniel Charles, publié par la revue *Traverses* en février 1982 – il y est question de musicologie)

Le titre *The Logical Song* de l'album *Breakfast in America* de Supertramp, paru en 1979, est ici un exemple populaire privilégié avec des bruits donnés par le chanteur lui-même ou encore des bruits de sonneries. Les vocalises hasardeuses de la chanteuse Solange avec le titre *Binz* de l'album *When I Get Home* paru en 2019 pourrait constituer un exemple plus contemporain.

Les bruits peuvent ainsi être partie prenante d'une communication, d'une autre communication, une communication piégée, à piéger. On comprend donc bien que cette métaphore du bruit se soit étendue pour rendre compte des « erreurs dans les données », ou dans notre cas particulier des erreurs d'OCR. Comme les bruits, les erreurs d'OCR engendrent d'abord une défiance, c'est-à-dire qu'on cherche à les éviter pour retrouver une harmonie originelle. En second lieu, les erreurs d'OCR peuvent aussi rendre compte d'un phénomène en amont ; par exemple, un passage bruité peut être le résultat d'un verso ayant déteint, les années faisant, sur le recto adjacent. Enfin, il semble aussi possible, comme c'est le cas dans le domaine musical, qu'on puisse « tirer parti des bruits » ; ainsi pourrait-on imaginer, grâce à l'observation fine du bruit dans les résultats d'OCR, retrouver les pages vides et les pages de reliures d'un ouvrage tant la segmentation peut être ardue pour ces exemples.

4.1.2 Qu'est-ce que *le* bruit ?

Si l'expression *le bruit dans les données* invoque un substantif singulier pour caractériser ce qui rompt avec l'harmonie originelle de ces données, il n'est pas clair que cela traduise une homogénéité de ce qui est considéré comme bruit. Très simplement, rappelons ici que le *bruit*, dans les résultats d'un système, est très souvent associé en complémentaire au *silence*, lequel correspond à tout ce que le système a omis. On a dès lors un faux sens apparent. D'un côté, nous disions pour les erreurs d'OCR que le bruit correspond à tous les mots ou caractères substitués, ajoutés ou omis ; de l'autre, nous disons que, pour un système de classification par exemple, le bruit correspond à tout ce qui est *en trop* (les faux positifs) alors que le silence correspond à tout ce qui *manque* (les faux négatifs).

Expressions concurrentes

Le terme *bruit* est largement utilisé dans la littérature. [Naji et al., 2011] utilisent dans leur titre l'expression « corpus bruité (OCR) ». [Hamdi et al., 2019,

2. Mouvement de développement que suit une cellule après sa création en vue de faire partie d'un organe spécialisé.

introduction] parlent quant à eux de « texte bruité » [3], comme [Franzini et al., 2018, section 2] qui utilisent l'expression « tolérance algorithmique au bruit » [4].

Plus spécifiquement, le qualificatif *sale* et sa connotation largement négative peut être rencontré. On parlera alors de « données sales », comme avec « OCR sale » [5] [Hill and Hengchen, 2019, titre].

Le terme erreur, qui implique la référence à une vérité, engendre aussi diverses expressions. [Taghva et al., 2004], dans leur titre, utilisent l'expression « erreurs d'OCR » [6], comme [van Strien et al., 2020, introduction] qui parlent simplement d'« erreurs » et en donnent certaines propriétés : « La sortie d'un logiciel d'OCR contient souvent des erreurs où le texte a mal été transcrit. Les erreurs vont d'un caractère incorrect, à un mot entier, et par voie de conséquence à des phrases mal transcrites. » [7]. Mais c'est oublier d'un côté le caractère itératif des erreurs d'OCR et les transcriptions de ce qui n'est pas du texte d'un autre côté.

Plutôt que de parler du « bruit » produit par l'OCR, [Mittendorf and Schäuble, 2000, introduction] introduisent le terme de « corruption » dans le cadre de la recherche d'information. Les auteurs continuent ainsi : « L'erreur d'OCR corrompt l'information, ce qui n'est pas souhaitable mais la recherche d'information peut y faire face. Une numérisation imprudente est cependant *responsable* de la perte d'information. » [8] [Mittendorf and Schäuble, 2000, section 2] [Hill and Hengchen, 2019, section 3.2] parlent aussi de corruption avec l'expression « corruption d'OCR » [9].

Enfin, le sigle *OCR* lui-même peut être employé en adjectif, parfois aussi par son développement comme verbe avec *océriser* ou comme nom avec *océrisation*, comme avec les exemples « texte OCR » [10] [Taghva et al., 1994, titre], « documents OCRisés » [11] [Hamdi et al., 2019, titre] ou encore « données OCRisées » [12] [Hamdi et al., 2019, résumé], renvoyant à une version par essence bruitée du « texte ».

Le bruit, une notion shannonienne

Dans sa *Théorie mathématique de la communication* [Shannon, 1949], Shannon se propose de définir les éléments intervenant dans le processus de communication – ce terme restant très général tant la prétention de la théorie est de s'appliquer à « tous les comportements humains » (Warren Weaver, dans la

3. *« noisy text »*
4. *« algorithmic tolerence to noise »*
5. *« dirty OCR »*
6. *« OCR Errors »*
7. *« The output of OCR software often contains errors where text has been incorrectly transcribed. Errors range from one character being incorrect, to entire words, and consequently sentences being incorrectly transcribed. »*
8. *« OCR error corrupt the information, which is not desirable but information retrieval can cope with it. Careless scanning is, however responsible for the loss of information. »*
9. *« OCR corruption. »*
10. *« OCR Text »*
11. *« OCRed Documents »*
12. *« OCRed Data »*

seconde partie de l'ouvrage intitulée *Contributions récentes à la théorie mathématique de la communication*). Le fameux schéma de Shannon en présente les éléments fondamentaux : la source d'information, le message, l'émetteur, le signal d'un côté, le signal reçu, le récepteur, le message et la destination d'un autre côté [Shannon, 1949, p. 12]. Un troisième élément est la source de bruit qui vient apporter une modification entre le signal émis et le signal reçu.

La *Théorie* est fondée sur l'observation que les messages peuvent (et doivent pour entrer dans cette théorie) se traduirent sous forme de probabilités (de fréquences, en l'occurrence). Ainsi, l'information d'un message est associée à la probabilité que la source fournisse ce message ; on a alors la formule de l'entropie :

$$H(x) = -\sum p_i log p_i$$

avec p_i la probabilité du symbole i [Shannon, 1949, p. 31]. Cette quantité d'information, aussi interprétable comme l'incertitude du message, est le fondement de toute la théorie. De là arrive le théorème fondamental qui fixe la limite théorique de la performance d'un codage enlevant toute redondance [13] à C/H (avec C la capacité du canal de transmission du message). D'abord envisagée pour des systèmes discrets sans bruit, les notions shannoniennes sont élargies en premier lieu aux systèmes discrets avec bruit, puis aux systèmes continus.

Dans la section *11. Représentation d'un canal discret bruité*, Shannon développe sa théorie pour les canaux bruités, et en explique la nature.

> Nous considérons maintenant le cas où le signal est perturbé par du bruit au cours de la transmission, ou a l'un ou l'autre des terminaux. Cela signifie que le signal reçu n'est pas nécessairement le même que celui envoyé par l'émetteur. [Shannon, 1949, p. 49]

Un cas particulier de cette différence entre le signal reçu et le signal émis est celui de la *distorsion*.

> Si un signal transmis particulier donne toujours le même signal à la réception, c'est-à-dire si le signal reçu est une fonction définie du signal transmis, alors l'effet peut être appelé distorsion. Si cette fonction est inversible – deux signaux transmis ne produisent jamais le même signal reçu – on peut corriger la distorsion, du moins en principe, simplement en appliquant la fonction inverse au signal reçu. [Shannon, 1949, p. 49]

Plus généralement :

> [l]e cas qui nous intéresse ici est celui où le signal ne subit pas toujours les mêmes changements lors de la transmission. [...] Dans ce cas, nous pouvons supposer que le signal reçu E est une fonction du signal transmis S et d'une seconde variable, le bruit N : $E = f(S, N)$ [Shannon, 1949, p. 49]

13. Les éléments redondants d'un message sont tout ce qui n'apporte pas de l'information ; il s'agit du « complément à 1 » de l'entropie de la source d'information. [Shannon, 1949, p. 38] La redondance est donc tout ce qui n'est pas porteur d'information, ce qui est, au fond, répété. L'introduction de redondance dans un message permet dès lors de vérifier l'intégrité de ce message.

Ceci permet à Shannon de considérer le bruit « comme une variable aléatoire au même titre que le message. En général, il peut être représenté par un processus stochastique convenable. » [Shannon, 1949, p. 49] Si la source de bruit peut être « représentée par un processus stochastique convenable », cela signifie que cette source de bruit est aussi une source d'information. Au fond, rien ne différencie fondamentalement la source d'information de la source de bruit, à part qu'on souhaite, par l'introduction de redondance, minimiser l'impact de la seconde.

Dans le contexte qui est le nôtre, le processus d'océrisation, une adaptation du schéma shannonien consisterait à mettre en source d'information l'item, c'est-à-dire l'ouvrage physique numérisé, et en destination le lecteur, ou bien l'utilisateur d'une bibliothèque numérique. Le signal émis serait la numérisation de l'ouvrage et le signal reçu le résultat de l'OCR, les données textuelles. Ainsi la source de bruit serait le module d'OCR de cette chaîne de traitements.

L'intérêt de procéder à cette analogie est qu'il devient possible, dans ce modèle, de calculer l'*entropie conditionnelle* de la sortie d'OCR, c'est-à-dire du signal reçu. On parle d'entropie *conditionnelle* et non pas simplement d'entropie car il y a intervention de la source de bruit : l'entropie conditionnelle est l'entropie de la source sachant une source de bruit. Si donc on est capable de calculer pour différents modèles d'OCR les entropies conditionnelles, il devient possible de quantifier l'information reçue et de quantifier le bruit généré par les erreurs d'OCR.

Puisqu'un i) l'entropie d'une source augmente à mesure que le signal est incertain [Shannon, 1949, p. 32] et que ii) la langue est régie par un ensemble de redondances [Shannon, 1949, p. 38], plus une sortie d'OCR sera bruitée, moins nous pourrons observer de redondance (si la source de bruit peut être modélisée par un processus stochastique, comme présupposé par modèle shannonien), plus donc l'entropie sera grande. Ces considérations peuvent être utiles dans une démarche d'évaluation rompant avec la nécessaire comparaison de la sortie d'OCR à une transcription de référence (voir Chapitre 5).

Erreurs d'OCR : des origines aux perceptions

Si l'on souhaite approcher un peu plus la notion de bruit relativement au processus d'océrisation, on peut chercher à comprendre quelles sont les origines de ces erreurs.

> L'OCR produit ses meilleures performances avec des documents bien imprimés et mordernes. Cependant, les documents historiques posent encore problème pour la reconnaissance de caractères et ainsi l'OCR de tels documents ne donne toujours pas de résultats satisfaisants. Parmi les raisons expliquant pourquoi les documents historiques posent toujours problème on a la variation des fontes entre différents matériaux, les mêmes mots orthographiés différemment, la qualité du matériel où certains documents peuvent présenter des déformations et l'indisponibilité de lexique de variants orthographiques historiques

connus [...]. [14] [Mutuvi et al., 2018, introduction]

À ces éléments communs posant problème à l'OCR s'ajoutent les « erreurs typiques d'OCR » [15] elles-mêmes. [Taghva et al., 2004, section 2] les analysent pour ensuite proposer des modules de post-traitement, de post-correction [16] donc. Sont identifiés comme erreurs typiques :

— les tirets des césures de fin de ligne, qui augmentent le nombre d'entrées et leurs fréquences dans l'index de mots qui, *de facto* n'existent pas (les « demi-mots ») ;
— les mots grammaticaux mal reconnus, qui normalement sont ignorés par leur système mais, étant mal reconnus, ils apparaissent dans l'index : et comme ce sont les mots les plus fréquents (cf. la « loi » de Zipf), ce sont aussi les erreurs les plus fréquentes ;

[Mutuvi et al., 2018, section 2.1] complète la liste en notant que :

> les erreurs communes d'OCR incluent les erreurs de ponctuations, de casse, de format de caractère, du sens des mots et de segmentation où les espaces entre différentes lignes, mots ou caractères mènent à de mauvaises reconnaissances des espaces blancs [...]. Les erreurs d'OCR peuvent aussi provenir d'autres sources telles que la variation de la fonte entre différents matériaux, la variation orthographique historique, la qualité du matériel ou le langage spécifique de certains textes médiatiques [...] [17]

[Hill and Hengchen, 2019, section 2.3] montrent enfin que, pour leur corpus, les erreurs d'OCR sont le plus souvent causées par l'ajout ou la suppression de s long ou des ligatures *ct*, *ff* et *ffl*.

Toutefois, le bruit n'est pas « neutre », mais bien dépendant du corpus qui le génère. On peut dès lors identifier ce qui cause le plus d'erreurs d'OCR, selon le corpus de l'étude.

> cette analyse nous permet de démontrer que les erreurs d'OCR ne sont pas neutres quand il s'agit de textes du dix-huitième siècle, mais plutôt, les long s et les ligatures ont statistiquement plus de chances d'être des erreurs dans mots des reconnus mais erronés. [18] [Hill and Hengchen, 2019, conclusion]

14. *« OCR produces its best results from well-printed, modern documents. However, historical documents still pose a challenge for character recognition and therefore OCR of such documents still does not yield satisfying results. Some of the reasons why historical documents still pose a challenge include font variation across different materials, same words spelled differently, material quality where some documents can have deformations and unavailability of a lexicon of known historical spelling variants [1]. »*

15. *« typical OCR errors »*

16. Un module de post-correction d'OCR prend en entrée une sortie d'OCR et procède à sa correction, qui peut être diplomatique, modernisée ou autres formes intermédiaires – il s'agit d'un continuum.

17. *« Common OCR errors include punctuation errors, case sensitivity, character format, word meaning and segmentation error where spacings in different line, word or character lead to mis-recognitions of white-spaces [22]. OCR errors may also stem from other sources such as font variation across different materials, historical spelling variations, material quality or language specific to different media texts [1]. »*

18. *« this analysis allowed us to demonstrate that OCR errors are not neutral when it comes*

Il est donc important de systématiquement nuancer les propos relativement au corpus de l'étude, ou, au moins, au type de texte de l'étude.

Enfin, la perception de ces erreurs est toute aussi fondamentale, dans la mesure où une perception amplificatrice peut avoir pour résultat la non utilisation de ces données (comme vu en Chapitre 2). [Traub et al., 2015] notent un écart entre la qualité réelle des données textuelles océrisées dont il est question et la qualité perçue de ces mêmes données. Ce qui était en jeu était « la faible qualité d'OCR perçue, et pas la vulnérabilité bien comprise des tâches de recherche des interviewés aux erreurs d'OCR. » [19] [Traub et al., 2015, section 2] Ce phénomène peut être expliqué en rappelant les résultats de [Chiron et al., 2017] : « Ils ont constaté qu'environ 15% des termes mal orthographiés représentent des entités nommées et que même 80% des 500 requêtes les plus fréquentes en contiennent au moins une. » [20] [Traub et al., 2018, section 2] Les entités nommées sont largement impactées par les erreurs d'OCR alors que ce sont des éléments particulièrement investigués, ce qui peut entraîner une sous-évaluation de la qualité réelle des textes océrisés.

En ces termes, les ressources océrisées peuvent être sous-exploitées, non-étudiées, ou longuement corrigées manuellement, alors qu'il serait possible de les exploiter en l'état – quoiqu'il faille définir ce qui est possible de réaliser avec de tels corpus océrisés. Le caractère « bruité » des corpus océrisés ne semble pas interdire toute étude.

4.1.3 Contre le terme « bruit »

Arrêtons-nous un instant sur le terme de bruit. Souvent, lorsqu'il est utilisé seul, il désigne toutes les *erreurs* introduites par le processus d'océrisation, c'est-à-dire (au fond comme avec le calcul du CER [21] ou WER [22]) toute insertion, substitution ou délétion. Le bruit ne fait que grand bruit. L'utilisation de ce terme porte alors à confusion, laissant à croire qu'il n'y a que des signaux *en plus* perturbants un signal originel malgré tout toujours présent dans son état le plus pur – qu'il s'agirait, par post-correction, de retrouver. La réalité est plus complexe.

Effectivement, le signal n'est pas que bruité *par-dessus*, mais aussi *en-dedans*. Un signal bruité *par-dessus* contiendrait le message originel et un ensemble aléatoire (pour se replacer dans le modèle shannonien) d'insertions. Or, on l'a vu, le bruit se caractérise aussi par des substitutions et des délétions qu'on ne peut pas toujours retrouver en post-correction.

Si donc le terme de « bruit » est commode par toutes les inférences métaphoriques qu'il permet, il biaise en même temps notre compréhension du phénomène

to eighteenth-century texts, and instead, the long-s and ligatures are statistically more likely to result in erroneously recognized words. »

19. *« the perceived low OCR quality, and the not well-understood susceptibility of the interviewees' research tasks to OCR errors. »*

20. *« They found that about 15% of the misspelled terms represent named entities and that even 80% of the top 500 queries contain at least a mention of one. »*

21. Taux d'erreur au caractère.

22. Taux d'erreur au mot.

mise en jeu. Ainsi est-il préférable de parler sobrement d'« erreurs d'OCR ».

4.2 À la recherche des impacts des erreurs d'OCR

Comme déjà montré *supra*, [Mutuvi et al., 2018, introduction] rappellent bien que

> [l'] OCR produit ses meilleurs résultats pour des documents modernes et correctement imprimés. Cependant, les documents historiques posent encore problème pour la reconnaissance optique de caractères et ainsi l'OCR de tels documents ne renvoie pas de résultats satisfaisants. Parmi les raisons pour lesquelles les documents historiques posent encore problème on peut inclure la variation des fontes entre matériaux, les mêmes mots écrits de manière différente, la qualité du matériel où certains documents peuvent avoir des déformations et l'inexistence de lexique de variants graphiques historiques [...] Ces facteurs réduisent la précision de la reconnaissance laquelle affecte le traitement des documents et, surtout, l'utilisation des bibliothèques numériques. [23]

Certaines propriétés inhérentes aux collections historiques et patrimoniales ont nécessairement une influence sur la qualité de la reconnaissance automatique de leur données textuelles par OCR. Si l'on comprend bien l'enjeu premier qui est la difficulté augmentée à procéder à d'autres traitements automatiques en aval, il en est un, plus direct, souligné justement par [Mutuvi et al., 2018, introduction]. Des données textuelles « bruitées », pour lesquelles on ne dispose pas toujours d'information sur l'origine et la qualité du processus d'océrisation, peuvent freiner la confiance que nous avons en elle et donc, peut-être surtout pour la communauté des chercheurs, freiner « l'utilisation des bibliothèques numériques ». [Traub et al., 2015, section 3] montrent qu'il existe des voies variées pour réduire le nombre d'erreurs dans une sortie d'OCR, sans toutefois qu'aucune d'elles ne puisse prétendre corriger *certainement* toutes les erreurs. D'où la nécessaire étude de leur impact. [Traub et al., 2018, section 2] montrent aussi qu'il existe peu de recherches sur cette question de l'impact des erreurs d'OCR et que

> même si le résultat de ces études peut aider à améliorer la qualité des données de manière plus efficace, on ignore encore comment cette correction affecte les recherches des chercheurs. [24]

23. *« OCR produces its best results from well-printed, modern documents. However, historical documents still pose a challenge for character recognition and therefore OCR of such documents still does not yield satisfying results. Some of the reasons why historical documents still pose a challenge include font variation across different materials, same words spelled differently, material quality where some documents can have deformations and unavailability of a lexicon of known historical spelling variants [1]. These factors reduce the accuracy of recognition which affects the processing of the documents and, overall, the use of digital libraries. »*

24. *« While the results from these studies can help improve data quality more efficiently, it remains unclear how this correction affects a scholar's research. »*

Car, corriger, c'est déjà transformer les données.

En guise d'introduction de cette section, citons quelques travaux présentant certains *seuils*.

> Les résultats jusqu'à présent démontrent qu'une OCR en-dessous du niveau 70-75% de précision a, sans surprise, un fort impact négatif sur nombre de méthodes analytiques. Cependant, ce qui peut être surprenant, c'est que les données OCR avec, ce qui peut sembler subjectivement être, des problèmes substantiels – des données dans lesquelles jusqu'à deux ou trois mots sur dix sont mal identifiés – sont encore potentiellement très utiles. [25] [Hill and Hengchen, 2019, section 4]

Ce même seuil de 70-80% est validé par [van Strien et al., 2020, section 2], citant [Hill and Hengchen, 2019] :

> De cette étude, un seuil de qualité d'OCR critique entre 70 et 80% a émergé, la plupart des tâches fonctionnent très mal en-dessous de ce seuil, les bons résultats sont atteints au-dessus, et des résulats fluctuants sont atteints en-dedans, selon la tâche à accomplir. [26]

Si certains travaux concluent, comme montré ci-dessus, en offrant aux lecteurs des *seuils* au-dessus et en-dessous desquels les performances des systèmes sont impactés, il reste que selon les corpus et les tâches, ces seuils eux-mêmes varient. L'affirmation de résultats généraux semblent dès lors périlleuse. C'est ainsi que [van Strien et al., 2020, section 5] concluent leur travail : « l'utilisation de texte océrisé a eu un impact sur toutes nos tâches, même si le degré varie. » [27] Nous déclinerons donc l'étude des impacts des erreurs d'OCR selon les tâches présentés dans les travaux.

Notons enfin que l'étude des impacts des erreurs d'OCR, si elle permet de conclure avec assurance pour une tâche donnée, peut servir de fondement à une évaluation extrinsèque de la qualité d'une OCR. D'où la circularité entre mesure de la qualité d'une document océrisé et étude de l'impact de ses erreurs d'OCR. Les sous-parties à venir s'articulent autour de la « complexité » de la tâche et de son moment de réalisation dans les chaînes de traitements. Ainsi trouvons-nous les prétraitements au début et, par exemple, la sémantique distributionnelle à la fin.

---

25. *« The results up until now demonstrate that OCR below the 70–75% accuracy level has, unsurprisingly, a strong negative impact on a number of analytical methods. However, what may be surprising is that OCR data with, what may appear subjectively to be, substantial issues—data in which as much as two or three words out of ten are misidentified—is still potentially very useful. »*

26. *« From this study, a critical OCR quality threshold between 70 and 80% emerged, where most tasks perform very poorly below this threshold, good results are achieved above it, and varying results are achieved within, according to the task at hand. »*

27. *« The use of OCR'd text has an impact on all of our tasks, though the degree varies. »*

4.2.1 Méthodes de comparaison

L'étude des impacts des erreurs d'OCR sur des tâches du TAL est fondée sur la variation en qualité de ce qui est proposé en entrée de ces tâches : le résultat de l'océrisation. La variable explicative est donc la qualité de l'OCR et la variable à expliquer les performances de la tâche de TAL étudiée. On cherche donc, toute chose égale par ailleurs, à faire varier la qualité de l'OCR et à observer les variations de performances en sortie.

Bien souvent, un corpus, disponible avec sa vérité de terrain, est océrisé et une métrique de qualité est calculée entre les résultats d'OCR et la vérité de terrain. [van Strien et al., 2020] calculent une distance de Levenshtein entre les pages du corpus océrisé et la vérité de terrain et ces pages sont réparties en quatre groupes : celles ayant une distance de Levenshtein supérieure à $0,9$, celles en ayant une supérieure à $0,8$, celles en ayant une supérieure à $0,7$ et celles en ayant une inférieure ou égale à $0,7$. Ces groupes sont appelés des *quality bands*. [Hill and Hengchen, 2019] procèdent de manière équivalente mais, au lieu de calculer la distance de Levenshtein, ce qui peut être coûteux pour de longues pages, ils répartissent les pages en huit sous-corpus selon la *F-mesure*[28] calculée au caractère : 60-65%, 65-70%, 70-75%, 75-80%, 80-85%, 85-90%, 90-95%, et 95-100%.

Une autre voie peut être de constituer un corpus numérique, autrement dit un ensemble de textes disponibles ne contenant pas d'erreur, et d'en générer automatiquement des images. [Hamdi et al., 2019, introduction] utilisent cette techniques et dégradent plus ou moins ces images avant OCR pour générer plusieurs versions du corpus, contenant plus ou moins d'erreurs d'OCR.[29] Ceci leur permet d'apprendre différents modèles de reconnaissance d'entités nommées et d'ainsi comparer leurs performances – leur corpus originel ayant été annoté en ce sens.

4.2.2 Pré-traitements des données

Même si le pré-traitement des données ne constitue pas souvent le point central des travaux de TAL, c'est l'étape inaugurale d'un processus souvent lui aussi en cascade. « L'OCR a un impact sur des tâches qui sont considérées comme "résolues", comme la segmentation en phrases. »[30] [van Strien et al., 2020, section 5] À propos de la syntaxe en dépendance, les auteurs notent que « l'analyse en dépendance est impactée plus sévèrement à mesure que la longueur augmente. Ceci suggère que nous devrions être particulièrement précautionneux lors de l'analyse de dépendance de textes de faible qualité OCR. »[31] [van Strien

28. Il s'agit d'une métrique combinant les deux mesures de précision et de rappel, généralement en tant que moyenne harmonique.

29. Aussi avons-nous ici une validation de l'importance de la numérisation.

30. *« OCR has an impact even on tasks which are considered "solved", such as sentence segmentation. »*

31. *« dependency parsing is impacted more severely as the length of the dependency grows. This suggests that we should be particularly cautious when applying dependency parsing on low quality OCR texts. »*

et al., 2020, section 5]

4.2.3 Collocations

[Hill and Hengchen, 2019, section 3.2] s'intéressent aux collocations, c'est-à-dire aux unités polylexicales qui apparaissent significativement ensemble dans un texte. Elles sont calculées selon le score de collocation *lambda*, avec une fréquence minimale de 10 et sans prendre en compte les mots vides. Ils comparent les listes de collocations obtenues à partir d'un corpus de référence (490 623 collocations) et à partir de sa version océrisée (605 569 collocations). 319 440 collocations ne sont pas partagées et 70% d'entre elles sont uniquement dans la liste issue du corpus d'OCR.

> Ainsi, le corpus OCR à la fois perd des collocations significatives statistiquement à cause de la corruption due à l'OCR, et gagne des collocations non existantes par l'introduction du bruit. [32] [Hill and Hengchen, 2019, section 3.2]

L'exemple des collocations contenant *public* et *publick* montre qu'environ la moitié des collocations (305) n'est pas partagée par les deux corpus et que des collocations manquantes dans le corpus OCR auraient intéressées les historiens (une liste est fournie en note 25).

Les corpus sont ensuite organisés selon des sous-corpus par tranche de F-mesure (entre 60 et 65%, entre 65 et 70, etc.) ; « les collocations OCR viennent le plus souvent des très mauvaises OCR » [33] Ceci atteste donc par l'expérience que les erreurs d'OCR sont itératives. Néanmoins, ce n'est parce qu'au-delà de 75% de F-mesure le nombre de collocations converge entre les deux corpus que le bruit n'est plus.

> le bruit reste un problème – en fait, la majorité des collocations du corpus OCR sont du bruit jusqu'à la tranche 65% de F-mesure, et même si le nombre total de collocations commencent à converger à partir de la tranche 75%, ce n'est pas avant la tranche 90% que les taux d'erreurs des collocations réelles arrivent à 10%. [34] [Hill and Hengchen, 2019, section 3.2]

L'étude des collocations semble donc périlleuse en contexte OCR, et une qualité dans les résultats des collocations semble nécessiter une qualité d'OCR au moins supérieure à 90% de F-mesure.

32. *« Thus, the OCR corpus both lost statistically significant collocations through OCR corruption, and gained non-existing collocations through the introduction of noise. »*
33. *« Thus, the OCR collocations were likely coming from the very bad OCR. »*
34. *« Nonetheless, noise remains a concern – in fact, a majority of collocations in the OCR corpora are noise until around the 65% F1 range, and although the total number of collocations begin to approximate each other from around the 75% mark, it is not until around the 90% mark that error rates in actual collocation matches are at the 10% level. »*

4.2.4 Reconnaissance d'entités nommées

La tâche de reconnaissance d'entités nommées consiste à repérer automatiquement dans un texte un ensemble d'unités (poly-)lexicales correspondant à des toponymes, des personnes physiques ou morales, des organisations, des dates, etc. Si la reconnaissance d'entités nommées à base de règles – utilisant des règles linguistiques et des dictionnaires, comme dans les années 1990 – ne peut supporter les erreurs d'OCR, la flexibilité de l'apprentissage automatique permet de nouvelles avancées [Hamdi et al., 2019, introduction]. Il semblerait toutefois que les erreurs d'OCR aient un impact marqué. [Hamdi et al., 2019, conclusion] concluent que

> la précision de la reconnaissance d'entités nommées chute de 90% à 60% quand les taux WER [taux d'erreur au mot] et le CER [taux d'erreur au caractère] de la sortie d'OCR augmentent de 1% à 7% et de 8% à 20% respectivement. À partir de ces taux, nous supposons que les algorithmes actuels de reconnaissance d'entités nommées état de l'art ne peuvent être invoqués que lorsque la qualité de l'OCR est suffisamment bonne. [35]

C'est aussi le résultat de [van Strien et al., 2020] qui précisent le phénomène en expliquant que l'impact n'est pas le même pour toutes les entités nommées, l'impact étant plus important pour les entités géopolitiques, les dates et les personnes. « Nous suggérons que cet impact inégal sur différents types d'entités soit pris en compte lors de l'utilisation de NER [36] sur du texte OCR. » [37] [van Strien et al., 2020, section 5]

L'étude de [Hamdi et al., 2019], à partir d'un corpus d'images générées automatiquement, montre que

> les résultats de la reconnaissance d'entités nommées sont diminués de $5,7$ points avec un CER de seulement $1,7\%$. Cela prouve que même avec un stockage et une numérisation parfaite [les images générées l'ont été à partir d'un corpus ne contenant pas d'erreur], la précision de la reconnaissance d'entité nommées peut être affectées par la qualité de l'OCR. [38] [Hamdi et al., 2019, section 2]

Toutefois, ils nuancent en rappelant que même avec des données océrisées, il est possible de trouver certains éléments :

> Les expériences ont montré que même si les entités nommées sont mal extraites par l'OCR, les systèmes de reconnaissance d'entités

35. *« NER accuracy drops from 90% to 60% when the WER and CER rates in the OCR outputs increase from 1% to 7% and from 8% to 20% respectively. From these rates, we assume that current state-of-the-art NER algorithms can only be relied upon when the OCR quality is sufficiently good. »*

36. Sigle désignant la tâche de reconnaissance d'entités nommées, *Named Entity Recognition* en anglais.

37. *« We suggest that this uneven impact on different entity types should be considered when using NER on OCR'd text. »*

38. *« NER results is lowered by 5.7 points with CER of only 1.7%. This proves even with perfect storage and digitization, NER accuracy may be affected by the OCR quality. »*

> nommées peuvent surpasser les erreurs d'OCR et reconnaître correctement une partie des entités nommées particulièrement quand les taux d'erreur ne sont pas trop élevés. [39] [Hamdi et al., 2019, section 2]
>
> L'analyse des erreurs montre aussi que 77% des entités nommées qui sont mal reconnues contiennent au moins une insertion de caractère, que 73% d'entre elles contiennent au moins une substitution et que 68% d'entre elles contiennent au moins une délétion. [40] [Hamdi et al., 2019, section 2]

4.2.5 Recherche d'information

La recherche d'information est la tâche qui consiste à trouver les documents d'une collection les plus pertinents face à une requête.

> En recherche d'information, la diminution de la qualité de l'OCR conduit à une divergence dans le classement des articles récupérés par rapport au texte corrigé par l'homme avec un nombre de résultats positifs en baisse et un nombre croissant de faux positifs. Nous constatons un impact moindre de l'amélioration de la qualité OCR sur la récupérabilité. [41] [van Strien et al., 2020, section 5]

Les erreurs d'OCR semblent donc avoir un effet important en recherche d'information, favorisant la perte de documents pertinents au profit de faux positifs.

Toutefois, ce phénomène doit être nuancé par la taille des documents du corpus étudié. Effectivement, [Traub et al., 2018, section 2], citant les travaux de [Mittendorf and Schäuble, 2000], rappellent que « nous pouvons attendre que l'impact des erreurs d'OCR sur de longs documents soit faible » [42]. Ceci est par ailleurs confirmé par les résultats de [Taghva et al., 1996], qui travaillent avec des modèles vectoriels, car ni la précision ni le rappel des documents ne sont impactés par les erreurs d'OCR – sur une collection de 674 documents, tous d'une longueur supérieure à 40 pages.

[van Strien et al., 2020, section 5] mettent néanmoins en garde à propos de la confiance que l'on peut accorder aux résultats obtenus sur des textes océrisés.

> Nos résultats concordent avec les recherches précédentes sur la récupérabilité et l'OCR et suggèrent la prudence dans la « confiance » des résultats de la récupérabilité sur du texte OCR. Ceci est particulièrement important lorsque les résultats de recherche sont directement

39. *« Experiments showed that even if named entities are wrongly extracted by the OCR, NER systems can overcome OCR errors and recognize part of the NEs correctly especially when error rates are not too high (cf. Figure 1). »*

40. *« The error analysis also shows that 77% of the named entities that are wrongly recognized contain at least one character insertion, that 73% of them contain at least one substitution and that 68% of them contain at least one deletion. »*

41. *« In information retrieval, decreasing OCR quality leads to a divergence in the ranking of retrieved articles compared to the human-corrected text with the number of hits declining and an increasing number of false positives. We find a smaller impact of improved OCR quality on retrievability. »*

42. *« [...] the effect of OCR errors on long documents can be expected to be very low. »*

> utilisés pour former des arguments, par exemple, en comptant les résultats de recherche pour un terme au fil du temps dans un corpus OCR, car la variation peut être dû à la la qualité OCR plutôt qu'à un changement de l'utilisation de ce terme. Cette mise en garde est particulièrement importante lorsque la qualité de l'OCR est inconnue. [43]

D'où l'importance de constituer une méthode non supervisée (c'est-à-dire n'exigeant aucune vérité de terrain) pour connaître, au moins, la qualité relative d'un texte océrisé.

4.2.6 Modélisation thématique

L'étude de [Mutuvi et al., 2018] montre que les erreurs d'OCR ont un impact réel sur la modélisation thématique, tâche consistant à détecter la présence d'un thème dans un texte et à lui associer un ensemble d'unités lexicales descriptives. Ils utilisent un corpus océrisé et transcrit manuellement de monographies et de périodiques anglais issus de la BNF et de la *British Library*. Ce corpus est disponible en trois versions : la vérité de terrain transcrite et de référence, la version océrisée et la version océrisée et alignée à la vérité de terrain. Ainsi, ils disposent de trois versions plus ou moins bruitées du corpus.

Pour la tâche de modélisation thématique, ils cherchent à évaluer i) d'une part la stabilité des modèles thématiques avec les métriques *Average Term Stability* (ATS) et *Pairwise Normalized Mutual Information* (PNMI) et ii) d'autre part la qualité des thèmes extraits en calculant la moyenne des distances cosinus des vecteurs *word2vec* des mots clefs de ces thèmes.

Deux méthodes sont utilisées pour apprendre les modèles thématiques : la *Latent Dirichlet Allocation* (LDA) et la *Non-negative Matrix Factorization* (NMF). Pour la stabilité, les résultats passent de $0,265$ à $0,252$ (LDA) et de $0,414$ à $0,383$ (NMF) ; pour la cohérence, ils passent de $0,362$ à $0,353$ (LDA) et de $0,475$ à $0,472$ (NMF).

> Il est évident par cette étude que les erreurs d'OCR peuvent avoir un impact négatif sur la modélisation thématique, ce qui affecte donc la qualité des thèmes découverts dans ces corpus de textes. Surtout, cela peut entraver l'exploration et l'exploitation de documents historiques précieux qui nécessitent l'utilisation de techniques OCR pour permettre leur digitalisation. [44] [Mutuvi et al., 2018, conclusion]

43. *« Our results accord with previous research on retrieval and OCR and suggest caution in "trusting" retrieval results on OCR'd text. This is particularly important when search results are directly used to make arguments, for example, by counting search results for a term over time in a OCR'd corpus, since the variation may be a proxy for OCR quality rather than a change in underlying usage of that term. This caution is particularly important when OCR quality is unknown. »*

44. *« It is evident from this study that OCR errors can have a negative impact on topic modeling, therefore affecting the quality of the topics discovered from text datasets. Overall, this can impede the exploration and exploitation of valuable historical documents which require use of OCR techniques to enable their digitization. »*

L'impact négatif des erreurs d'OCR peut néanmoins être difficilement perceptible s'il l'on ne dispose pas d'une vérité de terrain.

> Nous constatons que la détérioration de la qualité de l'OCR conduit à un impact croissant sur les modèles thématiques par rapport à ceux formés sur le texte corrigé par l'homme. Il convient de noter la manière subtile dont les modèles thématiques sont affectés par la qualité de l'OCR : les sujets ne perdent pas de sens, mais s'écartent de plus en plus de ceux formés sur du texte corrigé par l'homme. Cela signifie que cet effet ne sera pas facilement « repéré » lors de la formation des modèles thématiques sur du texte OCR de mauvaise qualité, d'autant plus que les évaluations intrinsèques ne capturent pas cet effet. D'après nos résultats, nous recommandons une préférence pour l'OCR de haute qualité, idéalement au-dessus de 90% et au moins au-dessus de 80%. [45] [van Strien et al., 2020, section 5]

Bien qu'il faille garder certaines réserves comme [van Strien et al., 2020, section 5], d'autres auteurs soulignent l'excellence de l'interprétabilité de ces modèles en présence d'OCR.

> Plus encore cependant, ces ajouts et soustractions n'ont pas d'impact sur l'interprétabilité des thèmes par les humains, ce qui rend les thèmes OCR à la fois d'extrêmement bonnes réplications des thèmes TCP [vérité de terrain], ainsi que qualitativement excellents. [46] [Hill and Hengchen, 2019, section 2.3]

Plus généralement enfin, [Mutuvi et al., 2018, section 2.1] concluent leur revue de littérature en soulignant que les erreurs d'OCR ont un faible impact sur les tâches supervisées du TAL, alors que les performances des tâches non supervisées comme la modélisation thématique en pâtissent plus largement.

4.2.7 Attribution d'auteurs

Pour la tâche d'attribution d'auteurs, [Hill and Hengchen, 2019, section 3.4] s'aperçoivent que pour un corpus océrisé avec une qualité de plus de 75% de F-mesure (ce qui est peu), la qualité de l'attribution d'auteur est plus ou moins la même. Pour leurs textes, « il y avait une amélioration négligeable des résultats due à des données plus propres que la tranche 75% » [47] [Hill and Hengchen, 2019, section 3.4]. Plus encore, ils constatent que

45. *« We find that worsening OCR quality leads to a growing impact on topic models when compared to those trained on the human-corrected text. Of note is the subtle way in which topic models are impacted by OCR quality : topics do not become meaningless, but instead increasingly diverge from those trained on human-corrected text. This means that this effect will not be easily "spotted" when training topic models on poor quality OCR'd text, particularly since intrinsic evaluations do not capture this effect. From our results we recommend a preference for high quality OCR ideally above 90% and at least above 80%. »*

46. *« Most importantly, however, these additions and subtractions do not impact topic interpretability by humans, making the OCR topics both extremely good replications of the TCP topics, as well as qualitatively excellent. »*

47. *« In our tests, there were negligible improvements to results due to cleaner data from the 75% range [...]. »*

> ce qui peut être surprenant, c'est que, entre les tranches de précision, la taille du corpus a un impact sur les résultats ce qui pourrait indiquer que, dans certains cas, un plus gros corpus peut être tout aussi, voire plus, avantageux qu'un texte plus propre. [48] [Hill and Hengchen, 2019, section 3.4]

[Franzini et al., 2018, section 4.2.3] montrent quant à eux, dans le cadre d'une classification de lettres entre Jacob et Wilhem Grimm sur une période de 70 ans, que « même si une transcription automatique [par HTR] augmente significativement le risque de mauvaise classification de texte comparé à l'OCR, une propreté d'environ 20 % est suffisante pour avoir une probabilité supérieure à la chance » [49] [Franzini et al., 2018, conclusion]. On a donc, selon cette étude, un système présentant un caractère « proportionnel » : « plus les textes sont propres, plus il est probable que l'attribution soit correcte. » [50] [Franzini et al., 2018, section 4.2.3]. On a donc intérêt, selon cette analyse, à nettoyer autant que possible les données avant de procéder à leur analyse.

En somme, pour la tâche d'attribution d'auteurs, les performances des systèmes semblent moins affectées que pour les tâches précédentes. Toutefois, il n'est pas clair entre les deux études précitées qu'il s'agisse de la qualité des données océrisées ou de longueur des textes qui est responsable de cette « robustesse ».

4.2.8 Apprentissage de modèles de langue vectoriels à partir de données océrisées

[van Strien et al., 2020, section 5] montrent que l'apprentissage de vecteurs avec *Word2Vec* est largement affecté lorsque les données océrisées sont utilisées en *fine-tuning*[51], alors que si de grands corpus de textes océrisés sont utilisés pour procéder à un apprentissage « direct », les résultats sont prometteurs.

> nos résultats suggèrent que les vecteurs de mots prédits par les modèles de langue *Word2Vec* peuvent être significativement affectés par les erreurs OCR lorsque les textes OCR sont utilisés en *fine-tuning*. Les modèles de langue directement entraînés sur de grands corpus OCR peuvent toujours produire des vecteurs de mots robustes bien que nous n'ayons pas entièrement testé cette hypothèse. L'impact de l'OCR sur les LM [modèles de langue] est un domaine avec des voies prometteuses pour une enquête plus approfondie. [52] [van Strien

48. *« However, what may be surprising is that, between accuracy ranges, corpus size has an impact on results which may indicate that, in some cases, a larger corpus may be as, or more, valuable than cleaner text. »*

49. *« even though automated transcription significantly increases the risk of text misclassification in comparison to OCR, a cleanliness of above 20% is already enough for it to have a higher-than-chance probability »*

50. *« the cleaner their texts are, the more probable correct attribution is. »*

51. C'est-à-dire que les données océrisées sont utilisées en second temps pour ajuster l'apprentissage du modèle OCR.

52. *« our results suggest that the word vectors predicted by Word2Vec LMs can be significantly affected by OCR errors when OCR'd texts are used for fine-tuning. LMs directly trained on large OCR'd corpora may still yield robust word vectors though we have not fully test*

et al., 2020, section 5]

Plus particulièrement, [Hill and Hengchen, 2019, section 3.3], en s'intéressant aux unités lexicales *king, west, princess, passion, public* et *religion*, regardent les 25 mots les plus similaires dans plusieurs sous-corpus à F1 variable.

> Même si des résultats pertinents peuvent apparaître immédiatement dans certains cas (*roi*), généralement ce n'est pas le cas. Les résultats vraiment sensés ont généralement été trouvés au-delà de la tranche 80% [de F-mesure] – à l'exception très importante des mots contenant un s long ou des ligatures. [53] [Hill and Hengchen, 2019, section 3.3]

Il semblerait donc que les erreurs d'OCR peuvent avoir un impact limité si les vecteurs sont appris sur de larges corpus. Cependant, si les vecteurs ont été appris sur des données ne contenant pas d'erreurs d'OCR et que des textes océrisés viennent en *fine-tuning*, les résultats semblent chuter. Lorsqu'il s'agit d'observer les vecteurs selon leur similarité, des résultats sensés exigent une F-mesure d'au moins 80%.

4.2.9 Conclusion

Le *bruit* apparaît comme rupture à l'harmonie, mais pas comme quelque chose *en plus* comme l'est la distorsion shanonnienne. Lorsqu'il y a introduction de bruit, on ne peut plus retrouver le message intinial.

En outre, le bruit offre des indications sur le processus d'OCR lui-même. Par exemple, de mauvaises numérisations se traduisent, ou plutôt engendrent, une OCR de piètre qualité.

Aussi est-il impossible de donner *a priori* un seuil minimal au-dessus duquel des tâches de TAL (par exemple) se dérouleraient sans accros. On observe une dépendance à la tâche et aux données.

Enfin, il manque un moyen de mesurer le bruit sans vérité de terrain, au moins pour être en mesure de comparer différents modèles d'OCR.

---

this assumption. The impact of OCR on LMs is an area with promising paths for further investigation. »

53. *« Although relevant results can be found in some cases immediately (king), these are generally outliers. Truly meaningful results were generally found after the 80% mark– with the very important exception of words containing the long-s and ligatures. In these cases, results remain very poor throughout. »*

Deuxième partie

Estimation automatique du taux d'erreur dans les données textuelles océrisées

Chapitre 5

Mesurer le « bruit » : motivation, ambition, réalisation

Sommaire

L'étude des conséquences des erreurs d'OCR contenues dans un corpus, quels que soient les phénomènes concrets qu'on veuille bien y associer, nécessite en premier lieu la définition d'une quantité mesurant ce « bruit ». Sans une telle mesure, on ne peut décrire la quantité d'erreurs présente dans un corpus : il est donc impossible d'en étudier les effets sur la création de corpus, les chaînes de traitement du TAL ou encore la linguistique de corpus. La question de la définition d'une quantité mesurant les erreurs présentes dans un corpus est donc centrale. D'une part, si les logiciels d'OCR offrent des transcriptions automatiques de qualité pour des documents contemporains générés électroniquement, ils sont nettement moins robustes face à des documents historiques (voir Partie I). D'autre part, comme le soulèvent [Springmann et al., 2014], la qualité d'un modèle d'OCR n'est pas stable d'un corpus à l'autre car ceux-ci sont particulièrement hétérogènes. Mesurer la qualité d'une sortie d'OCR nécessite alors, au moins pour un ensemble réduit de la collection à numériser et à océriser, une transcription manuelle et certaine à laquelle comparer les sorties d'OCR ; et ce, dès lors qu'une nouvelle collection est à océriser. Or cette vérité de terrain est coûteuse à constituer, ce qui parfois rend impossible toute évaluation. Par exemple, la transcription du *Premier courrier françois, traduit fidèlement en vers burlesques* (Moreau : 2 848) a pris 1 heure et 50 minutes et sa révision par un second transcripteur 30 minutes – ce document contient 15 pages constituées de 495 lignes.

Telle mesure est classiquement décrite par le *CER* (pour *Character Error Rate*, ou « taux d'erreur au caractère » en français). La différence entre une chaîne de caractères résultat d'OCR et sa version corrigée – autrement dit, de référence – est calculée en réalisant la somme des insersions, substitutions et délétions nécessaires pour passer de la première à la seconde ; la somme est ensuite divisée par le nombre total de caractères de la chaîne de référence. Cette transcription, en plus d'être onéreuse et fastidieuse, n'est pas toujours possible à mettre en œuvre ; sans oublier, comme dans [Taghva et al., 1994, section 2.1], que « dès qu'il y a intervention humaine, même pour une correction, certaines erreurs et inconsistances doivent être attendues. » [1] Si le CER mesure précisemment toutes les opérations nécessaires pour passer du résultat de l'OCR à sa version « corrigée et parfaite », il reste que parfois on doive se contenter d'autres mesures, tant le temps de la transcription est long.

1. *whenever human intervention is introduced, even for correction, some errors and inconsistencie should be expected.*

La question est donc : « Comment pouvons-nous estimer la qualité d'une OCR en l'absence de toute vérité de terrain ? » [2] [Springmann et al., 2016, section 4] L'enjeu est effectivement double tant cette estimation est importante « aussi bien pour le contrôle de la qualité des résultats d'OCR d'un unique modèle que pour la sélection du modèle le plus approprié à un imprimé donné » [3] [Springmann et al., 2016, section 4]. En clair, on cherche à réaliser un genre de « diagnostic » [Springmann et al., 2016, section 4], aux niveau macro (celui du corpus), meso (celui de l'œuvre) et micro (celui de la ligne). Car, « [a]voir une manière automatique de mesurer la qualité d'une OCR est une partie critique de tout projet de grande envergure de conversion de documents. » [4] [Taghva et al., 2004, section 3.3]

5.1 Évaluation non supervisée d'OCR

Procéder à une évaluation de sortie d'OCR sans vérité de terrain revient à mettre en place une *évaluation non supervisée* – au sens que cette évaluation n'est pas supervisée par la référence que constitue cette vérité de terrain acquise manuellement. L'enjeu d'une telle évaluation est évident : il devient possible d'évaluer les données textuelles extraites de vastes collections numériques. Toutefois, sans vérité de terrain, nombre d'autres indices peuvent être envisagés, rendant parfois trouble l'interprétation de ces mesures. On rappelle dans [Tanguy, 2020] que plusieurs voies d'« évaluation non supervisées » ont déjà été explorées, avec un succès plus ou moins franc.

5.1.1 Les valeurs de confiance des logiciels d'OCR

Une voie explorée par [Springmann et al., 2016, Hill and Hengchen, 2019] est de mettre à profit les valeurs de confiance des logiciels d'OCR. Ceux-ci renvoient en effet une valeur correspondant à l'intensité de la confiance que le logiciel associe au caractère qu'il propose. Dans le cas d'une hésitation entre deux caractères proches (par exemple, G et O), le conflit est traduit par deux valeurs de confiance similaires et plus faibles que dans le cas d'une certitude pour un caractère en particulier. [Springmann et al., 2016] supposent que « la somme des valeurs de confiance associées aux caractères de sortie doit ainsi être corrélée avec la précision de la sortie d'OCR » [5] et le vérifient très nettement. Une interprétation peut être la suivante : « On peut donc dire qu'une moyenne des confiances au caractère de 98 % pour une sortie d'OCR mène à un intervalle de

2. *How can we estimate OCR quality in the absence of any ground truth data ?*

3. *both for quality control of the OCR result from a single model and for the selection of the most appropriate model with respect to a given printed document.*

4. « Having an automated way of measuring OCR quality is a critical part of any large document conversion project. »

5. *« The sum of the confidences over all output characters should therefore correlate with the accuracy of the output. »*

précision de (95,26 %, 95,69 %) avec une probabilité de 95 %.» [6] Cette moyenne des taux de confiance est aussi utilisée pour repérer les pages et les lignes faisant le plus défaut. Par ailleurs, [Hill and Hengchen, 2019, section 4] montrent aussi qu'« [...] il y a une relation claire entre ces estimateurs [les valeurs de confiance] et la réelle F-mesure (quoique le logiciel surestime la précision). » [7] Plus encore, « [...] il est évident que l'on peut avoir confiance en eux. » [8]

Notons cependant que, comme le soulignent [Hill and Hengchen, 2019, section 2.2], « [...] il est difficile de savoir ce que cela mesure [...] », à tel point qu'ils n'utilisent pas cet indice dans leurs analyses [9].

5.1.2 Exploitation de ressources lexicales modernes

Par ailleurs, [Springmann et al., 2016] proposent d'estimer la qualité d'une sortie d'OCR en exploitant la « lexicalité » [10] de celle-ci. Cette lexicalité est calculée en faisant la moyenne, pour chaque *token* observé dans la sortie d'OCR, des distances de Levenshtein entre ces *tokens* et leur supposé équivalent moderne le plus proche (*supposé* car la relation entre deux formes de deux états différents d'une même langue n'est pas nécessairement bijective). La ressource utilisée n'est pas présentée mais il est montré que cette autre estimation est aussi corrélée à la précision du modèle d'OCR. [van Strien et al., 2020, section 3.2.2] propose une variante de la méthode précitée en calculant le ratio, par œuvre, des mots appartenant ou non à un lexique externe. Ils notent une « forte similarité » entre cette méthode et la distance de Levenshtein normalisée. [Taghva et al., 2004, section 3.3] le font en ne comptant pas les mots vides. Notons néanmoins, comme [van Strien et al., 2020, section 3.2.2], que ce genre de méthode « [...] a d'autres types de challenges, comme le choix du dictionnaire, le vocabulaire changeant, le vocabulaire spécialisé et la variation orthographique. Une requête directe à un dictionnaire proposera aussi un « score » égal entre un mot ayant un caractère erroné et un autre en contenant plusieurs. » [11]

5.1.3 La richesse lexicale

[Franzini et al., 2018, section 4.1.4] utilise le concept de « richesse lexicale » pour évaluer leur sortie d'HTR et d'OCR. Les métriques utilisées pour calculer la « richesse lexicales » sont les suivantes :

— l'entropie de Shannon, qui « [se] focalise sur la queue de la [courbe de] distribution des mots (les mots les moins fréquents comme les hapax),

6. *« We can therefore say that a mean character confidence of 98% for the OCR result leads to an accuracy intervalle of (95.26%, 95.69%) with 95% probability. »*

7. *« [...] we have shown that there is a clear relationship between these estimates and actual F1 scores (albeit, the software overestimated accuracy). »*

8. *« [...] there is evidence that one can be confident in them. »*

9. *« However, because it is difficult to know exactly what is being measured here, the scores were not used in our analysis. »*

10. *« lexicality »*

11. *« [...] has other potential challenges, including choice of dictionary, changing vocabulary, specialist vocabulary and spelling variations. »*

qui se stabilise lentement et la rend plus sensible aux petits changements à mesure que le texte progresse » [12] [Franzini et al., 2018, section 4.1.4] ;
— l'index de Simpson (ou *Inverse Participation Ratio*, IPR), qui « se focalise plus sur le cœur de la [courbe de] distribution des mots (les mots les plus fréquents), il se stabilise donc rapidement pour des textes de longueur N » [13] [Franzini et al., 2018, section 4.1.4].

Ils observent une corrélation entre l'augmentation du nombre d'erreurs d'HTR et d'OCR et ces deux métriques, donc deux versions de la richesse lexicale (avec des valeurs p de $1.04 \times 10e - 7$ et $8.03 \times 10e - 6$ pour les erreurs d'HTR pour l'IPR et l'entropie, respectivement).

5.1.4 Étude des *bounding boxes*

Avant de proposer un ensemble de caractères, les logiciels d'OCR segmentent les images proposées à l'océrisation. Ces segmentations (en colonnes, en lignes, en mots ou encore en caractères) apparaissent pour [Gupta et al., 2015] comme de bons indicateurs pour estimer la qualité d'une sortie d'OCR. En effet, s'agissant d'un processus en cascade, si la segmentation fait défaut, l'océrisation en pâtira largement. En recueillant les informations graphiques associées aux objets résultant de la segmentation (les *bounding boxes*), les auteurs proposent d'apprendre un modèle de classification permettant de distinguer deux types de *bounding boxes* : les *bounding boxes* pertinentes (*BBs*) et les *bounding boxes* non pertinentes (*noise BBs*) [14]. Le calcul de cet estimateur se réalise ensuite en comptant la proportion des *noise BBs*. Il est conclu que i) « [...] la proportion des *bounding boxes* non pertinentes fournies par le logiciel d'OCR tend à être représentative de la qualité du document » [15] et que ii) « [...] à mesure que la proportion de *boundind boxes* non pertinentes dans un document augmente, les différences entre l'OCR et les transcriptions manuelles augmentent également. » [16]

5.1.5 Les modèles de langue

Les modèles de langue, appris au grain mot, sont fréquemment utilisés en reconnaissance de la parole. [Chen et al., 1998] ont proposé d'utiliser les modèles de langue non pas pour corriger en post-traitement les sorties de reconnaissance d'un flux de parole mais pour estimer la qualité de cette sortie. La perplexité et ses dérivés (comme la log-perplexité) y apparaissent fortement corrélées au WER

12. *« [...] focuses on the tails of the word distribution (least frequent words like hapax legomena), which stabilize slowly and make it more sensitive to small changes as text progresses. »*
13. *« [...] it focuses more on the core of the word frequency distribution (the most frequent words), and so it stabilizes quickly with text length N[...] »*
14. Par exemple, une *bounding boxe* pertinente encadre une ligne ou un mot alors qu'une *bounding boxe* non pertinente encadre deux lignes juxtaposées mais appartenant à deux colonnes différentes.
15. *« [...] the proportion of noise BBs retourned by the OCR engine tends to be representative of the document's quality. »*
16. *« [...] as the proportion of noise BBs in a document increases, so do differences between OCR and manuel transcriptions also increase. »*

(*word error rate*, ou taux d'erreur au mot) avec, pour le premier jeu de données de leur étude, une relation presque parfaitement linéaire. Du côté de l'évaluation d'OCR, [Tanguy, 2020] entraîne des modèles de langue au grain caractère sur un corpus de référence (constitué d'imprimés du XVII^{e} siècle, le corpus de [Gabay, 2019]) puis récupère les probabilités d'apparition des caractères de la sortie d'OCR. Aucune corrélation avec les CER n'est observée tant les valeurs des perplexités des modèles de langue appris sont aberrantes.

5.1.6 Utiliser des pseudo-vérités de terrain

[Ul-Hasan et al., 2016] proposent d'utiliser la sortie d'un logiciel d'OCR (en l'occurrence, `Tesseract`) comme une pseudo-vérité de terrain sur laquelle est appris un premier modèle. Si l'objectif de ce travail n'est pas d'estimer la qualité d'une sortie d'OCR, les auteurs se soucient du manque de transcriptions à disposition et atteignent avec ces pseudo-vérités de terrain des précisions de l'ordre de 95% sur des documents imprimés du XVII^{e} siècle.

5.1.7 Évaluation extrinsèque par tâches en aval

Dans le cadre d'une étude sur l'impact des erreurs d'OCR sur les tâches « classiques » du TAL, [van Strien et al., 2020] évaluent des sorties d'OCR de qualité variable (organisées selon quatre intervalles de distances de Levenshtein entre la référence et la sortie d'OCR) à travers les performances de diverses tâches telles que les prétraitements linguistiques (segmentation en phrase, reconnaissance d'entités nommées et syntaxe en dépendance), la recherche d'information, la détection de thèmes et le *fine-tuning* de modèles de langue. Cela leur permet, à travers l'observation du comportement des différents systèmes en fonction de la qualité des données océrisées fournies, de proposer des recommandations. Ainsi, préconisent-ils, par exemple, pour la détection de thème une « [...] forte qualité d'OCR idéalement supérieure à 90% et au moins supérieure à 80% »[17], sans quoi les résultats obtenus ne seront pas interprétables. On peut donc raisonner inversement, non plus pour arriver à ces recommandations mais bien en partant de celles-ci et en observant, dans l'exemple pris ici de la détection de thèmes, l'interprétabilité des résultats obtenus. D'où la nécessaire circularité entre *mesure* et *étude* de l'impact de l'OCR.

5.1.8 Apprendre un modèle de prédiction

[Hill and Hengchen, 2019, section 4] cherchent à apprendre un modèle capable d'approcher la F-mesure d'une page océrisée en lui fournissant un ensemble d'indices relatifs à la page océrisée[18]. La première catégorie d'indices testés est la diversité lexicale ; les mesures MTLD, HDD et Maas ont été testées par les auteurs. Toutefois les auteurs constatent que, dans leur corpus, « [...] la diversité

17. *« [...] high quality OCR ideally above 90% and at least above 80% »*
18. *« [...] it may be possible to develop models which estimate OCR accuracy. »*

lexicale des mauvaises OCR semble ressembler aux textes propres [...] » [19]. En outre, ils testent aussi la distribution des lettres *s* et *ct* (ligature), mais notent que « [...] nous ne faisons que constater la cause d'une mauvaise OCR, pas une indication de celle-ci. » [20] Il conviendrait donc d'étendre la recherche sur ce point.

Les travaux précités exploitent d'autres types d'indicateurs que le CER pour juger de la qualité des sorties d'OCR, ces indicateurs nécessitant ou non des ressources linguistiques externes ou l'apprentissage de modèles connexes (modèles de classification ou modèles de langue). L'ensemble de ces indices de la qualité d'une OCR forme ce qu'on pourrait appeler un « faisceau informatif ». Nous cherchons donc à exploiter ce « faisceau » pour proposer, en fin de chapitre, un modèle (formé par apprentissage automatique) de prédiction du CER.

5.2 Reproduction et adaptation de l'expérience menée par [Springmann et al., 2016]

Selon les termes de [Cohen et al., 2018], nous cherchons dans cette section à reproduire (« reproducibility ») les conclusions, et non les découvertes ni les valeurs, de [Springmann et al., 2016], sans pour autant pouvoir répliquer (« replicability ») leur expérience – n'ayant pas accès aux mêmes données.

5.2.1 Taux de confiance et taux de lexicalité

Définissons formellement les deux mesures choisies pour procéder l'évaluation non supervisée d'OCR. La première, le taux moyen de confiance (T_{con}), correspond à la moyenne des taux de confiance fournis par les logiciels d'OCR. En plus des sorties en texte brut, ceux-ci proposent effectivement des formats enrichis en métadonnées contenant les taux de confiance que les logiciels associent aux caractères reconnus [21]. On peut donc calculer, pour une ligne ou une page océrisée, la moyenne de ces taux. Il s'agit donc d'une mesure indépendante de toute ressource externe. Avec C le nombre de caractères de la ligne ou la page océrisée, c_i le i-ème caractère et $conf(c_i)$ le taux de confiance donné par le logiciel d'OCR, on a :

$$T_{con} = \frac{1}{C} \sum_{i=0}^{i=C-1} conf(c_i) \tag{5.1}$$

La seconde mesure, le taux de lexicalité (T_{lex}), calcule le ratio des mots fournis par le logiciel d'OCR correspondant véritablement à des mots, en interrogeant un lexique. Il s'agit d'une adaptation de la mesure proposée par [Springmann

19. *« [...] lexical diversity of bad OCR seems to resemble clean texts [...] »*
20. *« [...] we may simply be witnessing the cause of bad OCR, rather than an indication of it. »*
21. Ces taux de confiance, aussi appelés précisions au caractères, sont les seuls indices dont nous disposons qui évite la transcription. [Taghva et al., 2004, section 3.3]

et al., 2016] car ceux-ci associent à chaque mot reconnu la plus petite distance de Levenshtein entre ce mot et sa forme moderne supposée. Comme nous disposons d'un lexique du français du XVII^{e} siècle, le LGeRM [Souvay and Pierrel, 2009], nous proposons d'uniquement calculer la part des mots reconnus appartenant à ce lexique.

Ajoutons qu'il existe bien la notion de *mot* dans une sortie d'OCR, quoique très réductrice car simplement entendue comme suite contiguë de caractères séparés par des blancs (le plus typique étant l'espace). Cette notion s'actualise d'ailleurs dans l'arborescence des balises des fichiers de sorties des logiciels. Par exemple, les fichiers HTML [22] donnés par le logiciel d'OCR `Kraken` correspondent à des pages composées d'une suite de lignes elles-mêmes composées d'une suite de mots eux-mêmes composés d'une suite de caractères :

```
<div class="ocr_page"...
    <span class="ocr_line"...
        <span class="ocrx_word"...
```

On peut donc, pour chaque mot (ou *ocr_word* pour `Kraken`) rassembler les caractères reconnus et interroger un lexique pour savoir si cette suite de caractères correspond effectivement à un mot. Précisons que le lexique doit être composé de formes fléchies. Reste finalement à calculer la part des suites de caractères correspondant à des mots du lexique, au niveau de la ligne ou de la page océrisée. Avec W le nombre de suites de caractères appartenant au lexique et $\not{W}$ le nombre de suites de caractères n'y appartenant pas, il suffit de calculer :

$$T_{lex} = \frac{W}{W + \not{W}} \tag{5.2}$$

Il s'agit donc d'une mesure dépendante d'une ressource linguistique, ce qui exclut d'entrée de jeu sa pertinence pour toutes les langues ou états de langue peu dotés.

Pour juger de la pertinence de ces deux mesures, une première voie consiste à étudier leur corrélation avec le CER, lequel nécessite une transcription des numérisations à évaluer. Une seconde voie possible consiste à étudier, manuellement et attentivement, le comportement de ces mesures sur un ensemble de numérisations ; dans ce cas, les transcriptions ne sont pas nécessaires.

5.2.2 Coefficient de corrélation et valeur p

Le coefficient de corrélation exprime à quel point deux variables sont liées. Ce coefficient étant une normalisation, il appartient à l'intervalle $[-1; 1]$: les valeurs positives indiquent que les deux variables évoluent dans le même sens et les valeurs négatives qu'elles évoluent dans un sens opposé. Plus une corrélation est proche des bornes 1 ou -1, plus le lien entre les deux variables est fort ; au contraire, plus la corrélation est proche de 0, plus ce lien se dissipe. Valider

22. Hypertext Markup Language.

la pertinence des taux T_{con} et T_{lex} revient donc à observer un coefficient de corrélation proche des bornes -1 ou 1. Précisons que le coefficient de corrélation calculé dans cette section est le coefficient (linéaire) de Pearson.

La relation de corrélation n'est toutefois significative que si la valeur p (*p-value* en anglais) est inférieure à un seuil, ceci traduisant que cette relation a peu de chances d'être due au hasard ; il s'agit d'un test de corrélation. Plusieurs seuils peuvent être admis, nuançant avec plus ou de moins de vigueur la présomption contre l'hypothèse nulle. Classiquement, on a :

— $p \leq 0,01$: très forte présomption contre l'hypothèse nulle ;
— $0,01 < p \leq 0,05$: forte présomption contre l'hypothèse nulle ;
— $0,05 < p \leq 0,1$: faible présomption contre l'hypothèse nulle ;
— $p > 0,1$: pas de présomption contre l'hypothèse nulle.

5.2.3 Le taux d'erreur au caractère (CER) et au mot (WER)

Le CER est calculé entre une suite de caractères de référence et une suite de caractères à tester. Son calcul réalise la somme des insertions (c_i), délétions (c_d) et substitutions (c_s) nécessaires pour passer de la chaîne de référence à la chaîne de test et la divise par le nombre total de caractères contenus dans la chaîne de référence (C).

$$CER = \frac{c_i + c_d + c_s}{C} \tag{5.3}$$

Remarquons qu'il peut être supérieur à 1 (ou à 100) si le nombre d'insertions est particulièrement élevé. Dans le cadre de cette étude, le CER a été calculé avec la librairie `asrtoolkit` (version 0.2.2) [23] en utilisant l'argument *char-level* (sans quoi c'est le WER qui est calculé).

Le WER est calculé de la même manière, mais au grain mot. On a donc, avec les insertions (w_i), délétions (w_d), substitutions (w_s) et le nombre de total de mots (W) :

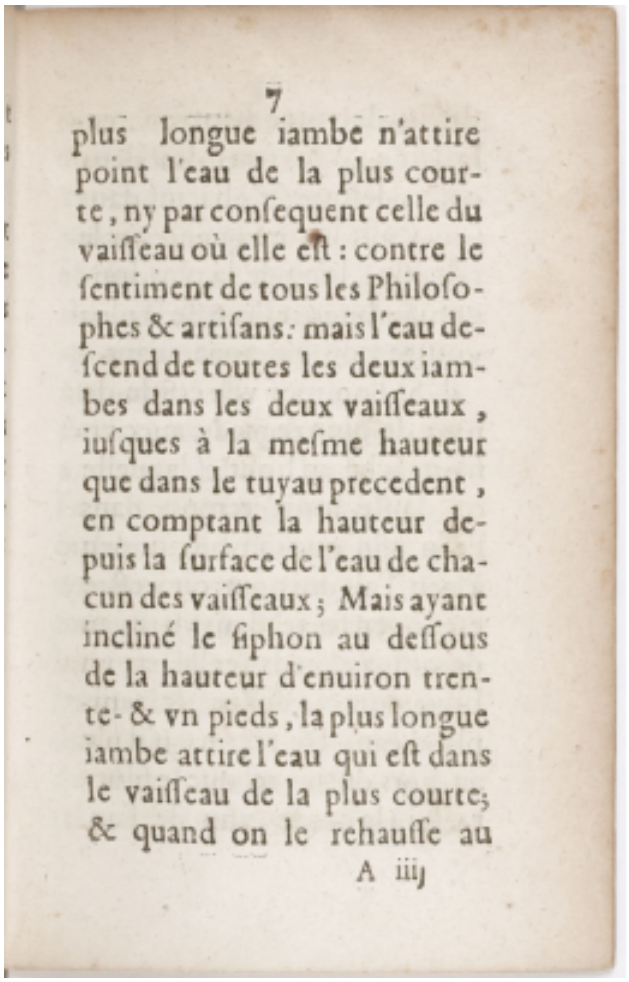
$$WER = \frac{w_i + w_d + w_s}{W} \tag{5.4}$$

5.2.4 Cadre expérimental

Un corpus d'œuvres françaises du XVII^{e} siècle

Rassemblé et transcrit par [Gabay, 2019], ce corpus est constitué d'une sélection d'œuvres françaises du XVII^{e} siècle – décrites dans le tableau 5.1 – et compte 228 pages numérisées [24] rendues disponibles avec leur transcription (environ 6 000 lignes, 30 000 mots et 150 000 caractères). Un exemple de numérisation et de transcription est proposé en figure 5.1.

23. https://pypi.org/project/asrtoolkit/0.2.2/ (consulté le 28 juin 2022).
24. Au format *jpg*.



7

plus longue iambe n'attire point l'eau de la plus courte, ny par conſequent celle du vaiſſeau où elle eſt : contre le ſentiment de tous les Philoſophes & artiſans: mais l'eau deſcend de toutes les deux iambes dans les deux vaiſſeaux, iuſques à la meſme hauteur que dans le tuyau precedent, en comptant la hauteur depuis la ſurface de l'eau de chacun des vaiſſeaux; Mais ayant incliné le ſiphon au deſſous de la hauteur d'enuiron trente- & vn pieds, la plus longue iambe attire l'eau qui eſt dans le vaiſſeau de la plus courte; & quand on le rehauſſe au

A iiij

7

plus longue iambe n'attire point l'eau de la plus cour¬ te, ny par conſequent celle du vaiſſeau où elle eſt: contre le ſentiment de tous les Philoſo¬ phes & artiſans: mais l'eau de¬ ſcend de toutes les deux iam¬ bes dans les deux vaiſſeaux, iuſques à la meſme hauteur que dans le tuyau precedent, en comptant la hauteur de¬ puis la ſurface de l'eau de cha¬ cun des vaiſſeaux; Mais ayant incliné le ſiphon au deſſous de la hauteur d'enuiron tren¬ te-& vn pieds, la plus longue iambe attire l'eau qui eſt dans le vaiſſeau de la plus courte; & quand on le rehauſſe au

A iiij

FIGURE 5.1 – Numérisation de la page 15 des *Experiences Nouvelles touchant le vide…* de Pascal (1647) présentée avec sa transcription diplomatique.

Titre	Auteur	Date	Domaine	Nb pages	Nb lignes	Nb mots
Oraisons funebres	Bossuet	1683	Théologie	27	770	4 128
La Pucelle…	Chapelain	1656	Poésie	28	753	4 735
Advis sur la peste	Ellain	1606	Science	22	618	3 168
Egalite des hommes et des femmes	Gournay	1622	Philosophie	31	825	4 284
La Maniere d'amolir les os…	Papin	1682	Science	23	548	2 230
Experiences Nouvelles…	Pascal	1647	Science	39	776	3 568
Introduction à la vie devote	Sales	1641	Théologie	25	618	3 915
Oeuvres completes (Tome II.)	Viau	1623	Poésie	33	852	4 055

TABLE 5.1 – Description des œuvres du corpus de [Gabay, 2019].

Océrisation des images

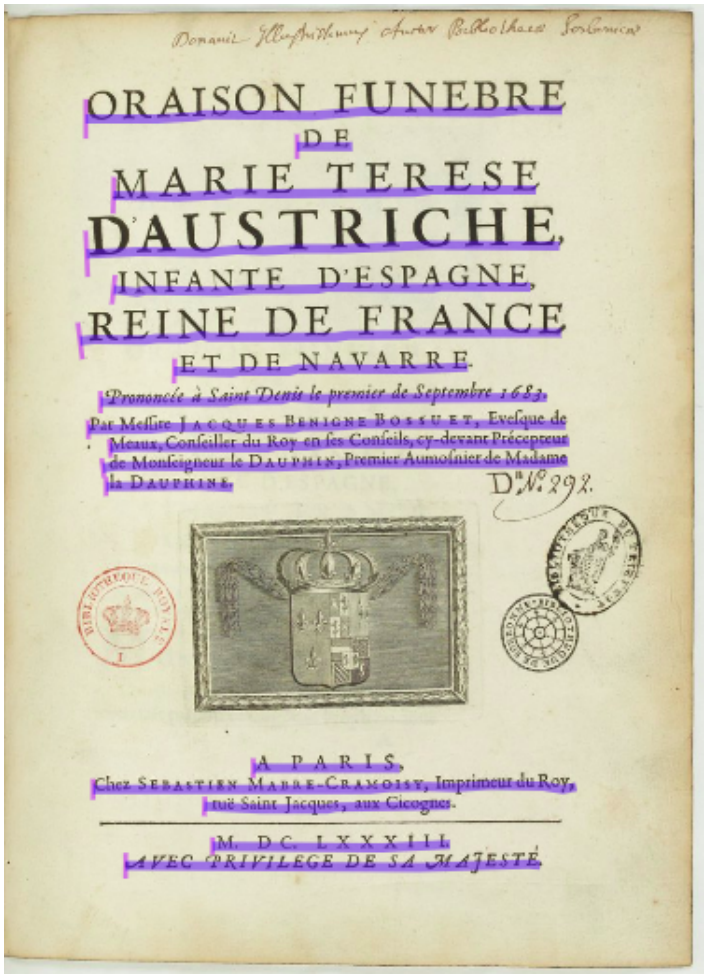
L'ensemble de ces numérisations a été océrisé en utilisant le logiciel `Kraken` (version 2.0.8) [25] (voir [Kiessling, 2019]).

Pour appliquer ses modèles de reconnaissance de caractères, `Kraken` prend en entrée des images binarisées ; un pixel ne peut être que blanc ou noir. Les numérisations ont donc été binarisées par le module *binarize* dédié. Ces images binarisées sont ensuite segmentées (entendons, découpées en colonnes, paragraphes, lignes, mots et caractères) en utilisant le module *segmente* de `Kraken`. Il reste finalement à appliquer le modèle d'OCR en utilisant le module *ocr*. Le modèle choisi [26] a été appris sur les mêmes données que le corpus décrit dans le tableau 5.1. Toutefois, lors de l'apprentissage, on fournit un ensemble de paires image-transcription à l'échelle de la ligne. On peut donc supposer raisonnablement que les sorties de ce modèle à qui on aura donné des pages n'atteindra pas à coup sûr 0% de CER.

La page : échelle de calcul des corrélations

Les CER et WER et les taux T_{con} et T_{lex} pourraient être calculés selon plusieurs échelles : celle de l'œuvre, celle de la page et celle de la ligne. On serait alors tenté de calculer les corrélations pour ces trois échelles. On chercherait ainsi à vérifier la stabilité des mesures T_{con} et T_{lex} face au changement d'échelle. Toutefois, l'échelle de l'œuvre et celle de la ligne ne peuvent être envisagées. Pour l'échelle de l'œuvre, la complexité du calcul du CER et du WER entre deux œuvres (une transcription et une sortie d'OCR) rend extrêmement coûteux son calcul. Ceci a pour conséquence qu'il faudrait procéder à la moyenne des CER et

25. http://kraken.re/ et https://pypi.org/project/kraken/ (consultés le 28 juin 2022).
26. Disponible ici : https://github.com/e-ditiones/OCR17 (consulté le 28 juin 2022) [Gabay, 2020].



ORAISON FUNEBRE
DE
MARIE TERESE
D'AUSTRICHE,
INFANTE D'ESPAGNE,
REINE DE FRANCE
ET DE NAVARRE.

Prononcée à Saint Denis le premier de Septembre 1683.

Par Messire JACQUES BENIGNE BOSSUET, Evesque de Meaux, Conseiller du Roy en ses Conseils, cy-devant Précepteur de Monseigneur le DAUPHIN, Premier Aumosnier de Madame la DAUPHINE.

A PARIS,
Chez SEBASTIEN MABRE-CRAMOISY, Imprimeur du Roy, ruë Saint Jacques, aux Cicognes.

M. DC. LXXXIII.
AVEC PRIVILEGE DE SA MAJESTÉ.

FIGURE 5.2 – Première page des *Oraisons funebres* de Bossuet, présentée avec les résultats de la segmentation manuelle. Cette visualisation est issue d'*e-scriptorium*.

ORAISON FUNEBRE
DE
MARIE TERESE
D'AUSTRICHE,
INFANTE D'ESPAGNE,
REINE DE FRANCE
ET DE NAVARRE.

Prononcée à Saint Denis le premier de Septembre 1683.

Par Messire JACQUES BENIGNE BOSSUET, Evesque de Meaux, Conseiller du Roy en ses Conseils, cy-devant Précepteur de Monseigneur le DAUPHIN, Premier Aumosnier de Madame la DAUPHINE.

D.M.2998.

A PARIS,
Chez SEBASTIEN MABRE-CRAMOISY, Imprimeur du Roy, ruë Saint Jacques, aux Cicognes.

M. DC. LXXXIII.
AVEC PRIVILEGE DE SA MAJESTÉ.

FIGURE 5.3 – Première page des *Oraisons funebres* de Bossuet, présentée avec les résultats de la segmentation automatique. Cette visualisation est issue d'*e-scriptorium*.

X, Y	$corr(X, Y)$	valeur p
CER, WER	0,981	4,58e-164

TABLE 5.2 – Corrélations (Pearson) et valeurs p entre le CER et le WER sur le corpus de [Gabay, 2019]

X, Y	$corr(X, Y)$	valeur p
CER, T_{con}	-0,557	5,30E-20
CER, T_{lex}	0,0342	0,607

TABLE 5.3 – Corrélations (Pearson) et valeurs p entre le CER et les taux T_{con} et T_{lex} sur le corpus de [Gabay, 2019]

WER par page ; ce qui n'est nullement plus informatif que la seule échelle de la page. Pour l'échelle de la ligne, on ne peut s'attendre à ce que la sortie d'OCR contienne le même nombre de lignes que la vérité de terrain (voir les figures 5.2 et 5.3) ; on ne peut donc pas savoir *a priori* à quelle ligne de la référence comparer une ligne de la sortie d'OCR. L'étude des corrélations du CER et du WER et des taux T_{con} et T_{lex} ne peut donc être faite qu'à l'échelle de la page.

5.2.5 Résultats

Remarquons en préambule que les taux T_{con} et T_{lex} étant calculés uniquement sur la sortie d'OCR, ils ne peuvent prendre en compte les délétions, information néanmoins contenue dans le CER et le WER.

Le tableau 5.2 montre que le CER et le WER sont très fortement corrélés, avec une valeur p très inférieure au seuil $0,01$. Calculer les corrélations avec le CER ou le WER n'a donc pas d'importance.

Avec le tableau 5.3, on observe que le T_{con} est corrélé négativement au CER, avec une valeur p bien inférieure à $0,01$. Que la corrélation soit négative signifie que les deux variables évoluent dans des sens opposés. En clair, quand le CER augmente, le T_{con} diminue ; quand le nombre d'erreurs au caractère augmente, le taux moyen de confiance au caractère diminue. La corrélation du CER et du T_{con} est plutôt [27] forte et la valeur p très inférieure à $0,01$ ce qui appuie l'affirmation de ce résultat. Il en va de même pour la corrélation $corr(WER, T_{con})$.

En outre, la représentation graphique du T_{con} en fonction du CER (échelle logarithmique de l'axe des abscisses) de la figure 5.4 permet de préciser la nature de cette corrélation, ou plutôt de la nuancer. On observe que, pour les valeurs faibles du CER (entendons inférieures à 10), le T_{con} reste autour de $0,99$, sans qu'il y ait une franche baisse de ce taux à mesure que le CER augmente. En

27. *Plutôt* car la corrélation est légèrement supérieure à $-0,5$.

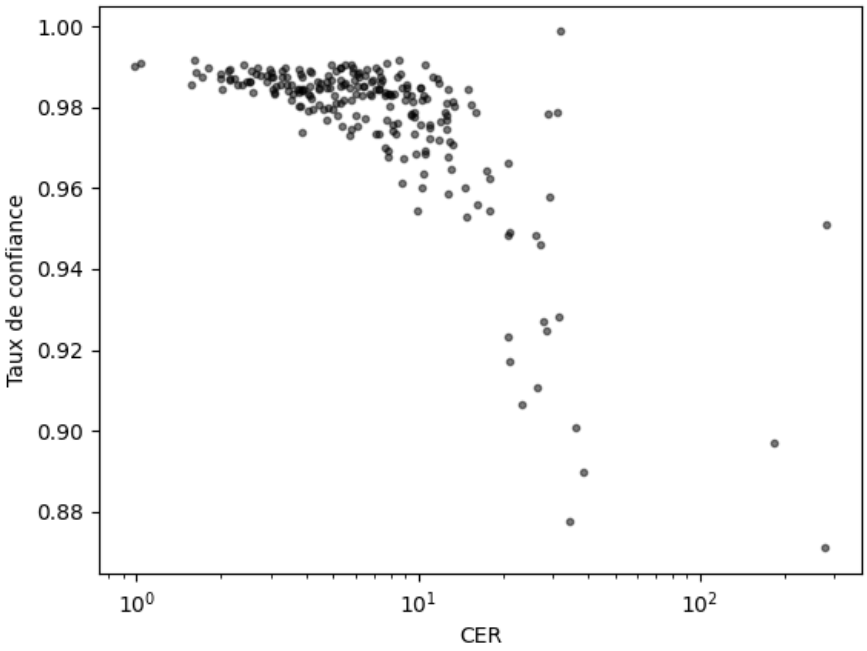


FIGURE 5.4 – Représentation graphique du CER et du T_{con}, échelle logarithmique en abscisses (corpus de [Gabay, 2019])

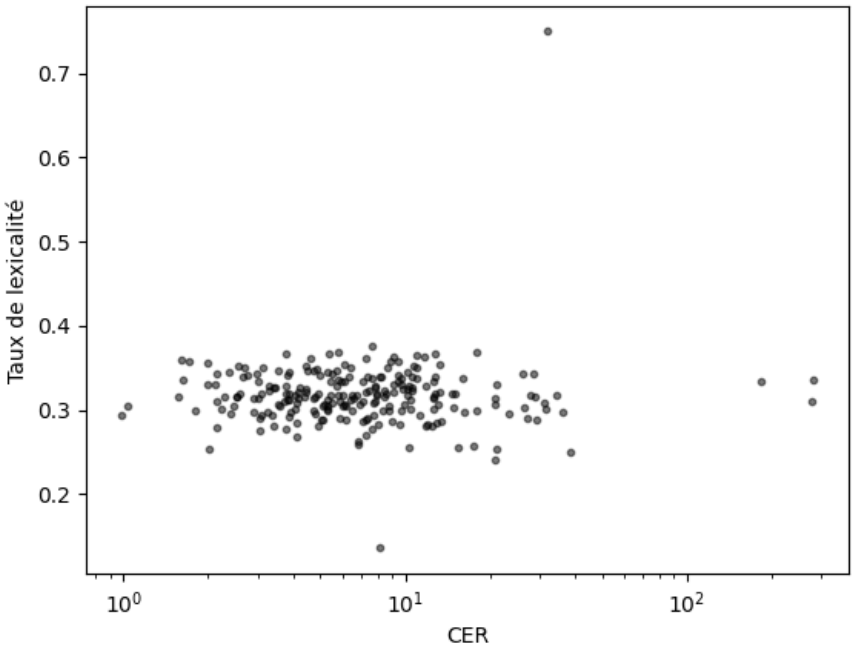


FIGURE 5.5 – Représentation graphique du CER et du T_{lex}, échelle logarithmique en abscisses (corpus de [Gabay, 2019])

revanche, pour les valeurs comprises en 10 et 50, on peut observer ce que suggère la corrélation de $-0,5$. Pour les quelques pages ayant un CER supérieur à 50, les valeurs du T_{con} sont erratiques ; la corrélation ne s'observe pas.

Enfin, pour le T_{lex}, le tableau 5.3 montre qu'il n'y aucune corrélation avec le CER ni le WER. La représentation graphique proposée en figure 5.5 le confirme : les valeurs de ce taux gravitent entre $0,25$ et $0,4$, quelque soit la valeur du CER.

L'étude des corrélations entre le CER ou le WER et les taux T_{con} et T_{lex} permet de reproduire la conclusion de [Springmann et al., 2016] pour le taux moyen de confiance T_{con} tout en réfutant la pertinence du T_{lex}. Le T_{con}, qui ne nécessite aucune ressource externe, semble donner une information fiable (même si incomplète) permettant de juger de la qualité d'une sortie d'OCR.

Notons ici que le T_{con} n'est pas homogène d'un logiciel à l'autre. Il peut être toujours très élevé pour un logiciel donné et beaucoup moins pour un autre. Il ne constitue donc pas, à lui seul, un moyen de comparer la performance de différents logiciels, mais seulement de comparer différents modèles.

5.3 Étude comportementale des T_{con} et T_{lex} sur un ensemble de documents non transcrits

Pour dépasser la seule étude des corrélations avec le CER, ces mesures peuvent être étudiées par le prisme de leur comportement. On cherche à observer les pages ou les lignes ayant un taux T_{con} et/ou T_{lex} faible. Il s'agit de voir, pour un ensemble de documents non transcrits, si ces mesures sont en pratique de bons indicateurs pour juger la qualité d'une sortie d'OCR.

5.3.1 Corpus

Contrairement à la section précédente, nous proposons ici une étude des taux T_{con} et T_{lex} sur un ensemble de numérisations non transcrites. Il s'agit de 180 mazarinades burlesques disponibles au téléchargement en ligne (libre de droit dans un usage non commercial) [28]. Les plateformes Gallica [29] et Google Livres [30] ont été utilisées, avec une priorité pour Gallica qui propose des numérisations de bien meilleure qualité, ainsi que des métadonnées plus riches et plus sûres. [Mittendorf and Schäuble, 2000] rappellent bien qu'« une numérisation imprudente est [...] responsable de la perte d'information ». Les numérisations disponibles sur Google Livres proviennent de diverses entreprises de numérisations sans que ses conditions et protocoles soient accessibles. Par exemple, sont proposées dans les figures 5.6 et 5.7 les pages 6 et 12 de *La nappe renversée, chez Renard, en vers burlesques*, téléchargé sur Google Livres [31]. La page 6 est très difficilement lisible

28. LIEN VERS LE CORPUS BURLESQUE.
29. https://gallica.bnf.fr/accueil/fr/content/accueil-fr?mode=desktop (consulté le 28 juin 2022).
30. https://books.google.fr/ (consulté le 28 juin 2022).
31. [32] (consulté le 8 octobre 2020).

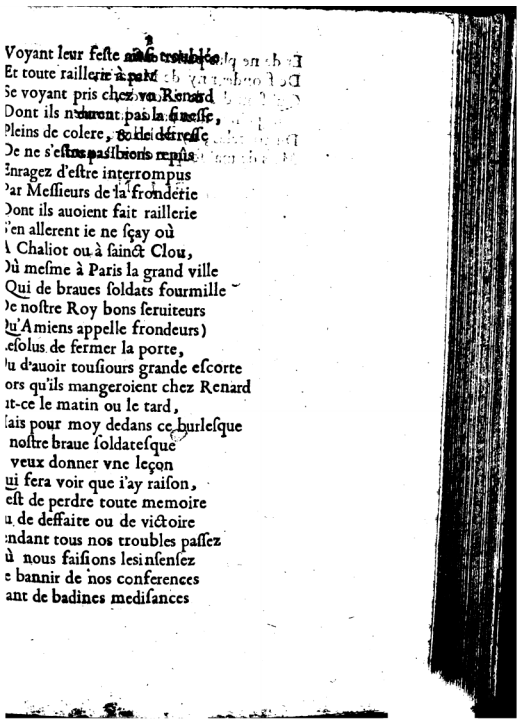


FIGURE 5.6 – Page 6 de *La nappe renversée, chez Renard, en vers burlesques*, 1649, Paris, Moreau 2525, téléchargé sur Google Livres (https://books.google.fr/books?id=fL1vR-ja5xgC&hl=fr).

FIGURE 5.7 – Page 12 de *La nappe renversée...*, téléchargé sur Google Livres.

	T_{con}		T_{lex}	
	page	ligne	page	ligne
Minimum	0	0,111	0	0
Premier quartile	0,953	0,913	0,291	0
Deuxième quartile	0,979	0,981	0,330	0,231
Troisième quartile	0,986	0,996	0,357	0,429
Maximum	0,999	1	1	1

TABLE 5.4 – Statistiques (minimum, quartiles et maximum) des taux T_{con} et T_{lex} à l'échelle de la page et de la ligne.

car on y devine par encre déteinte la page suivante ; la page 12 n'est en rien une numérisation de la mazarinade en question. Le téléchargement prioritaire des pièces sur Gallica a permis d'atteindre 78% (soit 142 pièces) de numérisations de qualité réalisées par la BNF.

Distribuées au format PDF, les numérisations ont été converties – en utilisant la librairie *pdftoppm* (version 0.62.0) – en une suite d'images PNG permettant par la suite de procéder à de la reconnaissance optique de caractères. Un tri parmi l'ensemble des images résultantes a écarté les pages générées par les plateformes (voir les images des figures 5.8, 5.9, 5.10). Ces images ont ensuite été binarisées, segmentées et océrisées en utilisant `Kraken` et son modèle pour le français du XVII^{e} siècle. On dispose finalement de 2 822 images PNG et du même nombre de sorties d'OCR au format HTML.

5.3.2 Calcul des T_{con} et T_{lex}

Les figures 5.11, 5.12, 5.13, 5.14, 5.15, 5.16, 5.17 et 5.18 montrent que, pour la majorité des pages et des lignes, le logiciel `Kraken` est confiant, avec un T_{con} presque systématiquement supérieur à $0,95$. [Springmann et al., 2016] ajoutent que « plus encore, tous les caractères ayant un taux de confiance supérieur à la moyenne (0,93) sont corrects. »[33] Les boîtes à moustache des figure 5.15, 5.16, 5.17 et 5.18 montrent, pour le T_{con}, des moyennes de $0,97$ à l'échelle de la page et de la ligne. On peut donc considérer que les pages ou les lignes ayant un T_{con} inférieur sont probablement mal océrisées et que, *a fortiori*, les pages et les lignes représentées comme des valeurs extrêmes minimales dans les boîtes à moustache sont purement du bruit.

Il en va de même pour l'analyse du T_{lex}, au moins pour ses valeurs à l'échelle de la page. Pour les valeurs calculées à l'échelle de la ligne, l'interprétation est beaucoup ardue tant les valeurs sont erratiques.

33. *more importantly, all characters with a confidence above average (0.93) are correct.*

Estrennes burlesques pour le premier jour de l'an 1650

FIGURE 5.8 – Exemple de page générée et incluse dans les PDF téléchargés sur Gallica.

FIGURE 5.9 – Exemple de page générée et incluse dans les PDF téléchargés sur Gallica.

A propos de ce livre

Ceci est une copie numérique d'un ouvrage conservé depuis des générations dans les rayonnages d'une bibliothèque avant d'être numérisé avec précaution par Google dans le cadre d'un projet visant à permettre aux internautes de découvrir l'ensemble du patrimoine littéraire mondial en ligne.

Ce livre étant relativement ancien, il n'est plus protégé par la loi sur les droits d'auteur et appartient à présent au domaine public. L'expression "appartenir au domaine public" signifie que le livre en question n'a jamais été soumis aux droits d'auteur ou que ses droits légaux sont arrivés à expiration. Les conditions requises pour qu'un livre tombe dans le domaine public peuvent varier d'un pays à l'autre. Les livres libres de droit sont autant de liens avec le passé. Ils sont les témoins de la richesse de notre histoire, de notre patrimoine culturel et de la connaissance humaine et sont trop souvent difficilement accessibles au public.

Les notes de bas de page et autres annotations en marge du texte présentes dans le volume original sont reprises dans ce fichier, comme un souvenir du long chemin parcouru par l'ouvrage depuis la maison d'édition en passant par la bibliothèque pour finalement se retrouver entre vos mains.

Consignes d'utilisation

Google est fier de travailler en partenariat avec des bibliothèques à la numérisation des ouvrages appartenant au domaine public et de les rendre ainsi accessibles à tous. Ces livres sont en effet la propriété de tous et de toutes et nous sommes tout simplement les gardiens de ce patrimoine. Il s'agit toutefois d'un projet coûteux. Par conséquent et en vue de poursuivre la diffusion de ces ressources inépuisables, nous avons pris les dispositions nécessaires afin de prévenir les éventuels abus auxquels pourraient se livrer des sites marchands tiers, notamment en instaurant des contraintes techniques relatives aux requêtes automatisées.

Nous vous demandons également de:

+ *Ne pas utiliser les fichiers à des fins commerciales* Nous avons conçu le programme Google Recherche de Livres à l'usage des particuliers. Nous vous demandons donc d'utiliser uniquement ces fichiers à des fins personnelles. Ils ne sauraient en effet être employés dans un quelconque but commercial.

+ *Ne pas procéder à des requêtes automatisées* N'envoyez aucune requête automatisée quelle qu'elle soit au système Google. Si vous effectuez des recherches concernant les logiciels de traduction, la reconnaissance optique de caractères ou tout autre domaine nécessitant de disposer d'importantes quantités de texte, n'hésitez pas à nous contacter. Nous encourageons pour la réalisation de ce type de travaux l'utilisation des ouvrages et documents appartenant au domaine public et serions heureux de vous être utile.

+ *Ne pas supprimer l'attribution* Le filigrane Google contenu dans chaque fichier est indispensable pour informer les internautes de notre projet et leur permettre d'accéder à davantage de documents par l'intermédiaire du Programme Google Recherche de Livres. Ne le supprimez en aucun cas.

+ *Rester dans la légalité* Quelle que soit l'utilisation que vous comptez faire des fichiers, n'oubliez pas qu'il est de votre responsabilité de veiller à respecter la loi. Si un ouvrage appartient au domaine public américain, n'en déduisez pas pour autant qu'il en va de même dans les autres pays. La durée légale des droits d'auteur d'un livre varie d'un pays à l'autre. Nous ne sommes donc pas en mesure de répertorier les ouvrages dont l'utilisation est autorisée et ceux dont elle ne l'est pas. Ne croyez pas que le simple fait d'afficher un livre sur Google Recherche de Livres signifie que celui-ci peut être utilisé de quelque façon que ce soit dans le monde entier. La condamnation à laquelle vous vous exposeriez en cas de violation des droits d'auteur peut être sévère.

À propos du service Google Recherche de Livres

En favorisant la recherche et l'accès à un nombre croissant de livres disponibles dans de nombreuses langues, dont le français, Google souhaite contribuer à promouvoir la diversité culturelle grâce à Google Recherche de Livres. En effet, le Programme Google Recherche de Livres permet aux internautes de découvrir le patrimoine littéraire mondial, tout en aidant les auteurs et les éditeurs à élargir leur public. Vous pouvez effectuer des recherches en ligne dans le texte intégral de cet ouvrage à l'adresse http://books.google.com

FIGURE 5.10 – Exemple de page générée et incluse dans les PDF téléchargés sur Google Book.

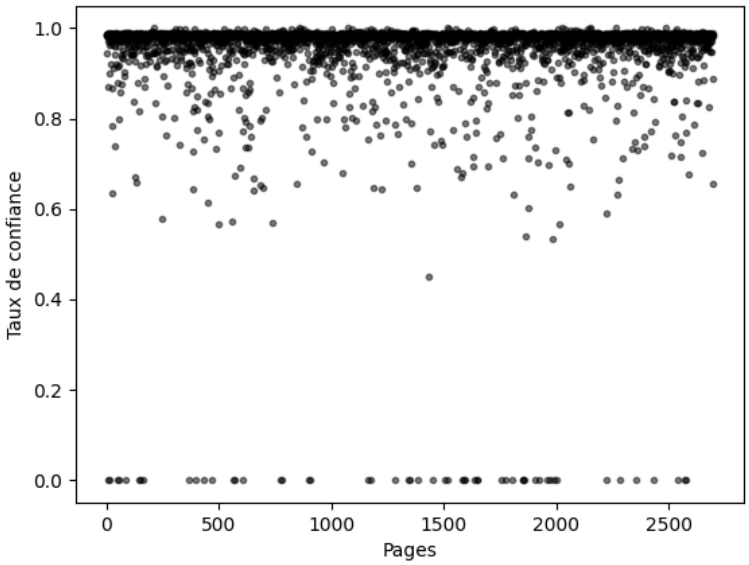


FIGURE 5.11 – T_{con} (page)

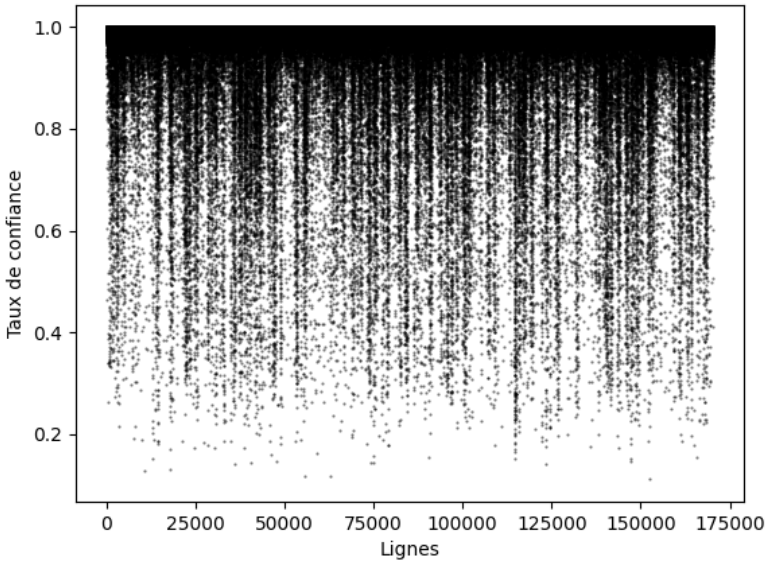


FIGURE 5.12 – T_{con} (ligne)

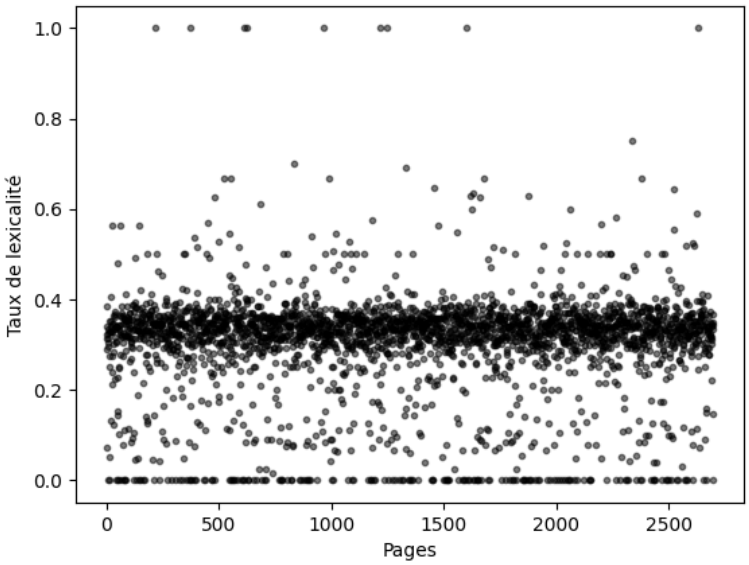


FIGURE 5.13 – T_{lex} (page)

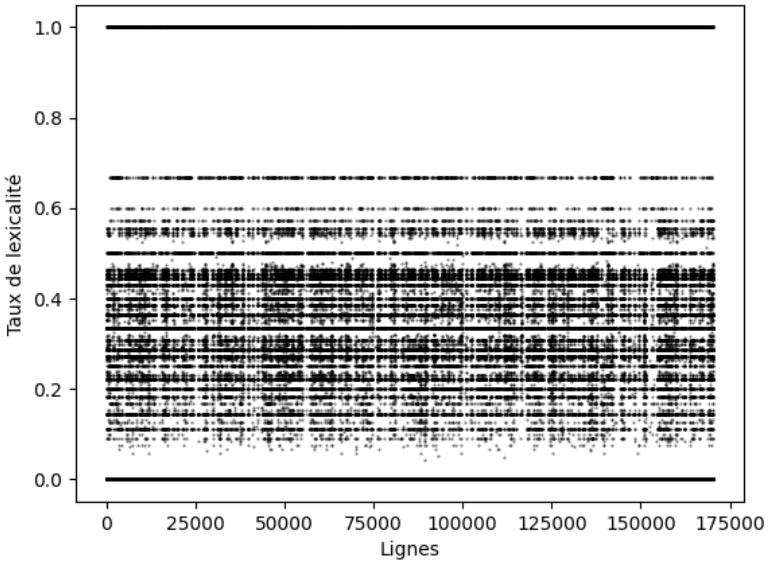


FIGURE 5.14 – T_{lex} (ligne)

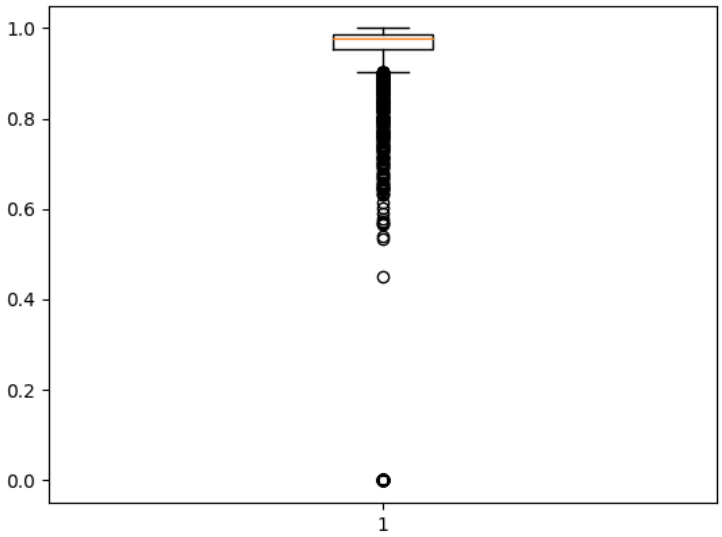


FIGURE 5.15 – T_{con} (page)

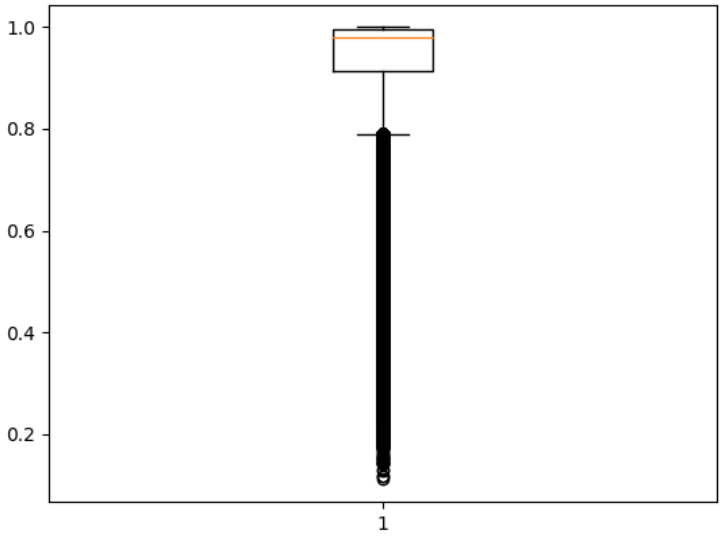


FIGURE 5.16 – T_{con} (ligne)

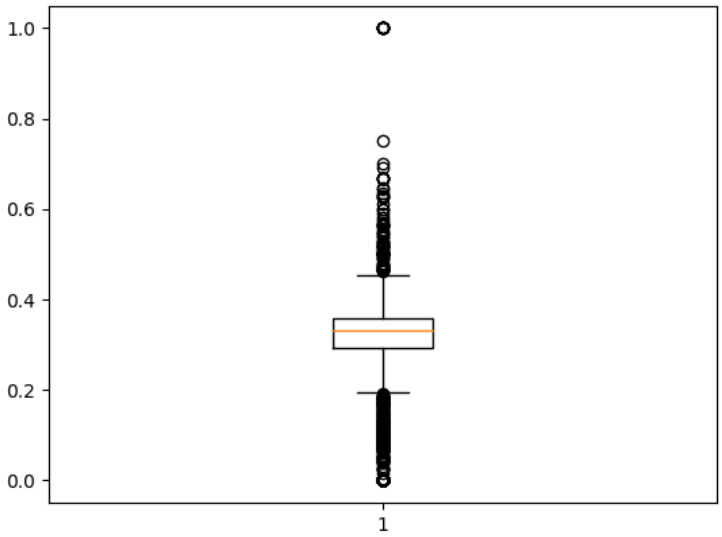


FIGURE 5.17 – T_{lex} (page)

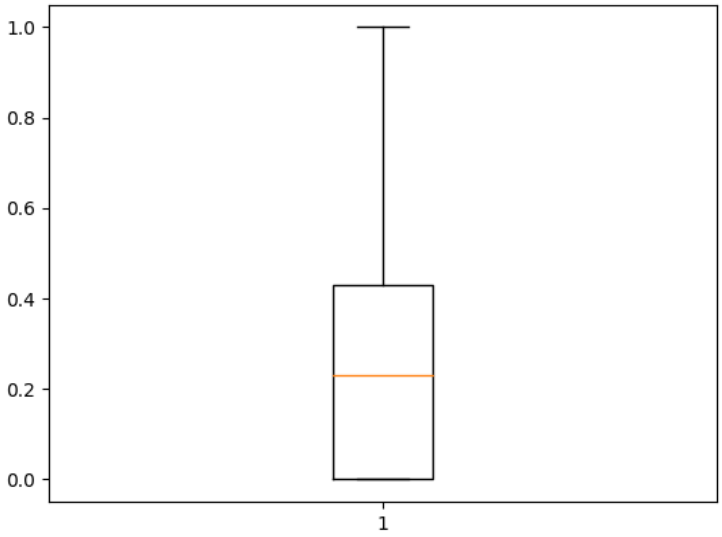


FIGURE 5.18 – T_{lex} (ligne)

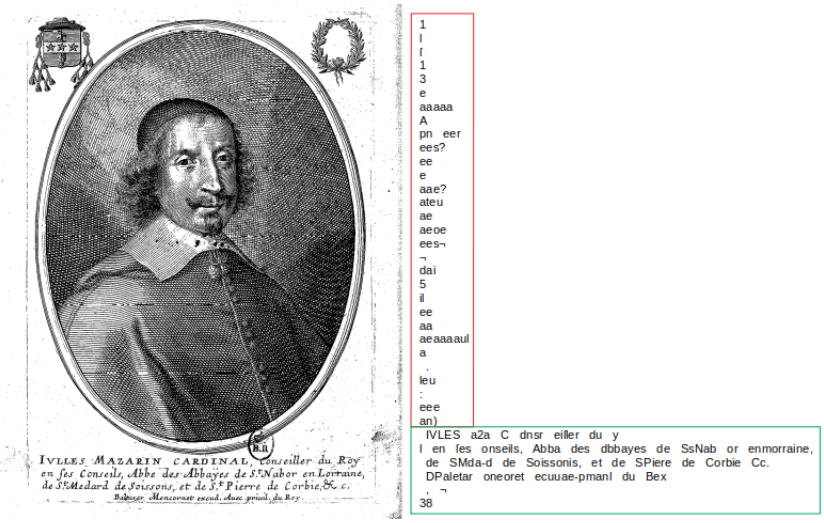


FIGURE 5.19 – Numérisation et transcription automatique de la septième page de *Apologie du cardinal, burlesque.* (Paris, 1649, Moreau 114), https://gallica.bnf.fr/ark:/12148/bpt6k5714613w.

5.3.3 Les valeurs extrêmes minimales sont-elles nécessairement du bruit ?

Un bruit, des bruits

On appelle généralement *bruit* toute réponse non pertinente d'un système, par opposition au *silence* qui est l'absence d'une réponse pertinente. Toutefois, dans notre cas, l'utilisation de l'unique terme *bruit* ne doit pas donner à penser que les causes impliquant le bruit dans les données textuelles sont uniformes. Deux types, qui correspondent à deux échelles, et à deux temps, peuvent être observés. D'un côté, au sein d'un passage contenant effectivement des données textuelles, le logiciel d'OCR peut faire localement une erreur (insertion, délétion ou substitution). Il s'agit d'un bruit qui ne rend pas forcément impossible la lecture. D'un autre côté, on a toutes les numérisations ou parties de numérisation qui ne donnent pas à voir de données textuelles (par exemple les pages de reliure ou encore les portraits imprimés dans le corps d'une page) ; le logiciel d'OCR peut malgré tout segmenter et transcrire ces zones. Nous qualifions ce bruit de *bruit pur* car l'exercice de lecture est impossible. En somme, le bruit peut surgir à deux moments du processus d'océrisation : lors de la segmentation ou lors de la transcription.

La figure 5.19 en illustre un exemple, pour la septième page de l'*Apologie du cardinal, burlesque.* (Paris, 1649), transcrite avec le même modèle d'OCR entraîné sur des œuvres du XVII^{e} siècle. Deux rectangles apparaissent sur le côté droit de la figure ; le rectangle rouge du haut identifie le *bruit pur* qui est le résultat d'une segmentation non attendue du portrait du Cardinal ; le

	T_{con} <0,97	T_{con} <0,94
Nb pages	146	10
Nb Google Book	37	9

TABLE 5.5 – Nombre de pièces téléchargées sur Google Livres ayant un Tcon inférieur à un certain seuil sur le corpus des mazarinades burlesques

	T_{lex} <0,34	T_{lex} <0,30
Nb pages	132	11
Nb Google Book	32	7

TABLE 5.6 – Nombre de pièces téléchargées sur Google Livres ayant un Tcon inférieur à un certain seuil sur le corpus des mazarinades burlesques

rectangle vert du bas identifie quant à lui le bruit résultat des segmentations et transcriptions du texte situé, dans la page, en-dessous du portrait.

Du bruit et des œuvres

Si les corrélations entre les T_{con} et T_{lex} et le CER n'ont pas été réalisées à l'échelle de l'œuvre, il est possible d'observer ces métriques à cette échelle en calculant simplement la moyenne de leurs valeurs par page. On disposera ainsi d'une valeur (moyenne) pour le T_{con} et T_{lex} pour chaque œuvre du corpus. Considérons les valeurs des T_{con} et T_{lex} inférieures à la moyenne :

— T_{con} : valeurs inférieures à 0,97 ;

— T_{lex} : valeurs inférieures à 0,34 ;

et les pièces du corpus téléchargées sur Gallica (142) et celles téléchargées sur Google Livres (38).

Les tableaux 5.5 et 5.6 montrent que, pour les pièces ayant un T_{con} inférieur à $0,97$, on a 37 des 146 pièces concernées qui ont été téléchargées sur Google Livres – ce qui représente presque la totalité des 38 pièces du corpus téléchargées sur cette plateforme. Plus encore, pour les dix pièces ayant un T_{con} inférieur à 0,94, neuf proviennent de *Google Livres*. On peut procéder au même constat, quoiqu'avec un peu plus de nuance, pour les pièces ayant un faible T_{lex}. Considérant les numérisations de Google Livres de moins bonne qualité que celles de Gallica, nous supposons que les sorties d'OCR de ces pièces devraient être de moins bonne qualité. Si cela est vrai, les précédentes observations permettent de dire que les T_{con} et T_{lex} sont en pratique de bons indicateurs de la qualité moyenne d'une OCR.

Du bruit et des pages

Considérons les valeurs extrêmes minimales suivantes représentées sur les boîtes à moustache des figures 5.17 et 5.18 :

— valeurs inférieures à $0,9$ pour le T_{con} ;
— valeurs inférieures à $0,2$ pour le T_{lex}.

Les premières valeurs sont associées à un ensemble de 288 pages et les secondes à un ensemble de 358, le tout représentant 540 pages. L'intersection de ces deux ensembles (c'est-à-dire les pages ayant à la fois T_{con} inférieur à $0,9$ et un T_{lex} inférieur à $0,2$) comprend 106 pages, soit 20 % des pages précitées. Après examen de ces 540 pages, 72 (13 %) contiennent en réalité des données textuelles. Si donc on ne peut pas voir dans ces évaluations un moyen sans faille d'isoler les pages contenant ou non des données textuelles, elles constituent de bons indices.

Du bruit et des lignes

Les taux T_{con} et T_{lex} peuvent aussi être calculés à l'échelle de la ligne ; il s'agit alors d'observer si l'exercice de lecture est possible ou non pour des taux T_{con} et T_{lex} inférieurs aux valeurs extrêmes minimales. Nous prenons les valeurs de $0,9$ et $0,2$ (pour les taux T_{con} et T_{lex}, respectivement) car ces valeurs extrêmes génèrent déjà beaucoup de lignes à évaluer (manuellement).

Pour le taux T_{con}, on ressort 5 491 lignes ayant un taux T_{con} inférieur à $0,9$, dont 2 192 ne contiennent qu'une espace. Pour le taux T_{lex}, on en ressort 17 472 avec un taux inférieur à $0,2$, dont 6 887 ne contiennent qu'une espace.

Avec les graphes des figures 5.20 et 5.21, on voit que les taux T_{con} et T_{lex} permettent de récupérer des lignes résultat d'une mauvaise segmentation : les lignes ne contenant qu'une espace (voir celles ne contenant qu'un caractère). On observe aussi des lignes contenant plusieurs dizaines de caractères. Très souvent (quoiqu'aucune statistique précise ne l'atteste encore), il s'agit de ligne du type *loolootlotopouoohpenl*. Il s'agit des « termes poubelle » [34] évoqués par [Taghva et al., 1994, section 3.4] : plus de 27 caractères, plus de 4 caractères identiques à la suite, grand ratio de caractères non alphabétiques ou encore mauvais ratio consonnes/voyelles. Notons enfin qu'entre ces lignes ne contenant qu'une espace et celles correspondant à des « termes poubelle », certaines lignes présentent des suites de caractères permettant la lecture, comme *friuole ?* .

Globalement donc, les taux T_{con} et T_{lex} sont de bons indicateurs de la qualité d'une sortie d'OCR, même si le second n'est pas corrélé au CER dans la première expérience traitant cette question. Par ailleurs, l'examen manuel du « comportement » de ces métriques sur un corpus pour lequel aucune vérité de terrain n'existe permet de valider par la pratique la pertinence, au moins relative, de ces métriques, tant à l'échelle de l'œuvre qu'à celle de la page et même de la ligne. Néanmoins, nuance est de rigueur : il s'agit d'indicateurs permettant de donner quelques indices sur la qualité des données, sans pour autant prétendre à rendre compte de la qualité *réelle* de celles-ci.

34. *junk terms*

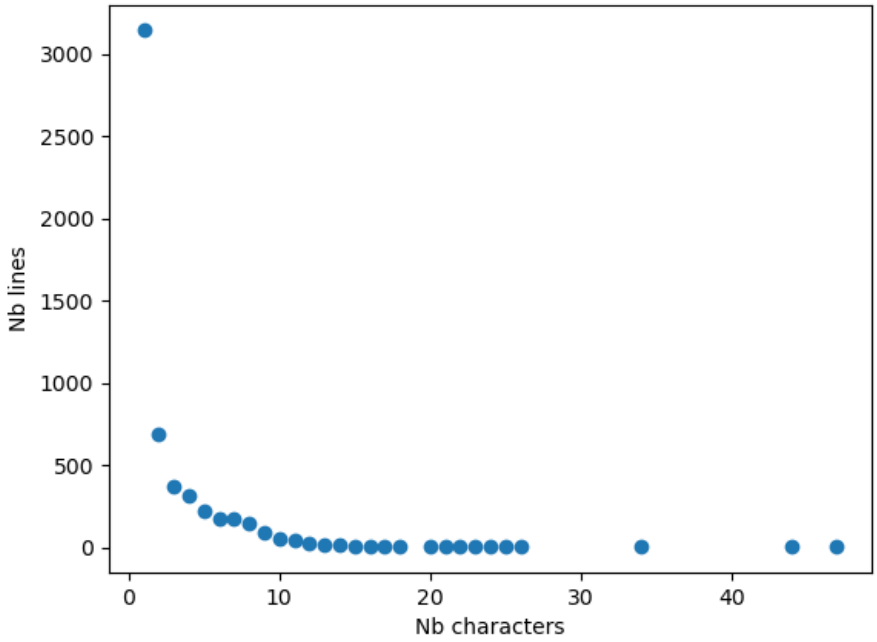


FIGURE 5.20 – Nombre de lignes en fonction du nombre de caractères contenus dans ces lignes, pour les lignes ayant une valeur T_{con} inférieure à la valeur extrême minimal $(0, 9)$

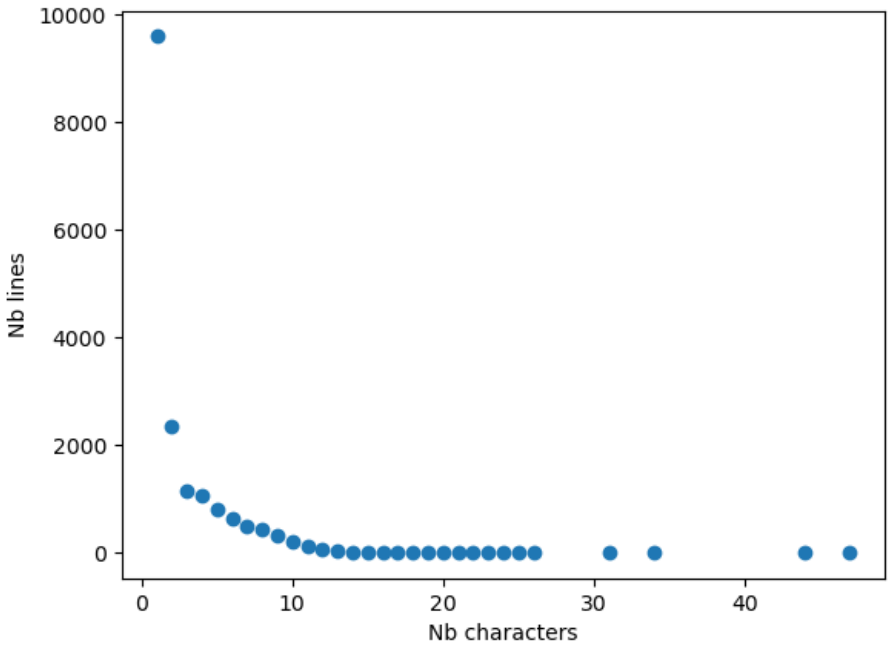


FIGURE 5.21 – Nombre de lignes en fonction du nombre de caractères contenus dans ces lignes, pour les lignes ayant une valeur T_{lex} inférieure à la valeur extrême minimal $(0, 2)$

5.4 L'équivocation shannonienne : mesure du « bruit dans les données » ?

Comme vu dans dans la sous-section 4.1.2, Shannon propose de définir les éléments intervenant dans le processus de communication. Sa *Théorie* est fondée sur l'observation que les messages peuvent (et doivent pour entrer dans cette théorie) se traduire sous forme de probabilité (de fréquence, en l'occurrence). Ainsi, l'information d'un message est associé à la probabilité que la source fournisse ce message ; on a alors la formule de l'entropie :

$$H(x) = \sum_i p_i \log \frac{1}{p_i} \Leftrightarrow H(x) = -\sum_i p_i \log p_i \tag{5.5}$$

avec x une source d'information (une variable aléatoire dans ce cadre mathématique) et p_i la probabilité que le symbole i apparaisse. Cette quantité d'information, aussi interprétable comme l'incertitude du message, est le fondement de toute la théorie. De là arrive le théorème fondamental qui fixe la limite théorique de la performance d'un codage enlevant toute redondance à C/H (avec C la capacité du canal).

Dans la *Théorie*, le bruit est, comme la source d'information, assimilé à un processus stochastique.

> Si un canal bruité est alimenté par une source, il y a deux processus stochastiques en œuvre : la source et le bruit. On peut donc calculer un certain nombre d'entropies. D'abord, il y a l'entropie $H(x)$ de la source ou de l'entrée du canal (qui seront égales si l'émetteur est non singulier). L'entropie de la sortie du canal, c'est-à-dire le signal reçu, sera noté $H(y)$. Dans le cas du canal sans bruit, $H(y) = H(x)$. L'entropie conjointe de l'entrée et de la sortie sera $H(x, y)$. Enfin, il y a deux entropies conditionnelles $H_x(y)$ et $H_y(x)$, l'entropie de la sortie quand l'entrée est connus, et réciproquement. Entre ces quantités, nous avons les relations
>
> $$H(x, y) = H(x) + H_x(y) = H(y) + H_y(x)$$
>
> Toutes ces entropies peuvent être mesurées par seconde ou par symbole. [Shannon, 1949, p. 50]

L'entropie conditionnelle $H_y(x)$, ou « équivocation » [Shannon, 1949, p. 51], correspond à l'entropie de la source d'information x sachant le bruit y. « Donc, en gros, $H_y(x)$ représente la quantité d'information supplémentaire qui doit être fournie par seconde au point de réception pour corriger le message reçu. » [Shannon, 1949, p. 53]

Dès lors, comme [Franzini et al., 2018], peut-on utiliser cette équivocation comme mesure du « bruit dans les données » ?

5.4.1 Calcul de l'équivocation d'une source discrète bruitée

Le calcul de l'équivocation, ou entropie conditionnelle de la source sachant le bruit, se réalise par l'observation des symboles émis par la source (*a posteriori*, donc les fréquences de ces symboles). Il suffit donc de décrire, en termes de fréquences, la succession de ces symboles. Comme proposé dans la section 3 *La série des approximations de l'anglais* [Shannon, 1949, p. 23], pour une suite de symboles correspondant à un énoncé langagier, ou plus simplement à du texte (Shannon préfère le terme « message »), une décomposition de ce texte en une suite de n-grammes semble adapté. Les fréquences de ces n-grammes constitueront les probabilités p_i de l'équation 5.4. Ainsi, si nous travaillons avec $n = 1$, Shannon parlera d'« approximation d'ordre 1 », avec $n = 2$, d'« approximation d'ordre 2 », et ainsi de suite.

L'équivocation peut être calculée pour toute source d'information conjuguée à une source de bruit ; il est donc possible de la calculer pour une sortie d'OCR. Cette quantité (voir équation 5.4) peut être calculée selon plusieurs ordres. Soit n un entier naturel (généralement compris entre 1 et 8). Il suffit de décomposer la chaîne de caractères résultat de la source bruitée en autant de n-grammes, de calculer la probabilité d'apparition de ces n-grammes et de poursuivre en appliquant la formule 5.4.

Soit l'énoncé suivant [35] : *Auec vn peu moins de careffes Que l'on n'enleue des Maiftreffes*. Pour $n = 1$, on a les unigrammes suivants : *A, u, e, c, , v, n, , p, e, u, , m, o, i, n, s, , d, e, , c, a, r, e, f, f, e, s, , Q, u, e, , l, ', o, n, , n, ', e, n, l, e, u, e, , d, e, s, , M, a, i, f, t, r, e, f, f, e, s*. Le calcul de la probabilité d'un symbole donné se réalise en divisant le nombre d'apparition de ce symbole par le nombre total de symboles. On a donc : $p(A) = \frac{1}{63}$, $p(u) = \frac{4}{63}$ et ainsi de suite. Il suffit enfin de parcourir tout l'énoncé, de multiplier chaque probabilité par le *log* de cette probabilité et de réaliser la somme inverse de l'ensemble de ces quantités. Dans notre exemple, on arrive à $H(x) = 12,18$.

5.4.2 Vers une corrélation de l'entropie conditionnelle au CER

De la même manière que dans la section 5.2.4, nous allons calculer, pour plusieurs OCR, la corrélation entre le CER associés à ces OCR et l'entropie conditionnelle calculée uniquement sur les sorties d'OCR – sans vérité de terrain ni autre information que le texte lui-même. Afin de calculer un CER, nous utilisons le même corpus constitué par [Gabay, 2019] (voir section 5.2.4).

Plusieurs OCR (logiciels et modèles) interviennent ici :

— `Kraken` – modèle XVII^{e} ;
— `Tesseract` :
  — modèle français... ;
  — ... + grayscale ;
  — ... + suppression du bruit ;

35. Issu de la *Lettre d'un inconnu envoyée à un sien ami à Saint-Germain-en-Laye, en vers burlesques*, Paris, Michel Mettayer, 1649, p. 3. (Moreau1833).

Entropies	Corrélations au CER	p-values
n = 1	-0,31	<0,01
n = 2	-0,21	<0,01
n = 3	-0,18	<0,01
n = 4	-0,20	<0,01
n = 5	-0,22	<0,01
n = 6	-0,23	<0,01

TABLE 5.7 – Corrélations des entropies (et valeur p) au CER pour n variant de 1 à 6 sur le corpus de [Gabay, 2019]

— ... + binarisation ;
— ... + dilation ;
— ... + redressement ;

Chaque page du corpus de [Gabay, 2019] a été océrisé par les sept modèles précités. Le corpus étant transcrit, il est possible de calculer les CER.

Résultats

Une fois calculé l'ensemble de ces entropies (pour chaque page océrisée du corpus et pour n variant de 1 à 6), on peut calculer les corrélations entre ces entropies et les CER.

Le tableau 5.7 montre des corrélations significatives (p-value $< 0,01$), allant de $-0,31$ pour $n = 1$, diminuant ensuite pour n compris entre 2 et 3 et se stabilisant pour n allant jusque 6 (et même 8). La corrélation la plus forte s'observe donc pour les unigrammes ce qui signifie que ce sont les caractères pris indépendamment les uns des autres qui sont le plus porteur de sens sur le *bruit* issu des erreurs d'OCR. D'autre part, plus un énoncé est bruité, plus il s'éloigne des « règles » de la langue et donc plus il est ardu de compresser ce message ce qui aboutit à une entropie plus élevée car porteuse de plus d'information (au sens de la succession des symboles, qui est donc plus erratiques donc plus difficilement compressible).

Que les corrélations soient toutes négatives avec le CER signifie donc que le *message* (l'énoncé) a une entropie qui augmente lorsque le CER diminue ; le *bruit*, dans ce corpus océrisé, est donc plus redondant que les données textuelles, ce qui en soi est un résultat de recherche allant contre l'intuition, mais validant encore le caractère répétitif des erreurs d'OCR.

L'entropie est donc un bon moyen d'approcher le CER (pour $n = 1$), même si la corrélation n'est pas très forte. D'abord, le test statistique de la valeur p montre que cette relation n'est en aucun cas due au hasard ; l'entropie est donc porteuse d'information. Ensuite, le calcul de l'entropie est presque gratuit car il ne nécessite aucune ressource, que ce soit une ressource linguistique comme le T_{lex} ou les informations données par un logiciel donné comme le T_{con}.

Enfin, on peut imaginer que cette quantité peut être une bonne manière

d'approcher le bruit dans les données textuelles pour tout type de texte car seul le texte suffit (en fait les unigrammes). Pour tenter de démontrer cela, nous allons procéder en sous-section 5.4.3 à une estimation du bruit dans un corpus radicalement différent : le corpus DAnIEL.

5.4.3 Étude de l'entropie sur un autre corpus (le corpus DAnIEL)

Construit en 2011 et 2012 [Lejeune et al., 2012] et utilisé jusque [Barbaresi and Lejeune, 2020], le corpus DAnIEL est composé de 1 694 documents (en chinois, anglais, grec, polonais et russe) disponible dans sa version originale HTML et dans sa version nettoyée contenant uniquement le contenu textuel et quelques balises. [Barbaresi and Lejeune, 2020] précise qu'il s'agit actuellement du plus grand corpus multilingue d'évaluation d'extraction de contenu du web. Par ailleurs, on dispose aussi, pour chaque document, de ses versions nettoyées par neuf outils : Html2Text, Inscriptis, Boilerpy3, Dragnet, Goose3, JusText, NewsPaper, News-Please et Python-Readability.

Mesurer le bruit dans le corpus DAnIEL

Les types mesures proposées par [Barbaresi and Lejeune, 2020] pour juger de la qualité de l'extraction des données textuelles contenues dans des pages HTML sont les suivantes :

- *clean_eval*, laquelle consiste à repérer l'ensemble des sous-chaînes communes entre les données textuelles de référence (ci-après appelé référence) et les données extraites automatiquement par les outils (ci-après appelé test) ;
- *voc_eval*, laquelle s'intéresse à l'union des vocabulaires de la référence et du test ;
- *occ_eval*, laquelle compare les nombre d'occurrences des entrées des vocabulaires de la référence et du test ;
- *KL_eval*, laquelle mesure l'« écart » entre les vecteurs des fréquences relatives des mots entre la référence et le test.

Selon *voc_eval* et *occ_eval*, JusText offre les meilleures performances en F-mesure, alors qu'il s'agit de Python-Readability pour *KL_eval* et de Goose3 pour *clean_eval*.

L'équivocation est-elle corrélée à une mesure du bruit dans Daniel ?

Le tableau 5.8 montre des corrélations entre l'entropie (calculée pour n variant de 1 à 6) et les métriques évaluant le bruit dans le corpus Daniel plutôt forte (car autour de 0,50). À la différence du précédent corpus de [Gabay, 2019], on observe que les corrélations sont positives, ce qui signifie que plus l'entropie augmente, plus les métriques augmentent à leur tour. Et pour cause, ces métriques ne mesurent pas ce qui a trait au bruit dans le corpus mais précisemment l'inverse, comme les sous-chaînes répétés (donc tout sauf le bruit) pour *clean_eval*.

	n = 1	**n = 2**	**n = 3**	**n = 4**	**n = 5**	**n = 6**
clean_eval	0,57	0,53	0,53	0,54	0,52	0,51
voc_eval	0,61	0,60	0,60	0,57	0,56	0,57
KL_divergence	0,62	0,63	0,68	0,68	0,65	0,62
occ_eval	0,57	0,63	0,65	0,63	0,60	0,59

TABLE 5.8 – Corrélation de l'entropie aux métriques évaluant le bruit dans le corpus DANIEL (p-values inférieures à 0,01)

5.4.4 Apprentissage de modèles de langues

Cette sous-section à fait l'objet d'une publication pour la conférence RECITAL 2020 : [Tanguy, 2020].

Sous le même principe que [Chen et al., 1998], nous proposons dans cette partie i) d'apprendre des modèles de langue sur un corpus en français pré-classique (XVII^{e} siècle), ii) de parcourir des sorties de logiciels d'OCR par fenêtre glissante en récupérant les probabilités de chaque modèle de langue de rencontrer une telle séquence de caractères pour enfin iii) estimer la qualité de ces sorties d'OCR. Différentes méthodes d'agrégation des probabilités précitées sont proposées pour estimer la qualité globale des *pages* océrisées. L'étude des corrélations entre ces estimations et les *CER* (pour chaque page du corpus) permettra de valider ou réfuter la pertinence de ces estimateurs sur le corpus de l'étude.

Deux types de modèles de langue (au grain caractère) ont été appris sur le sous-corpus dédié qui compte 121 caractères différents :

— des modèles de langue à probabilités conditionnelles, appris comme la probabilité d'observer un caractère sachant une séquence de caractères (un historique) ;
— des modèles de langue appris par des réseaux de neurones (LSTM et biLSTM).

Le premier type de modèles constitue une *baseline* puisque ces modèles sont simplement construits en comptant, par fenêtre glissante sur le corpus d'apprentissage, le nombre d'occurrences du caractère suivant la séquence de caractères contenue dans la fenêtre glissante. Ces occurrences absolues sont ensuite divisées par la somme des occurrences de tous les caractères suivants cette séquence et sont utilisées comme des probabilités, puisque contenues dans l'intervalle $[0; 1]$.

Les modèles de langue LSTM et biLSTM ont été choisis pour confronter aux modèles de la *baseline* des modèles appris par réseaux de neurones, en l'occurrence des réseaux de neurones récurrents. Ces modèles ont été appris en utilisant les modèles séquentiels de la librairie Python *keras*. Un *mapping* du vocabulaire est d'abord réalisé en prétraitement [36]. Les réseaux LSTM et biLSTM contiennent tous une couche *LSTM* et aux réseaux biLSTM est ajoutée une couche *Bidirectional* ; l'hypothèse étant ici que i) tout caractère ne peut suivre

36. À chaque élément du vocabulaire (entendu comme l'ensemble des caractères différents) est associé un entier dans une table.

tout autre caractère et que ii) tout caractère ne peut être précédé de tout autre caractère. Enfin, la fonction *softmax* est utilisée comme fonction d'activation. Ces modèles de langue ont été appris sur des séquences de n caractères, pour n variant de 2 à 10 et le nombre d'époques pour chaque apprentissage est 100. Finalement, on dispose donc de $3 * 9 = 27$ modèles de langue pour tester l'estimation de la qualité des sorties des logiciels d'OCR. Le nombre de caractères dans le vocabulaire de ces modèles de langue est de 121.

Probabilités des modèles de langue Les modèles de langue permettent de disposer de la probabilité qu'un caractère donné suive une certaine séquence de caractères. Si une sortie de logiciel d'OCR est parcourue par une fenêtre glissante à partir de laquelle est renvoyée une séquence de caractères et le caractère suivant cette séquence, pour une sortie d'OCR on dispose d'une suite d'au plus $C - n$ probabilités, avec C le nombre total de caractères et n la taille de la fenêtre glissante en caractères. *Au plus* car il est possible que certains caractères fournis par le modèle d'OCR n'aient pas été rencontrés dans le corpus d'apprentissage du modèle de langue [37].

On cherche donc à agréger ces probabilités, pour chaque document du corpus océrisé, dans l'objectif que ces agrégats soient corrélés au *CER* qu'on peut calculer grâce aux transcriptions. Il s'agit de calculer d'autres métriques ne nécessitant pas de vérité de terrain (à partir des probabilités fournies par les modèles de langue) et de valider ou réfuter la pertinence de leur estimation de la qualité d'une sortie d'OCR face à une métrique de référence, le *CER*.

Agrégations des probabilités des modèles de langue

La somme des probabilités Une première métrique peut être la somme des probabilités renvoyées par les modèles de langue. Sous réserve que les modèles de langue sont bien des distributions de probabilités, la somme des probabilités d'une suite de caractères correspondant à du texte est de 1 alors qu'elle ne peut l'être dans le cas contraire. Ainsi, pour une sortie d'OCR, on a :

$$S = \sum_{i=n+1}^{C-n} P_{LM}(c_i|h_{n,i})$$

Avec P_{LM} la probabilité renvoyée par un modèle de langue LM, n la taille de la fenêtre glissante en caractères, C le nombre total de caractères de la sortie d'OCR, c_i le i^{e} caractère de la sortie d'OCR et $h_{n,i}$ l'historique de n caractères du caractère c_i.

Le produit des probabilités Le produit des probabilités peut aussi constituer une autre métrique d'estimation. Il rend compte de la probabilité d'une suite

37. Par exemple, Kraken (anglais) et Tesseract (français), appris sur des documents contemporains, ont dans leur vocabulaire le symbole euro et peuvent le proposer dans leur océrisation. Pour un modèle de langue appris sur des données textuelles françaises du XVII$^{\text{e}}$ siècle, ce symbole n'existe pas.

de caractères selon l'hypothèse, ici réductrice, de l'indépendance. Il est défini comme :

$$Pr = \prod_{i=n+1}^{C-n} P_{LM}(c_i|h_{n,i})$$

La perplexité Plus couramment utilisée pour juger de la qualité d'un modèle de langue, la perplexité est la probabilité inverse de la sortie d'OCR normalisée par son nombre de caractères. Puisqu'elle mesure la distance entre la fonction de probabilité et les données de l'ensemble de test, elle semble pertinente à tester comme métrique d'estimation. Elle est définie comme :

$$PP = \frac{1}{(\prod_{i=n+1}^{C-n} P_{LM}(c_i|h_{n,i}))^{\frac{1}{C-n}}}$$

La log-perplexité Enfin, la log-perplexité peut aussi constituer une métrique d'estimation ; [Chen et al., 1998] montrent qu'elle aussi est corrélée au *WER* dans le domaine de la reconnaissance de la parole.

Échelles de calcul des agrégations

Les agrégations des probabilités précitées peuvent être calculées à plusieurs échelles : celle de l'œuvre, de la page, de la ligne ou encore du mot. Puisque la perplexité PP est calculée comme l'inverse d'une racine n-ième, elle tend vers 1 à mesure que le nombre total de caractères $C - n$ grandit. Plus le nombre de caractères sur lesquels elle est calculée grandit, moins elle est informative. Les estimateurs de qualité d'océrisation des pages sont donc calculés comme la moyenne des agrégations des probabilités calculées à l'échelle du mot [38].

Notons que réaliser une *moyenne* constitue un biais important. Les données textuelles issues d'OCR n'ont pas une qualité homogène pour une même œuvre ou une même page ; la moyenne efface ces disparités pourtant essentielles à soulever. D'autre part, certains mots comportent trop peu de caractères pour que le modèle de langue puisse leur calculer une probabilité. Certains passages sont ignorés et la moyenne ne le traduit pas.

Expérimentation et résultats

Pour calculer les métriques d'estimation de la qualité de l'OCR sur une page du corpus, on calcule ces métriques pour chaque mot de la page et on en fait la moyenne. Afin de confirmer ou réfuter l'intérêt de ces métriques, un calcul de corrélation Pearson [39] est réalisé avec le *CER*. Si une ou plusieurs métriques est

38. La tokenisation est réalisée par une simple segmentation par l'espace des chaînes de caractères.

39. En tant que normalisation des covariances, le coefficient de corrélation exprime à quel point deux variables sont liées. Ce coefficient étant une normalisation, il appartient à l'intervalle $[-1; 1]$; les corrélations positives indiquent que les deux variables évoluent dans le même sens et les corrélations négatives qu'elles évoluent dans un sens opposé. Plus une corrélation est

Métriques	Variations	Signes des corrélations avec le *CER*
CER	$\nearrow$	
S	$\searrow$	-
Pr	$\searrow$	-
$Pr^{\frac{1}{C-n}}$	$\searrow$	-
$PP = \frac{1}{Pr^{\frac{1}{C-n}}}$	$\nearrow$	+
$\log(PP)$	$\nearrow$	+

TABLE 5.9 – Variations et signes des corrélations avec le *CER* des métriques d'estimation pour un nombre d'erreurs qui augmente.

corrélée-s significativement au *CER*, on peut conclure que l'apprentissage d'un modèle de langue sur des données du français du XVII^{e} siècle et l'utilisation de ses probabilités pour estimer la qualité d'une sortie d'OCR sur des documents de la même période sont justifiés.

Préambule à l'analyse des corrélations La table 5.9 expose la variation des métriques d'estimation et le signe de leur corrélation avec le *CER* pour un nombre d'erreurs d'océrisation qui augmente. L'hypothèse est que les modèles de langue fournissent des probabilités (voir le paragraphe 5.4.4) plus élevées face à une sortie d'OCR sans erreur (du *texte*) et des probabilités plus faibles face à une sortie d'OCR avec erreurs (du *non-texte*). Ainsi, pour valider les métriques comme estimateurs pertinents, les corrélations entre le *CER* et la somme et le produit des probabilités doivent être négatives alors qu'elles doivent être positives entre le *CER* et la perplexité et la log-perplexité.

Corrélations entre le *CER* et les métriques d'estimation Le sous-corpus dédié à l'OCR est composé de 118 pages. On peut donc, pour les sorties des trois modèles d'OCR utilisés et pour les trois types de modèles de langue, calculer les métriques d'estimation et le *CER*, et ce pour chaque page du corpus.

OCR : Kraken (français contemporain). ML : probabilités conditionnelles.

	S		**Pr**		**PP**		**log(PP)**	
	corr.	**p-value**	**corr.**	**p-value**	**corr.**	**p-value**	**corr.**	**p-value**
n=2	-0,063	0,496	0,111	0,226	-0,016	0,862	-0,013	0,884
n=3	-0,098	0,287	0,073	0,426	0,005	0,955	0,021	0,820
n=4	-0,073	0,428	-0,043	0,642	-0,010	0,913	-0,014	0,879
n=5	-0,137	0,135	-0,043	0,638	-0,015	0,868	-0,026	0,780
n=6	-0,093	0,314	0,000	0,996	0,067	0,466	0,059	0,522
n=7	-0,035	0,708	-0,032	0,728	0,130	0,157	0,117	0,205
n=8	-0,064	0,485	-0,074	0,420	0,043	0,643	0,054	0,560
n=9	-0,057	0,538	-0,012	0,898	0,018	0,846	0,021	0,821
n=10	-0,046	0,615	-0,023	0,806	0,024	0,794	0,026	0,780

OCR : Tesseract (anglais contemporain). ML : probabilités conditionnelles.

	S		**Pr**		**PP**		**log(PP)**	
	corr.	**p-value**	**corr.**	**p-value**	**corr.**	**p-value**	**corr.**	**p-value**
n=2	-0,004	0,968	0,158	**0,086**	0,113	0,221	0,006	0,952
n=3	-0,130	0,156	0,009	0,920	-0,003	0,976	0,056	0,540

proche de 1 ou -1, plus le lien entre les deux variables est fort ; au contraire, plus la corrélation est proche de 0, plus ce lien se dissipe.

n=4	-0,124	0,178	-0,005	0,960	0,016	0,866	0,060	0,518
n=5	-0,158	**0,084**	-0,070	0,449	0,134	0,143	0,158	**0,085**
n=6	-0,138	0,133	-0,054	0,556	0,180	**0,049**	0,188	**0,040**
n=7	-0,100	0,278	-0,027	0,773	0,093	0,313	0,084	0,359
n=8	-0,055	0,547	-0,008	0,930	-0,006	0,949	-0,008	0,928
n=9	-0,054	0,554	-0,083	0,366	0,096	0,299	0,095	0,300
n=10	-0,024	0,796	-0,212	**0,020**	0,228	**0,012**	0,187	**0,041**

OCR : Kraken (français XVII^{e}). ML : probabilités conditionnelles.

	S		**Pr**		**PP**		**log(PP)**	
	corr.	**p-value**	**corr.**	**p-value**	**corr.**	**p-value**	**corr.**	**p-value**
n=2	-0,052	0,572	0,129	0,162	0,238	**0,009**	0,040	0,663
n=3	-0,080	0,384	0,087	0,343	-0,003	0,970	0,030	0,742
n=4	-0,041	0,654	-0,010	0,916	0,012	0,894	0,031	0,739
n=5	-0,111	0,225	-0,072	0,437	0,039	0,670	0,069	0,452
n=6	-0,135	0,142	-0,055	0,551	0,138	0,133	0,143	0,120
n=7	-0,086	0,348	-0,030	0,746	0,063	0,496	0,075	0,414
n=8	-0,096	0,296	-0,030	0,743	-0,045	0,622	-0,045	0,625
n=9	-0,052	0,574	-0,017	0,857	-0,015	0,874	-0,029	0,757
n=10	-0,021	0,817	-0,039	0,669	0,121	0,189	0,097	0,291

OCR : Kraken (français contemporain). ML : LSTM.

	S		**Pr**		**PP**		**log(PP)**	
	corr.	**p-value**	**corr.**	**p-value**	**corr.**	**p-value**	**corr.**	**p-value**
n=2	-0,059	0,520	0,094	0,305	0,046	0,615	0,000	0,999
n=3	-0,078	0,395	0,078	0,397	-0,031	0,739	-0,011	0,902
n=4	-0,094	0,306	-0,070	0,446	0,113	0,221	-0,111	0,227
n=5	-0,108	0,240	-0,048	0,600	-0,061	0,511	-0,077	0,404
n=6	-0,039	0,675	0,184	0,045	0,055	0,552	-0,092	0,320
n=7	0,198	**0,030**	-0,049	0,595	-0,051	0,578	0,035	0,702
n=8	-0,034	0,709	0,026	0,781	-0,017	0,854	0,058	0,530
n=9	-0,055	0,549	-0,016	0,860	-0,062	0,499	0,076	0,407
n=10	0,063	0,492	-0,036	0,695	-0,055	0,550	-0,061	0,506

OCR : Tesseract (anglais contemporain). ML : LSTM.

	S		**Pr**		**PP**		**log(PP)**	
	corr.	**p-value**	**corr.**	**p-value**	**corr.**	**p-value**	**corr.**	**p-value**
n=2	-0,010	0,911	0,138	0,131	-0,032	0,730	-0,081	0,377
n=3	-0,132	0,151	-0,046	0,621	-0,004	0,962	0,133	0,148
n=4	-0,085	0,354	-0,053	0,567	0,003	0,972	-0,009	0,926
n=5	-0,096	0,298	-0,080	0,383	-0,022	0,808	0,053	0,568
n=6	-0,107	0,245	0,034	0,716	-0,010	0,915	-0,038	0,683
n=7	-0,116	0,208	-0,079	0,390	0,106	0,250	0,007	0,938
n=8	0,043	0,640	-0,026	0,776	-0,034	0,711	-0,025	0,788
n=9	-0,044	0,636	0,024	0,791	-0,032	0,732	0,029	0,756
n=10	-0,025	0,788	-0,072	0,432	0,012	0,895	0,073	0,426

OCR : Kraken (français XVII^{e}). ML : LSTM.

	S		**Pr**		**PP**		**log(PP)**	
	corr.	**p-value**	**corr.**	**p-value**	**corr.**	**p-value**	**corr.**	**p-value**
n=2	-0,055	0,552	0,168	**0,066**	-0,006	0,944	-0,049	0,596
n=3	-0,065	0,482	0,114	0,215	-0,023	0,804	-0,062	0,503
n=4	-0,058	0,526	-0,030	0,741	-0,026	0,777	-0,024	0,793
n=5	-0,069	0,455	-0,052	0,576	-0,012	0,898	-0,005	0,953
n=6	-0,083	0,366	0,045	0,623	-0,008	0,930	0,017	0,856
n=7	-0,067	0,465	-0,049	0,595	-0,027	0,773	-0,056	0,541
n=8	-0,104	0,258	-0,041	0,653	-0,030	0,745	-0,029	0,751
n=9	-0,023	0,805	-0,036	0,697	-0,022	0,809	-0,034	0,710
n=10	0,181	**0,048**	-0,059	0,520	-0,006	0,947	-0,012	0,898

OCR : Kraken (français contemporain). ML : biLSTM.

	S		**Pr**		**PP**		**log(PP)**	
	corr.	**p-value**	**corr.**	**p-value**	**corr.**	**p-value**	**corr.**	**p-value**
n=2	-0,042	0,645	0,115	0,213	-0,040	0,661	-0,043	0,641
n=3	-0,098	0,289	0,160	**0,081**	-0,096	0,295	-0,077	0,404
n=4	-0,091	0,320	-0,087	0,346	0,085	0,357	-0,126	0,170
n=5	-0,019	0,837	-0,085	0,354	-0,049	0,595	-0,013	0,891
n=6	-0,076	0,411	0,131	0,154	-0,040	0,662	-0,087	0,345
n=7	0,010	0,914	-0,105	0,255	-0,058	0,529	0,009	0,925
n=8	-0,053	0,564	-0,085	0,357	0,623	**0,001**	0,053	0,563
n=9	-0,070	0,446	-0,060	0,517	-0,024	0,794	-0,054	0,560
n=10	0,084	0,361	-0,028	0,758	-0,033	0,722	-0,059	0,521

	ML probabilités conditionnelles	ML LSTM	ML biLSTM
n=2	90	14721	257646757092
n=3	126	1010690	235913940342
n=4	426	318251055	221055920422
n=5	1091	723946838	211044617070
n=6	1978	690749546	204520506752
n=7	2801	669397958	200184237186
n=8	3510	655634987	1161841181775
n=9	3940	647905538	13807745026062
n=10	4205	643364471	14481238375005

TABLE 5.11 – Moyennes des perplexités des modèles de langue sur le sous-corpus de test.

OCR : Tesseract (anglais contemporain). ML : biLSTM.

	S		Pr		PP		log(PP)	
	corr.	**p-value**	**corr.**	**p-value**	**corr.**	**p-value**	**corr.**	**p-value**
n=2	0,013	0,886	0,184	**0,044**	0,000	1,000	-0,117	0,203
n=3	-0,137	0,137	-0,022	0,816	0,117	0,203	0,157	**0,087**
n=4	-0,040	0,663	0,011	0,909	-0,025	0,788	-0,051	0,577
n=5	-0,079	0,389	-0,055	0,547	-0,006	0,946	0,023	0,800
n=6	-0,058	0,530	-0,012	0,893	-0,045	0,627	-0,098	0,287
n=7	-0,064	0,485	-0,013	0,885	-0,007	0,943	-0,093	0,314
n=8	0,036	0,696	-0,023	0,801	0,051	0,580	-0,036	0,694
n=9	-0,053	0,566	-0,033	0,724	0,007	0,936	0,050	0,586
n=10	-0,029	0,754	0,029	0,756	-0,024	0,797	0,027	0,773

OCR : Kraken (français XVII^{e}). ML : biLSTM.

	S		Pr		PP		log(PP)	
	corr.	**p-value**	**corr.**	**p-value**	**corr.**	**p-value**	**corr.**	**p-value**
n=2	-0,061	0,508	0,146	0,113	0,002	0,985	-0,060	0,513
n=3	-0,075	0,413	0,101	0,273	-0,026	0,779	-0,111	0,228
n=4	-0,048	0,601	-0,045	0,622	-0,011	0,909	-0,078	0,396
n=5	-0,040	0,661	-0,106	0,250	-0,038	0,681	-0,014	0,878
n=6	-0,045	0,622	-0,010	0,917	-0,048	0,605	-0,041	0,660
n=7	-0,106	0,251	0,064	0,489	-0,005	0,955	-0,021	0,823
n=8	-0,086	0,350	-0,057	0,539	0,022	0,815	-0,076	0,410
n=9	-0,017	0,855	-0,033	0,721	-0,007	0,940	0,011	0,908
n=10	0,202	**0,027**	-0,025	0,786	-0,027	0,767	0,012	0,898

TABLE 5.10 – Corrélations et *p-values* calculées entre les métriques d'estimation et le *CER*.

La table 5.10 montre les corrélations et les *p-values* calculées entre les métriques d'estimation et les *CER*. Une corrélation n'est toutefois significative que si la *p-value* est inférieure à un seuil, ceci traduisant que la relation de corrélation a peu de chances d'être due au hasard ; il s'agit d'un test de corrélation. Nous choisissons ici le seuil de $0,1$ qui rend compte d'une faible présomption contre l'hypothèse nulle. Les résultats de la table 5.10 montre des *p-values* presque toutes supérieures à ce seuil, ce qui signifie que s'il y a corrélation, elle n'est pas significative. Ces résultats semblent donc réfuter l'hypothèse initiale selon laquelle les probabilités des modèles de langue auraient pu être agrégées pour se substituer à un *CER* exigeant une vérité de terrain.

Les modèles de langue sont-ils inadaptés ? Les résultats précédents suggèrent que i) soit les modèles de langue sont de mauvaise qualité, ii) soit le

corpus de l'étude présente des spécificités particulières ou iii) soit les deux raisons précitées concourent à cette impasse.

Les modèles de langue ont été appris sur les œuvres de Bossuet, Chapelin, Ellain et Gournay. On peut donc les évaluer en calculant leur perplexité sur le sous-corpus des œuvres de Papin, Pascal, Sales et Viau. La table 5.11 présente les moyennes des perplexités des modèles de langue de l'étude calculées sur les vérités de terrain. Les modèles de langue LSTM et biLSTM présentent des perplexités aberrantes (ils sont non adaptés à la tâche) alors que seuls les modèles de langue à probabilités conditionnelles, pour $n \in [2; 4]$, présentent une meilleure qualité. Nous concluons donc que la mauvaise qualité des modèles de langue explique la non corrélation entre les estimateurs et le *CER*.

5.4.5 Conclusion

La mauvaise qualité des modèles de langue ne permet pas de valider ou réfuter notre hypothèse, selon laquelle agréger les probabilités des modèles de langue permettrait d'estimer la qualité d'une sortie d'OCR. Pour en faire l'expérience, il s'agirait de renouveler ces tests avec un ensemble plus vaste de transcriptions d'imprimés du XVII^{e} siècle. Cette dernière sous-section cherchait à proposer une alternative au manque de vérités de terrain mais nous constatons qu'un ensemble de 108 pages (16 315 mots) est insuffisant. Si cela ne contredit pas l'éventuelle pertinence des estimateurs envisagés, un ensemble conséquent de données textuelles en français du XVII^{e} siècle reste nécessaire au bon apprentissage des modèles langue.

A contrario, l'entropie apparaît à l'issue de ces deux expériences comme porteuse de sens dans l'évaluation du bruit dans les données textuelles bruitées ; quoique cette pertinence est incomplète car les corrélations sont autour de 0,30 pour le corpus des œuvres du XVII^{e}siècle et autour de 0,65 pour le corpus Daniel (en valeurs absolues).

De plus, cette métrique semble montrer que le bruit dans les données textuelles est, contrairement ce que l'intuition en dirait, plus redondant que les données textuelles. Ceci pousse à penser que, dans le cas de l'océrisation de numérisation ou de l'extraction du contenu textuel de pages web, le bruit est moins stochastique que la langue elle-même. En clair, il est plus « facile » de prédire ce que le bruit sera plutôt que de prédire la suite d'un énoncé. Ceci est attesté par le signe des corrélations calculées.

Enfin, on remarque que les petites valeurs de n (unigrammes, bigrammes et trigrammes) sont les meilleurs descripteurs, ce qui pour le temps de calcul et l'espace de description est un avantage certain. Nous choisissons les unigrammes comme descripteurs à partir d'ici.

5.5 Apprentissage d'un modèle de prédiction du CER

Si nous avons vu que les taux T_{con} et T_{lex} et l'entropie peuvent constituer de « bons » estimateurs de la qualité d'une sortie d'OCR, ils ne sont pas les seuls. Ainsi pourrait-on penser aux suites trop longues de voyelles (ou de consonnes, ou de symboles, etc.). Constatant qu'il existe une pléthore d'indices, disons, un *faisceau*, on peut se proposer d'apprendre un modèle d'apprentissage automatique apprenant sur ces indices et tentant d'approcher la valeur du CER – calculée, rappelons-le encore ici, sur une référence transcrite manuellement.

Il s'agit d'un problème de régression linéaire car les valeurs prédites appartiennent à un ensemble continu, ou infini – contrairement à la régression logistique, laquelle a un ensemble de prédiction fini. Toutefois, dans cette section, nous n'allons pas chercher à estimer *précisemment* la *vraie* valeur du taux d'erreur. Rappelons que nous cherchons à évaluer des modèles d'OCR, ce qui signifie les comparer les uns aux autres pour tenter d'élir le plus pertinent. Dès lors, il faut et il suffit que le taux d'erreur prédit par le modèle soit *corrélé* au CER, sans pour autant que les CER prédits soient proches des CER réels.

Implémentation

Cette régression linéaire a été implémentée en utilisant la classe *LinearRegression* de la librairie *sklearn* (version 0.23.2) en Python 3. Les données utilisées sont celles présentées en section 5.2.4. L'ensemble des pages a servi à l'apprentissage et au test du modèle, avec un ratio de 20% pour réaliser le test. Cette partition a été réalisée en utilisant *train_test_split* de la libraire *sklearn* (avec le paramètre suivant : *random_state=5*).

L'ensemble des attributs (ou, *features* en anglais) calculés pour toutes les pages océrisées du corpus est décrit dans le tableau 5.12

Métriques d'évaluation

Erreur quadratique moyenne (MSE) L'erreur quadratique moyenne est la moyenne des carrés des différences entre les valeurs prédites et les valeurs attendues.

Coefficient de détermination (R2) La documentation dit que le coefficient de détermination R2 donne une « [...] indication de la qualité de l'ajustement et, par conséquent, une mesure de la façon dont les échantillons inconnus sont susceptibles d'être prédits par le modèle, à travers la proportion de variance expliquée. Comme cette variance dépend du jeu de données, R2 peut ne pas être véritablement comparable entre différents ensembles de données. Le meilleur score possible est de 1,0 et il peut être négatif (car le modèle peut être arbitrairement plus mauvais). Un modèle constant qui prédit toujours la valeur attendue de y

stylo	nbChars	Nombre de caractères
	nbWords	Nombre de mots
	nbSpaces	Nombre d'espaces
	nbMAJ	Nombre de majuscules
	nbMin	Nombre de minuscules
	nbPunct	Nombre de ponctuations
long-s	nbs	Nombre de s
	nbS	Nombre de f
prop	propSpaces	Proportion d'espaces
	propMAJ	Proportion de majuscules
	propMin	Proportion de minuscules
	propPunct	Proportion de ponctuations
prop-long-s	props	Proportion de s
	propS	Proportion de f
	valCon	Moyenne des taux de confiance
	valLex	Taux de lexicalité
	valEnt	Entropie

TABLE 5.12 – Description des attributs utilisés pour procéder à la régression linéaire cherchant à prédire le CER

[la valeur prédite], sans tenir compte des caractéristiques d'entrée, obtiendrait un score R2 de 0,0. »[40]

Coefficient de corrélation (et valeur p) Voir la section 5.2.2.

Résultats

Sur l'ensemble d'apprentissage, les résultats du tableau 5.13 montrent que, à l'exception de la valeur de lexicalité, tous les descripteurs sont bénéfiques au modèle (augmentation de R2 et diminution de la MSE).

Pour le modèle appris avec les descripteurs *stylo+long-s+prop+valCon+valLex+valEnt*, pour l'ensemble de test, on a $MSE = 10,50$, ce qui signifie qu'en moyenne, le carré des distances entre le CER prédit et le CER attendu est égal à 10,50. Ce même modèle donne une corrélation Pearson de 0,87 (avec $p = 1,9e^{-15}$, ce qui signifie qu'on a une très forte présomption contre l'hypothèse nulle) : le CER prédit et le CER attendu sont donc fortement corrélés.

40. *« [...] indication of goodness of fit and therefore a measure of how well unseen samples are likely to be predicted by the model, through the proportion of explained variance. As such variance is dataset dependent, R2 may not be meaningfully comparable across different datasets. Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a R2 score of 0.0. »* voir https://scikit-learn.org/stable/modules/model_evaluation.html#r2-score, consulté le 28 juin 2022.

	train (80 %)		*test (20 %)*		
	MSE	R2	MSE	R2	Pearson
stylo	34,00	0,31	16,18	-2,31	0,77 (2,7e-10)
stylo+long-s	33,71	0,32	16,36	-2,38	0,76 (7,1e-10)
stylo+prop	30,00	0,46	13,99	-1,48	0,93 (3e-20)
stylo+long-s+prop	38,69	0,51	11,21	-0,59	0,85 (1,2e-13)
stylo+long-s+prop+valCon	27,88	0,54	12,95	-1,12	0,88 (1,2e-15)
stylo+long-s+prop+valLex	28,69	0,51	11,20	-0,59	0,85 (7,5e-14)
stylo+long-s+prop+valCon+valLex	27,87	0,54	12,94	-1,12	0,87 (6,6e-15)
stylo+long-s+prop+valCon+valLex+valEnt	27,86	0,54	10,50	-0,39	0,87 (1,9e-15)
stylo+long-s+prop+valEnt	28,74	0,51	9,59	-0,16	0,85 (5,8e-14)

TABLE 5.13 – Résultats des évaluations des régressions linéaires estimants le CER

Par ailleurs, on remarque que le modèle *stylo+prop* atteint le maximum de corrélation avec $0,93$ (p-value inférieure à 0,01). C'est dire que les attributs stylométriques d'une part et leur proportion d'autre part sont de forts indices pour comparer les modèles d'OCR entre eux. Plus encore, ce constat montre que des attributs *simples* sont plus efficaces que des attributs calculés comme le taux de confiance moyen (qui n'est pas stable d'un logiciel a un autre), le taux de lexicalité (qui nécessite une ressource) ou encore l'entropie de Shannon.

On trouve en figure 5.22 une représentation graphique des CER prédits en fonction des CER attendus où l'on observe clairement cette corrélation (modèle *stylo+prop*).

Remarquons toutefois que le modèle contenant tous les attributs (*stylo+long-s+prop+valCon+valLex+valEnt*) constitue le meilleur compromis entre corrélation, R2 et MSE.

Résultats (cross-validation)

Les résultats présentés au-dessus sont très limités dans la mesure où les ensembles d'apprentissage et de test sont très limités : 182 et 46, respectivement. Pour améliorer ce point, on peut se proposer de mener la même expérience en procédant à une validation croisée[41]. Il s'agit d'une technique d'échantillonage consistant à diviser l'ensemble de toutes les données en k parties et à utiliser $k-1$ parties pour l'apprentissage et la partie restante pour le test ; cette opération est réalisée k fois en changeant la partie de test à chaque tour.

La comparaison des tableaux 5.14 et 5.15 montre que le modèle *stylo+long-s+prop+valCon+valLex+valEnt* est nettement plus stable au changement de données que le modèle *stylo+prop*. Nous concluons donc que : i) les attributs simples *stylo* et *prop* fournisse une quantité importante d'information dans la prédiction du taux d'erreur au caractère, ii) mais il est néanmoins informatif d'ajouter lors de l'apprentissage des attributs calculatoire nécessitant ou pas des ressources externes. Pour ces raisons, nous choisissons de garder le modèle

41. *cross-valisation*, en anglais.

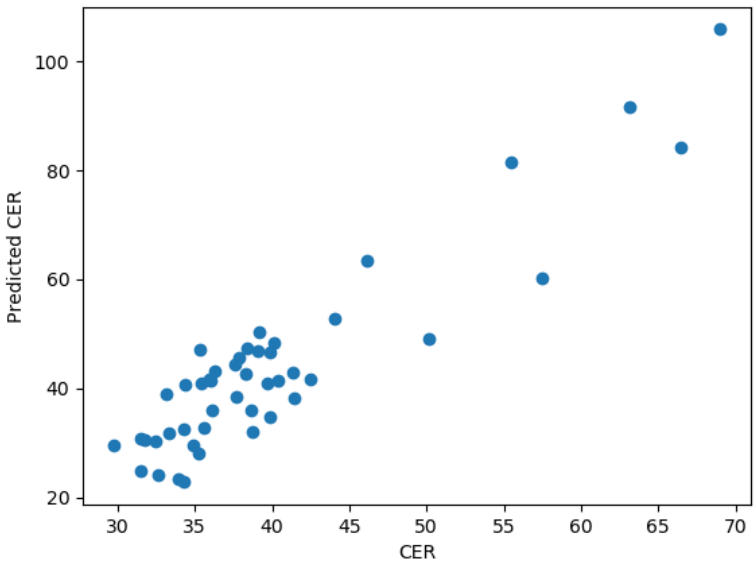


FIGURE 5.22 – Représentation graphique des CER prédit en fonction des CER attendus

Partitions	Corrélations (Pearson)
1	0,77
2	0,76
3	0,66
4	0,78
5	0,94

TABLE 5.14 – Corrélation Pearson pour cinq validations croisées sur l'ensemble des données (modèle *stylo+prop*)

Partitions	Corrélations (Pearson)
1	0,79
2	0,91
3	0,76
4	0,92
5	0,86

TABLE 5.15 – Corrélation Pearson pour cinq validations croisées sur l'ensemble des données (modèle *stylo+long-s+prop+valCon+valLex+valEnt*)

stylo+long-s+prop+valCon+valLex+valEnt pour comparer, sur le corpus des mazarinades, différents logiciels et modèles d'OCR.

Étude comportementale

Appliquons le modèle appris sur toutes les données du corpus précédent avec les descripteurs *stylo+long-s+prop+valCon+valLex+valEnt* au corpus des mazarinades burlesques.

Observons les pages qui ont un CER prédit supérieur à 100. Nous prenons cette valeur arbitrairement mais selon deux logiques : d'une part, un CER élevé limite le nombre de pages à étudier et, d'autre part, on s'attend à observer des pages avec un fort nombre d'insertions (cf. l'équation 5.3).

Parmi les 295 pages concernées, on compte :

— 220 pages contenant des données textuelles ;
— 145 premières pages, avec plusieurs fontes, plusieurs tailles, des gravures ;
— 136 pages contenant une gravure ou un tampon ;
— 65 pages vides ;
— 18 pages à double colonnage ;
— 11 pages sur lesquelles on observe la page précédente qui a déteint ;
— 6 pages de reliure ;
— 4 pages où l'on observe une main ;
— 3 pages contenant un grand portrait.

En outre, pour les dix pages ayant un CER prédit compris entre 0 et 10, l'exercice de lecture est non seulement possible mais aussi presque aussi aisé qu'avec une transcription manuelle.

Nous donnons trois exemples de pages (accompagnées de leur OCR), prises au hasard, respectant les contraintes suivantes :

— CER prédit inférieur à 10 (figure 5.23) ;
— CER prédit aux alentours de 50 (figure 5.24) ;
— CER prédit supérieur à 100 (figure 5.25).

5.6 Conclusion

L'étude des indices, qu'on qualifiera de « faisceaux », offrant de l'information sur la quantité d'erreurs issues de l'OCR permet, de manière affirmative, d'approcher automatiquement les valeurs relatives des CER. *Approcher* car il s'agit d'un modèle prédictif et non du calcul lui-même ; et valeurs *relatives* car ce n'est pas tant la valeur même de la prédiction qui est porteuse d'information mais bien cette valeur relativement à d'autres valeurs. Ainsi le modèle appris permet-il de répondre à la question : entre le modèle *ML1* et le *ML2*, lequel est le plus fidèle à la numérisation ?

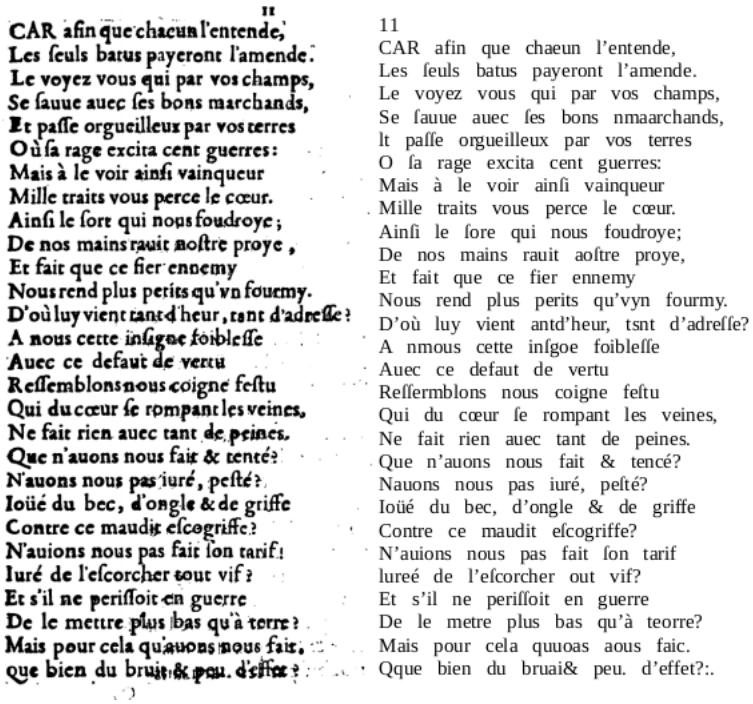
Il convient donc de proposer à l'étude l'océrisation d'une petite collection de mazarinades, de les océriser selon plusieurs modèles d'OCR et de définir celui qui est le plus performant pour de telles numérisations anciennnes. Ceci est l'objet du chapitre suivant.

de la Guerre de Paris. 9

Sçauoir que pour leuer ſoldats,
Tant de pied comme ſur dadas,
L'on taxéreit toutes les portes,
Petites, grandes, foibles fortes.
Que la Cochere fournisoit
Tant que le bloeus dureroit,
Vn bon cheual auec vn homme,.
Ou qu'elle donneroit la ſomme
De quinze piſtolles de poids,
Payables pour la premiere fois:;
Les petites, vn Mouſquetaire,
Ou trois piſtoles pour en faire:
Hommes chez le Marchand ſortans
Et tout fins neufs, & tous battans.
Ce iour en leuant ſa bequille
Le Gouuerneur de la Baſtille,
Qu'on ñmoit Monſieur du Tremblay,
Luy qui iamais n'auoit tremblé,
Vieil Soldat & vieil Gentil-homme,
A Monſieur d'Elbeuf qui le ſomme
De luy remettre ce Chaſteau,
Reſpoudit tres-bien & tres beau
Qu'il ne luy plaiſoit de le rendre,
Et qu'il pretendoit le deffendre.
Mais il ne fut pas ſi méchant.
Queſix canons deſſus le champ
Ne nous ouurirent cette place.
Sans auoir touché la ſurface:¬
Ce n'eſt pas qu'ils ne fiſſent pouf,¬
Que la Garniſon ne diſt ouf,
Qu'elle ne paruſt ſur la breche,
Qu'elle n'employaſt poudre & mécha,
Que maint coup ne fuſt entendu;
Mais c'eſt qu'il eſtoit deffendu
Que dans ce beau ſiege de balle
Aucun coſté chargeaſt à balle
Qu'il neuſt cré, Retirez vous,
Autant pour eux comme pourdous,
Sur les meſmes peines quion donne
Au meurtrier d'vne perſonne,
Car quiconque euſt fait autrement
Auroit, peché mortellement
Tout autant qu'en va homicide.
Vn homme moins vaillant qu'Alcide,.
Mais certes plus homme d'honneur,.
Brouſſel, en fut fait Gouuerneur,A
Et ſon fils en cette occurence
Fut pourueu de la Lieutenance..
9

Le Mercredy mis ſur pied futs. Ian¬¬
Le premier Regiment qu'on eut:
Sur pied, non i'apperçoy que i'erre,
Les pieds n'en touchoent point à terre;
Nos guerriers eſtoient ſur cheuaux
Preſts à fuyr deuant les Royaux.
Ce fut cette meſme iournée Arriu'e
ds De
Qu'vne petite haquenée
Apporta de noſtre coſté
Alxandre reſſuſciné,
de Beaau
fort.
Ce grand Beaufort dont la preſence
Nous rendit beaucoup d'aſſeurance,
Ce Heros, ce fils de Henry
Ce braue, ce Prince aguerry,
Iuſques chez Renard, redoutable.
Ennemy juré de la table,
Ce fameux fleau des lerzais
Quand ils caufent comme des iais,
Ce mars qui bat, qui rompt, qui frappe,
Et perce tout iuſqu'à la nappe;
Ce Prince plus blond qu'vn baſſin
Et plus deuot qu'vn Capucin,
Qui mit en rut toutes nos femmes
Les honneſtes & les infames,
Baiſa toujours & rebaiſa,.
Car iamais il ne refuſa
Ny Harangere ny Marchande,
leune, vieille, laide, galande,
Qui luy cryoit à qui plus fort
Baiez my Monſieur de Biaufort,
L'vne tendoit vn vilain mouſle:.
L'autre ne montroit que trois dents,
Dont l'ebenne eſtoit ſuffiſante
Pour en faire plus de ciaquante.
Il en baiſa prés detrois cent
Toutes d'vn baiſer innocent,
Forsvne ieune femme groſſe
Qui deſcendit de ſon Caroſſe,
Diſnt, mon fruict feroit marqué;
Car dans le baiſer applicqué
Au milieu de ſa belle bouche,
Il eut vn deſir de ſa couche,
Et luy demanda rendezvous,
En la baiſant deux autres coups:
Mais il fut depuis à confeſſe:
Enfin ayant baiſé ſans ceſſe
¬ ¬¬¬

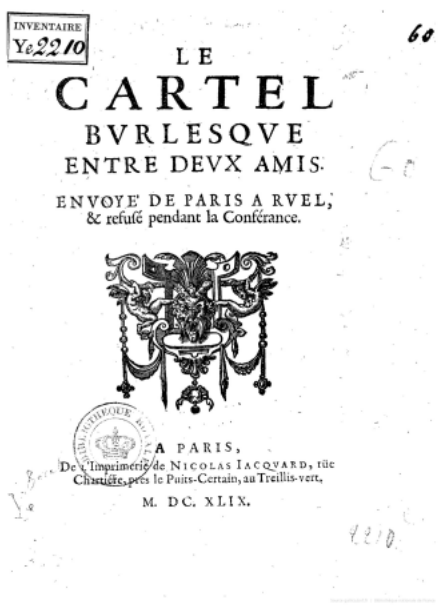
FIGURE 5.23 – Page 16 du *Courrier burlesque de la guerre de Paris, envoyé à monseigneur le prince de Condé...*, Moreau 814, 1650, Anvers, avec $CER_{predit} = 5,5$



11

CAR afin que chaeun l'entende,
Les ſeuls batus payeront l'amende.
Le voyez vous qui par vos champs,
Se ſauue auec ſes bons nmaarchands,
lt paſſe orgueilleux par vos terres
O ſa rage excita cent guerres:
Mais à le voir ainſi vainqueur
Mille traits vous perce le cœur.
Ainſi le ſore qui nous foudroye;
De nos mains rauit aoſtre proye,
Et fait que ce fier ennemy
Nous rend plus perits qu'vyn fourmy.
D'où luy vient antd'heur, tsnt d'adreſſe?
A nmous cette inſgoe foibleſſe
Auec ce defaut de vertu
Reſſermblons nous coigne feſtu
Qui du cœur ſe rompant les veines,
Ne fait rien auec tant de peines.
Que n'auons nous fait & tencé?
Nauons nous pas iuré, peſté?
Ioüé du bec, d'ongle & de griffe
Contre ce maudit eſcogriffe?
N'auions nous pas fait ſon tarif
lureé de l'eſcorcher out vif?
Et s'il ne periſſoit en guerre
De le metre plus bas qu'à teorre?
Mais pour cela quuoas aous faic.
Qque bien du bruai& peu. d'effet?:.

FIGURE 5.24 – Page 12 du *Stratagème, ou le Pour et le contre du départ de Mazarin, en vers burlesques.*, Moreau 3720, 1652, s.l., avec $CER_{predit} = 50,4$


INVENTAIRE
Ye 2210

60

LE
CARTEL
BVRLESQVE
ENTRE DEVX AMIS.
ENVOYE' DE PARIS A RVEL,
& refuſé pendant la Conférance.

A PARIS,
De l'Imprimerie de NICOLAS IACQVARD, tüe
Chartiére, prés le Puits-Certain, au Treillis-vert,
M. DC. XLIX.

m7
le
LE
CARTE L
D ¬
4
6
BVRLESQVE
ENTRE DEVX AMIS.
ENVOYE' DE PARIS A RVEL,
& refuſé pendant la Conférance.
An
e8
A PA1S,
n iruneislil e e ls l l e oe e no in
Chartiete, preés le Puits-Certain, au Treillis-vei
M. DC. XLIX.

FIGURE 5.25 – Page 3 du *Cartel burlesque entre deux amis, envoyé de Paris à Ruel et refusé pendant la conférence.*, Moreau 643, 1649, Paris, avec $CER_{predit} = 183,9$

Chapitre 6

Océriser des documents historiques

Sommaire

Ce chapitre se concentrera sur un corpus constitué non pas par cohérence mais par nécessité, lequel étant le résultat de la première livraison de numérisation du prestataire engagé. Il s'agit de 425 mazarinades qui ont été sélectionnées par les bibliothécaires de la Mazarine en respectant deux critères : i) les items numérisés ne l'ont pas déjà été et ii) un jeu complet de métadonnées leur est associé.

Nous proposons ensuite une comparaison de plusieurs modèles d'OCR sur un ensemble d'environ 500 mazarinades. Ceci permettra d'observer *en pratique* le caractère « bruité » de ce type de données.

6.1 Océrisation de mazarinades

6.1.1 Données

Les données océrisées dont il est question dans cette section constitue un ensemble de numérisations qui n'est pas un corpus au sens linguistique du terme, notamment au sens de Rastier. Il s'agit du premier lot de mazarinades envoyé par la Bibliothèque Mazarine au groupe *Puce & Plume*. Ces pièces ont été sélectionnées par la Mazarine car elles étaient référencées dans la *Mazarinum* et disposaient ainsi de métadonnées complètes.

425 items ont été numérisés et rendus disponibles par le groupe *Puce & Plume* dans plusieurs formats : PDF, JPEG et TIFF. Un dossier ALTO est aussi disponible, lequel contient les sorties des OCR de Tesseract 4.0. Au total, on dispose de 425 PDF, 11 087 JPEG, TIFF et ALTO.

6.1.2 Méthode

Dans cette section, nous souhaitons procéder à une océrisation raisonnée de cet ensemble de mazarinades. *Raisonnée* car nous proposons à l'étude la comparaison de deux logiciels d'OCR – Kraken et Tesseract – selon plusieurs paramétrages. L'océrisation étant un processus faisant intervenir beaucoup de paramètres pouvant influencer la qualité des données textuelles en sortie, nous comparons et contrôlons étape par étape la qualité – estimée – des sorties d'OCR.

Nous souhaitons procéder à une étude du genre de celle menée par [Burgy et al., 2020]. Dans ce projet, les auteurs océrisent une collection de documents historiques (collection *De Bry*, XVIe et XVIIe siècle, 29 livres). Ils comparent quatre logiciels libres d'OCR (Tesseract, Kraken, Calamari et OCR4all) selon les modèles standard, mais aussi en paramétrant les modèles et les paramètres et enfin en faisant des prétraitements des images (comme la binarisation et le rognage des bords).

6.1.3 Évaluation

L'évaluation de la qualité relative des différents modèles et paramétrages est réalisée à deux niveaux, un niveau macroscopique et un niveau microscopique. Le niveau macroscopique correspond à l'ensemble des pages du lot 1, c'est-à-dire les 11 087 images. Néanmoins nous ne disposons pas de transcription de référence, ou, vérité de terrain, pour ces pages numérisées. Nous allons donc procéder à une estimation de la qualité des OCR comme présenté en chapitre 5 : une prédiction de taux d'erreur au caractère par un modèle d'apprentissage automatique.

Niveau microscopique ensuite, en procédant à la transcription de vingt pages du corpus sélectionnées aléatoirement, avec la contrainte de contenir des données textuelles. Ces pages ont été transcrites diplomatiquement, c'est-à-dire en restant fidèle à ce qui apparaît sur la numérisation (pas de modernisation par exemple). L'objectif est de construire une vérité de terrain, certes très restreinte, mais permettant le calcul réel du taux d'erreur au caractère pour valider ou réfuter

les phénomènes apparaissant au niveau macroscopique. En moyenne, 13 minutes ont été nécessaires pour transcrire ces pages, avec un écart-type très grand car cette moyenne cache bien des disparités, notamment :

— certaines pages contiennent beaucoup plus de données textuelles que d'autres (temps de transcription allant de 3 minutes à 26) ;
— deux pages sont en latin, et la transcription d'une langue qu'on ne maîtrise pas est beaucoup plus longue.

Une remarque, nuançant la qualité de cette vérité de terrain, doit être faite : un seul annotateur l'a produit, sans qu'aucun accord inter-annotateur ne puisse être calculé. Néanmoins, un second annotateur a corrigé les transcriptions, en passant en moyenne dix minutes par pages. Les erreurs étaient très restreintes, de l'ordre de un ou deux caractères par page.

6.1.4 Océrisation

Si [Burgy et al., 2020] testent quatre logiciels libres, nous proposons ici une réduction des logiciels testés à Kraken et Tesseract car le premier dispose d'un modèle entraîné sur des données du XVII^e^ siècle en français et le second est le logiciel libre le plus utilisé actuellement. Comme les auteurs ne sélectionnent pas Calamari (tous leurs output étaient vides, aucune reconnaissance n'avait été réalisée) ni OCR4all (qui présentait de bons résultats mais avait un défaut – le même que Kraken – qui était de segmenter à l'infini les pages ne contenant pas de donnée textuelle), nous ne les testons pas ici.

Kraken

Pour le logiciel Kraken (version 2.0.8) [Kiessling, 2019], deux modèles ont été testés : un très adapté *a priori* car appris sur des données textuelles en français du XVII^e^ siècle [1] et un second non adapté car appris pour l'anglais contemporain [2]. Les fichiers de sortie de Kraken sont des fichiers HTML.

Comme [Burgy et al., 2020], on remarque que le segmenteur de Kraken tente de segmenter toutes les pages, même les pages de reliures ou les pages vides. N'ayant pas de donnée textuelle dans ces pages, le segmenteur ne s'en sort pas faisant ralentir extrêmement le processus (boucle infinie). Pour cela, nous avons dû fixer un *timeout* [3] à une minute pour dépasser ces pages dans le processus en chaîne d'océrisation du Lot 1.

Avec le modèle pour le français du XVII^e^ siècle, on dispose à la fin du processus de 7 323 fichiers de sortie ; avec le modèle par défaut, on dispose de 7 137 fichiers de sortie. On comprend donc bien que cette méthode du *timeout* est largement sous-optimale dans la mesure où l'on ne peut savoir avec assurance si les pages qui prendront plus d'une minute de traitement seront les vides ou les pages de reliure. Preuve en est, l'utilisation d'un modèle différent implique un

1. [Gabay, 2020, Gabay, 2019]
2. Modèle par défaut.
3. Un *timeout* est une commande qui permet d'arrêter un processus si celui-ci dépasse le temps renseigné en paramètre.

nombre de fichiers de sortie différent, ce qui montre qu'il est très probable que des pages contenant des données textuelles n'ont pas été océrisées ou bien que des pages vides ou de reliure l'ont été. Avant même de regarder les performances de ces modèles sur le corpus qui est le notre à ce moment de l'étude, le logiciel KRAKEN ne semble pas la meilleure option car très chronophage (plusieurs jours pour les 425 items du corpus) et peu robuste aux pages vides et de reliure.

Tesseract

À confronter à KRAKEN, nous choisissons le logiciel TESSERACT dans sa version 4.0. Celui-ci dispose de plusieurs modèles, dont deux que nous avons sélectionnés : le modèle *eng* (anglais) et le modèle *fra* (français). Les formats de sortie sont en ALTO-XML.

TESSERACT donne 10 966 fichiers de sortie pour les deux modèles, le reste seulement étant selon lui des pages vides. Le temps d'océrisation pour ces quelques 10 000 pages est de quelques heures ; TESSERACT n'a donc pas le probème du segmenteur de KRAKEN et ne boucle pas à l'infini, ce qui est un réel avantage.

6.1.5 Résultats

Évaluation sur l'échantillon de vingt pages

Ces vingt pages ayant été transcrites manuellement, il est possible de calculer le taux d'erreur au caractère moyen par page, le CER. Ceci nous donne une idée très précise de la qualité des sorties d'OCR sur l'échantillon en question.

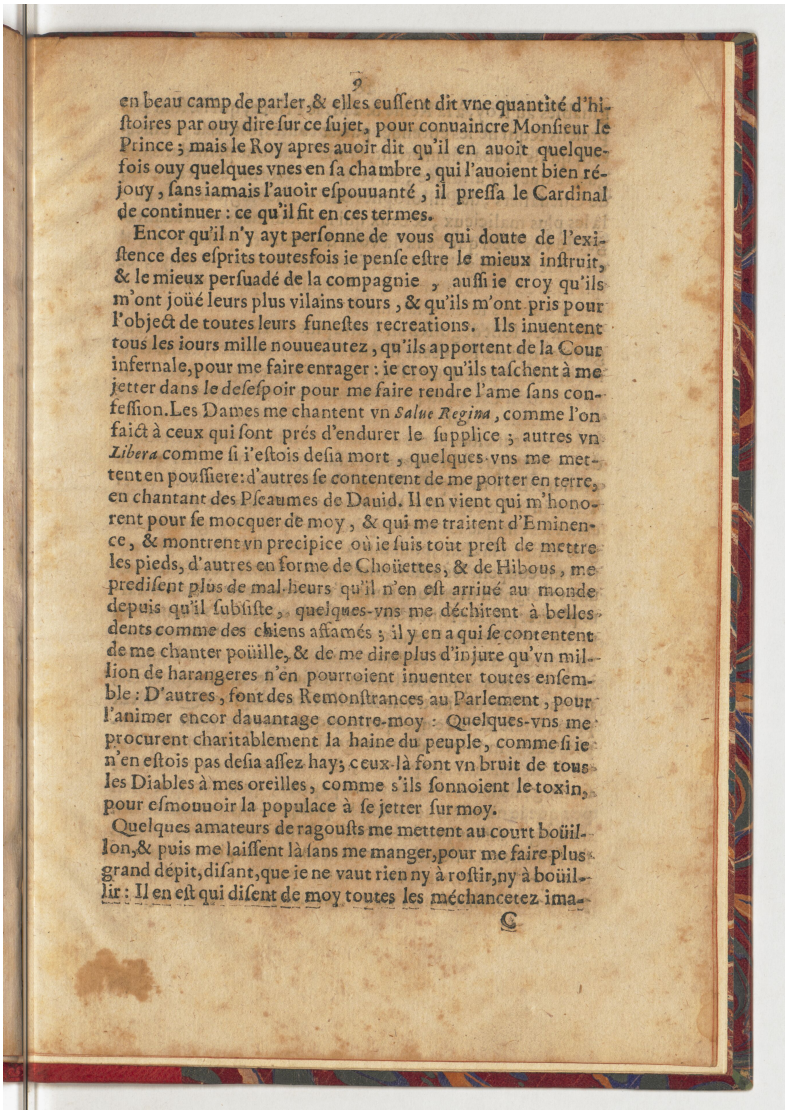
Pour KRAKEN, deux modèles ont été testés (le modèle par défaut appris sur des données en anglais contemporain et le modèle appris sur des données en français du XVII^{e} siècle). Comme pour chaque page deux formats sont disponibles, le JPEG et le TIFF, nous réalisons quatre tests avec le logiciel KRAKEN.

Pour TESSERACT, nous testons aussi deux modèles sur les JPEG et les TIFF, le modèle anglais et le modèle français (tous les deux ont été appris sur des données textuelles contemporaines). Nous testons aussi plusieurs pré-traitements des images (voir les figures 6.1, 6.2, 6.3, 6.4, 6.5 et 6.6), réalisés avec la librairie Python *CV2* (version 4.5.1) :

1. *grayscale*, nuances de gris (figure 6.2) ;
2. *thresholding*, binarisation (figure 6.3) ;
3. *remove noise*, suppression du bruit dans l'image par floutage (figure 6.4) ;
4. *opening*, érosion et dilatation des formes (figure 6.5) ;
5. *deskew*, redressement de l'image (figure 6.6).

Les tableaux 6.1 et 6.2 montrent les taux moyens d'erreur au caractère par page ainsi que les moyennes calculées sur l'ensemble des pages de l'échantillon. Plusieurs remarques.

1. D'abord, l'utilisation des images TIFF est souvent récompensée par de plus faibles CER (ce qui correspond à une meilleure qualité), sans toutefois que cela soit systématique. Parfois même, l'utilisation des JPEG



9

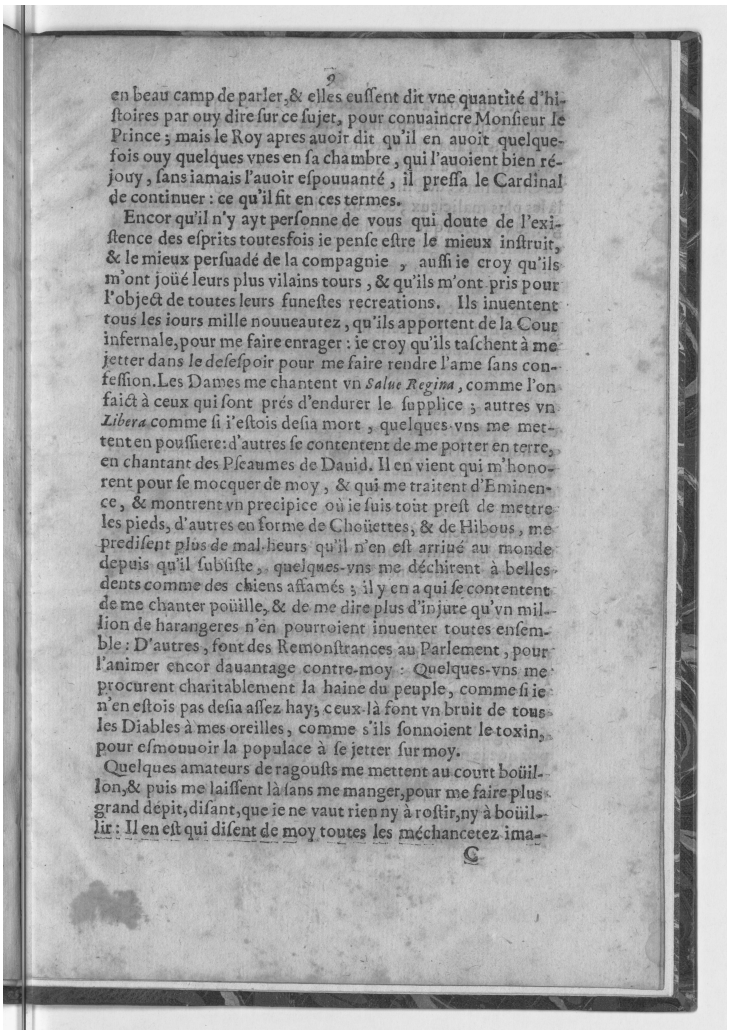
en beau camp de parler, & elles euſſent dit vne quantité d'hiſtoires par ouy dire ſur ce ſujet, pour conuaincre Monſieur le Prince ; mais le Roy apres auoir dit qu'il en auoit quelquefois ouy quelques vnes en ſa chambre, qui l'auoient bien réjouy, ſans iamais l'auoir eſpouuanté, il preſſa le Cardinal de continuer : ce qu'il fit en ces termes.

Encor qu'il n'y ayt perſonne de vous qui doute de l'exiſtence des eſprits toutesfois ie penſe eſtre le mieux inſtruit, & le mieux perſuadé de la compagnie, auſſi ie croy qu'ils m'ont joüé leurs plus vilains tours, & qu'ils m'ont pris pour l'object de toutes leurs funeſtes recreations. Ils inuentent tous les iours mille nouueautez, qu'ils apportent de la Cour infernale, pour me faire enrager : ie croy qu'ils taſchent à me jetter dans le deſeſpoir pour me faire rendre l'ame ſans confeſſion. Les Dames me chantent vn *Salue Regina*, comme l'on faict à ceux qui ſont prés d'endurer le ſupplice ; autres vn *Libera* comme ſi i'eſtois deſia mort, quelques-vns me mettent en pouſſiere : d'autres ſe contentent de me porter en terre, en chantant des Pſeaumes de Dauid. Il en vient qui m'honorent pour ſe mocquer de moy, & qui me traitent d'Eminence, & montrent vn precipice où ie ſuis tout preſt de mettre les pieds, d'autres en forme de Choüettes, & de Hibous, me prediſent plus de mal-heurs qu'il n'en eſt arriué au monde depuis qu'il ſubſiſte, quelques-vns me déchirent à belles dents comme des chiens affamés ; il y en a qui ſe contentent de me chanter poüille, & de me dire plus d'injure qu'vn million de harangeres n'en pourroient inuenter toutes enſemble : D'autres, font des Remonſtrances au Parlement, pour l'animer encor dauantage contre-moy : Quelques-vns me procurent charitablement la haine du peuple, comme ſi ie n'en eſtois pas deſia aſſez hay ; ceux-là font vn bruit de tous les Diables à mes oreilles, comme s'ils ſonnoient le toxin, pour eſmouuoir la populace à ſe jetter ſur moy.

Quelques amateurs de ragouſts me mettent au court boüillon, & puis me laiſſent là ſans me manger, pour me faire plus grand dépit, diſant, que ie ne vaut rien ny à roſtir, ny à boüillir : Il en eſt qui diſent de moy toutes les méchancetez ima-

C

FIGURE 6.1 – *Image originale*



9

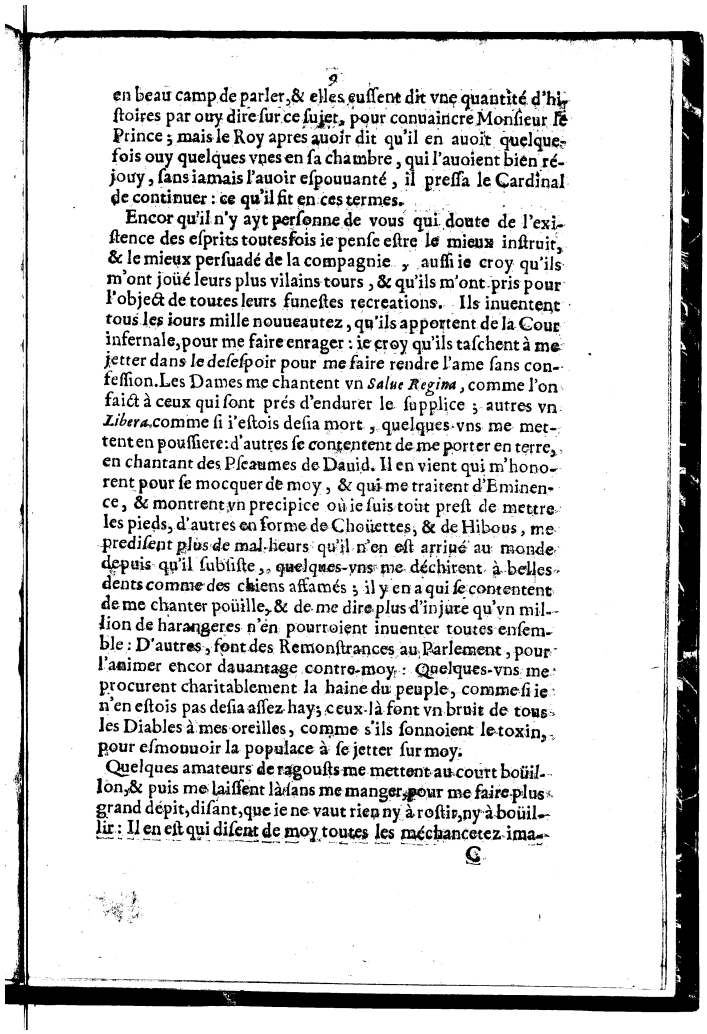
en beau camp de parler, & elles euſſent dit vne quantité d'hiſtoires par ouy dire ſur ce ſujet, pour conuaincre Monſieur le Prince ; mais le Roy apres auoir dit qu'il en auoit quelquefois ouy quelques vnes en ſa chambre, qui l'auoient bien réjouy, ſans iamais l'auoir eſpouuanté, il preſſa le Cardinal de continuer : ce qu'il fit en ces termes.

Encor qu'il n'y ayt perſonne de vous qui doute de l'exiſtence des eſprits toutesfois ie penſe eſtre le mieux inſtruit, & le mieux perſuadé de la compagnie, auſſi ie croy qu'ils m'ont joüé leurs plus vilains tours, & qu'ils m'ont pris pour l'object de toutes leurs funeſtes recreations. Ils inuentent tous les iours mille nouueautez, qu'ils apportent de la Cour infernale, pour me faire enrager : ie croy qu'ils taſchent à me jetter dans le deſeſpoir pour me faire rendre l'ame ſans confeſſion. Les Dames me chantent vn *Salue Regina*, comme l'on faict à ceux qui ſont prés d'endurer le ſupplice ; autres vn *Libera* comme ſi i'eſtois deſia mort, quelques-vns me mettent en pouſſiere : d'autres ſe contentent de me porter en terre, en chantant des Pſeaumes de Dauid. Il en vient qui m'honorent pour ſe mocquer de moy, & qui me traitent d'Eminence, & montrent vn precipice où ie ſuis toût preſt de mettre les pieds, d'autres en forme de Choüettes, & de Hibous, me prediſent plus de mal-heurs qu'il n'en eſt arriué au monde depuis qu'il ſubſiſte, quelques-vns me déchirent à belles dents comme des chiens affamés ; il y en a qui ſe contentent de me chanter poüille, & de me dire plus d'injure qu'vn million de harangeres n'en pourroient inuenter toutes enſemble : D'autres, font des Remonſtrances au Parlement, pour l'animer encor dauantage contre-moy : Quelques-vns me procurent charitablement la haine du peuple, comme ſi ie n'en eſtois pas deſia aſſez hay ; ceux-là font vn bruit de tous les Diables à mes oreilles, comme s'ils ſonnoient le toxin, pour eſmouuoir la populace à ſe jetter ſur moy.

Quelques amateurs de ragouſts me mettent au court boüillon, & puis me laiſſent là ſans me manger, pour me faire plus grand dépit, diſant, que ie ne vaut rien ny à roſtir, ny à boüillir : Il en eſt qui diſent de moy toutes les méchancetez ima-

C

FIGURE 6.2 – *grayscale*



9

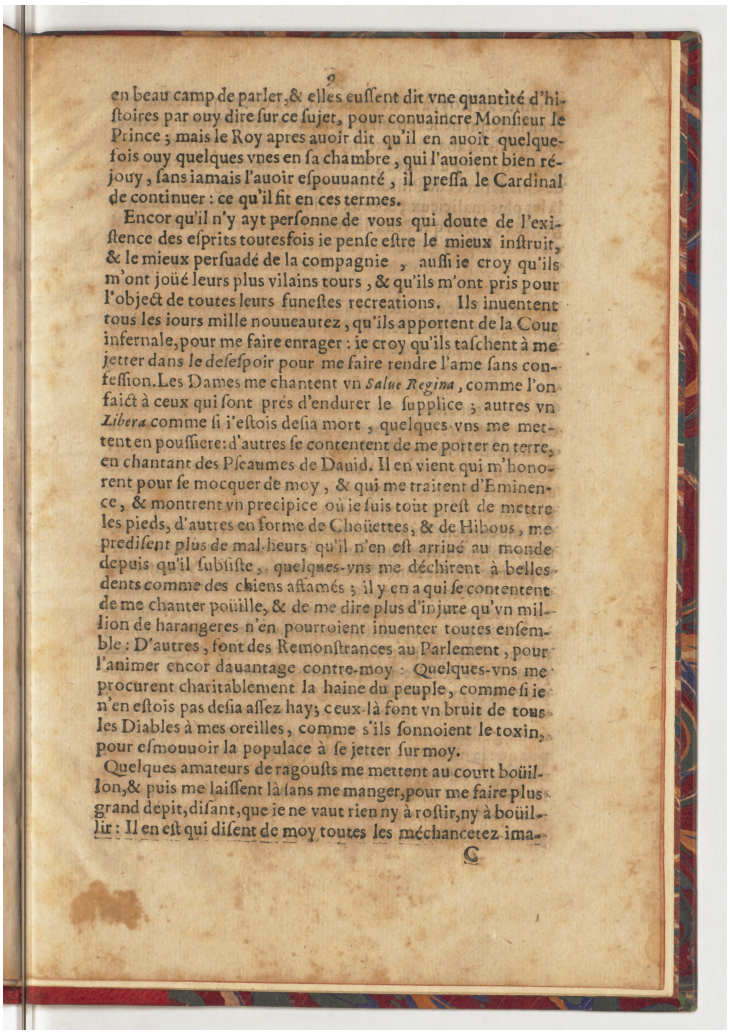
en beau camp de parler, & elles euſſent dit vne quantité d'hiſtoires par ouy dire ſur ce ſujet, pour conuaincre Monſieur le Prince ; mais le Roy apres auoir dit qu'il en auoit quelquefois ouy quelques vnes en ſa chambre, qui l'auoient bien réjouy, ſans iamais l'auoir eſpouuanté, il preſſa le Cardinal de continuer : ce qu'il fit en ces termes.

Encor qu'il n'y ayt perſonne de vous qui doute de l'exiſtence des eſprits toutesfois ie penſe eſtre le mieux inſtruit, & le mieux perſuadé de la compagnie, auſſi ie croy qu'ils m'ont joüé leurs plus vilains tours, & qu'ils m'ont pris pour l'object de toutes leurs funeſtes recreations. Ils inuentent tous les iours mille nouueautez, qu'ils apportent de la Cour infernale, pour me faire enrager : ie croy qu'ils taſchent à me jetter dans le deſeſpoir pour me faire rendre l'ame ſans confeſſion. Les Dames me chantent vn *Salue Regina*, comme l'on faict à ceux qui ſont prés d'endurer le ſupplice ; autres vn *Libera*, comme ſi i'eſtois deſia mort, quelques-vns me mettent en pouſſiere : d'autres ſe contentent de me porter en terre, en chantant des Pſeaumes de Dauid. Il en vient qui m'honorent pour ſe mocquer de moy, & qui me traitent d'Eminence, & montrent vn precipice où ie ſuis tout preſt de mettre les pieds, d'autres en forme de Choüettes, & de Hibous, me prediſent plus de mal-heurs qu'il n'en eſt arriué au monde depuis qu'il ſubſiſte, quelques-vns me déchirent à belles dents comme des chiens affamés ; il y en a qui ſe contentent de me chanter poüille, & de me dire plus d'injure qu'vn million de harangeres n'en pourroient inuenter toutes enſemble : D'autres, font des Remonſtrances au Parlement, pour l'animer encor dauantage contre-moy : Quelques-vns me procurent charitablement la haine du peuple, comme ſi ie n'en eſtois pas deſia aſſez hay ; ceux-là font vn bruit de tous les Diables à mes oreilles, comme s'ils ſonnoient le toxin, pour eſmouuoir la populace à ſe jetter ſur moy.

Quelques amateurs de ragouſts me mettent au court boüillon, & puis me laiſſent là ſans me manger, pour me faire plus grand dépit, diſant, que ie ne vaut rien ny à roſtir, ny à boüillir : Il en eſt qui diſent de moy toutes les méchancetez ima-

C

FIGURE 6.3 – *thresholding*



9

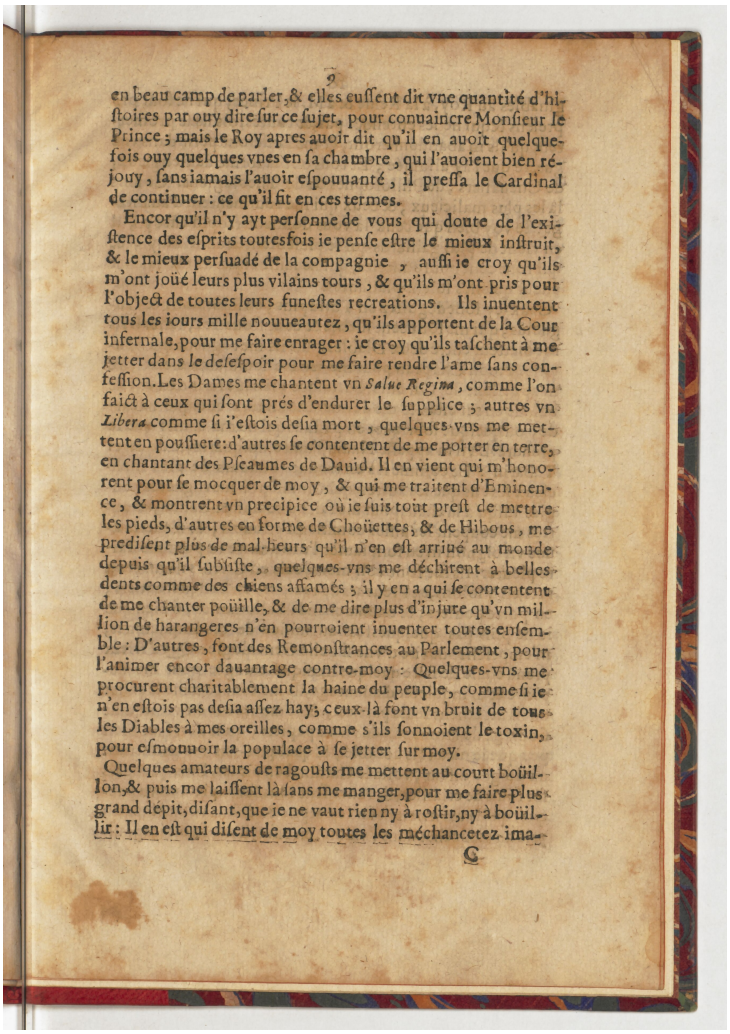
en beau camp de parler, & elles euſſent dit vne quantité d'hiſtoires par ouy dire ſur ce ſujet, pour conuaincre Monſieur le Prince ; mais le Roy apres auoir dit qu'il en auoit quelquefois ouy quelques vnes en ſa chambre, qui l'auoient bien réjouy, ſans iamais l'auoir eſpouuanté, il preſſa le Cardinal de continuer : ce qu'il fit en ces termes.

Encor qu'il n'y ayt perſonne de vous qui doute de l'exiſtence des eſprits toutesfois ie penſe eſtre le mieux inſtruit, & le mieux perſuadé de la compagnie, auſſi ie croy qu'ils m'ont joüé leurs plus vilains tours, & qu'ils m'ont pris pour l'object de toutes leurs funeſtes recreations. Ils inuentent tous les iours mille nouueautez, qu'ils apportent de la Cour infernale, pour me faire enrager : ie croy qu'ils taſchent à me jetter dans le deſeſpoir pour me faire rendre l'ame ſans confeſſion. Les Dames me chantent vn *Salue Regina*, comme l'on faict à ceux qui ſont prés d'endurer le ſupplice ; autres vn *Libera* comme ſi i'eſtois deſia mort, quelques-vns me mettent en pouſſiere : d'autres ſe contentent de me porter en terre, en chantant des Pſeaumes de Dauid. Il en vient qui m'honorent pour ſe mocquer de moy, & qui me traitent d'Eminence, & montrent vn precipice où ie ſuis toût preſt de mettre les pieds, d'autres en forme de Choüettes, & de Hibous, me prediſent plus de mal-heurs qu'il n'en eſt arriué au monde depuis qu'il ſubſiſte, quelques-vns me déchirent à belles dents comme des chiens affamés ; il y en a qui ſe contentent de me chanter poüille, & de me dire plus d'injure qu'vn million de harangeres n'en pourroient inuenter toutes enſemble : D'autres, font des Remonſtrances au Parlement, pour l'animer encor dauantage contre-moy : Quelques-vns me procurent charitablement la haine du peuple, comme ſi ie n'en eſtois pas deſia aſſez hay ; ceux-là font vn bruit de tous les Diables à mes oreilles, comme s'ils ſonnoient le toxin, pour eſmouuoir la populace à ſe jetter ſur moy.

Quelques amateurs de ragouſts me mettent au court boüillon, & puis me laiſſent là ſans me manger, pour me faire plus grand dépit, diſant, que ie ne vaut rien ny à roſtir, ny à boüillir : Il en eſt qui diſent de moy toutes les méchancetez ima-

C

FIGURE 6.4 – *remove noise*



9

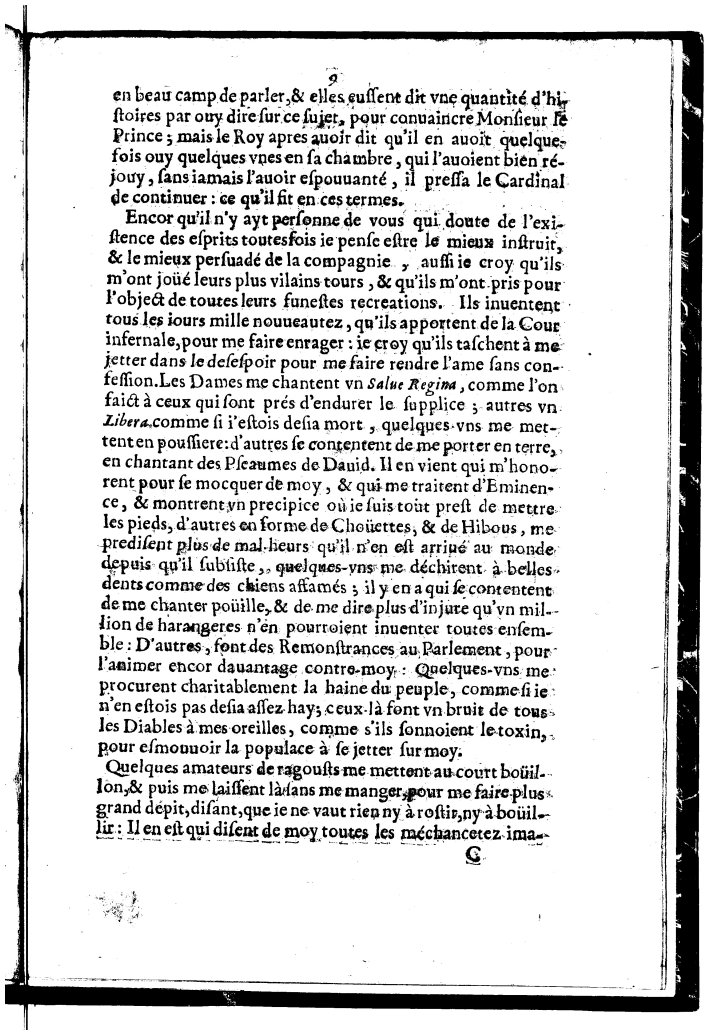
en beau camp de parler, & elles euſſent dit vne quantité d'hiſtoires par ouy dire ſur ce ſujet, pour conuaincre Monſieur le Prince ; mais le Roy apres auoir dit qu'il en auoit quelquefois ouy quelques vnes en ſa chambre, qui l'auoient bien réjouy, ſans iamais l'auoir eſpouuanté, il preſſa le Cardinal de continuer : ce qu'il fit en ces termes.

Encor qu'il n'y ayt perſonne de vous qui doute de l'exiſtence des eſprits toutesfois ie penſe eſtre le mieux inſtruit, & le mieux perſuadé de la compagnie, auſſi ie croy qu'ils m'ont joüé leurs plus vilains tours, & qu'ils m'ont pris pour l'object de toutes leurs funeſtes recreations. Ils inuentent tous les iours mille nouueautez, qu'ils apportent de la Cour infernale, pour me faire enrager : ie croy qu'ils taſchent à me jetter dans le deſeſpoir pour me faire rendre l'ame ſans confeſſion. Les Dames me chantent vn *Salue Regina*, comme l'on faict à ceux qui ſont prés d'endurer le ſupplice ; autres vn *Libera* comme ſi i'eſtois deſia mort, quelques-vns me mettent en pouſſiere : d'autres ſe contentent de me porter en terre, en chantant des Pſeaumes de Dauid. Il en vient qui m'honorent pour ſe mocquer de moy, & qui me traitent d'Eminence, & montrent vn precipice où ie ſuis tout preſt de mettre les pieds, d'autres en forme de Choüettes, & de Hibous, me prediſent plus de mal-heurs qu'il n'en eſt arriué au monde depuis qu'il ſubſiſte, quelques-vns me déchirent à belles dents comme des chiens affamés ; il y en a qui ſe contentent de me chanter poüille, & de me dire plus d'injure qu'vn million de harangeres n'en pourroient inuenter toutes enſemble : D'autres, font des Remonſtrances au Parlement, pour l'animer encor dauantage contre-moy : Quelques-vns me procurent charitablement la haine du peuple, comme ſi ie n'en eſtois pas deſia aſſez hay ; ceux-là font vn bruit de tous les Diables à mes oreilles, comme s'ils ſonnoient le toxin, pour eſmouuoir la populace à ſe jetter ſur moy.

Quelques amateurs de ragouſts me mettent au court boüillon, & puis me laiſſent là ſans me manger, pour me faire plus grand dépit, diſant, que ie ne vaut rien ny à roſtir, ny à boüillir : Il en eſt qui diſent de moy toutes les méchancetez ima-

C

FIGURE 6.5 – *opening*



9

en beau camp de parler, & elles euſſent dit vne quantité d'hiſtoires par ouy dire ſur ce ſujet, pour conuaincre Monſieur le Prince ; mais le Roy apres auoir dit qu'il en auoit quelquefois ouy quelques vnes en ſa chambre, qui l'auoient bien réjouy, ſans iamais l'auoir eſpouuanté, il preſſa le Cardinal de continuer : ce qu'il fit en ces termes.

Encor qu'il n'y ayt perſonne de vous qui doute de l'exiſtence des eſprits toutesfois ie penſe eſtre le mieux inſtruit, & le mieux perſuadé de la compagnie, auſſi ie croy qu'ils m'ont joüé leurs plus vilains tours, & qu'ils m'ont pris pour l'object de toutes leurs funeſtes recreations. Ils inuentent tous les iours mille nouueautez, qu'ils apportent de la Cour infernale, pour me faire enrager : ie croy qu'ils taſchent à me jetter dans le deſeſpoir pour me faire rendre l'ame ſans confeſſion. Les Dames me chantent vn *Salue Regina*, comme l'on faict à ceux qui ſont prés d'endurer le ſupplice ; autres vn *Libera*, comme ſi i'eſtois deſia mort, quelques-vns me mettent en pouſſiere : d'autres ſe contentent de me porter en terre, en chantant des Pſeaumes de Dauid. Il en vient qui m'honorent pour ſe mocquer de moy, & qui me traitent d'Eminence, & montrent vn precipice où ie ſuis tout preſt de mettre les pieds, d'autres en forme de Choüettes, & de Hibous, me prediſent plus de mal-heurs qu'il n'en eſt arriué au monde depuis qu'il ſubſiſte, quelques-vns me déchirent à belles dents comme des chiens affamés ; il y en a qui ſe contentent de me chanter poüille, & de me dire plus d'injure qu'vn million de harangeres n'en pourroient inuenter toutes enſemble : D'autres, font des Remonſtrances au Parlement, pour l'animer encor dauantage contre-moy : Quelques-vns me procurent charitablement la haine du peuple, comme ſi ie n'en eſtois pas deſia aſſez hay ; ceux-là font vn bruit de tous les Diables à mes oreilles, comme s'ils ſonnoient le toxin, pour eſmouuoir la populace à ſe jetter ſur moy.

Quelques amateurs de ragouſts me mettent au court boüillon, & puis me laiſſent là ſans me manger, pour me faire plus grand dépit, diſant, que ie ne vaut rien ny à roſtir, ny à boüillir : Il en eſt qui diſent de moy toutes les méchancetez ima-

C

FIGURE 6.6 – *deskew*

Kraken		**Tesseract**						
en_best	*model17*	*eng*	*fra*	*+greyscale*	*+remove noise*	*+thresholding*	*+opening*	*+deskew*
23,36 %	12,97 %	20,21 %	19,69 %	21,15 %	11,56 %	11,00 %	12,55 %	12,55 %
16,05 %	7,42 %	14,84 %	16,50 %	11,73 %	13,17 %	10,14 %	11,43 %	11,43 %
27,91 %	12,90 %	64,13 %	63,64 %	37,49 %	25,02 %	24,38 %	22,48 %	21,56 %
52,80 %	11,60 %	60,71 %	56,75 %	21,42 %	29,33 %	21,28 %	19,10 %	19,10 %
32,47 %	7,15 %	24,03 %	30,69 %	22,34 %	9,73 %	10,23 %	13,21 %	13,21 %
38,09 %	10,78 %	13,04 %	9,39 %	9,39 %	4,52 %	4,52 %	5,91 %	5,91 %
41,34 %	12,83 %	59,35 %	58,66 %	30,49 %	11,63 %	8,44 %	9,48 %	9,48 %
50,64 %	15,01 %	62,17 %	63,91 %	38,82 %	40,79 %	41,40 %	42,46 %	42,46 %
32,54 %	20,47 %	16,38 %	16,16 %	80,82 %	18,32 %	17,24 %	9,27 %	9,27 %
25,02 %	16,49 %	27,14 %	24,52 %	15,90 %	8,41 %	8,99 %	9,87 %	9,91 %
18,88 %	10,09 %	19,21 %	18,74 %	22,98 %	11,16 %	11,67 %	13,12 %	18,33 %
20,32 %	6,53 %	12,41 %	8,43 %	6,99 %	6,67 %	5,93 %	6,48 %	6,48 %
24,52 %	12,96 %	62,96 %	62,30 %	25,11 %	11,71 %	10,24 %	10,75 %	10,16 %
18,47 %	8,48 %	28,23 %	24,79 %	28,94 %	12,87 %	12,02 %	11,36 %	15,55 %
64,94 %	15,71 %	15,82 %	14,31 %	9,71 %	8,89 %	8,23 %	8,85 %	9,33 %
25,82 %	11,35 %	18,11 %	20,36 %	10,33 %	13,38 %	11,86 %	10,18 %	10,18 %
24,31 %	7,54 %	32,41 %	31,60 %	20,52 %	9,84 %	8,71 %	10,17 %	10,17 %
80,97 %	34,18 %	14,86 %	12,28 %	15,16 %	7,48 %	9,66 %	12,21 %	11,78 %
22,68 %	12,11 %	39,16 %	38,54 %	17,78 %	18,93 %	18,24 %	18,01 %	14,18 %
51,54 %	31,28 %	43,40 %	45,75 %	13,56 %	14,11 %	10,49 %	4,88 %	11,94 %
34,63 %	13,89 %	32,43 %	31,85 %	23,03 %	14,38 %	13,23 %	13,09 %	13,65 %

TABLE 6.1 – CER des pages selon les modèles d'OCR appliqués aux images TIFF. En dernière ligne, les moyennes.

Kraken		Tesseract						
en_best	*model17*	*eng*	*fra*	*+greyscale*	*+remove noise*	*+thresholding*	*+opening*	*+deskew*
21,81 %	13,63 %	25,99 %	22,60 %	21,38 %	12,78 %	10,86 %	9,68 %	11,98 %
14,53 %	7,04 %	15,29 %	13,85 %	11,66 %	13,25 %	10,22 %	11,20 %	11,20 %
27,91 %	10,92 %	73,43 %	71,81 %	34,46 %	24,74 %	23,26 %	23,82 %	23,04 %
51,30 %	12,14 %	72,99 %	74,49 %	26,88 %	22,92 %	18,55 %	16,64 %	22,37 %
32,77 %	7,45 %	34,56 %	41,51 %	23,24 %	10,73 %	11,32 %	13,90 %	13,90 %
37,22 %	10,09 %	12,17 %	10,61 %	20,70 %	4,87 %	5,22 %	5,74 %	6,26 %
39,54 %	11,37 %	58,48 %	58,83 %	21,62 %	11,03 %	8,61 %	8,96 %	13,01 %
50,64 %	13,72 %	53,15 %	50,80 %	40,64 %	35,86 %	37,98 %	44,20 %	44,58 %
31,03 %	16,16 %	35,56 %	36,42 %	32,33 %	17,03 %	17,89 %	9,70 %	10,99 %
24,44 %	15,03 %	15,07 %	12,99 %	11,49 %	9,87 %	8,99 %	9,08 %	9,41 %
17,54 %	9,67 %	19,26 %	19,16 %	15,58 %	10,74 %	11,58 %	13,16 %	18,14 %
17,55 %	6,34 %	14,58 %	11,11 %	7,45 %	6,57 %	6,48 %	7,08 %	7,08 %
23,56 %	11,93 %	54,57 %	56,19 %	12,59 %	10,97 %	11,12 %	11,05 %	11,12 %
18,57 %	6,65 %	42,22 %	42,55 %	24,13 %	15,93 %	11,45 %	11,64 %	16,59 %
64,60 %	13,86 %	17,02 %	14,55 %	9,47 %	9,33 %	8,10 %	9,26 %	9,13 %
24,73 %	10,62 %	20,22 %	18,11 %	13,24 %	12,36 %	10,98 %	9,09 %	9,09 %
23,60 %	7,06 %	27,71 %	24,77 %	16,67 %	11,01 %	8,68 %	12,40 %	12,40 %
81,57 %	30,94 %	20,09 %	17,41 %	12,94 %	7,55 %	10,23 %	15,26 %	13,47 %
24,22 %	10,96 %	41,00 %	39,54 %	20,77 %	15,63 %	22,45 %	15,17 %	20,23 %
43,40 %	24,96 %	63,29 %	69,44 %	12,48 %	11,21 %	7,23 %	5,24 %	9,40 %
33,53 %	12,53 %	35,83 %	35,34 %	19,48 %	13,72 %	13,06 %	13,11 %	14,67 %

TABLE 6.2 – CER des pages selon les modèles d'OCR appliqués aux images JPEG. En dernière ligne, les moyennes.

Modèles		Tcon	Tlex	CER prédit
Kraken	enbest	0,959	0,278	145,92
	model17	0,943	0,326	140,70
Tesseract	eng	0,596	0,333	422,52
	fra	0,612	0,365	431,55
	fra+grayscale	0,635	0,376	425,69
	fra+grayscale+removenoise	0,642	0,356	457,28
	fra+grayscale+removenoise+thresholding	0,661	0,378	424,21
	fra+grayscale+removenoise+thresholding+opening	0,649	0,360	448,19
	fra+grayscale+removenoise+thresholding+opening+deskew	0,640	0,354	472,28

TABLE 6.3 – Moyennes Tcon, Tlex et CER (prédits) des pages selon les modèles d'OCR appliqués aux images JPEG

(images moins bonne qualité donc) semble bénéfique. C'est le cas pour les deux modèles de KRAKEN (environ 1 point de différence) et le modèle *fra+grayscale* (+3,5%). Ce résultat contrecarre l'idée selon laquelle une numérisation de meilleure qualité est récompensée par une hausse des performances de l'OCR.

2. Ensuite, on remarque que le modèle TESSERACT *fra+grayscale+remove_noise+thresholding+opening* descend à un $CER = 13,09\%$ (sur les images TIFF) alors que le modèle *a priori* très adapté *model17* de KRAKEN est à $13,89\%$. Non seulement KRAKEN pose problème car il est très chronophage en raison de son incapacité à gérer les pages vides et les pages de reliure, mais en plus le modèle précité de TESSERACT est un modèle par défaut fourni lors du téléchargement du logiciel et ne nécessite que quelques pré-traitements des images. L'économie de l'apprentissage d'un modèle adapté aux documents français du XVII^{e} siècle semble donc justifiée.
3. On remarque très clairement l'importance du pré-traitement des images. Effectivement, sans pré-traitement, on est autour de 30% de CER avec TESSERACT (modèles *eng* et *fra*). La binarisation (réalisée nécessairement avec KRAKEN) permet de faire baisser le CER à 13%, et ce pour KRAKEN et TESSERACT. La binarisation est donc particulièrement efficace, d'où son utilisation obligatoire avec KRAKEN.
4. Enfin, les gains associés à chaque pré-traitement semblent cumulables, c'est ce que montrent les CER des différents modèles TESSERACT (car chaque pré-traitement est ajouté aux précédents).

Evaluation non supervisée du lot entier

On peut ensuite proposer une évaluation non supervisée en prédisant par regression linéaire le CER (voir section 5.5 du chapitre précédent). L'objectif est d'observer, pour un lot entier de numérisations, quel modèle d'OCR et quels prétraitements des images sont le plus adaptés.

Le tableau 6.3 donne les valeurs des T_{con}, des T_{lex} et des CER prédits en fonction des modèles et des prétraitements effectués sur les images. Ce test n'a été réalisé que sur les images JPEG car les images TIFF sont très lourdes (près de dix huit fois plus lourdes que les JPEG) et la sous-section précédente a

montré qu'il n'y a pas de claire influence positive sur l'OCR lors de l'utilisation des images TIFF. Pour KRAKEN, on observe que le modèle appris sur des données du XVII^{e} siècle semble plus performant que le modèle anglais par défaut. Pour TESSERACT, on observe que les T_{con} sont nettement plus faibles, ce qui implique un CER prédit largement plus élévé. Il apparaît impossible donc de comparer les CER prédits entre les deux logiciels, mais il reste possible pour un même logiciel de comparer ses modèles et les prétraitements réalisés. Pour TESSERACT donc, les meilleurs taux T_{lex} et CER prédits sont atteints par le modèle *fra+grayscale+removenoise+thresholding*, comme le montre le tableau 6.2 sur l'échantillon des vingt pièces transcrites.

De deux choses l'une.

1. D'une part, ces résultats montrent que le CER prédit semble être un bon moyen d'évaluer la sortie d'un logiciel d'OCR si l'on ne dispose pas de vérité de terrain. Il est toutefois difficile de comparer, avec le CER prédit mais pas avec le T_{lex} qui lui est indépendant des logiciels, les valeurs des CER prédits entre logiciels différents car les T_{con} ne peuvent être comparé et qu'ils impliquent, s'ils varient beaucoup, des valeurs de CER prédit largement différents.
2. D'autre part, le modèle KRAKEN appris sur les données du XVII^{e}siècle semble être le plus performant avec le modèle TESSERACT *fra+grayscale+removenoise+thresholding*. Plus encore, si l'on s'en réfère au T_{lex}, c'est ce dernier qui présente le plus de mots « connus », c'est-à-dire ne comportant pas d'erreur et appartenant au lexique LGeRM.

6.2 Observation des données bruitées

Les neuf modèles présentés *supra* permettent d'obtenir neuf versions de ce premier lot de mazarinades : le résultat de l'OCR de ces neuf modèles. On l'a vu, ces neuf versions du « corpus » [4] ne sont pas équivalentes en terme de qualité, et on peut les ordonner selon celle-ci, ou le « bruit » qu'elles contiennent selon le CER prédit [5] :

— TESSERACT *eng*
— TESSERACT *fra+grayscale+removenoise+thresholding*
— TESSERACT *fra+grayscale*
— TESSERACT *fra*
— TESSERACT *fra+grayscale+removenoise+thresholding+opening*
— TESSERACT *fra+grayscale+removenoise*
— TESSERACT *fra+grayscale+removenoise+thresholding+opening+deskew*
— KRAKEN *model17*
— KRAKEN *enbest*

4. Nous utilisons ce terme par commodité, bien qu'il ne s'agisse pas à proprement parler d'un corpus qui présenterait une unité sensée.

5. Bien que KRAKEN présente des CER prédits supérieurs à ceux de TESSERACT, nous le plaçons en dernière position car cela est dû aux taux T_{con} largement supérieurs à ceux de TESSERACT, alors même que les taux T_{lex} sont inférieurs.

6.2.1 Statistiques descriptives

Comme proposé dans [Taghva et al., 2004, section 2.1], sont proposées à l'étude quelques statistiques descriptives des neuf versions du corpus :

— Taille du corpus compressé en MB ;
— Moyenne du nombre de caractères par page ;
— Moyenne du nombre de mots par page ;
— Nombre de caractères uniques ;
— Nombre de mots uniques ;
— Nombre d'hapax ;
— Nombre de non hapax ;
— Nombre de mots appartenant au LGeRM.

		Taille (MB)	Nb caractères	Nb caractères / page	Nb mots	Nb mots / page	Nb caractères uniques	Nb mots uniques	Nb hapax	Non non hapax	Nb mots corrects
Tesseract	*eng*	8,6	9 434 942	692	1 853 727	169	110	327 910	263 443	64 467	841 549
Tesseract	*fra* *+grayscale* *+removenoise* *+thresholding*	9,2	9 458 365	690	1 819 147	164	139	287 532	228 189	59 343	979 301
Tesseract	*fra* *+grayscale*	9,1	10 016 906	726	1 980 224	179	139	303 140	241 982	61 158	1 014 366
Tesseract	*fra*	6,4	9 143 545	668	1 834 300	167	139	297 162	240 584	56 578	912 660
Tesseract	*fra* *+grayscale* *+removenoise* *+thresholding* *+opening*	6,3	9 130 070	666	1 760 214	159	139	294 106	236 056	58 050	935 759
Tesseract	*fra* *+grayscale* *+removenoise*	6,8	8 883 652	656	1 709 853	156	144	286 625	229 330	57 295	901 469
Tesseract	*fra* *+grayscale* *+removenoise* *+thresholding* *+opening* *+deskew*	9,2	8 714 142	635	1 682 475	152	139	284 075	228 276	55 799	896 445
Kraken	*model17*	7,8	11 817 884	1 189	3 117 189	426	106	232 863	172 506	60 357	942 959
Kraken	*enbest*	6,6	9 155 062	974	2 213 810	310	85	334 761	281 157	53 604	617 612

TABLE 6.4 – Statistiques descriptives des corpus selon les modèles d'OCR

Plusieurs constats peuvent être faits après lecture du tableau 6.4 :

— Les corpus les moins « bruités » contiennent plus de caractères et plus de mots, ce qui suggère que les corpus « bruités » le sont aussi par le silence. Certaines portions de texte ne sont pas reconnues. Ceci peut être dû au module de segmentation mais aussi au modèle d'OCR lui-même car le phénomène exposé *supra* se remarque aussi entre les deux modèles KRAKEN qui partagent la même segmentation.

— Les corpus les plus « bruités » ont tendance à avoir plus de caractères uniques et moins de mots uniques, ce qui conforte l'intuition.

— Contrairement à [Taghva et al., 2004, section 2.1], on n'observe pas une augmentation de la taille des corpus à mesure que ceux-ci sont « bruités ». Par exemple, pour KRAKEN, le corpus résultat du modèle *enbest* est moins lourd que le corpus résultat du modèle *model17* ($7,8 > 6,6$) alors que ce dernier contient moins de mots uniques, moins d'hapax et plus de non hapax ; on aurait supposé que pour ces raisons la compression aurait été plus efficace.

— Comme dans [Taghva et al., 2004, section 2.1], on remarque – pour TESSERACT, ce n'est pas vrai pour KRAKEN – paradoxalement que les corpus les moins bruités contiennent moins de mots correctement orthographiés (au sens qu'ils appartiennent au LGeRM). Par exemple, le meilleur modèle de TESSERACT contient 841 549 mots appartenant au LGeRM alors que le moins bon en compte 896 445. Toutefois, il n'y a pas de « proportionnalité » entre le caractère bruité des corpus et le nombre de mots correctement orthographié : le modèle *fra+grayscale* (qui n'est ni le plus ni le moins bruité) contient 1 014 366 unités lexicales correctement orthographiées.

— La remarque précédente n'est pas non plus corrélée au nombre de mots que contient le corpus (une explication aurait été que tel corpus contient plus de mots ce qui aurait pu impliquer mécaniquement plus de mots correctement orthographiés). Pour TESSERACT, le meilleur modèle contient moins de mots appartenant au LGeRM que le modèle le moins performant, alors que ce dernier contient moins de mots que le meilleur modèle.

— Toutefois, les meilleurs modèles ont tendance à compter plus de mots non hapax que les modèles les moins performants. C'est le cas pour les deux modèles de KRAKEN mais aussi pour les modèles TESSERACT. Par exemple, le meilleur modèle de TESSERACT compte 64 467 non hapax alors que le moins bon en compte 55 799.

— De manière surprenante, pour TESSERACT on peut aussi constater ce phénomène pour les hapax : ils sont moins nombreux pour les modèles les moins performants.

— Néanmoins, le modèle TESSERACT *fra+grayscale+removenoise+thresholding* (le deuxième dans le classement du CER prédit) est celui qui compte le moins d'hapax mais c'est aussi le second en terme de mots appartenant au LGeRM. D'ailleurs, c'est aussi ce modèle qui maximisait les valeurs de T_{con} et T_{lex} (voir tableau 6.3) ; ces constats permettent de l'ériger au rang de modèle TESSERACT le plus performant.

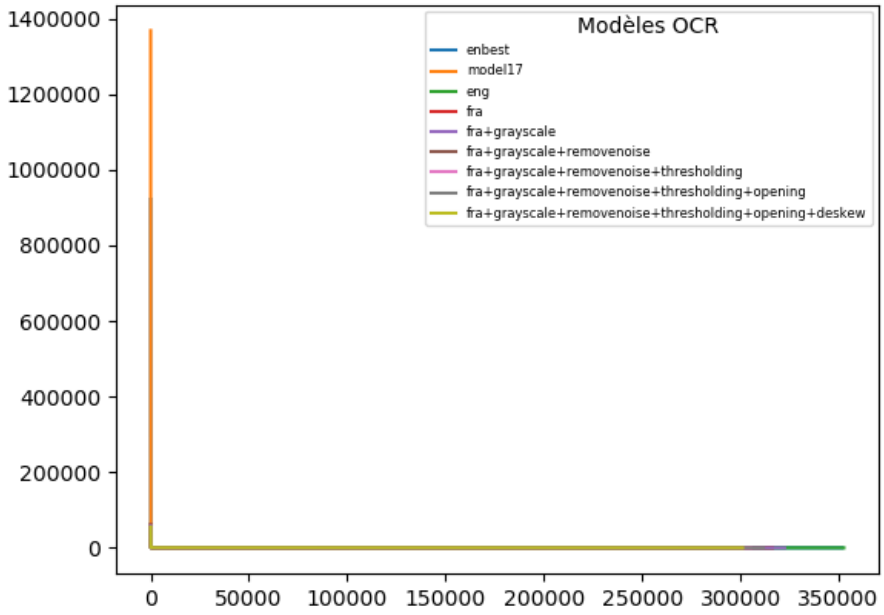


FIGURE 6.7 – *Fréquences des mots des corpus océrisés selon les neuf modèles OCR*

6.2.2 Distribution des mots et loi de Zipf

Les proportions de mots hapax et de non hapax peuvent être observées au prisme d'une représentation graphique : en abscisses, les mots ordonnés par ordre de fréquence, en ordonnées les dites fréquences. Si l'on en croit [Hill and Hengchen, 2019, section 2.1], les résultats, en terme de distribution, devraient être très similaires à l'attendu, c'est-à-dire une courbe du type loi de Zipf. Ils vont même jusqu'à dire que « [...] l'introduction d'erreurs d'OCR ne modifie pas la distribution attendue » [6] [Hill and Hengchen, 2019, section 2.1]. Avec les neuf versions d'OCR du lot 1, nous devrions avoir neuf courbes très similaires les unes aux autres, des courbes de type hyperboliques (loi de Zipf [7]).

La figure 6.7 montre bien que, comme [Hill and Hengchen, 2019, section 2.1], l'introduction d'erreurs d'OCR n'influence pas la distribution des mots ; on observe, pour les neuf corpus, une distribution du type loi de Zipf. Néanmoins, à y regarder de plus près – grâce à une échelle logarithmique, figures 6.8 et 6.9 –

6. *« [...] the introduction of OCR errors does note distort the expected distribution. »*
7. Une minorité de mots sont très fréquents alors que la majorité des mots sont très peu fréquents : la fréquence d'un mot est inversement proportionnelle à son rang.

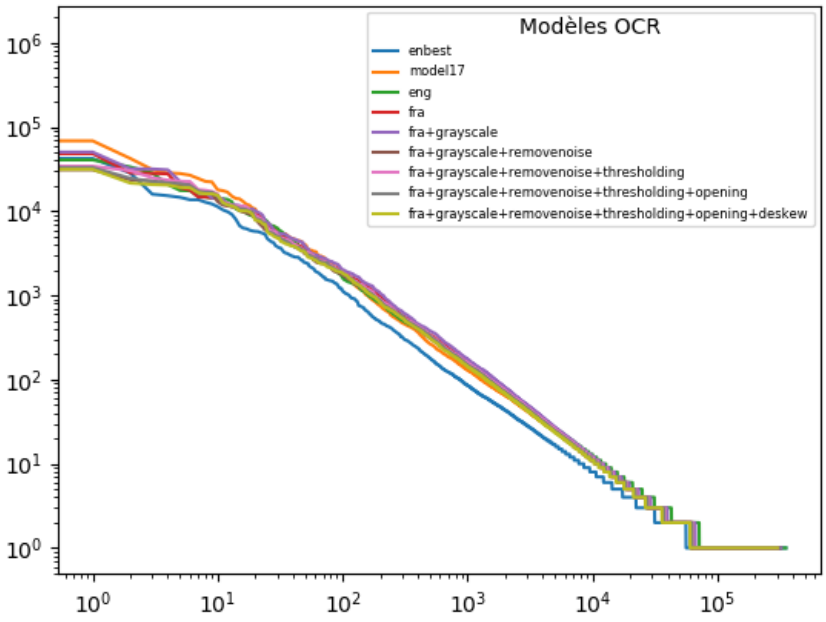


FIGURE 6.8 – *Fréquences des mots des corpus océrisés selon les neuf modèles OCR – échelles logarithmiques*

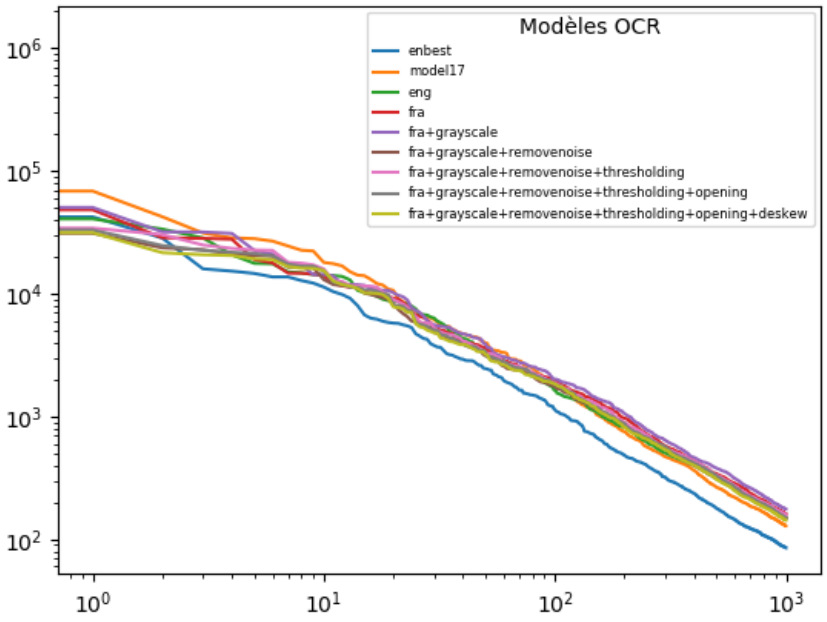


FIGURE 6.9 – *Fréquences des 1 000 premiers mots des corpus océrisés selon les neuf modèles OCR – échelles logarithmiques*

on observe certaines nuances :

— le modèle Kraken *model17* présente bien une minorité de mots plus représentés dans les hautes fréquences, ses non hapax sont plus représentés ;
— le modèle Kraken *enbest* présente plus de mots moins représentés, le tableau 6.4 montre d'ailleurs qu'il s'agit du modèle contenant le plus d'hapax ;
— la courbe du modèle Tesseract *fra+grayscale+removenoise+thresholding+opening+deskew* est celui qui présente des fréquences moins élevées pour les mots les plus représentés et des fréquences plus élevées pour les mots les moins représentés, ce qui suggère plus de « bruit » dans les données ;
— le modèle *fra+grayscale* apparaît comme le modèle Tesseract présentant, pour les mots les plus représentés, des fréquences plus élevée.

En ces termes, le modèle Kraken *model17* semble présenter une distribution plus « zipfienne » ce qui peut donner confiance en ce modèle, d'autant plus qu'il s'agit du troisième modèle contenant le plus de mots appartenant au LGeRM (voir tableau 6.4).

6.2.3 Observation des mots les plus fréquents

Les figures 6.7, 6.8 et 6.9 montrent qu'il existe des différences dans les termes les plus représentés. On peut donc observer, pour chaque sortie d'OCR, les mots les plus représentés dans le corpus et les fréquences associées. Sans surprise, on observe que les mots grammaticaux apparaissent dans les hautes fréquences et que les « mots pleins » apparaissent dans les basses fréquences. Néanmoins, les résultats entre les corpus sont assez inégaux, comme le suggère [Hill and Hengchen, 2019, section 2.1].

— les modèles de Tesseract présentent beaucoup de caractères « spéciaux » tels que ©, ¥ ou encore % alors qu'ils sont nettement moins présents dans les hautes fréquences des modèles de Kraken, ce qui est appuyé par le tableau 6.4 ;
— dans les modèles Tesseract il y a aussi plus de mots « termes poubelles » [8] comme, par exemple : *ss, Li, ut, c, ï*, etc. ;
— seul le modèle Kraken *model17* utilise le s long ;
— enfin, lorsque les mots grammaticaux et les mots « termes indésirables » sont enlevés, les mots plein restant sont partagés entre tous les modèles (on trouve par exemple : *Roy, Monfieur, contre, Prince, Cardinal*).

Ces observations poussent elles aussi à ériger le modèle Kraken *model17* comme meilleur modèle OCR.

6.2.4 Erreurs fréquentes

Une analyse de l'échantillon des vingts pages transcrites permet de repérer les erreurs récurrentes des modèles.

8. Ou, *junk terms* [Taghva et al., 1994, section 3.4].

— Le double colonnage rend difficile la segmentation de la page et l'OCR s'en trouve détériorée : deux lignes adjacentes sont segmentées en une unique ligne.
— Les tâches d'encre ou encore les gravures sont souvent segmentées et océrisées ce qui ajoute des faux positifs.
— Les lettrines [9] ne sont parfois pas reconnues.
— Les diacritiques se situent souvent à la suite des lettres les supportant.
— Certaines lettres sont souvents confondues :
  — *f* et *f* ;
  — *e* et *c* ;
  — *c* et *G* ;
  — *t* et *r* ;
  — *1* et *r* ;
  — *I* et *l* ;
  — *J* et *l* ;
  — *r* et *t* ;
  — *r* et *s* ;
  — *c* et *O* ;
  — *ll* et *l* ;
  — *ff* et *f* ;
  — *cts* (avec ligature) et *Ets* ;
— Les espaces peuvent être omis (mots agglutinés) ou ajoutés (mots découpés).
— Les premières et dernières lettres d'une ligne peuvent être omises.
— Lorsque le *z* est en italique, il est souvent pris pour une *Z* en capitale.

6.2.5 Plongements lexicaux

En 2013, [Mikolov et al., 2013] proposent une nouvelle façon de représenter les données textuelles : les vecteurs « denses » qu'on appelle plongements lexicaux (ou *embeddings* en anglais). Il s'agit d'apprendre, pour chaque unité lexical d'un corpus, un vecteur contenant une ou quelques centraine(s) de dimension(s). [Mikolov et al., 2013] proposent deux modèles (CBOW et *skip-gram*), qu'on désigne souvent par l'unique terme *Word2Vec*.

Pour les neuf versions océrisées du Lot 1, nous pouvons donc apprendre des vecteurs denses en utilisant le modèle *Word2Vec* du package Python *gensim.models*. Ce modèle prend en entrée un corpus représenté sous forme d'une liste de listes – le premier niveau correspond aux phrases et le second aux mots [10]. Mis à part les ponctuations qui ont été remplacées par des espaces, aucun prétraitement des données textuelles n'a été réalisé.

Une fois le modèle appris, on dispose d'autant de vecteurs que d'unités lexicales. Par Analyse en Composantes Principales (ACP), nous pouvons représenter graphiquement ces vecteurs sur deux dimensions. Selon cette représentation,

9. Premières lettres gravées en grand au début d'un texte.
10. La tokenisation est réalisée par la simple utilisation de la fonction *.split()* car c'est ainsi que les logiciels d'OCR segmentent les lignes en mots ; par les espaces blancs.

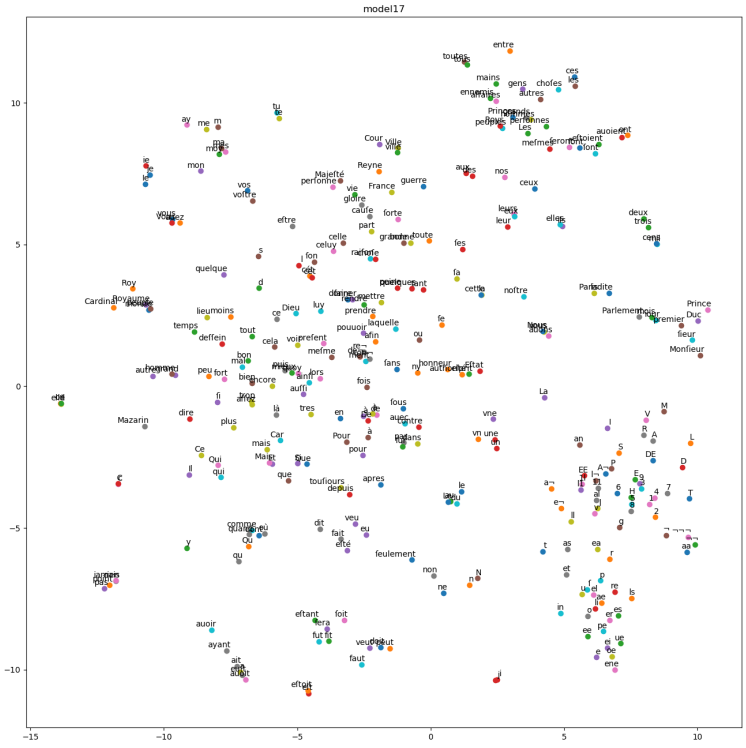


FIGURE 6.10 – ACP du corpus océrisé par le modèle KRAKEN *model17*

si deux points[11] sont proches, c'est que les vecteurs associés à ces points sont aussi proches, donc qu'ils partagent (d'un point de vue distributionnel) une sémantique. Afin de rendre lisible les figures, nous ne représentons que les unités apparaissant au moins 500 fois dans le corpus.

7

Les figures 6.10, 6.11, 6.12, 6.13, 6.14, 6.15, 6.16, 6.17 et 6.18 montrent très clairement une proximité entre les unités lexicales correspondant à des erreurs de segmentation et/ou d'OCR. Pour les ACP des corpus océrisés par TESSERACT, on observe une dichotomie presque parfaite entre les unités lexicales correspondant à du bruit et celle étant de véritables mots.

On peut clairement conclure qu'il y a, d'un point de vue distributionnel, une similarité entre les unités lexicales correspondant à du bruit « pur », c'est-à-dire

11. Un point représente dans le plan un vecteur.

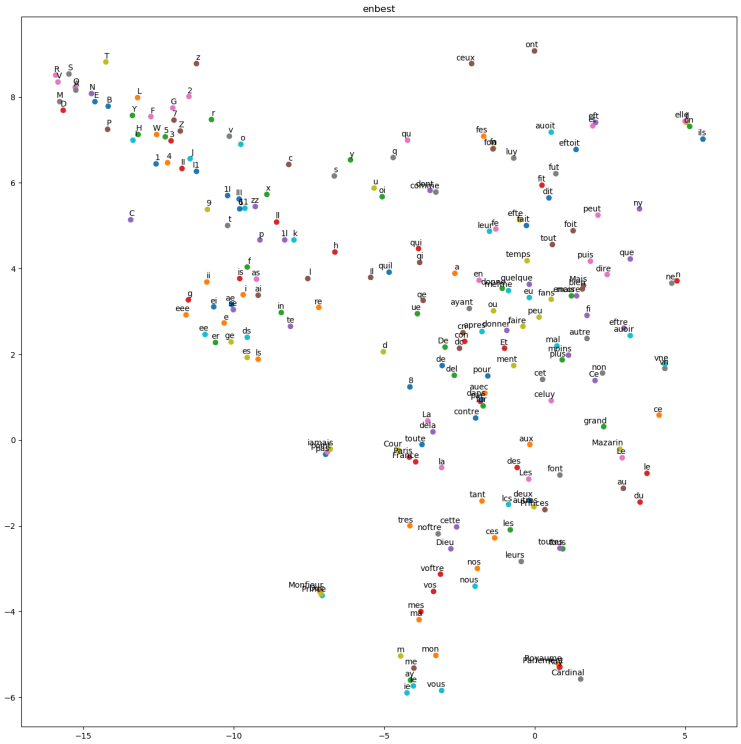


FIGURE 6.11 – ACP du corpus océrisé par le modèle KRAKEN *enbest*

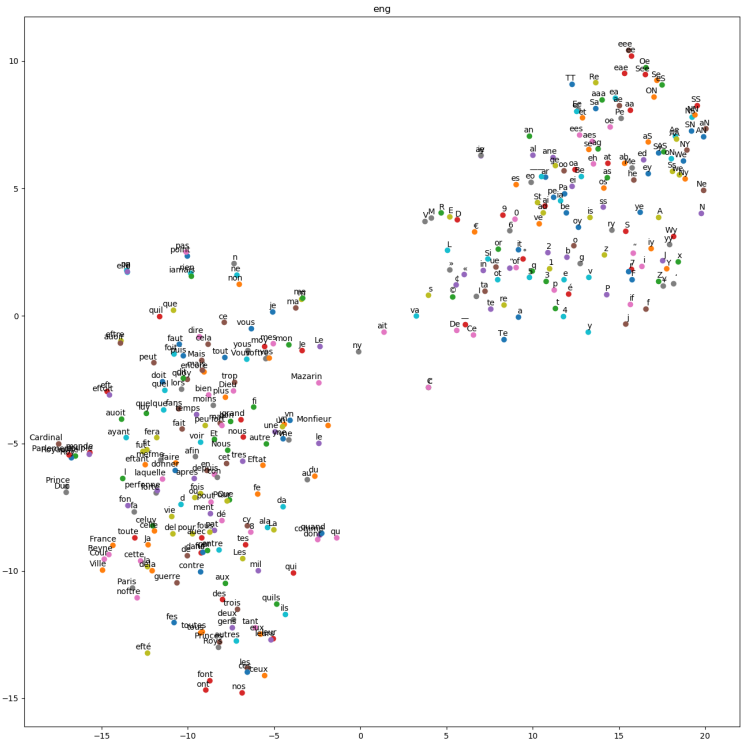


FIGURE 6.12 – ACP du corpus océrisé par le modèle TESSERACT *eng*

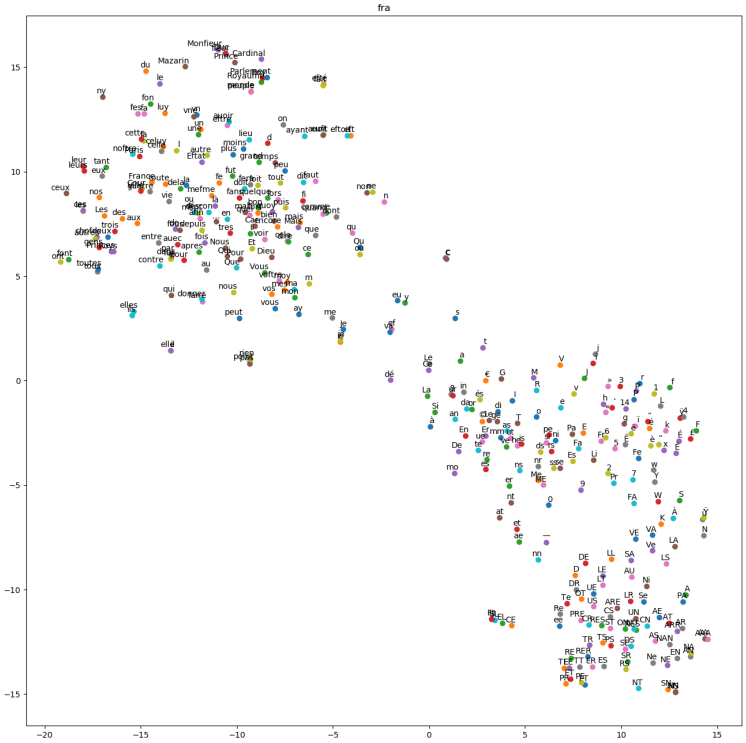


FIGURE 6.13 – ACP du corpus océrisé par le modèle TESSERACT *fra*

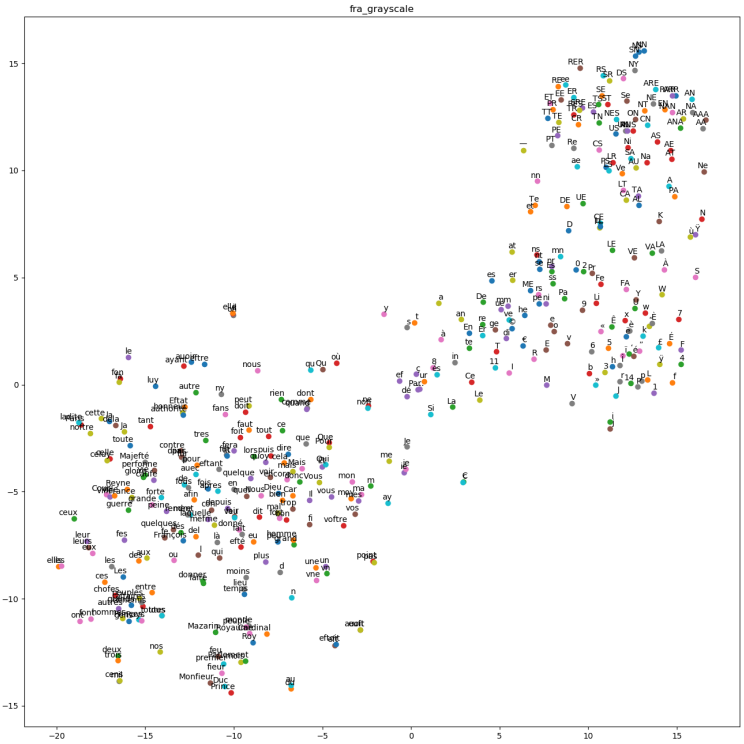


FIGURE 6.14 – ACP du corpus océrisé par le modèle TESSERACT *fra+grayscale*

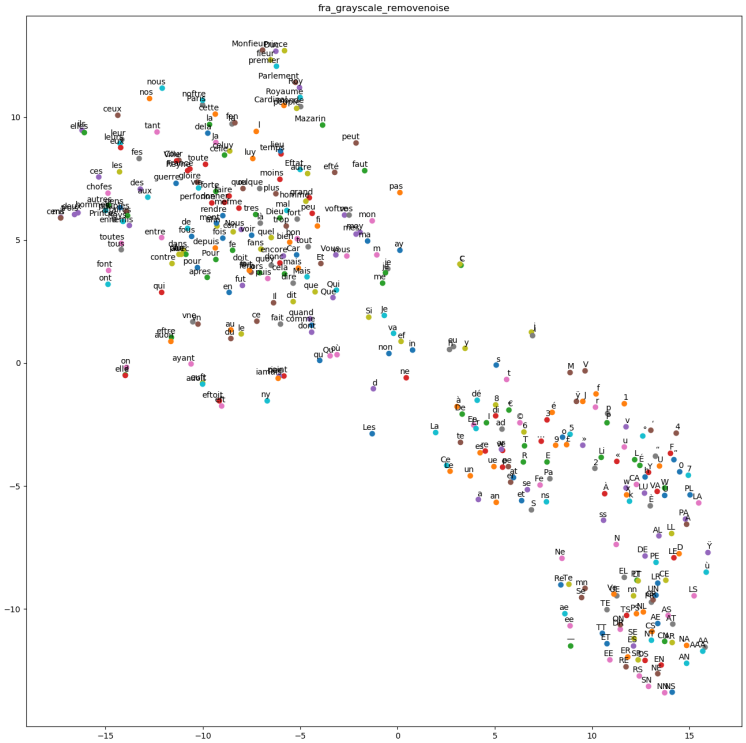


FIGURE 6.15 – ACP du corpus océrisé par le modèle TESSERACT *fra+grayscale+removenoise*

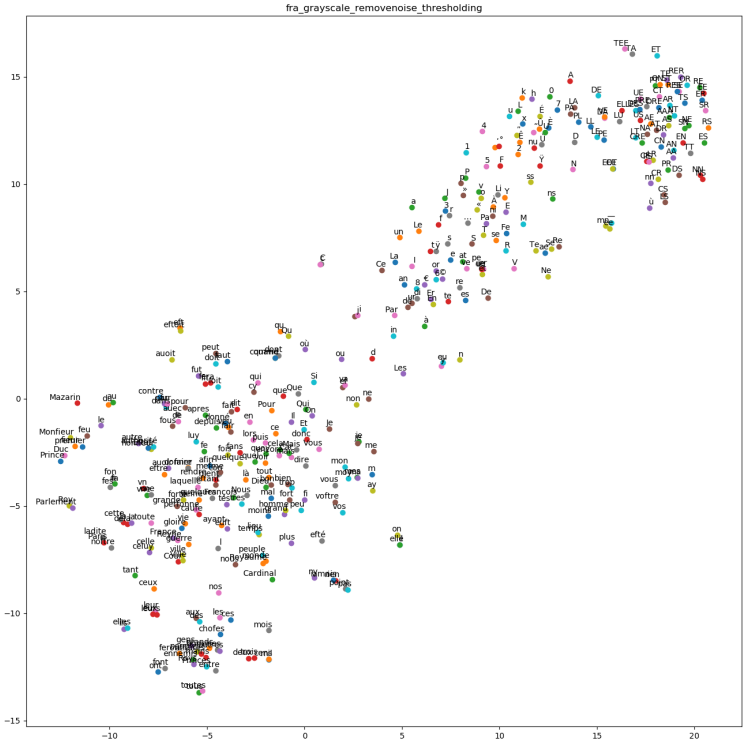


FIGURE 6.16 – ACP du corpus océrisé par le modèle TESSERACT *fra+grayscale+removenoise+thresholding*

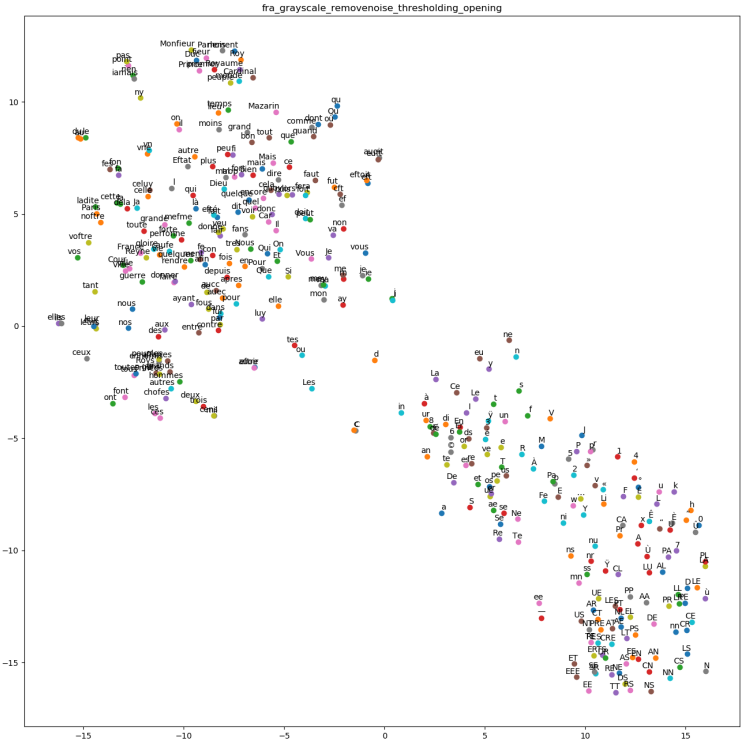


FIGURE 6.17 – ACP du corpus océrisé par le modèle TESSERACT *fra+grayscale+removenoise+thresholding+opening*

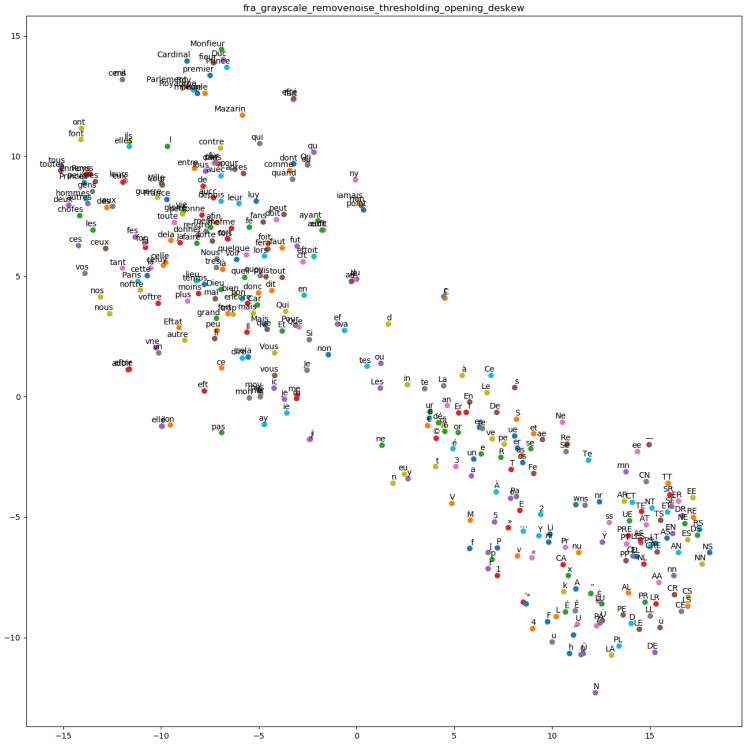


FIGURE 6.18 – ACP du corpus océrisé par le modèle TESSERACT *fra+grayscale+removenoise+thresholding+opening+deskew*

les insertions dues à une segmentation erronée. Cette dernière affirmation se justifie par le fait que ces unités i) ne correspondent à aucune entrée de lexique (et on ne peut les lire), ii) elles apparaissent dans des contextes similaires faisant que les vecteurs associés sont proches (c'est ce que nous montrent les ACP) et iii) l'unité entière est bruit, il ne s'agit pas de mots contenant un ou plusieurs caractère(s) bruité(s) (dans ce cas, le mot bruité sera très probablement un hapax et n'apparaîtra pas sur l'ACP car la fréquence minimale a été fixée à 500).

Les modèles de plongements lexicaux sont donc capables de modéliser et reconnaître les unités correspondant à du « bruit pur ». Ceci renforce une fois de plus l'idée selon laquelle les erreurs d'OCR sont moins stochastiques qu'on l'imagine.

6.3 Conclusions

L'élection entre Tesseract et Kraken n'est pas aisée. D'un côté, Kraken semble plus adapté (cf. sous-sections 6.2.2 et 6.2.3). D'un autre côté, Kraken est beaucoup plus long à exécuter et le CER prédit érige Tesseract avec tous les prétraitements en modèle le plus performant.

Nous concluons donc que le modèle XVII^{e} de Kraken doit être utilisé pour les petites collections quand le modèle français avec tous les prétraitements de Tesseract doit être utilisé pour les collections plus larges.

Troisième partie

Erreurs d'OCR et évaluations extrinsèques : textométrie et similarité textuelle

Chapitre 7

Stylistique sur corpus bruité : les mazarinades burlesques

Sommaire

7.1 Des données au corpus

Ce chapitre a fait l'objet d'une publication portée par Karine Abiven et co-signée avec Gaël Lejeune. La référence est la suivante : [Abiven et al., 2021].

Parmi les quelques cinq milles documents de ce qu'on appelle le corpus des mazarinades, il est possible d'explorer un ensemble d'écrits dits « burlesques » (style, registre, ou « genre d'écrire », comme on dit à l'époque qui nous intéresse [1]). On peut estimer à plusieurs centaines les documents relevant de cette pratique d'écriture ([Carrier, 1996, p. 93] va même jusqu'à dire qu'ils représentent un quart des mazarinades), qu'on peut résumer comme un mélange stylistique et un jeu avec les normes, qui s'exercent alors surtout en vers dans des poésies en octosyllabes.

Toujours dans le but de montrer que les données bruitées issues d'OCR sont exploitables, et puisque lire de près quelques centaines de documents s'avère en pratique impossible (contrainte temporelle de la recherche), il est possible d'éloigner un peu la focale et d'étudier ces données textuelles par le biais d'outils statistiques, pour – peut-être – faire émerger des caractéristiques encore mal connues, notamment métriques et lexicales.

Néanmoins, telle étude fait courir le risque d'isoler les écrits de leurs contextes de production ; la mise en série de faits de langue congruents a en effet pour revers de les désindividualiser, lesquels ne sont pourtant interprétables qu'en tant qu'ils sont situés, dans le temps et en tant qu'actions d'écriture [Jouhaud, 2009]. Pour pallier ce défaut, il conviendra de revenir dès que possible aux contextes des occurrences, pour chercher à affiner leurs coordonnées énonciatives premières.

Au plan épistémologique, nous cherchons à dépasser l'étude par induction à partir de « quelques coups de sonde » [2]. Cette dernière méthode, si elle est parfaitement légitime, manque peut-être de systématicité, ce que nous proposons de résoudre grâce à l'outillage textométrique. Il s'agira i) de vérifier des hypothèses préalables, en les confrontant à la réalité des données (approche *corpus-based* [Pincemin, 1999]) mais aussi ii) de « laisser parler les données », en extrayant ce qui est statistiquement plus présent, par exemple, dans les écrits burlesques de la Fronde (approche *corpus-driven*). Nous avons adopté une démarche plutôt *corpus-driven*, notamment en raison de la définition en extension de notre objet.

Plusieurs difficultés sont alors à résoudre : quelles sont les mazarinades en « style burlesque » ? Comment en dresser une liste de titres exploitable au format numérique ? Où en trouver des représentations numériques de qualité ? Comment les transformer en mode texte ? Comment ces textes peuvent-ils alors être fouillés ?

1. [Ménage, 1650, p. 159]. Sur le débat sur la définition « du » burlesque comme style, registre ou genre, voir [Nédélec, 2004, p. 269].
2. [Nédélec, 2004, p. 432-437]

7.2 À la recherche des traits d'écriture burlesque

Tant la production « en vers burlesques » aura été un véritable phénomène éditorial, les mazarinades sont parfois réduites à ce « style ». Celui-ci est caractérisé globalement par une discordance entre le référent et le signifiant. Sa visée est souvent dite ludique, parfois railleuse (ce qui rapproche le burlesque de la satire), car fondée sur la dérision de tout sans distinction ni privilège (notamment pour le pouvoir en place), et critique.

Lors de sa première éclosion en France dans les années 1630-1640, le burlesque est souvent considéré comme une forme de parodie car une de ses formes privilégiées est le « travestissement » d'épopée. Les écrits burlesques de la Fronde recourraient en fait assez peu à ce procédé [Nédélec, 2004], ce que nous tenterons de vérifier. Au reste, ce « genre d'écrire » est réputé indéfinissable, de sorte qu'il est souvent décrit plutôt en extension, c'est-à-dire par ses manifestations concrètes, qu'en intension, par ses traits définitoires, qui restent donc à dessiner.

Le burlesque est souvent associé à une innovation lexicale ([Bar, 1960] consacre 300 pages sur 400 au lexique dans sa caractérisation du « genre burlesque ») : les deux points principaux sont l'usage massif des variations du français et les néologismes. Concernant les variations, elles sont aussi bien diachronique, diatopique, que diastratique (c'est-à-dire sociolectale : jargons techniques, tours populaires et argot, principalement). S'agissant plus précisément des écrits burlesques suscités par la Fronde, ils sont réputés relever d'un style particulièrement cru et obscène ([Nédélec, 2004, p. 385] met en doute cette idée en rappelant que cette variation du français n'est nullement propre au burlesque). Il serait ainsi intéressant de mesurer la spécificité des écrits de la Fronde en la matière.

7.3 Fabriquer le corpus des écrits burlesques de la Fronde

7.3.1 Lister des titres : des bibliographies papier à leurs numérisations

Si aucune liste exhaustive recensant l'ensemble des « mazarinades » n'existe, la *Bibliographie des Mazarinades* de Célestin Moreau et les suppléments successifs apportés par Moreau lui-même mais aussi Philippe Van Der Haeghen, Emile Socard et Ernest Labadie permettent d'atteindre précisément 5 059 références [3]. Ces recueils bibliographiques n'étant pas disponibles en format numérique, nous les avons océrisés en utilisant le logiciel `Kraken` [4]. L'ensemble de ces données

3. Deux remarques permettent à la fois de revoir à la hausse ou à la baisse ce nombre. D'une part, à la hausse si l'on inclut, comme Célestin Moreau, des « textes fantômes (dont on sait par les sources qu'ils ont existé mais dont on n'a pas conservé d'exemplaires) ». D'autre part, à la baisse car Hubert Carrier comme la Bibliothèque Mazarine notent que Célestin Moreau a pris en compte les titres, et non les éditions distinctes, ce qui multiplie les entrées dès lors que le titre diffère légèrement. Hubert Carrier produit ainsi une estimation de 5 200 écrits distincts (modulo 5% d'erreur) [Carrier, 1989, p. 72].

4. Version 2.0.8

textuelles a ensuite été structuré afin de pouvoir les utiliser dans un tableur : à chaque titre (et son numéro interne dans le recueil bibliographique) sont associés une date et un lieu d'édition (quand ces données sont disponibles), un nombre de pages ainsi que l'ensemble de la notice conçue par le bibliographe Moreau. L'ensemble des titres ainsi acquis a été corrigé manuellement [5].

Parmi ces 5 059 titres donc, seuls ceux comportant « burlesque » – ou une variante graphique de ce terme (« bvrlesque », « bvrleſque » ou « burleſque ») – sont rassemblés pour constituer le corpus des pièces burlesques de la Fronde [6]. Supposer qu'une pièce est burlesque si son titre l'indique peut sembler accorder une confiance excessive aux étiquettes endogènes. Mais la dimension commerciale de ces écrits (peu coûteux, parfois même distribués dans la rue) implique justement un étiquetage explicite. Cela est notamment vrai de la plus grosse masse du corpus constitué par la poésie et très souvent sous-titré « en vers burlesques » : ce sous-titre est si fréquent que *burlesque* peut y signifer *poésie*, par opposition à *prose* [7], ou plus spécifiquement *octosyllabe*, ce mètre étant presque systématique dans les « vers burlesques » [8]. Ceci explique deux biais de notre méthode de sélection des titres par l'étiquetage explicite. Soit il peut s'agir d'un faux titre, qui ne renvoie pas au contenu, mais seulement à l'octosyllabe – procédé utilisé au moins deux fois [9], sans doute pour des raisons commerciales, car le vers burlesque était, à une époque où la vente de petits imprimés sauvait bien des imprimeries, un argument de vente de choix. Enfin, la pièce peut être dite burlesque selon les critères généralement admis et rappelés plus haut, sans que cette étiquette ne figure dans le titre, auquel cas nous avons ajouté par un repérage manuel les pièces repérées. Cela est vrai notamment des quelques pièces en prose du corpus (par exemple la *Lettre de remerciment envoyee au Cardinal Mazarin... avec la harangve de dame Denise* ou l'*Entretien de Fanchon, Toinon et Nichon sur l'arrivée de leurs galands. Piece morale*), mais aussi bien en vers (*Le Médecin politique, qui donne un souverain remède, pour guérir la France malade à l'extrémité*).

Cette liste s'élève, en l'état, à 250 titres [10], dont 179 sont librement disponibles

5. Voir le fichier *Moreau_bib_mazarinades.ods* à l'adresse https://github.com/jbtanguy/revue_HN_burlesque_fronde.git (consulté le 28 juin 2022).

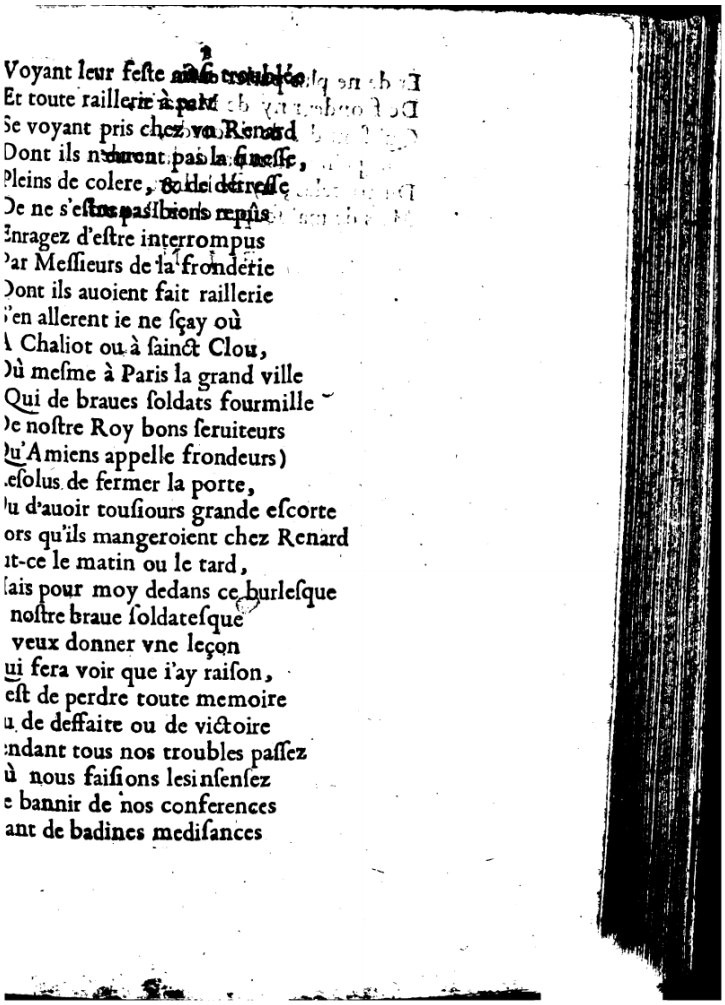
6. Il reste alors à laisser de côté les faux positifs que cette démarche peut comporter, notamment les titres dont le commentaire du bibliographe contient « burlesques », mais au sujet d'un autre ouvrage, et à supprimer les doublons quand le numéro d'un supplément bibliographique constitue un commentaire à un item déjà connu et non un ajout d'item.

7. « [...] D'eſtre voſtre ancien Seruiteur,/Autant en Burleſque qu'en Proſe » (*Théologien traduit en vers burlesques*, p. 3).

8. Voir toutefois la *Ballade burlesque des partisans* en décasyllabes.

9. Dans deux poèmes qui ont toutes les apparences de textes de piété : *La Passion de nostre Seigneur, en vers burlesques. Dédiée aux âmes dévotes*, Paris, chez la veuve Jean Rémy, 1649 ; *Extase de la France Mourant d'amour pour Jesus-Christ crucifié en vers burlesques*, Paris, Claude Morlot, 1649.

10. Ce qui amène à relativiser l'estimation à 1 300 libelles en vers burlesques, le quart des mazarinades, selon Carrier (qui n'a pas pu publier la liste des titres concernés, laquelle reste donc à constituer). Certes, le mot n'est pas toujours dans le titre, mais significativement il l'est dans tous les exemples cités par Carrier, et il serait étonnant qu'il ne le fût pas dans plus d'un millier de ces écrits.



Voyant leur feſte mal troublée
Et toute raillerie à part
Se voyant pris chez vn Renard
Dont ils n'auoient pas la fineſſe,
Pleins de colere, & de triſteſſe
De ne s'eſtre pas mieux repûs
Enragez d'eſtre interrompus
Par Meſſieurs de la frondérie
Dont ils auoient fait raillerie
S'en allerent ie ne ſçay où
A Chaliot ou à ſainct Clou,
Où meſme à Paris la grand ville
Qui de braues ſoldats fourmille
De noſtre Roy bons ſeruiteurs
Qu'Amiens appelle frondeurs)
Reſolus de fermer la porte,
Du d'auoir touſiours grande eſcorte
ors qu'ils mangeroient chez Renard
it-ce le matin ou le tard,
lais pour moy dedans ce burleſque
noſtre braue ſoldateſque
veux donner vne leçon
ui fera voir que i'ay raiſon,
eſt de perdre toute memoire
u de deffaite ou de victoire
ndant tous nos troubles paſſez
ù nous faiſions les inſenſez
e bannir de nos conferences
ant de badines mediſances

FIGURE 7.1 – Page 6 de *La nappe renversée, chez Renard, en vers burlesques*, Moreau 2525, 1649, Paris, https://books.google.fr/books?id=fL1vR-ja5xgC&hl=fr, téléchargé sur Google Livres

FIGURE 7.2 – Page 12 de *La nappe renversée...*, téléchargé sur Google Livres

au téléchargement en ligne. Les plateformes Gallica [11] et Google Livres [12] ont été utilisées, avec une priorité pour Gallica qui propose des numérisations de bien meilleure qualité, ainsi que des métadonnées plus riches et plus sûres. Les numérisations disponibles sur Google Livres proviennent de diverses entreprises de numérisations sans que leurs conditions et protocoles soient accessibles.

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 [13]. La page 6 (figure 7.1) est très difficilement lisible car on y devine par encre déteinte la page suivante. La page 12 (Figure 7.2) n'est en rien une numérisation de la mazarinade en question.

Le téléchargement prioritaire des pièces sur Gallica a permis d'atteindre 79% (soit 142 pièces) de numérisations de qualité réalisées par la BNF.

7.3.2 Océrisation

Distribuées au format PDF, les numérisations ont été converties – en utilisant la librairie *pdftoppm* (version 0.62.0) – en une suite d'images PNG permettant par la suite de procéder à de la reconnaissance optique de caractères. Un tri parmi l'ensemble des images résultantes a écarté les pages générées par les plateformes. Ces images ont ensuite été binarisées, segmentées et océrisées en utilisant `Kraken` et son modèle entraîné pour le XVII^{e} siècle. On dispose finalement de 2 822 images PNG et du même nombre de sorties d'OCR au format HTML.

7.3.3 Taux de reconnaissance estimé (Gallica)

La plateforme Gallica propose, pour certaines de ses numérisations, une transcription automatique résultat d'OCR. À cette sortie d'OCR est associé un « taux de reconnaissance estimé », lequel correspond au pourcentage de mots supposés corrects par le logiciel d'OCR – ce taux est proposé à l'échelle du document en tant que moyenne des taux de reconnaissance par page. « Par exemple, une qualité estimée de 98% signifie que deux mots sur cent sont potentiellement erronés [14]. » Toutefois, on ne trouve pas de complément d'information sur ce « taux de reconnaissance estimé ». Plus encore, l'investigation des fichiers ALTO (disponibles *via* l'API de Gallica) montre que, parmi les 1 980 pages du corpus pour lesquelles une océrisation de la BNF existe, 1 734 l'ont été par *Isako* [15], 168 par *ABBYY* [16] et 78 ne renseignent pas l'outil. Cette pluralité d'outils traduit donc une diversité de campagnes d'océrisation – comme le suggère justement l'attribut *ID* de la balise *ocrProcessing*, prévoyant ainsi, au sein même de la structure de données, la possibilité d'une succession de plusieurs campagnes d'OCR. Bref, derrière le « taux de reconnaissance estimé » se cachent divers

11. gallica.bnf.fr – BNF (consulté le 28 juin 2022)
12. https://books.google.fr/ (consulté le 28 juin)
13. https://tinyurl.com/nappe-renversee-renard (consulté le 28 juin 2022)
14. Voir https://gallica.bnf.fr/edit/und/consulter-les-documents#Mode~texte~et~OCR, section *Qualité de l'OCR* (consulté le 28 juin 2022).
15. Isako © - Document Conversion System [DCS]®
16. ABBYY FineReader Engine

outils, diverses campagnes d'OCR, réalisées à divers moments... rendant vaine la comparaison de deux résultats d'OCR par cette mesure. [Traub et al., 2015, section 4.1] remarque le même problème dans l'interprétation des valeurs de confiance de sorties d'OCR pour un corpus de journaux hollandais : « [...] il nous faut comprendre ce que ces valeurs signifient et/ou comment elles ont été calculées. » [17]

En plus de l'utilisation d'un modèle libre et appris sur des données du XVII^{e} siècle, ce constat justifie notre volonté d'océriser le corpus selon un protocole unique et de l'évaluer par des métriques clairement exposées [18].

7.4 Données bruitées, statistiques et interprétation

Résultat d'OCR, les données dont nous disposons sont imparfaites, au sens où elles ne rendent pas compte parfaitement du texte présent dans les documents d'origine, et ce de manière irrégulière. En effet, certaines pages sont presque parfaitement transcrites quand d'autres sont tout à fait illisibles. L'anticipation des biais de l'étude permet d'emblée de mettre en évidence des limites au propos tout en préfigurant ce que ces données bruitées donnent à voir.

7.4.1 Taux d'erreur au caractère

Le modèle d'estimation du taux d'erreur au caractère (voir Chapitre 5) donne un taux d'erreur estimé de $136,32$, ce qui n'a pas de sens en soi mais bien en comparaison avec l'utilisation d'un autre modèle. À titre indicatif, l'océrisation du corpus avec le modèle français de `Tesseract` donne un taux d'erreur estimé de $306,78$. Le modèle d'OCR utilisé est donc bien plus performant.

7.4.2 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 d'entrée de jeu les deux problèmes du bruit et du silence. Le premier, pluriel en tant que résultat d'une océrisation d'une portion de texte inexistante [19] mais aussi en tant que confusion entre caractères « proches » [20], a pour double effet d'ajouter de fausses entrées dans les tables de fréquences et de diminuer les fréquences des entrées mal reconnues. Le second, en tant qu'absence de reconnaissance de portions de texte pourtant

17. *« [...] we need to find out what these values mean and/or how they have been computed. »*
18. Notons toutefois que la BNF a internalisé les campagnes de d'océrisation de ses fonds documentaires depuis le mois de mars 2019, offrant ainsi une unité dans la procédure et dans l'évaluation. Cette hétérogénéité de Gallica est par ailleurs vouée à disparaître avec le montage d'un projet pilote au DataLab de la BNF : Galli(Corpor)a.
19. Par exemple, la fausse reconnaissance de texte dans un portrait.
20. Par exemple, le *G* et le *O*.

présentes dans le document numérisé, implique une diminution des fréquences des formes non reconnues – et certains hapax deviennent des nullax. L'analyse de ces tables de fréquences ne peut donc être réalisée que dans l'observation de phénomènes saillants. Autrement dit, cette étude sur corpus bruité ne peut rendre compte de la rareté des phénomènes.

De la donnée textuelle au fac-similé numérique

Le bruit peut faire émerger de fausses hypothèses, et en particulier dans une étude s'intéressant de près au lexique et à ses innovations (ou néologismes). Par exemple, l'occurrence *bougrene* repérée dans un des textes n'est pas un mot-valise agglomérant habilement la *gangrène* au *bougre* suggérant ainsi l'envahissement inévitable de la France par le « goût italien ». Il ne s'agit pas d'agglomérer le bougre (au sens de « gay ») à la métaphore de la gangrène qui viendrait, en l'occurrence, de l'Italie avec le Cardinal Mazarin. Il s'agit simplement d'une confusion entre les graphèmes *ri* et *n*. Un retour à l'image du livre numérisé permet de lire *bougrerie*. Ainsi gardons-nous la notion de page dans nos corpus, en ne concaténant pas l'ensemble des pages d'une même œuvre, assurant dès lors un retour aisé à la première source que constitue le document numérisé.

Il en va de même pour des unités lexicales apparaissant dans les hautes fréquences. Citons par exemple les nombreuses répétitions, notamment de noms propres, dues au péritexte (numérisé et océrisé) :

- le dispositif du titre courant, dans le corpus de contraste : le mot *Polyandre* apparait comme très fréquent, mais c'est en raison du titre courant en pages paires de l'édition de 1648 de *Polyandre* (roman de Sorel, souvent considéré comme burlesque) (voir figure 7.3) ;
- les didascalies nominatives des textes de théâtre multiplient les noms de personnages ;
- une édition du XIX^{e} siècle, choisie parce qu'elle regroupe plusieurs numéros successifs d'un périodique titré la *Muze Historique* (de Loret), introduit une haute fréquence de la graphie *muze* parce que le titre est reproduit en titre courant des pages paires (sur 601 pages).

Notons que ces problèmes d'identification du péri-texte ne sont ni spécifiques aux documents anciens ni même aux documents non-nativement numériques, voir par exemple [Giguet et al., 2020] sur la segmentation de page dans des documents PDF contemporains.

Des spécificités aux fréquences absolues

En outre, le calcul des spécificités lexicales (qu'on décrira dans la sous-section sur l'outil Antconc page 165) est sensible aux variations graphiques et orthographiques. Si le *s* long (*ſ*) est absent du corpus de référence mais présent dans le corpus d'étude, alors le *s* long sera fortement représenté dans la liste des formes spécifiques. Dès lors, il nous paraît prudent – sinon nécessaire – de procéder à un retour systématique aux tables de fréquences. Si une forme

l'on parloit aſſez d'Aduocats eſ-
coutans & promenans, & de plai-
dans, & de conſultans, mais que
l'on ne faiſoit point mention d'a-
uocats beuuans, & mangeans, ce
qui luy ſembloit fort neceſſaire
apres tant de trauail & d'exercice.
Il s'eſtonnoit comment l'on ſe
pouuoit tenir ſi long-temps en
vn lieu, & meſme en ſi grande
compagnie, ſans y faire quelque
bon repas, & il diſoit que l'on
auoit eu tort d'apeller vn Palais
le lieu où l'on rendoit la Iuſtice,
s'il ne ſeruoit à y manger effecti-
uement ; & ſi par ce moyen il
n'auoit quelque raport au Palais
de la bouche. Quant à eſtre Me-
decin, les ragouſts que l'on pre-
ſente aux gens de cette condi-
tion luy faiſoient mal au cœur ; il

FIGURE 7.3 – Exemple de péritexte générant du bruit (titre courant en haut de page)

apparaît spécifique à un corpus, nous comptons, pour ce corpus et le corpus contrastif, l'ensemble des fréquences absolues de cette forme et de certaines variantes possibles. Il s'agit de vérifier par les fréquences absolues la tendance observée au sein des spécificités.

7.4.3 De la nécessaire contextualisation des occurrences

Sans les négliger tout à fait, nous avons traité avec précaution les thématiques saillantes ou les isotopies qui semblaient fréquentes, notamment parce que notre fouille est limitée aux mots et aux caractères et que notre corpus n'est ni traité ni enrichi. Par ailleurs, la prudence à l'égard d'une « thématique du mot clé » [Rastier, 2011, p. 31] est de mise, puisqu'on ne peut exclure en droit que « là où le thème revê[t] sa plus grande efficacité narrative, il perd sa lexicalisation privilégiée ». Ensuite et plus précisément s'agissant d'écrits d'actualité et polémiques, l'interprétation de séries thématiques semble hasardeuse, en l'absence de regroupements en séries cohérentes (par exemple d'un même parti ou camp politique). Un repérage purement lexical revient à ne pas distinguer les usages contrastés de telle ou telle isotopie : « à décomposer ainsi les discours [par thèmes], on finit par ne plus rencontrer qu'un immense entrelacs d'éléments de sens dispersés sur toute l'étendue d'une aire culturelle, par effacer toute discontinuité », a-t-on pu écrire au sujet de la sémantique de corpus polémiques [Maingueneau, 1983, p. 18]. La recherche de discontinuités, autrement dit de formes locales de spécificité, supposerait ainsi une annotation et des partitions de corpus très fines, qui sont un horizon mais non l'objet du présent travail.

7.5 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 [Rastier, 2011, p. 33]. L'établissement de tels ensembles doit répondre à des critères explicites, mais, lorsque tout est à construire, il faut parfois faire des compromis entre ce qui est possible et ce qui est désirable [Hunston, 2008, p.156]. Deux difficultés se posent alors en l'espèce : la délimitation de ce que sont les textes burlesques (question en partie aporétique comme vu plus haut), et la disponibilité numérique des textes identifiés. Pour le premier problème, on s'est fondé par exemple sur le large corpus cité par Francis Bar, qui a une définition pour partie datée des notions littéraires (« genre » burlesque, « grands écrivains », etc.). De plus, et cette limite est très importante, nous avons parfois importé sans distinction des recueils de textes variés (*Oeuvres complètes* de Saint-Amant, pour le Corpus Burlesque hors Fronde par exemple), alors qu'il aurait peut être fallu trier et ôter les écrits relevant d'autres pratiques.

Concernant l'obtention des données elles-mêmes, elle a été contrainte par la disponibilité numérique. Nous avons choisi des éditions de textes littéraires de la première modernité libres de droits (arbitrairement délimitée de 1 500 à

Corpus	Oeuvres	Occurrences	Formes
Non burlesque	1 092	27 873 298	231 283
Burlesque	235	3 525 879	256 909
Burlesque hors Fronde	56	3 123 098	230 889
Burlesque Fronde (corpus d'étude)	179	402 781	50 642
Burlesque Vers	202	1 733 044	170 243
Burlesque Prose	33	1 792 835	134 855

Table 7.1 – Description synthétique des corpus : nombre d'œuvres, nombre d'occurrences et nombre de formes (calculés avec Antconc)

1 700, englobant ce que Frantext étiquette comme *Français pré-classique* et, en partie, *Classique*). Au-delà de l'arbitraire de ces choix, il s'agissait d'obtenir un corpus de référence d'une taille suffisante pour être statistiquement représentatif. S'il n'existe pas de taille idéale minimale, on considère souvent comme le fait [Sinclair, 2003] qu'un million de *tokens* pour chaque sous-corpus à comparer est une taille raisonnable.

Le caractère opportuniste [McEnery and Hardie, 2011] de nos corpus induit un défaut de calibrage générique. Ainsi le théâtre y a-t-il sans doute une place plus importante en raison du corpus en ligne *Théâtre Classique*, qui a constitué une partie de nos sources. Ensuite, la qualité et l'enrichissement de ces diverses données sont hétérogènes, depuis la pure sortie d'OCR (notre corpus d'étude, Corpus Burlesque Fronde), jusqu'à des données lemmatisées, POS-taggées et rendues disponibles en TEI (Presto et Frantext), utilisées pour les corpus de contraste. Le projet E-ditiones, qui rendra disponible en 2021 un corpus de textes du XVII^{e} siècle présentant tous les genres de discours alors pratiqués (genres littéraires, droit, sciences, correspondances, etc.) pourrait permettre à l'avenir de pallier ces défauts de calibrage (voir [Gabay et al., 2020] [21] (consulté le 28 juin 2022)).

7.5.1 Description des corpus contrastifs

Cherchant à caractériser les écrits burlesques parus pendant la Fronde, 56 œuvres burlesques ont été rassemblées pour constituer un corpus burlesque contrastif (hors Fronde, donc). Deux sources principales ont permis la constitution de ce corpus : d'une part la bibliographie constituée par Francis Bar dans *Le genre burlesque en France au XVIIème siècle, Etude de style.* et d'autre part le sous-ensemble des textes identifiés comme burlesques dans le sous-ensemble *1500-1700* du corpus *Frantext* [22]. Notons enfin que ce corpus a pu être divisé en deux sous-corpus, les textes en vers et ceux en prose, offrant ainsi une nouvelle modalité de contraste.

21. https://github.com/e-ditiones
22. https://www.frantext.fr/ (consulté le 28 juin 2022).

Parallèlement, 1 092 œuvres non burlesques ont été rassemblées pour constituer un second corpus contrastif ; il s'agit d'une compilation des œuvres non burlesque figurant dans les corpus *Frantext*, *Sermo*[23] et *Théâtre Classique*[24]. La volumétrie des différents corpus utilisés dans notre étude est présentée dans le tableau 7.1[25].

7.5.2 Antconc

Nous avons choisi de réaliser des expériences avec l'outil Antconc pour montrer que l'exploitation d'un corpus bruité ne nécessitait pas d'outillage complexe. Cet outil présente l'avantage de poser peu de contrainte sur le format des données d'entrée[26]. De plus, il est multi-plateformes et fonctionne sans installation préalable[27], permettant ainsi de juger rapidement de l'utilisabilité d'un corpus[28]. C'est une manière d'évaluer à quel point le bruitage du corpus, sensible sur une analyse locale, gène réellement une analyse statistique menée au niveau global. Enfin, il exploite des mesures statistiques bien connues et validées par la communauté scientifique, ce qui nous permet de vérifier s'il est possible de détecter des traits saillants dans un corpus bruité sans avoir recours à des mesures complexes ou totalement *ad hoc*.

Antconc utilise l'information mutuelle [Kullback, 1959] pour mesurer les collocations (onglet *collocates*). Pour ce qui est de l'utilisation de corpus contrastifs (*reference corpus*), il permet de vérifier la spécificité d'un terme (*keyness*) au moyen de la *log-likelihood*[Fischer, 1922] ou du $\mathcal{X}^2$[Chernoff and Lehmann, 1954] Ces mesures présentent l'avantage d'être communément utilisées en statistiques en général et en statistiques textuelles. Bien entendu, se pose la question de la qualité des observables que l'on peut tirer d'un tel corpus bruité mais les patrons de tokenisation utilisés par Antconc ne semblent pas affectés outre-mesure par le bruitage de l'entrée[29].

7.5.3 Les lexies référant à l'actualité

D'abord, une simple confrontation entre les écrits burlesques produits pendant la Fronde et ceux produits avant ou après (Burlesque Fronde *vs.* Burlesque hors Fronde) montre une fréquence supérieure de noms propres dans notre corpus d'étude, ce qui est sans doute régulier pour des textes d'actualité. Ces résultats d'ordre descriptif permettent de vérifier que le bruit (dans les données) peuvent certes produire du silence (dans les résultats), mais que cela ne gène que

23. http://sermo.unine.ch/SERMO/ (consulté le 28 juin 2022)
24. http://www.theatre-classique.fr/pages/programmes/PageEdition.php (consulté le 28 juin 2022)
25. Le fichier *Corpus_metadata.ods*, que nous fournissons sur https://github.com/jbtanguy/revue_HN_burlesque_fronde.git, renseigne l'ensemble des métadonnées des corpus précités.
26. De simples fichiers textes éventuellement regroupés dans un même dossier.
27. Contrairement à des outils plus évolués comme TXM par exemple.
28. Par exemple pour des explorations par des étudiants.
29. Sachant que la tokenisation automatique n'est pas un problème résolu que ce soit pour le français moderne ou le français classique

marginalement le repérage des unités lexicales saillantes. Apparaissent ainsi en haute fréquence les principaux anthroponymes attendus (ceux des personnalités impliquées dans les événements : *Mazarin*, *Condé*, *Beaufort*, *Parlement*) ainsi que les appellatifs ou les désignateurs associés : « Messieurs » (désignant les parlementaires) ou « partisans » (pour les financiers enrichis et associés à Mazarin par la vindicte des contemporains). On obtient aussi pour les toponymes plus fréquents qu'ailleurs des résultats attendus : *Espagne*, la Fronde se déroulant sur fond de guerre avec le voisin espagnol, ou *Germain* pour *Saint-Germain* où fut signée la paix mettant fin au blocus de Paris fin mars 1649.

Nous avons mené une expérience sur un sous-corpus de travail qui est une partition de notre corpus d'étude : les 66 pièces de notre corpus justement parues pendant le blocus de Paris (siège de la capitale par les armées royales pour tenter d'étouffer le début de la révolte frondeuse, entre le départ de la Cour du palais du Louvre le 6 janvier 1649 et la paix de Saint-Germain le 1er avril 1649). Dans ce sous-corpus apparaissent nettement des spécificités lexicales (obtenues par confrontation au corpus des autres pièces burlesques) relevant de l'isotopie de la guerre (obsidionale). Ainsi trouve-t-on *trouppes* (+162,79), *generaux* (+135,1), *guerre* (+116,99), *foldats* (+105,43), *mouſquet* (+75,97), *arme* (+72,55), *mareſchal* (+70,88), *regiment* (+63,53), *famine* (+59,55), *cauallerie* (+54,48), *ſiege* (+49,31), *pillage* (+38,41), *canon* (+36,35), *trouppe* (+32,59), *deſtruire* (+31,44), *caualiers* (+30,5), *miſeres* (+27,98), *piſtolets* (+26, 69), *armes* (+26,32), *ſentinelles* (+26,02) et *attentat* (+25,6).

Conformément aux précautions prises ci-dessus, nous cantonnons cet examen d'isotopie à un sous-corpus de travail (voir sous-section sur la contextualisation des occurrences, p. 163) pour approfondir à présent sur le corpus d'étude entier deux exemples de résultats situés à un niveau plus local (stylistique puis énonciatif).

7.5.4 Usage polémique du motif de la « Muse burlesque »

Rappelons ici que l'approche est *corpus driven* car on « laisse parler les données », c'est-à-dire que les interprétations stylistiques sont menées à partir des résultats du logiciel Antconc et pas à partir de connaissances externes. Cette dernière affirmation n'est peut-être pas si vraie dans la mesure où l'on ne peut écarter toute connaissance du domaine...

Le mot *muse* apparaît dans les hautes fréquences du corpus Burlesque Fronde, au rang 520 sur les 50 642 formes du corpus, ce qui représente 63 occurrences, et 77 en incluant la flexion du mot au pluriel. Le mot apparaît également dans les spécificités positives du corpus Burlesque face au corpus Non Burlesque, sans que ce résultat soit original[30], puisque le motif de la muse est très fréquent en poésie (il l'est donc dans le corpus Burlesque, majoritairement poétique) alors le corpus Non Burlesque mélange prose et poésie. Les statistiques confirment cette distribution générique des corpus,

30. Mais pouvoir retrouver certaines caractéristiques évidentes malgré le bruitage des données n'est pas totalement dénué d'intérêt d'un point de vue expérimental.

d'étude et de référence : globalement, les variantes du mot *muse* sont spécifiques au corpus Burlesque.

Les indices énonciatifs les plus associés au mot dans les écrits burlesques (repérés grâce à la fonction *Clusters/N-grams* d'Antconc) indiquent une invocation récurrente à cette allégorie de l'inspiration, grâce aux marques de l'interlocution : « ma muze », « ma muse », « chere muſe », « notre muse », « toy muſe » ; le « o » apparaît également dans les cooccurrents fréquents, en emploi « invocatif » [Grinshpun, 2006]. L'invocation à une muse « burlesque » ou « grotesque » a déjà été repérée comme typique du burlesque dès avant la Fronde [Nédélec, 2004, p. 432-433]. La fréquente invocation à la « Muse » est attendue dans un corpus poétique du XVII^e^ siècle, le motif étant peut-être celui qui est le plus repris en poésie depuis le siècle précédent [31], et en particulier d'une manière parodique (antiépique et anticlassique) ou transgressive, notamment chez les satiriques [Debailly, 2012, p. 131-459-710]. Depuis lors convoquées souvent pour être récusées [Debailly, 2012, p. 667-671], les muses cristallisent une attitude métapoétique qui fait par ailleurs partie des caractéristiques de l'écriture burlesque [Cronk, 1987]. L'invocation-révocation de la Muse n'est pas ainsi essentiellement la marque du travestissement épique – qui a pu être vu un peu trop rapidement comme le « style burlesque le plus typique de la Fronde » [Carrier, 1993, p. 101] –, car ce motif de l'invocation à la muse n'indique pas une réécriture triviale du modèle épique, mais se réduit à « une vague allusion » et à un « emprunt isolé » [Nédélec, 2004, p. 433].

De fait, on ne trouve pas dans les corpus burlesques ici fouillés d'autres mots en hautes fréquences qui renverraient à la mythologie, par contraste avec les autres textes de l'époque (corpus Non Burlesque), où la référence antique est très vivace. Les adjectifs caractérisants, que fait apparaître la recherche de cooccurences, sont les suivants : *marmote* (au sens de « grotesque »), *boufonne*, *falote* (au sens de « impertinent » et « drôle »), *grotefque*, *badine*. On vérifie ainsi que ces écrits des années 1648-1653 poursuivent la déconstruction de la figure mythologique de l'inspiration : sans doute cette allégorie de l'inspiration est-elle un symbole efficace à déboulonner, en disqualifiant les antiques muses comme « vieilles et camuses », en renvoyant dos-à-dos merveilleux païen et chrétien [Nédelec, 2020, p.86] et en faisant par plaisanterie de leurs avatars des réalités triviales (dans le corpus d'étude, *Pégaze* rime avec *Aze*, c'est-à-dire « âne »). D'autres cooccurrents comme « rire », « petite », ou « pauvre » pointent vers cette démystification de la figure du poète que l'écriture burlesque met en jeu : non plus d'origine transcendante, la poésie prend désormais sa source dans l'expérience du poète, ou une vision « humorale », où la muse se voit remplacée par l'humeur, ou la verve [Folliard, 2020, au sujet de Viau]. Le mot *humeur* est d'ailleurs fréquemment présent (61^e^ rang de fréquence), notamment dans un sens métapoétique, comme en attestent les cooccurrents adjectivaux postposés : *folle, badaude, crotesque, debonnaire, insolente, mutine, plaisante, soldatesque* ; les *n*-grammes les plus fréquents à gauche permettent d'accéder aux caractérisants

31. Pour un bilan, y compris bibliographique, sur ce point, voir [Galand, 2016] *Introduction* note 12.

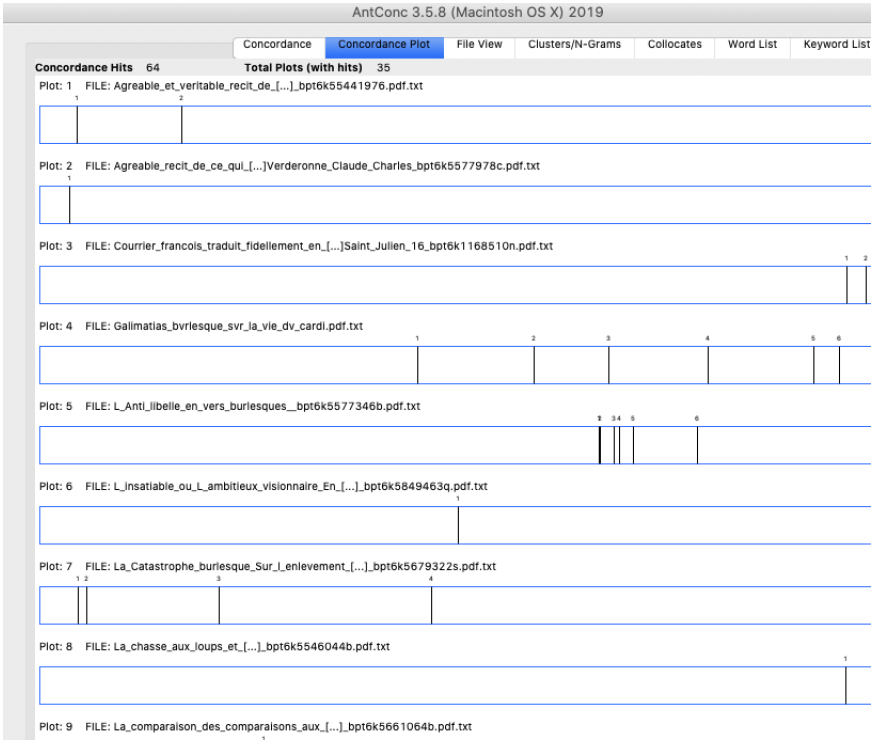


FIGURE 7.4 – *Concordance plots* donnés par ANTCONC, pour l'entrée *muse*

adjectivaux antéposés (car à réalisation monosyllabique) : *bonne*, *gaillarde* ou *belle* : l'expression « belle humeur » est en effet quasi-synonyme de *burlesque* à l'époque, comme en témoigne Dassoucy ([Assoucy, 1650]).

L'allusion à une muse commune ou triviale fonctionnerait ainsi comme un signal stylistique de l'écriture burlesque plus que comme déclencheur d'une parodie de l'épique ou comme révocation transgressive de l'antique. La position du mot dans l'économie textuelle de certains écrits peut conforter cette intuition. On peut ainsi observer de loin la présence du mot *Muse* dans chacun des textes grâce à une fonctionnalité du logiciel ANTCONC nommée *concordance plot*, qui permet de spatialiser schématiquement les zones d'emploi d'un mot dans chaque texte (voir figure 7.4). Le mot n'apparaît ainsi pas seulement au début des pièces, ce qui suggère qu'il n'est pas cantonné à un pur emploi invocatif.

Cette visualisation nous indique par exemple une forte fréquence du mot à la fin de l'écrit intitulé *L'antilibelle en vers burlesques*, p. 9 :

```
N'en   doutez   point   eſprit   de crotte,
En   qui quelque   muſe   marmote,
Muſe   eſprit   de   rebellion [...]
Muſe   fourde   borgne   &   boiteuſe
Qui   porte   des   haillons   de   gueuſe,
```

Corpus d'étude	Corpus de référence	Spécificité
Burlesque	Non Burlesque	+822,35
Burlesque hors Fronde	Non Burlesque	+370,34
Burlesque Fronde	Burlesque hors Fronde	+333,32
Burlesque Fronde	Burlesque + Non Burlesque	+570,79

Table 7.2 – Spécificité du *On* avec différentes configurations de contraste

Muſe au nez ſans ceſſe morueux
A qui chaſtie outre les yeux, [...]
Muſe dont les cheueux en teſte,
Sont beſte d'vne pire beſte [...]
Car cette noire miluſine
Eſt ſuiette de proſerpine,
Elle eſt engendrée en Enfer,
Son grand pere c'eſt Lucifer
Et n'eſtoit montée ſur terre,
que pour y fomenter la guerre,
C'eſt là bourruë (sic) & fols eſprits
Le genie de vos eſcrits [...]

On observe une délocution de l'objet « muse », qui n'est plus l'allégorie à qui on s'adresse, mais bien un attribut des poètes qu'il s'agit ici d'attaquer. Cet écrit imprimé après la paix de Saint-Germain apparaît comme une réponse au flot de pièces burlesques parues pendant le blocus de Paris ; cette disqualification de la « muse burlesque » permet de qualifier l'écrit comme anti-frondeur. Le mot *Muse* est ici non plus allégorique mais métonymique, utilisé pour viser collectivement les poètes ayant publié pendant le blocus pour le dénoncer.

7.5.5 Le « burlesque *On* »

Les spécificités lexicales obtenues en opposant le corpus Burlesque Fronde au corpus Burlesque hors Fronde montrent une surreprésentation du pronom personnel indéfini *on* – d'ailleurs non sujet à la flexion ni à la variation orthographique, ce qui rend ce résultat d'autant plus fiable. Dans notre corpus d'étude, on recense 3 452 occurrences de ce pronom, utilisé avec marquage stylistique dans deux types d'emploi : la personnification et l'anaphore.

Propagation et personnification de la rumeur

Ce pronom 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) permet de compter un nombre important d'occurrences associées au verbe de discours rapporté *dire* : *dit-on* (73 occ.) [dont *ce dit-on* (18 occ.), *(dit-on)* (1 occ.) et *(ce dit-on)* (3 occ.)], *on dit que* (36 occ.), *qu'on dit* (18 occ.) [dont *ce qu'on dit* (8 occ.), dont *a ce qu'on dit* (4

occ.)], *on dit qu'* (14 occ.), *comme on dit* (8 occ.), *on m'a dit* (7 occ.) [dont *à ce qu'on m'a dit* (2 occ.)], *diroit-on* (3 occ.), *difoit-on* (3 occ.), *comme l'on dit* (3 occ.), *de ce que l'on m'en dit* (2 occ.), *on dit par tout* (2 occ.) et *dirit-on* (*sic*) (1 occ.). Notons les incises *(dit-on)* et *(ce dit-on)* qui utilisent les parenthèses, soulignant par la typographie le phénomène majeur qu'induisent ces emplois : le désengagement énonciatif du scripteur, et la plaisante mise en scène de celui-ci. Ces imprimés éphémères de la Fronde ont ainsi diffusé à grande échelle un des procédés typiques des écrits d'actualité depuis la fin du XVI^{e} siècle : colporter, inventer, mettre en scène la rumeur.

Parmi les 3 452 occurrences, 115 [32] sont en capitales (*ON*). Ce genre d'indice typographique est intéressant car ces écrits avaient une vocation commerciale très certaine, et devaient adopter des stratégies pour attirer l'attention. Cet mise en relief par le haut de casse apparaît dans les pièces suivantes :

— *Extraordinaire arrivée du burlesque On de ce temps qui sait, qui fait et qui dit toutes les particularités du siége de Cambray, avec un sommaire de l'ordre du festin fait aux généraux et Parlement d'Angleterre par les communes.* (17)
— *Le buflesque On de ce temps renouvelé, qui sait, qui fait et qui dit tout ce qui s'est passé depuis la guerre. Première partie.* (90)
— *L'On du temps tout nouveau, en vers burlesque (sic), par C. D. B. M.* (7)
— *Le qu'en dira-t-on de Mazarin, burlesques (sic).* (1)

Cette capitalisation massive est surtout le fait des deux premiers de ces écrits (dans lesquels sont rassemblées plus de 90% de ces occurrences en capitales), mais les deux autres aussi présentent une surreprésentation du pronom, en minuscules. Les trois premiers titres constituent en fait une sorte de mini-périodique, qui a connu plusieurs livraisons (six en tout [33]) et qui exploite, comme un filon apparemment rentable, la figuration personnifiée de la rumeur. Celle-ci est d'ailleurs alors proverbiale, comme l'atteste le *Dictionnaire* de Furetière, où sujet et verbe sont nominalisés ensemble : « *On dit* est un grand menteur, pour dire, que les bruits du commun sont souvent faux ». Cette série d'écrits recourt donc à une personnification alors figée en catachrèse.

Cette substantivation, d'origine étymologique – *on* provenant du substantif *homme* –, est ici remotivée par la figure de personnification. Dans les 4 pièces sus-citées, elle est attestée (par les moyens qui sont les nôtres) 7 fois (contre 208 occurrences d'emploi standard du pronom) :

— Dans l'*On du temps tout nouveau, en vers burlesque (sic), par C. D. B. M.* avec : *cet on* (1 occ.), *cet ON* (2 occ.), *Cét ON* (1 occ.), *cét On* (1 occ.) et *C'eft ON* (1 occ.) ;
— Dans le *Qu'en dira-t-on (le) de Mazarin, burlesques (sic)* avec :*mon ON* (1 occ.).

L'usage figuré reste donc relativement exceptionnel, mais sa mise en valeur est toutefois notable par la place en mot-titre et l'usage des capitales : cette figure certes un peu éculée présente sans doute un intérêt certain, dès lors qu'il s'agit

32. Soit 6% du nombre d'occurrences totales de *on* dans le corpus Burlesque Fronde.
33. Le *Burlesque On de ce temps* connait trois autres suites.

d'appâter le chaland. Le repérage automatique permis par Antconc est ici d'une grande utilité, puisqu'il autorise la distinction des casses.

Les anaphores en *on*

L'observation des *concordance plots* (avec un *concordance plot* par mazarinade, et non par page) permet de révéler les passages du corpus où le *on* est surreprésenté (disons plutôt, *concentré*). Les anaphores en *on* apparaissent récurrentes, le pronom étant sujet de verbes dénotant des procès, du type *on court, on s'auance, l'on fuit*[34]. On compte 29 mazarinades présentant ce type d'énumérations[35]. Ce stylème a plusieurs effets : il tend à mettre sur le même plan plusieurs groupes différents, troublant la compréhension claire du schéma actantiel ; par suite il participe d'une écriture figurative, en particulier pour évoquer des personnages non individués et des mouvements de foule (rapprochant le style burlesque du style satirique en ce point). Il s'agit d'un emploi de *on* qu'on pourrait dire « de multitude[36] ». C'est ce que montre l'extrait de l'*Agréable récit de ce qui s'est passé aux dernières barricades de Paris, décrites en vers burlesques* (Paris, Nicolas Bessin, 1649, p. 10) :

Le peuple fait les Barricades;
Les pourſuiuant auec brauades
De tous coſtez on fait grand bruit,
On court, on s'auance, l'on fuit,
Mac ons, Charpentiers, Eſtuuiſtes
Imprimeurs, Relieurs, Copiſtes,
[...]

On voit ici comment le repérage d'une simple unité lexicale comme le *on* peut embrayer sur le repérage d'autres stylèmes, comme ici l'énumération, voire l'accumulation de noms (la liste des noms de métiers a été abrégée, mais elle se poursuit sur 19 vers). On repère ainsi un outil d'amplification caractéristique, qui occasionne les innovations poétiques typiques du burlesque (comme l'usage de jargons techniques en poésie). Il ne s'agit pas toutefois d'écraser toutes les occurrences à un seul type d'emploi, le pronom *on* ayant précisément cette caractéristique de pouvoir renvoyer à tout sujet animé indéfini. Le flou dans la référence des pronoms avait déjà été notée par [Jouhaud, 2009, p. 109], comme une manière de chercher l'adhésion du plus grand nombre en maintenant l'ambiguïté ou d'insinuer des assertions mensongères. Ainsi l'extrait suivant, tiré de l'*Agréable et véritable récit de ce qui s'est passé devant et depuis l'enlèvement du roi hors de la ville de Paris par le conseil de Jules Mazarin, en vers burlesques.*

34. *Agréable récit de ce qui s'est passé aux dernières barricades de Paris, décrites en vers burlesques.*, page 13.

35. Les pièces suivantes sont exemplaires : *Premier (le) courrier françois, traduit fidèlement en vers burlesques*, *Courrier (le) burlesque de la guerre de Paris, envoyé à monseigneur le prince de Condé, pour divertir Son Altesse durant sa prison : ensemble tout ce qui se passa jusqu'au retour de Leurs Majestés* et *Courrier (le) burlesque de la paix de Paris.*

36. Pour les autres contextes d'emploi de ce *on* de multitude, qu'on retrouve dans les mémoires de Rets, dans les chansons, etc., voir [Abiven, 2019].

(p. 10), vise à insister sur le rejet unanime de Mazarin, pour mieux rendre scandaleuse sa carrière sans entraves (évoquée après l'adversatif « mais ») :

On commence de le cognoiſtre,
On s'appercoit que c'eſt vn traiſtre,
Que pour attraper le teſton,
Il pippe comme vn beau demon.
On quitte là ſa compagnie,
On le fuit comme vne Harpye,
On le laiſſe comme vn trompeur,
Où pluſtoſt comme vn fin voleur,
Contre luy on crie, on murmure,
Mais le pis, il va ſe mocquant,
Car il a gaigné leur argent,

Le repérage de ces deux dernières séquences permet de voir l'orientation argumentative divergente des écrits qui les contiennent, le premier (l'*Agréable récit*) évoquant les émeutes du « peuple » avec une distance certaine[37], alors que le second (l'*Agréable et véritable récit*) est clairement hostile à Mazarin. La proximité des titres suggère d'ailleurs que le second répond au premier, pour lui faire pièce. Ce dernier exemple montre que le pronom se met aisément au service du style pamphlétaire, où les cibles plurielles sont englobées dans une même attaque *ad hominem* (devenant alors *ad homines*). Cet extrait du *Portrait des favoris, en vers burlesques* (p. 18) assimile les partisans (hommes d'affaires enrichis sous Mazarin) à des hommes de main sans scrupule du ministre :

Cette ame fourbe nous deuore,
Et nous fait des maux plus cuiſans,
Par la main de ſes Partiſans,
On viole, on nous aſſaſſine,
On fuit tout ce qu'on s'imagine
Par la plus grande liberté
Où la guerre ait iamais eſté.

Ces deux exemples (allégorie de la « muse burlesque » devenue métonymie des poètes du blocus, et emploi énumératif ou pamphlétaire du *on* de multitude) montrent les atouts du repérage statistique de spécificités contrastives, surtout s'il est combiné à des fonctionnalités comme la visualisation de la textualité, l'examen des cooccurents et le tri des occurrences par casses.

37. « Quand le peuple sera pour nous/Sans doute qu'on filera doux », p. 7) ; de fait on attribue généralement cet écrit à un familier de Gaston d'Orléans, du côté de la Cour à cette date.

7.6 Exploration non-contrastive des écrits burlesques de la Fronde

L'autre série d'expériences considère le corpus d'étude de l'intérieur, sans passer par le contraste – sauf comparaisons particulières rendant plus claire l'interprétation. Nous examinerons les rimes, puisque le corpus présente une véritable homogénéité de ce point de vue (poésies étiquetées comme *vers burlesques* à 97 %). De même, l'homogénéité historique du corpus d'étude permettra enfin l'étude des spécificités sémantiques de quelques termes choisis.

7.6.1 Des rimèmes typiques du burlesque ?

L'expérience consistant à relever les n-grammes de caractères les plus fréquents en fin de ligne a pu montrer que les noms propres ou descriptions définies référant aux événements d'actualité servent très souvent de mots à la rime. En comparant le corpus Burlesque Fronde aux textes du corpus Burlesque hors Fronde, on s'aperçoit que *-dinal* est la 23e finale la plus fréquente, contre le rang 487 hors mazarinades, de même pour *-zarin* (16e *vs.* 783e rime la plus fréquente). Le début du dernier mot des vers est fréquemment : *franc-*, *paris-*, *ville-*, *parle-*, *mazar-*, *cardi-*, *bvrle-*..., et la fin : *-ment*, *-ance*, *-ille*, *-aris*, *dinal*, *zarin*... Il semble donc que les mots de l'actualité (France, Paris/Ville, Parlement, Cardinal, Mazarin) sont non seulement plus fréquents, ce qui est trivial pour des textes d'actualité, mais aussi servent de matrice à la fabrication du vers, puisqu'ils constituent manifestement une contrainte à partir de laquelle s'élaborent les autres rimes.

Plus avant, cette première approche des rimes pourrait mettre au jour des rimèmes [Beaudouin, 2004, p. 123] typiquement burlesques. Pour juger de la fréquence des rimes dans nos données non enrichies, on doit combiner les formes fléchies des finales (singulier et pluriel notamment). Par exemple la rime en *-ade* apparaît 275 fois, ce qui fait partie des hautes fréquences, après toutes les rimes de mots d'actualité. Les mots-rimes sont fréquemment les suivants : *bravades, rodomontades, barricades, estocades, embuscades, capilotades, algarades, boutade, mascarade, boutade, camarade, mousquetade, saluades, gourmades* (« coups de poing »), *grenade* (au sens de « petite bombe »), *incartades, rebuffades, fanfaronnades, mascarades, salade* – au sens technique de « casque », souvent réservé à l'évocation de guerres antiques selon le *Dictionnaire de l'Académie* (1694). Au plan lexical, l'ensemble exemplifie bien le mélange sociolectal, avec des termes techniques (*capilotades, salade*) à côté de mots à connotation familière ou péjorative : le sémantisme du suffixe *-ade* (puisqu'en l'occurrence la rime est un morphème dérivationnel, les mots-rimes étant essentiellement des noms) renvoie à l'idée d'action ou de collectif, souvent avec cette connotation, d'après le *Trésor de la Langue Française*. Cette rime suscite par ailleurs des néologismes : « Dès les premieres barricades/sans recommencer les frondades » (*Tableau raccourci des courtisans*, p. 3). Les titres d'œuvres néologiques (*Juliade, Miliade, et Mazarinade*) exploite la polysémie du suffixe : à la fois formateur de titres

d'œuvres épiques (le sens collectif renvoyant alors au grand nombre de parties composant de telles oeuvres) et péjoration de noms d'action ou de collectif. Les mots-rimes appartiennent d'ailleurs pour plusieurs à une isotopie de la guerre, décrite plus comme une « bastonnade » que comme un conflit organisé.

Cet élément d'explication générique (le burlesque comme parodie d'épopée) n'est néanmoins par le seul : un dictionnaire des rimes de l'époque, particulièrement sensible aux italianismes [La Noue, 1623, p. 34] cite les rimes en *-ade* comme presque toutes issues de noms italiens en *-ata*, tout en précisant qu'« on en pourra adjouter encore icy de nouveaux [...] quand la pratique les aura un peu adoucis ». Ces mots-rimes semblent donc d'une part être encore sentis comme des emprunts (conformément à la tendance anti-puriste du burlesque) et d'autre part apparaître comme relativement rares puisque La Noue indique que sa liste pourrait être augmentée à mesure que l'usage acclimaterait ces mots italiens au français.

Enfin, une recherche de la spécificité de cette rime peut être menée par une confrontation avec le rimarium de [Beaudouin, 2000] [38]. Les rimèmes en *-ade(s)* y sont au nombre de 26 occurrences (sur 80 000 vers étudiés), avec 19 paires de mots-rimes différentes, c'est-à-dire parmi les plus rares de ce corpus. Aussi, ce timbre pourrait sonner comme spécifique du burlesque, et donnerait à cette pratique d'écriture et de lecture sa couleur, et sa voix – car ces écrits étaient vraisemblablement souvent oralisés.

Le même travail pourrait être mené sur des rimes qui semblent intuitivement rares ailleurs et qui sont fréquentes dans le corpus d'étude, comme la rime en *-ique* (qui apparaît au 10e rang du calcul de spécificité du corpus Burlesque Fronde) ou la rime en *-if* ; ce sont des rimes qui sont d'ailleurs utilisées comme contraintes de poèmes monorimes (*L'envoy de Mazarin au Mont Gibel, ou l'etique de Mazarin*, rimé entièrement sur *-ique*) ou birimes (*Virelay sur les vertus de Sa Faquinance*, rimant sur *-if* et *-aire* exclusivement). Il faudrait pour affermir ces constats bénéficier d'un corpus de référence spécifiquement poétique et finement partitionné par genres (sur l'importance du genre pour la constitution des corpus de référence, voir [Rastier, 2011, p.34]).

7.6.2 Plongements lexicaux et similarité sémantique : deux exemples

[Mikolov et al., 2013] ont proposé, en 2013, deux modèles (*CBOW* et *skip-gram*) généralement désigné par l'unique terme *word2vec*. Il s'agit d'associer à tous les *tokens* (ou les occurrences) un vecteur « dense » – entendons, avec un nombre réduit de dimensions, de l'ordre de la centaine. On appelle ces vecteurs des *word embeddings* (ou plongements lexicaux en français). Ces modèles apprennent ces vecteurs, après une initialisation aléatoire, sur un corpus donné à l'apprentissage, selon un processus de prédiction itératif. L'hypothèse distributionniste est ici invoquée, supposant que des mots apparaissant dans des contextes similaires partagent une certaine sémantique. Nous proposons deux exemples d'utilisation –

38. annexe 3 : https://tel.archives-ouvertes.fr/tel-00377348

en sémantique calculatoire – de ce type de vecteurs. La phase d'apprentissage a été réalisée dans les conditions suivantes :

— Librairie : *gensim* (3.8)
— Modèle : *gensim.word2vec* (CBOW)
— Nombre de dimensions : 100
— Fenêtre d'apprentissage : 5
— Fréquence minimale d'apparition d'un mot pour son apprentissage : 5

Nous avons donc cherché à observer la robustesse de cette technologie face à des données (i) sujettes à la variation graphique et (ii) bruitées par l'OCR. Pour cela, nous avons cherché, pour un ensemble fini d'unité lexicale, les vecteurs les plus proches au sens de la similarité cosinus, calculée entre deux vecteurs ainsi :

$$cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{||\, a \,|| \times ||\, b \,||}$$

Nous proposons ici une analyse de certains mots choisis en fonction de notre connaissance du corpus. Il s'agit donc cette fois d'une approche plus *corpus based*.

Parmi les vecteurs les plus proches du vecteur de *paille*, – la paille étant, dans ce conflit, un signe de ralliement des frondeurs (du parti de Condé), qu'ils posaient sur leur chapeau – on trouve : *arme*, *infame*, *iustice*, *rumeur*, *fronde*, *nuict*. À titre de comparaison, avec un apprentissage réalisé sur le corpus Non Burlesque (de la section précédente), on trouve : *farine*, *poix*, *cheminee*, *chaux*, *graisse*, *decoction*, qui renvoie à une attestation de *paille* dans des contextes agricoles et domestiques. Au rebours, on observe bien une proximité sémantique de *paille* et de l'isotopie */politique/* dans le corpus d'étude Burlesque Fronde. Une interprétation de ce résultat consisterait à dire que le mot *paille* apparaît dans des contextes similaires aux mots *arme*, *facon*, *infame*, *iustice*... Les *word embeddings* permettent ainsi de montrer un usage et une sémantique particulière pour le mot *paille* dans les écrits de la Fronde.

Le mot *poëte* offre aussi des résultats intéressants. Dans le corpus d'étude, les vecteurs proches sont : *peintre*, *brutal*, *coquin*, *vieillard*, *métier*, *voleur*, *balet*, *mensonge*, *paysan*, *larron*, *spectacle*, *magicien*, *marchand*, *demon*. À titre de comparaison, un apprentissage des vecteurs réalisé sur le corpus Non Burlesque donne : *philosophe*, *peintre*, *roman*, *personnage*, *traducteur*, *medecin*. On retrouve des noms de métier, ainsi que des substantitfs relevant de l'isotopie */Belles-Lettres/* alors qu'avec notre corpus d'étude on observe plutôt des figures négatives comme *voleur*, *larron*, *vieillard* ou encore *demon*. L'apprentissage des *word embeddings* sur le corpus d'étude Burlesque Fronde suggère ainsi que la figure du poète crotté, topique dans la poésie, notamment satirique, depuis le début du siècle [39] (consulté le 28 juin 2022), y est dans des contextes axiologiquement connotés.

Cette section sera publiée dans la revue « Le Verger ».

39. Pour une mise au point synthétique et une bibliographie sur le sujet, voir : http://www.parnassereforme.fr/fiches/poetecrotte.php

7.7 De l'intérêt des données imparfaites

Ces premières explorations permettent d'assurer l'intérêt de notre hypothèse inaugurale (la fécondité des données bruitées utilisées en corpus) en prouvant leur exploitabilité (existence de phénomènes fréquents) et en en dessinant certaines limites (dont notamment l'étude impossible de la rareté). Toujours est-il que, même si elles nous privent de certaines voies d'étude, les données océrisées – et donc imparfaites au sens où elles sont « bruitées » [40] – permettent l'étude textométrique. En outre, le processus d'OCR (qu'on cherche à maîtriser par ses évaluations) contrecarre la nécessaire transcription manuelle, longue, coûteuse et fastidieuse ; on dispose donc virtuellement d'une quantité supérieure de textes étudiables.

Reste une question en suspens : le « bruit » des données océrisées relève-t-il plus véritablement du *bruit* ou du *silence* ? En clair, les données océrisées contiennent-elles (i) de faux éléments ajoutés ou (ii) de véritables éléments ôtés ? Les termes obscènes, qu'on associe généralement aux écrits burlesques, n'apparaissent pas dans les spécificités calculées entre le corpus BURLESQUE *vs.* NON BURLESQUE ; faut-il y voir la conséquence d'un silence trop important dans les données ? Certaines recherches [Nédélec, 2004] mettent en doute le fait que l'écriture burlesque soit la seule caractérisée par un tel lexique. Le débat reste ouvert.

7.8 Conclusion

Les données océrisées permettent donc l'étude textométrique. Le *distant reading* est effectivement de mise quand l'OCR cherche à acquérir une masse de données ; les phénomènes fréquents le sont d'autant plus. Dans le contexte de l'OCR, l'acquisition de la masse prévaut donc à l'étude de la rareté. Enfin, la critique de l'impossible *close reading* sur de telles données ne tient pas. Effectivement l'on ne peut étudier la rareté avec des textes océrisés, mais comme on l'a dit, l'intérêt de l'OCR est d'acquérir de la masse de données pour étudier ce qui est fréquent. Le *close reading* doit dès lors s'en tenir à la lecture occulaire des œuvres originales.

40. Cette expression condensée inclut les différents types de bruit (dus à la segmentation ou à la transcription) et les silences.

Chapitre 8

Similarité textuelle en contexte bruité

Sommaire

La motivation première de ce chapitre est de proposer une expérience pratique informatique utilisant des textes bruités par OCR, et d'étudier leur exploitabilité. L'idée d'utiliser la similarité textuelle vient de la présence, dans le « corpus » des mazarinades, des recueils éditoriaux. Ceux-ci rassemblent plusieurs mazarinades publiées par ailleurs les unes des autres. Par exemple, on trouve des recueils éditoriaux en cherchant dans la *Bibliographie des mazarinades* de Célestion Moreau les entrées comportant :

— « gazette » ;
— « recueil » ;
— « périodique » (Moreau 811 à 835 et Moreau 2847-2848 – les titres commençant par « courrier » sont des périodiques, souvent avortés, mais pouvant malgré tout avoir connu des éditions en recueils rassemblant les quelques livraisons du périodique paru) ;
— « journaux » du Parlement (Moreau1741, Moreau1742, Moreau1743, Moreau 1750, Moreau1762, Moreau2537, Moreau3054, Moreau3097, Moreau3728, Socard 3-48).

Nous avions d'abord souhaité une vectorisation de toutes les pages de tous les documents du « corpus » pour ensuite calculer une similarités entre toutes les pages pour retrouver les passages équivalents (section 8.1).

Toutefois, cette expérience s'est avérée non seulement difficile mais presque impossible en pratique tant le coup calculatoire est élevé. En effet, ce calcul des similarités est coûteux en termes de complexité : le cosinus a une complexité égale à $O(n2)$, il est quadratique en le nombre de dimensions. Aussi, le nombre de pages du corpus est environ égal à 28 000 ce qui représente plus de 800 millions de paires de similarités à calculer. Pour ces raisons d'ordre technique, nous avons choisi de passer à l'échelle du document pour calculer les similarités.

Par ailleurs, nous disposons d'une ressource qui organise les documents en corpus. Il s'agit des *épisodes* relevés par Célestin Moreau dans sa *Bibliographie des mazarinades*. Il s'agit de certains évévenements marquant de la Fronde auxquels ont été rattachés par ce bibliographe certains les documents qui y font référence comme par exemple l'arrestation des Princes le 18 janvier 1650 (58 mazarinades pour cet épisode). Ils sont au nombre de 112. Il s'agira, après avoir reconnu cette difficulté de repérer automatiquement les passages alignés des recueils éditoriaux automatiquement, de proposer une étude de ces épisodes en partant des hypothèses que les documents d'un même épisode sont plus proches les uns des autres que des documents d'épisodes différents et aussi qu'il peut exister des épisodes plus proches que d'autres (notamment ceux proches sur l'échelle temporelle).

Les figures 8.1, 8.2, 8.3 et 8.4, réalisées par Gaël Lejeune, montrent une organisation graphique chronologique des épisodes de Célestion Moreau.

Au plan épistémologique, ce chapitre propose une expérience non supervisée avec le calcul des similarité entre page de la première partie, et une expérience « pseudo-supervisée » avec l'étude des épisodes de Moreau. Nous chercherons à montrer, par l'exemple, l'importance de la supervision dans le cadre de l'exploitation de données textuelles bruitées. Car, et c'est évident dans une certaine mesure, ajouter de la non supervision à l'étude de données bruitées revient à ajouter une couche de difficultés dans les démonstations. Nous montrons dans ce chapitre l'impossibilité du travail non supervisé en contexte bruité, pour la tâche de similarité textuelle.

8.1 Un repérage automatique des recueils éditoriaux ?

Si la première constatation qui a motivé ce travail a été l'existence des recueils éditoriaux [1], il apparaît que, au sein du corpus des mazarinades, des pages peuvent être, sinon similaires, aussi strictement égales. On peut dès lors se proposer de vectoriser les pages du corpus selon les caractères qu'elles contiennent [2]. Autrement dit, on peut vectoriser les pages selon ses n-grammes de

1. Pour une revue sur ce phénomène éditorial, voir [Bordes, 2022].
2. Sur le débat entre vectorisation par mots ou vectorisation par caractères, voir le chapitre 6 ou les travaux de Claveau.

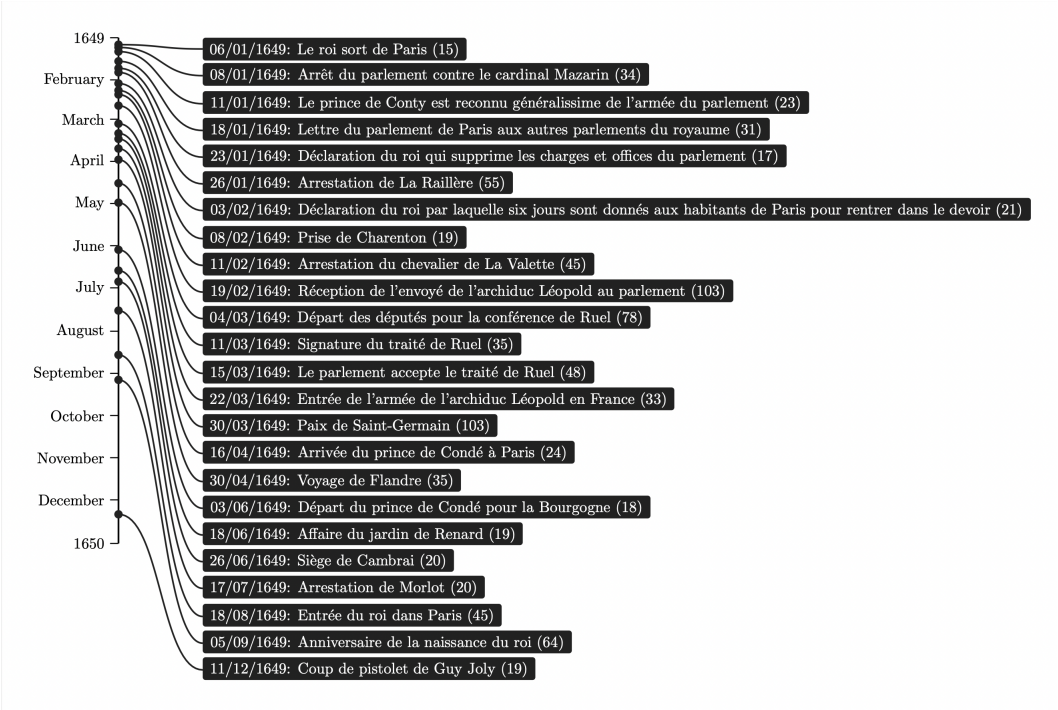


FIGURE 8.1 – Organisation chronologique des épisodes relevés par Célestion Moreau (1649)

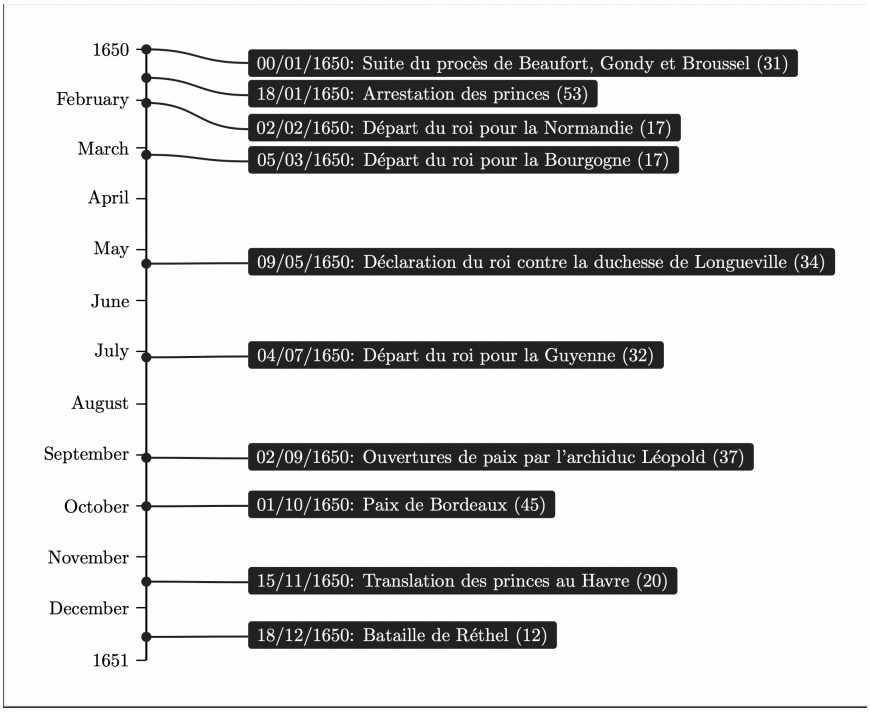


FIGURE 8.2 – Organisation chronologique des épisodes relevés par Célestion Moreau (1650)

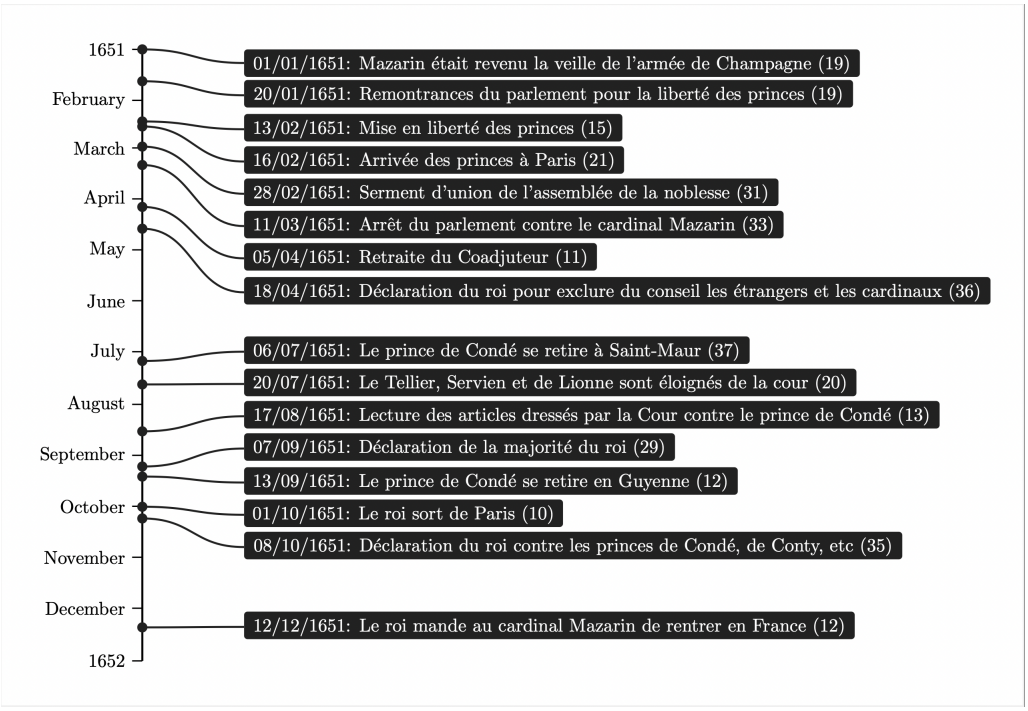


FIGURE 8.3 – Organisation chronologique des épisodes relevés par Célestion Moreau (1651)

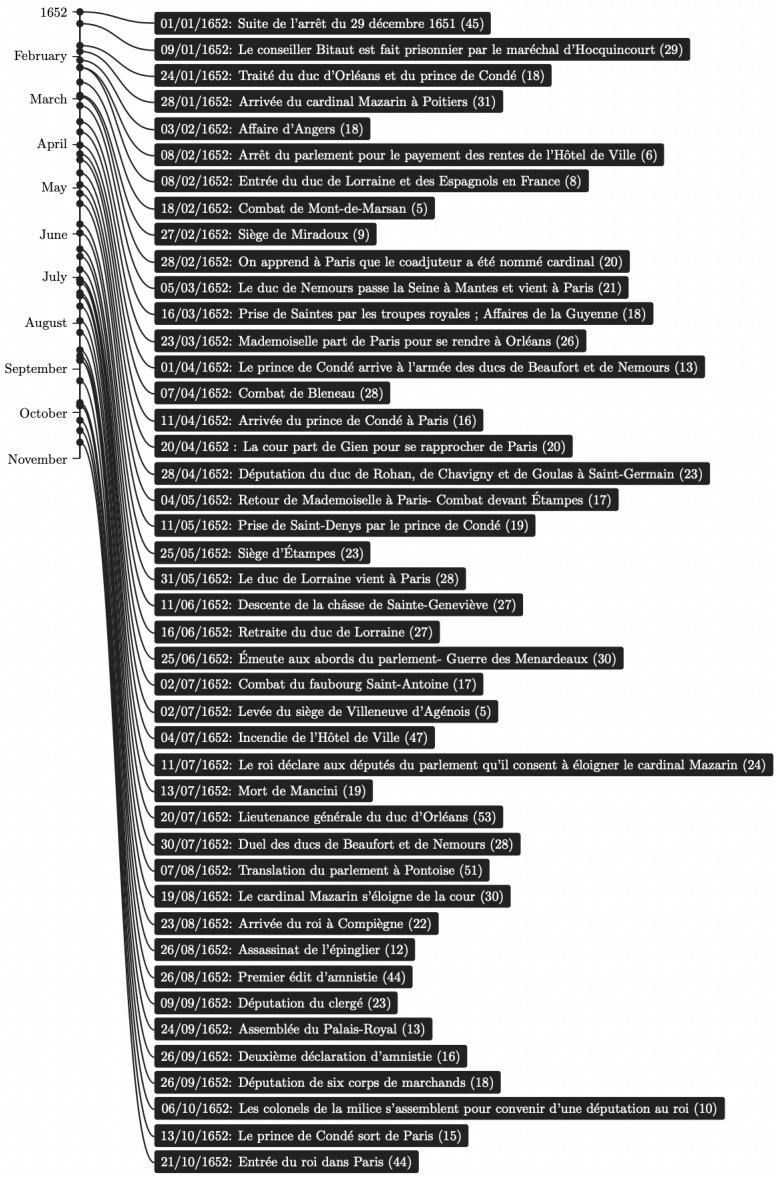


FIGURE 8.4 – Organisation chronologique des épisodes relevés par Célestion Moreau (1652)

caractères. Si la communauté précosine plutôt un n variant de 2 à 3 (notamment [Grefenstette, 1995, Beesley, 1988] pour l'identification de la langue), nous cherchons ici à démontrer le résultat du chapitre 5 qui, lui, préconise un $n = 1$.

Ensuite, il devient possible de réaliser des calculs mathématiques sur ces vecteurs, et en particulier un calcul de distance, notamment cosinus. Si le calcul de la distance cosinus est largement utilisé dans le domaine pour procéder au calcul de distance, il est toutefois de rigueur de noter dès à présent qu'il existe d'autres types ou familles de distance qu'il aurait peut-être été intéressant de tester – notamment Dice ou encore Jaccard.

Dans cette section, nous chercherons à observer, à travers le calcul de similarités et l'alignement textuel, les passages textuels similaires de notre corpus. Nous cherchons *in fine* à procéder à de l'alignement textuel. L'alignement textuel consiste à repérer, dans deux textes différents, disons A et B, les passages qui sont repris dans le texte A ainsi que dans le texte B. Si, pour des textes correctement formés, cette tâche consiste à procéder à un alignement de séquences de caractères au sens strict, pour des données bruitées l'enjeu est de dépasser les erreurs de transcription pour réaliser ledit alignement, en plus des éventuelles variations orthotypographiques.

Si le chapitre précédent (chapitre 7) a montré que le *distant reading* est pertinent sur des textes résultats d'OCR car il est possible de reconstruire des connaissances connues et d'en créer de nouvelles [3], cette section cherche à appuyer ce résultat par une tâche particulièrement délicate en contexte bruité : l'alignement de séquences. Il conviendra de conclure à la fin de cette section de l'impossible alignement en contexte bruité selon la méthode proposée. Toutefois, force est de constater d'entrée de jeu que la tâche proposée n'est peut-être pas la plus pertinente même si le corpus l'imposerait presque ; peut-être est-ce en réalité le travail d'outils déjà existant comme *Text Pair*.

8.1.1 État de la question

Alignement textuel : au carrefour des disciplines

Dans [Saignol, 2021, section 1] est retracée brièvement l'histoire de l'alignement textuel au prisme d'une contextualisation épistémologique. Cette histoire commence en 1965 avec la formalisation de la distance d'édition par Vladimir Levenshtein. Cette quantité mesure le nombre d'étapes minimales nécessaires au passage d'une chaîne de caractères de départ à une chaîne de caractères d'arrivée (avec les trois opérations que l'on connait déjà : l'insertion, la suppression et le remplacement). L'algorithme de Needleman Wunsch poursuit cette recherche pour l'alignement textuel, en 1970, en permettant l'alignement de macromolécules biologiques. Il s'agit d'un « acte de naissance » de la bio-informatique qui « met au jour de nouvelles perspectives de recherche en distinguant les alignements « globaux » (sur l'ensemble de la longueur des séquences) ou « locaux » (limités à certaines régions) ». Dans les années 1980 se construisent de

3. Par exemple, retrouver des rapprochements déjà observés par la critique érudite sur le sujet.

vastes corpus multilingues dans le cadre de projet internationaux. C'est alors le « passage à l'échelle » (capacité d'un algorithme à fonctionner suivant un changement de taille ou de volume conséquent) qui constitue un obstacle, lequel sera en partie résolue durant la décennie suivante grâce aux développement de la bioinformatique. Le logiciel BLAST (Basic Local Alignment Search Tool) est ici exemplaire de cette avancée pour le domaine, comme le programme (intégré au système d'exploitation GNU) DIFF. [Saignol, 2021, section 1] constate même l'intégration native de systèmes de comparaison de deux versions d'un même document dans les suites bureautiques telles que WORD ou LIBREOFFICE. Les années 2000 marquent la réunion de la bioinformatique et de la recherche littéraire, notamment avec le logiciel MEDITE, conçu conjointement par le laboratoire d'informatique Lip6 de l'Université Pierre et Marie Curie et l'Institut des Textes et Manuscrits Modernes (ITEM). Les années 2010 voient finalement s'engager une réflexion sur la visualisation et l'interopérabilité des aligneurs.

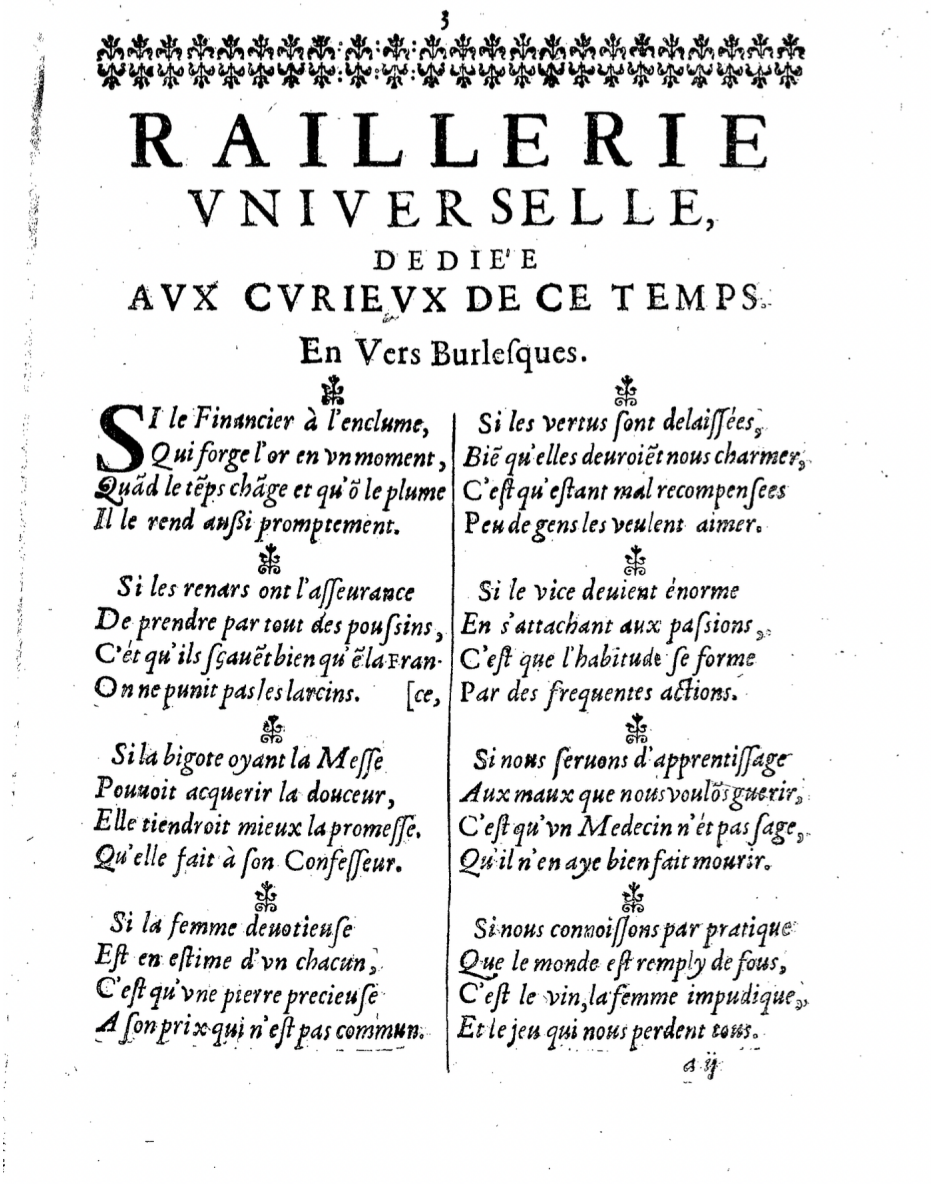
Les auteurs de COLLATEX (*T.I.D. Group*), cités par [Saignol, 2021, section 2], identifient les différentes étapes généralement effectuées pour procéder à « mise en correspondance de séquences textuelles par le biais d'un traitement computationnel » :

— Tokenisation
— Normalisation : « correction des textes en supprimant les informations peu pertinentes et génératrices de bruit (la ponctuation par exemple) » ;
— Alignement
— Analyse
— Visualisation

Au plan du fonctionnement des algorithmes, [Brixtel, 2007, section 2] rappelle les deux méthodes principales d'alignement phrastique. Celle de [Kay et al., 1994] consiste à aligner la première et la dernière phrase des deux versions du document et ensuite aligner les phrases intermédiaires avec comme hypothèse que deux phrases sont alignées si elles partagent un maximum de mots. L'autre méthode, celle de [Gale and Church, 1991], se base sur la longueur en caractères des phrases à aligner, avec comme postulat que deux phrases alignées sont deux phrases qui partagent le même nombre relatif de caractères.

8.1.2 Alignement endogène

Une première étape de ce travail consiste à réaliser l'alignement textuel sur un petit corpus de textes dans le but d'établir la meilleure *pipeline*. Il s'agit de ne travailler que sur les mazarinades océrisées (voir Chapitre 6) pour i) apprendre différents types de vecteurs associés à chaque page du corpus, ii) de calculer différents types de mesure de similarités entre tous ces vecteurs, iii) de trouver quelles similarités permettant d'associer des pages candidates à l'alignement textuel et iv) de procéder à l'alignement à proprement parler. Puisque les vecteurs sont appris sur le même corpus qui servira aussi à tester l'alignement, cette expérience est qualifiée d'*endogène*.



3

RAILLERIE VNIVERSELLE,

DEDIE'E

AVX CVRIEVX DE CE TEMPS.

En Vers Burleſques.

SI le Financier à l'enclume,
Qui forge l'or en vn moment,
Quãd le tẽps chãge et qu'õ le plume
Il le rend auſſi promptement.

Si les renars ont l'aſſeurance
De prendre par tout des pouſsins,
C'ét qu'ils ſçauẽt bien qu'ẽ la France
On ne punit pas les larcins.

Si la bigote oyant la Meſſe
Pouuoit acquerir la douceur,
Elle tiendroit mieux la promeſſe.
Qu'elle fait à ſon Confeſſeur.

Si la femme deuotieuſe
Eſt en eſtime d'vn chacun,
C'eſt qu'vne pierre precieuſe
A ſon prix qui n'eſt pas commun.

Si les vertus ſont delaiſſées,
Biẽ qu'elles deuroiẽt nous charmer,
C'eſt qu'eſtant mal recompenſees
Peu de gens les veulent aimer.

Si le vice deuient énorme
En s'attachant aux paſsions,
C'eſt que l'habitude ſe forme
Par des frequentes actions.

Si nous ſeruons d'apprentiſſage
Aux maux que nous voulõs guerir,
C'eſt qu'vn Medecin n'ét pas ſage,
Qu'il n'en aye bien fait mourir.

Si nous connoiſſons par pratique
Que le monde eſt remply de fous,
C'eſt le vin, la femme impudique,
Et le jeu qui nous perdent tous.

FIGURE 8.5 – Page 6 de la *Raillerie universelle dediee aux curieux de ce temps. En Vers Burlesques.* (Moreau : 2 960)

Un exemple de textes à aligner

En guise d'introduction, nous donnons l'exemple de deux documents contenant le même texte. Il s'agit de la *Raillerie universelle dediee aux curieux de ce temps. En Vers Burlesques.* (indice dans la bibliographie de Célestin Moreau : 2 960) et des *Veritez absolues, et sans contredit. En Vers Burlesques.* (deuxième supplément de Célestin Moreau, indice : 207), 1652, Paris.

En figures 8.5 et 8.6 on trouve les pages 6 des documents précités correspondant aux début de ces vers burlesques. On observe que le même texte apparaît dans les deux pages présentées, lequel commence en ces termes : *Si la femme deuotieufe / Eft en eftime d'vn chacun / C'eft qu'vne pierre precieufe / A fon prix qui n'eft pas commvn.*

L'objectif est donc de proposer une méthodologie et une méthode permettant de rapprocher les pages du même type que l'exemple, et ce automatiquement.

Un premier filtre : le taux d'erreur estimé des pages

Comme nous le développerons dans la sous-section suivante, un des objectifs est de filtrer les pages non pertinentes à l'alignement textuel. Un premier filtre peut donc être de supprimer du corpus les pages ayant un taux d'erreur au caractère estimé inférieur aux pages contenant effectivement du contenu textuel.

La figure 8.7 montre un grand nombre de pages vides ou presque vides. On observe aussi que certaines pages contiennent beaucoup de caractères (plus de 3 500). Cet histogramme permet de filtrer les pages contenant moins de 500 caractères et celles en contenant plus de 3 500 (valeurs aberrantes).

La figure 8.8 permet de raisonnablement n'accepter que les pages ayant un T_{lex} compris entre $0,4$ et $0,8$.

La figure 8.9 montre que la majorité des pages ont un CER prédit entre 0 et 750 ; nous filtrons donc selon cet intervalle.

La figure 8.9 montre clairement un taux d'erreur prédit qui augmente lorsque le nombre de caractères augmente lui aussi. Cela vient appuyer les décisions prises avec les graphes 8.7 et 8.9 : il faut filtrer les pages contenant moins de 10 caractères (pas assez d'observables), celles en contenant plus de 3 500, celles ayant un CER prédit hors des bornes 0 et 750.

Enfin, la figure 8.11 ne permet pas de discriminer certaines pages selon la double approche T_{lex} et nombre de caractères. Nous gardons donc sans l'appuyer plus encore le constat permis grâce à la figure 8.8 : nous ne gardons les pages ayant un T_{lex} compris entre $0,4$ et $0,8$.

Après suppression des pages correspondant aux critères établis ci-dessus, le corpus contient 26 869 pages contenant très probablement du contenu textuel. Nous procédons ensuite au calcul des similarités entre ces pages pour réaliser un deuxième filtre précédant l'alignement textuel.

Similarité entre textes Le calcul des similarités entre textes se réalise à l'échelle de la page pour permettre d'identifier les passages potentiellement

LES VERITEZ ABSOLVES, ET *ſans contredit.*

EN VERS BVRLESQVES.

CES veritez que ie vous donne
S'accommodent à noſtre tẽps:
Et ie crois n'offencer perſonne
Si ie parle des mal contens.
Chacun croit bien dedãs ſon ame
Qu'il n'a rien qui le peut picquer,
Soit Gentilhõme ou biẽ ſoit Dame,
Et pourtant veux-ie en attaquer.
Ce n'eſt pas pourtãt que ie touche
La vertu ny les vertueux,
Qui ſe ſẽt morueux qu'il ſe mouche
De moy ie ne dis rien pour eux.
Si la femme deuotieuſe
Eſt en eſtime d'vn chacun,
C'eſt qu'vne pierre precieuſe
A ſon prix qui n'eſt pas commun.
Si les vertus ſont delaiſſées,
Biẽ qu'elles deuroiẽt nous charmer,
C'eſt qu'eſtant mal recompenſées
Peu de gens les veulent aimer.
Si le vice deuient enorme
En s'attachant aux paſſions,
C'eſt que l'habitude ſe forme
Par des frequentes actions.
Si nous ſeruons d'apprentiſſage
Aux maux que nous voulons guerir,
C'eſt qu'vn Medecin n'eſt pas ſage,
Qu'il n'en aye bien fait mourir.
Si nous cognoiſſons par pratique
Que le monde eſt remply de fous,
C'eſt le vin, la femme impudique,
Et le ieu qui nous perdent tous.
Si la perſonne bien accorte
Se regle en ſa condition,
C'eſt crainte d'vne chaine forte
Que traine l'obligation.
Si l'on met de la difference
Aux hommes par l'exterieur,
C'eſt qu'on s'attache à l'apparence,
Quand on ne connoiſt pas vn cœur.
Si le ſage aime mieux vn Liure
Qu'vn homme pour ſon entretien,
C'eſt qu'vn liure aprẽd à biẽ viure,
Et l'homme ſouuent ne vaut rien.
Si la choſe a plus de merite
En cachant les defaurs qu'elle a,
L'eau qui dort & l'hõme hypocrite
Ont de l'auantage en cela.

FIGURE 8.6 – Page 6 des *Veritez absolues, et sans contredit. En Vers Burlesques.* (Moreau, supplément 2 : 207)

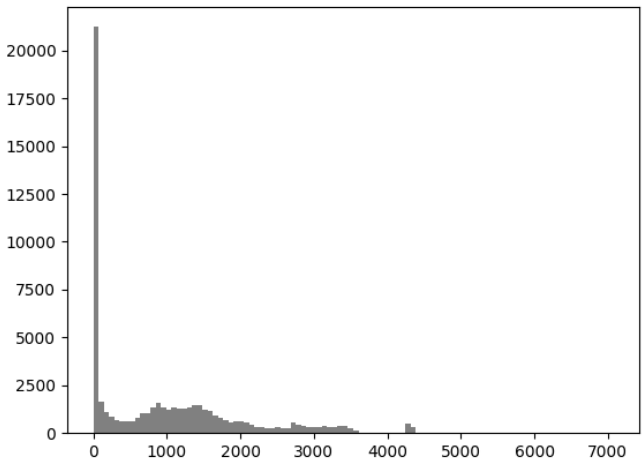


FIGURE 8.7 – Histogramme des nombres de caractères sur l'ensemble des pages du corpus des mazarinades

empreints de reprises. Pour réaliser ce calcul de similarité, il convient de proposer une méthode de vectorisation du contenu textuel des pages numérisées. Ces vectorisation peuvent être de plusieurs types, se réaliser au grain mot ou au grain caractère ; elles peuvent aussi être creuses ou denses. Ensuite, le calcul des similarités peut lui aussi être réalisé de plusieurs manières tant le nombre de distances (le complémentaire à 1 de la similarité) est conséquent. Il s'agit ici de trouver la meilleure combinaison vectorisation-similarité permettant de discriminer les pages pertinentes à l'alignement textuel, contenant possiblement des reprises donc, de celles n'en contenant probablement pas.

Il est nécessaire de réaliser cette étape préalable car le temps de calcul nécessaire pour procéder à l'alignement de séquences est particulièrement coûteux (cf. le coût quadratique de son calcul). Pour un MacBook Pro 2021, processeur Apple M1 Pro, 32Go mémoire unifiée, on a :

— avec 10 documents : la similarité entre paires prend 1,45 secondes et l'alignement en prend 0,41 ;
— avec 100 documents : la similarité entre paires prend 2,71 secondes et l'alignement en prend 3,84 ;
— avec 1 000 documents : la similarité entre paires prend 16,93 secondes et l'alignement en prend 34,69.

Il n'y a donc pas de passage à l'échelle, ni pour les similarités ni pour l'alignement. Mais ce dernier se révèle plus chronophage que le premier.

Outre ce coût computationnel, si l'alignement était réalisé sur toutes les pages du corpus (qui contient environ 60 000 pages, ce qui fait environ $3,5$ milliards de combinaisons et environ $3,5/2$ milliards de combinaisons différentes) il générerait

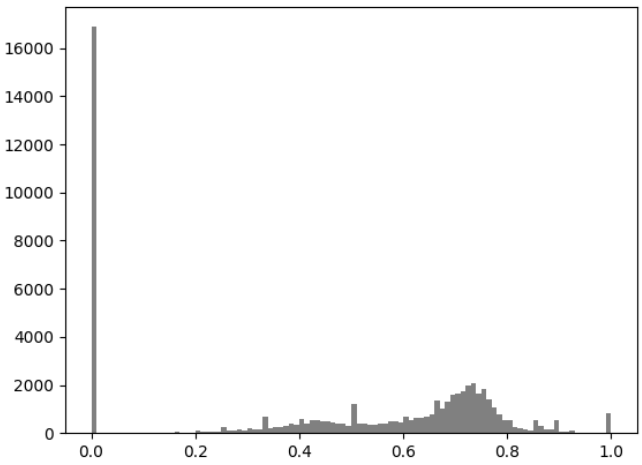


FIGURE 8.8 – Histogramme des Tlex sur l'ensemble des pages du corpus des mazarinades

tant de résultats que leur analyse sera bien impossible. Le calcul des similarités vient donc jouer un rôle de filtre préalable au calcul et à l'analyse manuelle ; on vient attaquer le corpus non pas dans son entiereté mais sur une partie qui, *a priori* semble pertinente.

Vectorisation Si la vectorisation de documents peut se faire de plusieurs manières [4], il en est des plus simples et des plus complexes. Dans le contexte qui est le nôtre, nous choisissons une vectorisation simple et peu coûteuse (selon le temps computationnel mais aussi selon le temps de la compréhension de la construction de ces vecteurs). Nous choisissons donc de réaliser une vectorisation par compte absolu et relatif des caractères (unigrammes) présents dans le document courant. La dimension de ces vecteurs est donc égale au vocabulaire du corpus, c'est-à-dire à l'alphabet de celui-ci. Il contient 61 entrées.

Calcul des (dis-)similarités Le calcul des distances, qui est équivalent à celui du calcul des similarités, peut être fait de plusieurs manières. Nous choisissons trois représentants état de l'art des trois « familles » de distances (voir par

4. D'abord, elle peut être *endogène* ou *exogène*, c'est-à-dire que les données servant à apprendre les vecteurs sont les mêmes que celles qu'il faut vectoriser, ou pas. Ensuite, pour réaliser la vectorisation, on peut opter pour des vectorisations creuses ou denses, c'est-à-dire que le nombre de dimensions des vecteurs est soit égal au nombre d'entrées du vocabulaire du corpus ou bien seulement quelques centaines. Cette différence est le résultat de la différence de l'apprentissage des vecteurs. Enfin, les vecteurs peuvent être appris au grain mot ou au grain caractère.

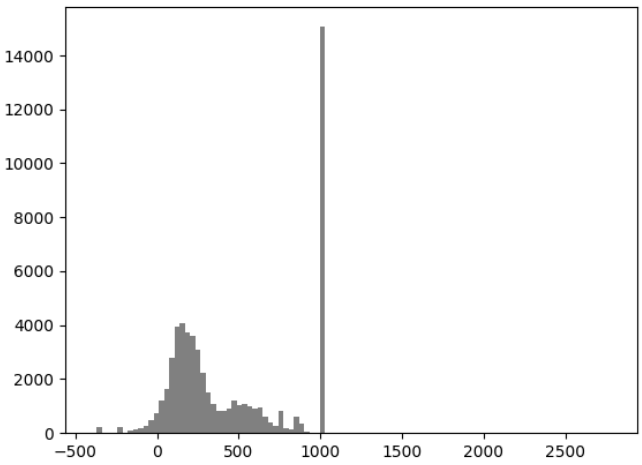


FIGURE 8.9 – Histogramme des CER prédits sur l'ensemble des pages du corpus des mazarinades

exemple : https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.spatial.distance.html consulté le 28 juin 2022) : *cosinus* (sur valeurs réelles), *Dice* (sur valeurs booléennes) et *euclidienne* (sur valeurs réelles). Notons d'entrée de jeu que la distance *Dice* se mesurant sur des valeurs booléennes (ensemble de Vrais et de Faux), elle perd de son sens sur des vecteurs denses car tous les vecteurs seraient considérés comme des vecteurs ne contenant que la valeur booléenne Vrai, ce qui, dans le cas de nos deux vectorisations, ne pose toutefois pas de problème. Ces calculs ont été réalisés avec la librairie *Python 3* nommée *scipy*.

On dispose donc, pour chaque paire de pages à analyser, de deux types de vecteurs (compte absolu et compte relatif) et de trois types de mesure de distance (cosinus, dice et euclidienne). Il convient donc de mener une étude liminaire pour établir quel est le type de vectorisation et la mesure de distance la plus efficace en terme de discrimination des pages les plus « proches » (au sens qu'elles contiennent possiblement des reprises textuelles).

Analyse par graphes Dans cette sous-sous-section, nous cherchons à définir quelle est la meilleure mesure de similarités nous permettant de discriminer les pages contenant potentiellement un passage de reprise ou de réécriture. Nous travaillons d'abord avec les distances. S'agissant de proposer des rapprochements de pages (faible distance), nous devons définir i) une mesure adaptée discriminant plusieurs groupes de pages et ii) un seuil (ou deux : un minimum et un maximum) permettant de trouver les pages potentiellement rapprochables. Pour cela, nous

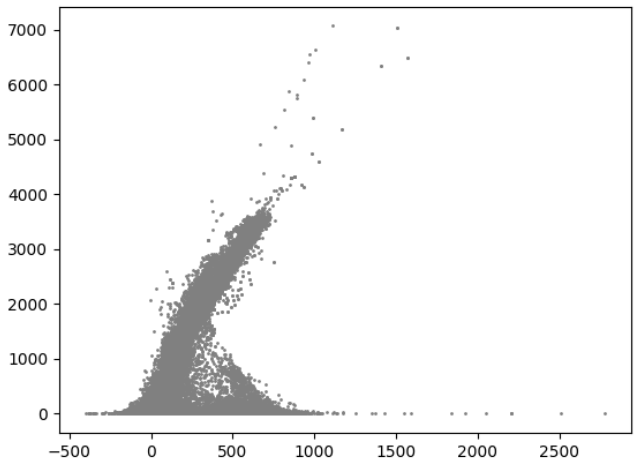


FIGURE 8.10 – Nuage de points des taux d'erreur prédits (abscisses) en fonction du nombre de caractères (ordonnées), pour toutes les pages du corpus des mazarinades

proposons une étude en *DataViz*[5].

Les cartes de chaleur représentent les valeurs sous forme de carrés colorés. Plus un carré est clair (ou foncé, cela dépend du réglage) plus la valeur sous-jacente est importante (ou faible, donc). Dans notre cas, celui de la distance entre pages de mazarinades, nous aurons en axe des ordonnées toutes les pages des mazarinades, et en abscisses aussi, de sorte qu'une matrice carré apparaisse ; les valeurs représentées par couleur sont donc les valeurs des distances calculées entre les pages en abscisses et en ordonnées.

En première hypothèse, nous pouvons supposer qu'une diagonale foncée apparaîtra car, sur la diagonale, sont représentées les valeurs des distances calculées sur les mêmes pages (donc, pas sur deux pages différentes). Ensuite, nous pouvons aussi et raisonnablement supposer que certaines pages du corpus seront ou très similaires ou très distantes (selon le cas, mais le principe est le même) à toutes les autres pages du corpus.

En outre, nous cherchons ici, par l'observation des cartes de chaleur, à trouver quelle est la distance qui permettrait de distinguer plus facilement les pages similaires de celles qui ne le sont probablement pas. En clair, nous cherchons, avec comme hypothèse raisonnable que les pages contenant des reprises sont minoritaires, des cartes de chaleur ou les zones foncées (correspondant à de faibles distances, donc de forte similarités) sont elles aussi minoritaires.

5. *DataViz*, contraction de *Data Vizualisation* signifie « visualisation de données ». Cela signifie que nous allons analyser les mesures de similarités et les seuils à définir en nous appuyant uniquement sur des graphes.

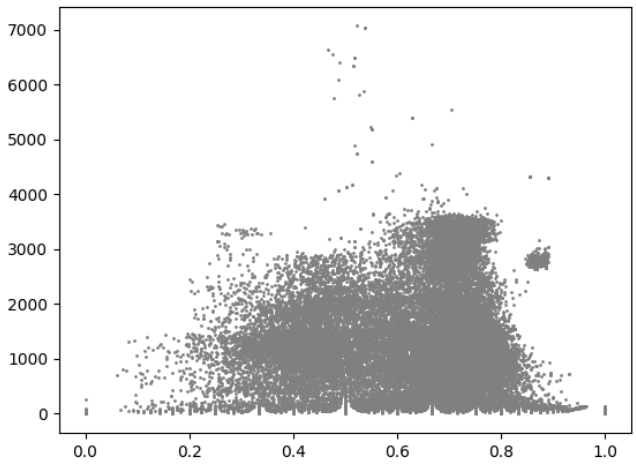
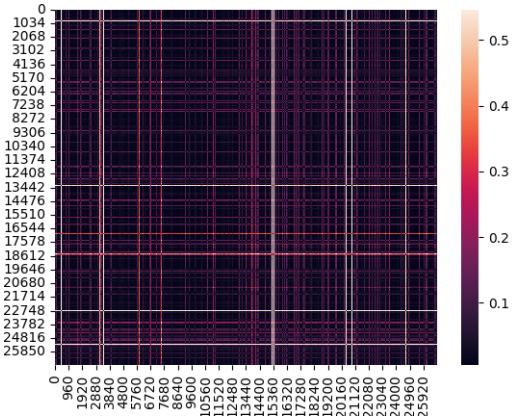


FIGURE 8.11 – Nuage de points des taux de lexicalité (abscisses) en fonction du nombre de caractères (ordonnées), pour toutes les pages du corpus des mazarinades

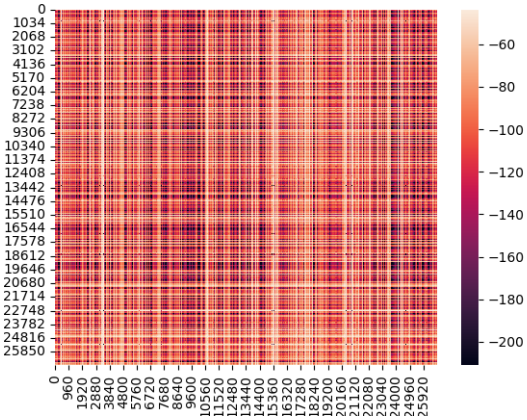
Les cartes de chaleur des figures 8.12a, 8.12b et 8.12c (compte absolu) et des figures 8.13a, 8.13b et 8.13c (compte relatif) indiquent que la distance cosinus n'est pas la plus pertinente pour notre problématique, tant les cartes sont foncées (ce qui signifie que la majorité des pages sont considérées similaires entre elles, ce qui ne nous aide pas à les discriminer). En revanche, la mesure Dice, pour ces deux vectorisations, semble plus adaptée, quoique son sens reste encore obscure (voir sous-sous-section 8.1.2). On observe aussi que la vectorisation à compte relatif offre des distances globalement plus faibles entre les vecteurs. En outre, et ce pour les trois mesures de similarités, on observe des lignes blanches (et pas de lignes noires) ce qui signifie que certaines pages du corpus sont très différentes des autres et ce quelle que soit la page présentée à la comparaison.

Les figures 8.14a, 8.14b, 8.15a, 8.15b, 8.16a et 8.16b montrent, avec une plus grande franchise de ce résultat pour la vectorisation par compte absolu, qu'on a une propension plus importante à avoir une faible distance lorsqu'il y a plus de caractères dans les deux pages. Cela signifie donc que plus il y a d'observables et plus ceux-ci sont communs, plus la distance est généralement faible ; ce qui va bien dans le sens de l'intuition qu'on en a. Il est donc pertinent de n'observer, pour l'alignement à venir, que les paires de pages ayant un nombre de caractères conséquent et une faible distance. Ce résultat s'observe plus aisément avec la distance cosinus.

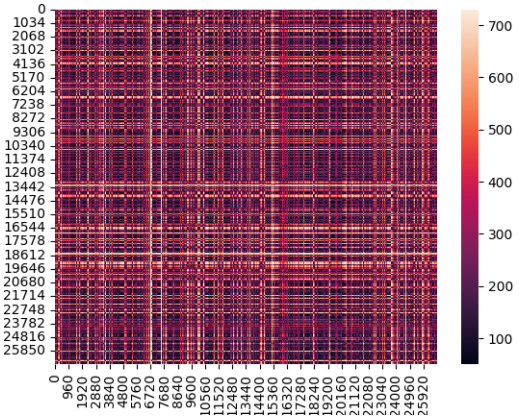
Les figures 8.17a, 8.17b, **??**, 8.18b, 8.19a et 8.19b montrent quant à elles que les faibles distances (qu'elles soient cosinus, dice ou euclidiennes) ont plus de taux de lexicalité différents (faibles ou forts, donc) que les fortes distances.



(a) Distance cosinus

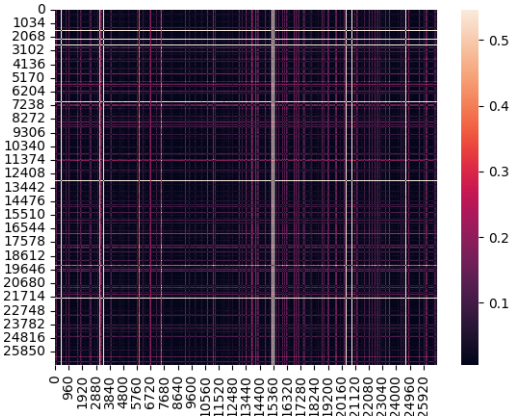


(b) Indice de Dice

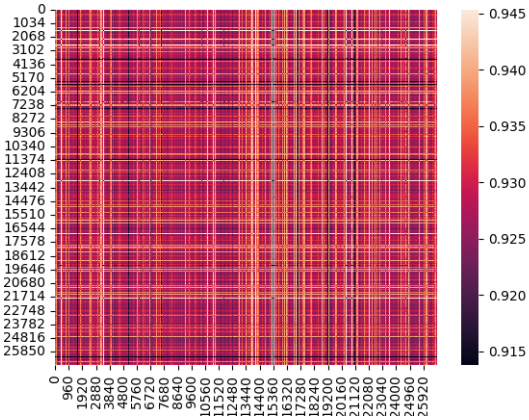


(c) Distance euclidienne

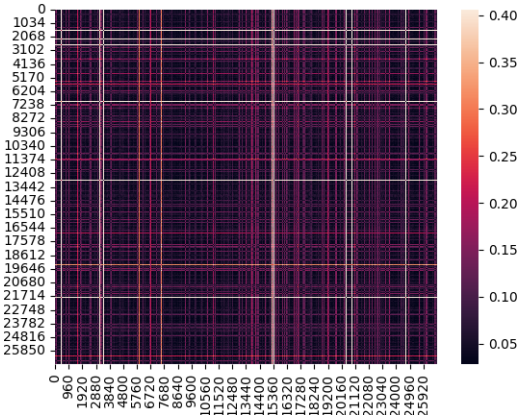
FIGURE 8.12 – Distance entre les pages du corpus (vectorisation en valeurs absolues)



(a) Distance cosinus

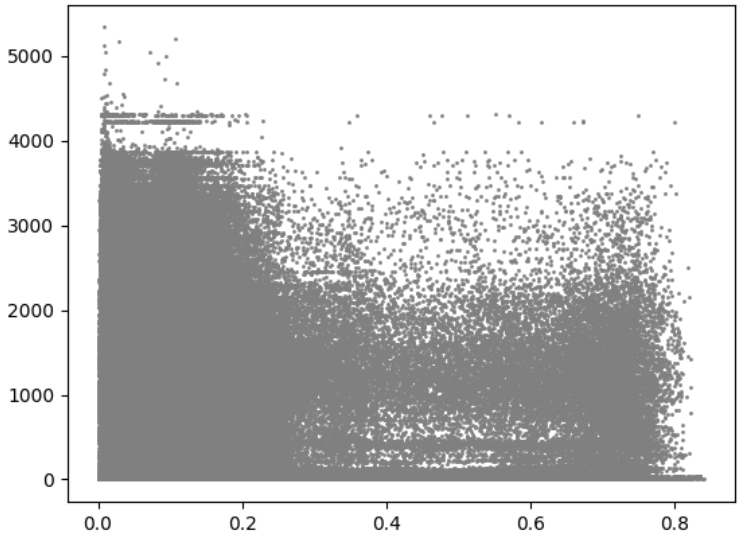


(b) Indice de Dice

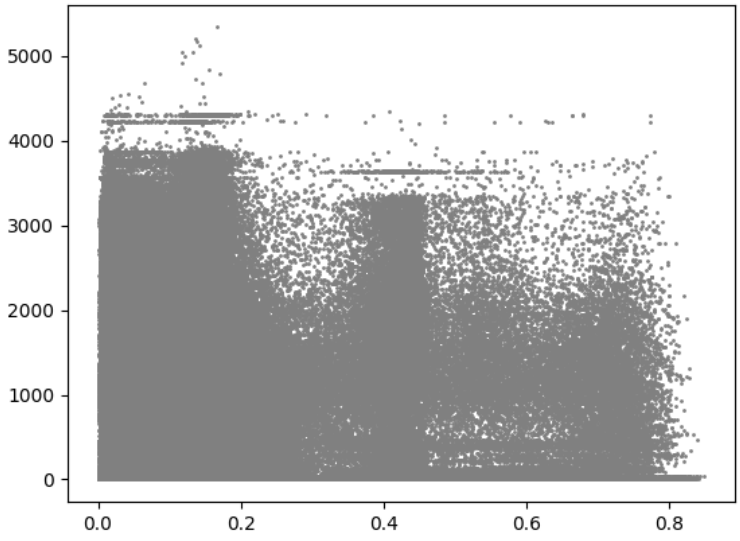


(c) Distance euclidienne

FIGURE 8.13 – Distance entre les pages du corpus (vectorisation en valeurs relatives)

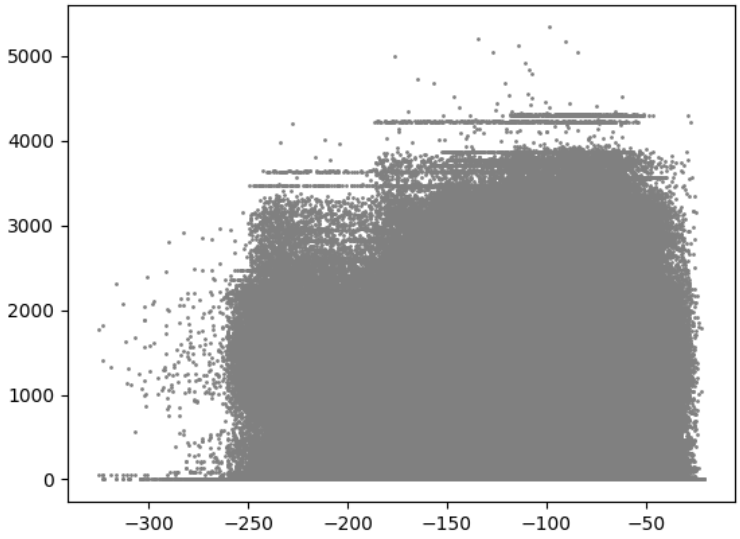


(a) Vectorisation en valeurs absolues

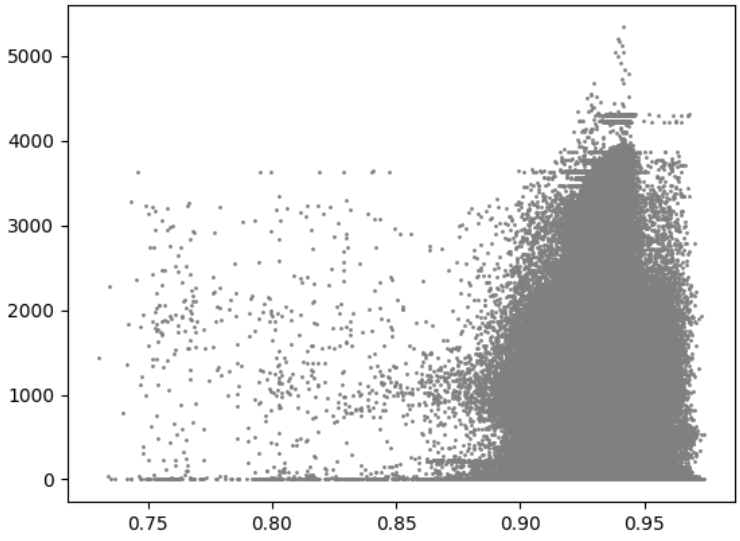


(b) Vectorisation en valeurs relatives

(c) Nuages de points des distances cosinus en fonction de la moyenne harmonique des nombres de caractères sur le corpus des mazarinades

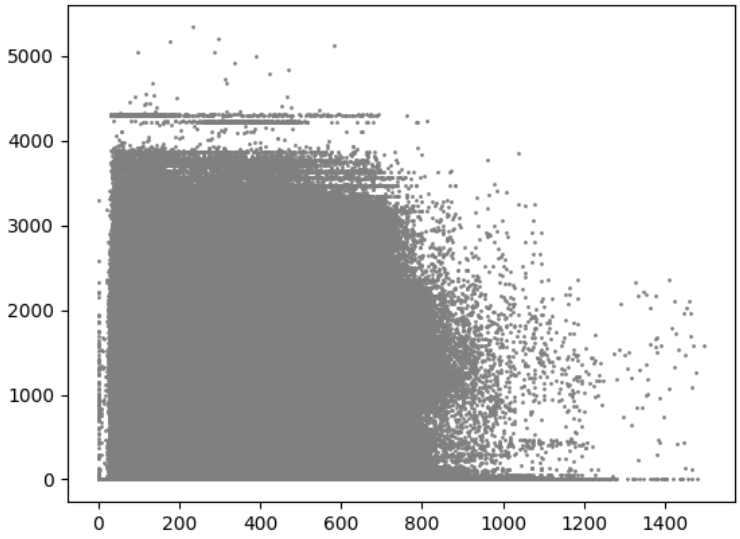


(a) Vectorisation en valeurs absolues

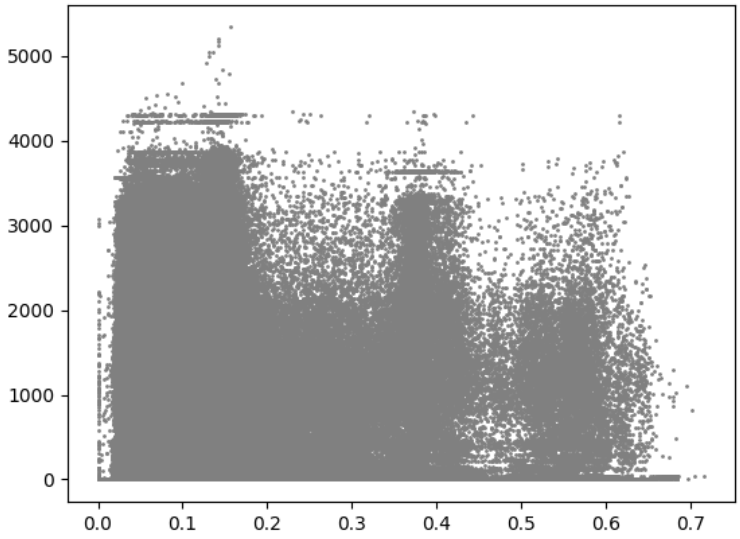


(b) Vectorisation en valeurs relatives

FIGURE 8.15 – Nuages de points des distances Dice en fonction de la moyenne harmonique des nombres de caractères sur le corpus des mazarinades

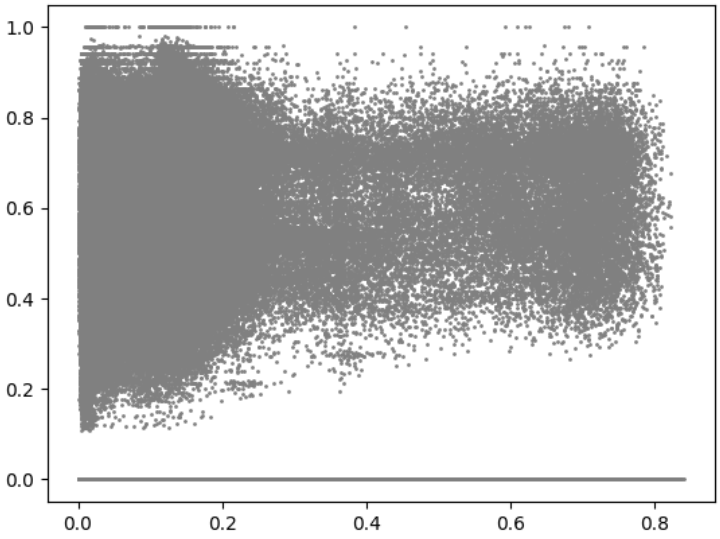


(a) Vectorisation en valeurs absolues

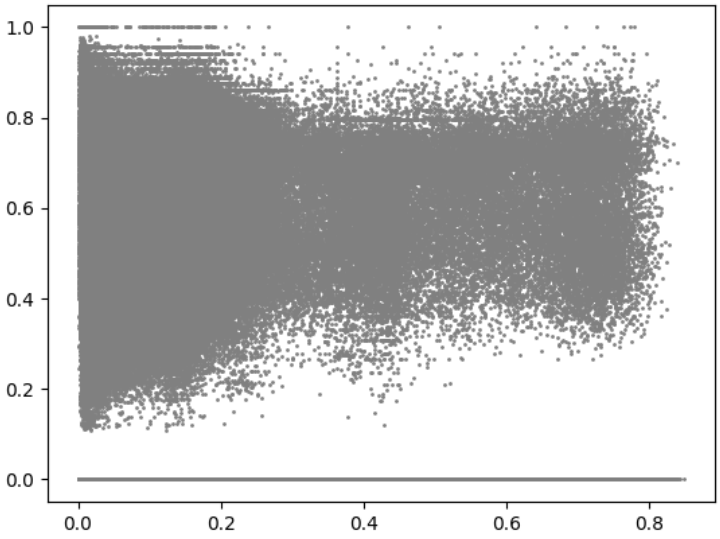


(b) Vectorisation en valeurs relatives

FIGURE 8.16 – Nuages de points des distances euclidiennes en fonction de la moyenne harmonique des nombres de caractères sur le corpus des mazarinades

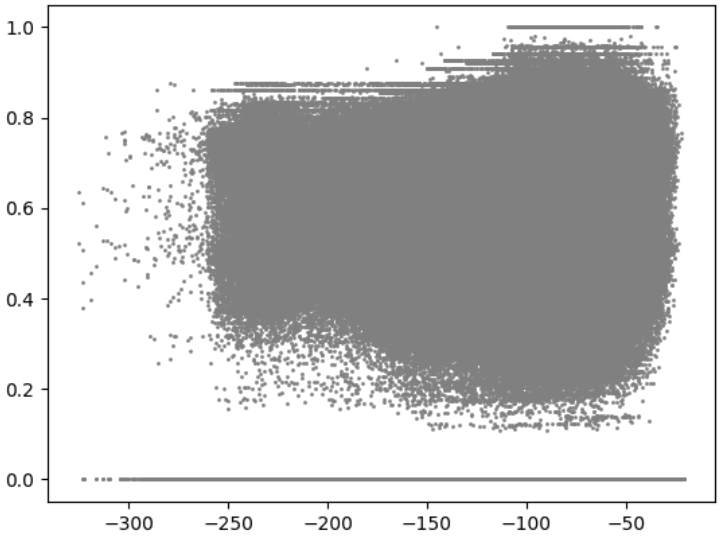


(a) Vectorisation en valeurs absolues

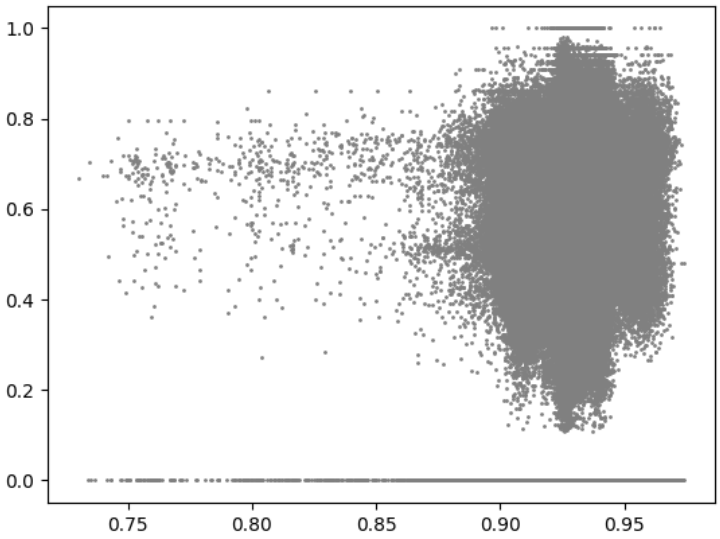


(b) Vectorisation en valeurs relatives

FIGURE 8.17 – Nuages de points des distances cosinus en fonction de la moyenne harmonique des taux de lexicalité sur le corpus des mazarinades

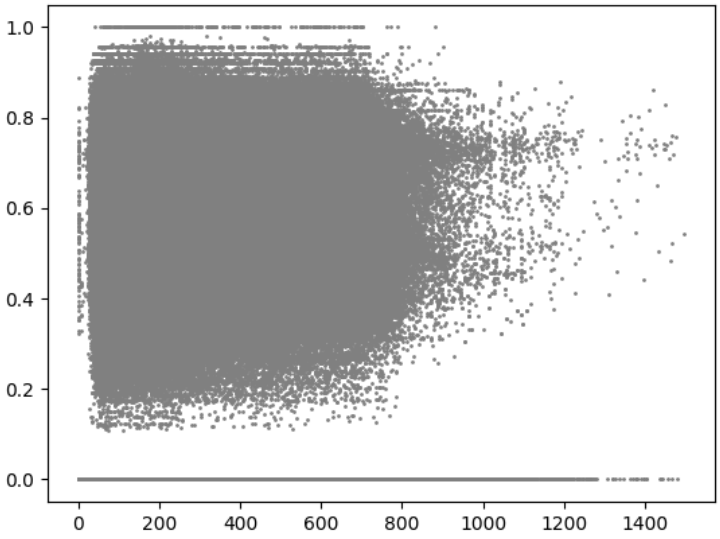


(a) Vectorisation en valeurs absolues

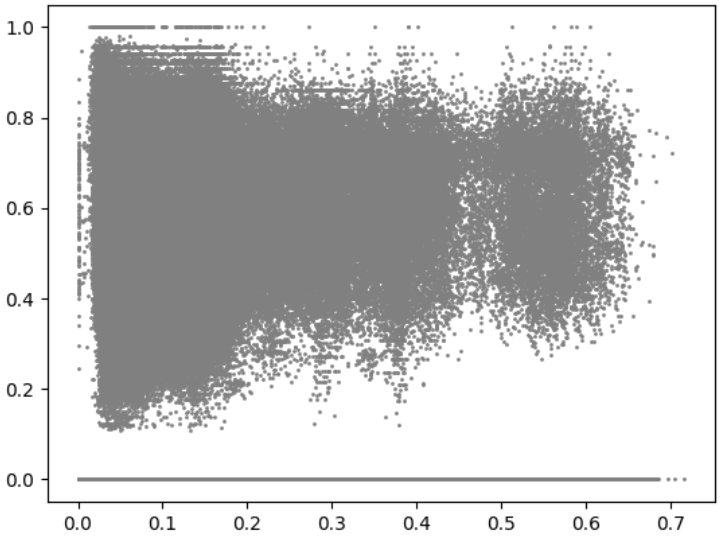


(b) Vectorisation en valeurs relatives

FIGURE 8.18 – Nuages de points des distances Dice en fonction de la moyenne harmonique des taux de lexicalité sur le corpus des mazarinades



(a) Vectorisation en valeurs absolues



(b) Vectorisation en valeurs relatives

FIGURE 8.19 – Nuages de points des distances euclidiennes en fonction de la moyenne harmonique des taux de lexicalité sur le corpus des mazarinades

On n'observe donc pas une tendance qui permettrait de discriminer les pages propices à l'alignement. Ce qui signifie que l'on ne peut pas filtrer selon le critère du taux de lexicalité, lequel peut être faible ou fort pour les faibles distances.

Il en va de même pour les représentations graphiques en fonction du CER estimé (voir les figures **??**, 8.20b, **??**, 8.21b, **??** et 8.22b). En effet, on n'observe pas une tendance selon laquelle à mesure que le CER est faible on aurait plus de chances d'observer des similarités fortes et donc pertinentes à l'alignement.

En conclusion, nous retenons la vectorisation des pages par compte absolu car elle propose moins de concentration des individus (en termes de distance entre ceux-ci) ce qui permet de mieux les discriminer. La mesure cosinus couplée au nombre de caractères contenus dans les pages semblent être la meilleure combinaison pour discriminer les pages alignables ; pour les faibles distances cosinus, on observe que sur le nuage de points on a plus d'individus à fort nombre de caractères (moyenne harmonique).

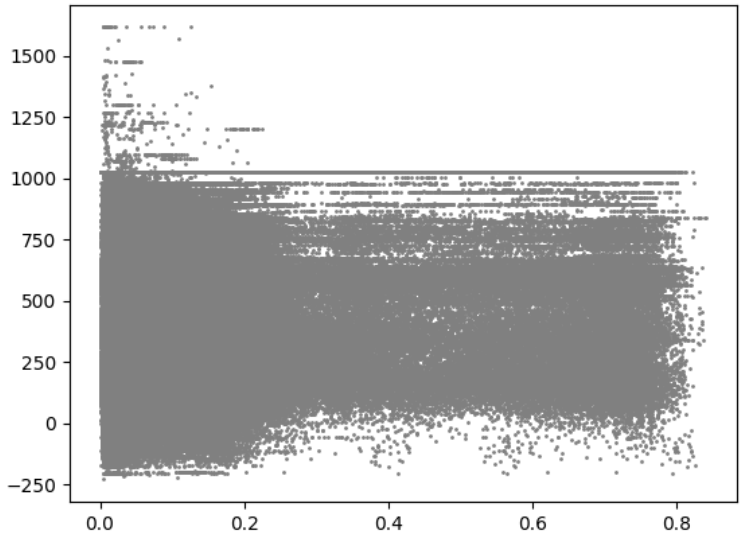
Le 8.23 montre que les scores d'alignement vont de, en gros, 500 à 1500, pour une similarité entre page d'au moins 0.99 (similarité cosinus). On observe un petit mont autour de 1400. Aussi proposons-nous dans la sous-section suivante une étude des pages concernées par une forte similarité et un fort score d'alignement global.

Alignement de séquences

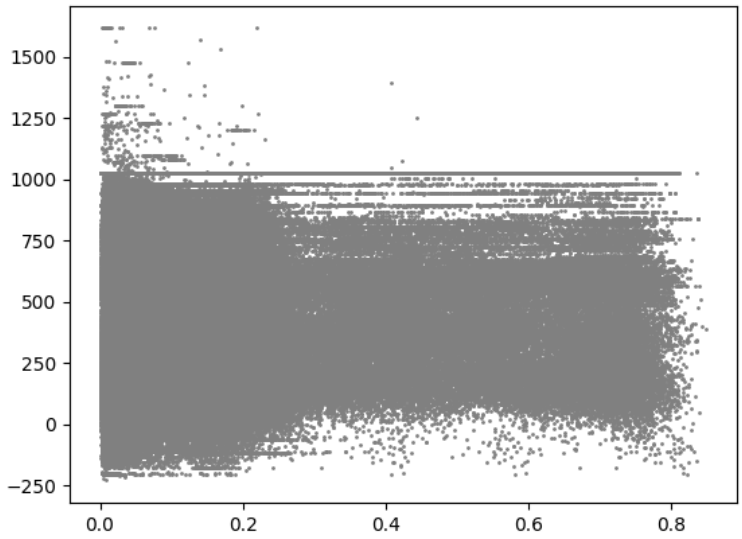
Si certaines pages ont une faible distance (nous parlerons désormais de forte similarité), cela ne signifie pas nécessairement qu'elles contiennent des reprises textuelles, du plagiat ou qu'il s'agit de rééditions. Il apparaît dès lors nécessaire d'ajouter une brique d'alignement de séquences à l'édifice pour mettre en lumière les passages véritablement « similaires » au sens que la lecture des deux pages similaires permet au lecteur d'identifier des passages qui sont repris.

Le processus Pour cela, nous utilisons une librairie d'alignements de séquences d'ADN : *Bio* en Python 3. L'intérêt de cette librairie, et de l'alignement de séquence d'ADN en général, est qu'elle ne cherche pas au sens stricte et local la plus longue chaîne de caractères égale dans deux textes différents. Cette recherche est fait au niveau global, ou macroscopique, pour tenter de dépasser les variations locales ; dans notre cas, on en formule l'hypothèse, celles du bruit. Il y a donc une flexibilité : les séquences résultats de l'alignement ne sont pas forcément égales, il peut y avoir quelques insertions, substitutions ou délétions. Et cela n'est pas sans rappeler les erreurs de transcription dues à l'OCR. Nous faisons donc l'hypothèse que ce type d'alignement sera non sensible au bruit dû à l'OCR.

Nous disposons donc d'un moyen de rapprocher des pages candidates (vectorisation par compte absolu des unigrammes et calcul de la distance cosinus). Ceci nous permettra de ne garder qu'une minorité de pages à présenter au module d'alignement – les pages pour lesquelles on observe une faible distance en même temps qu'un grand nombre de caractères dans les deux pages.

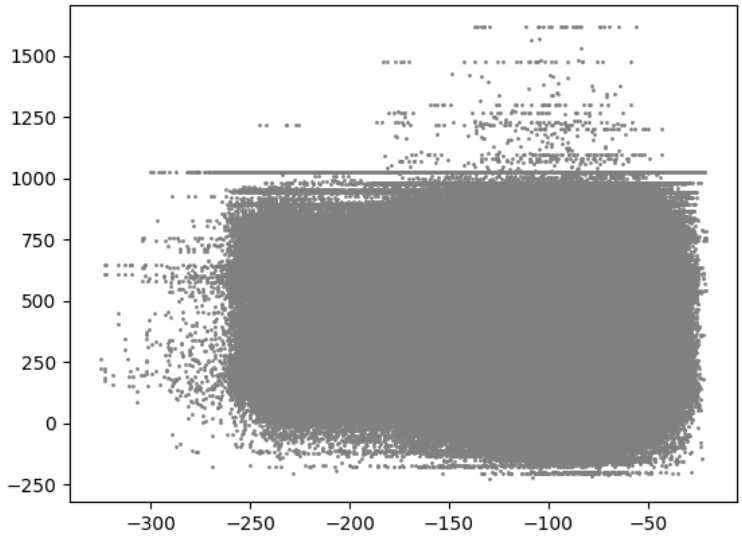


(a) Vectorisation en valeurs absolues

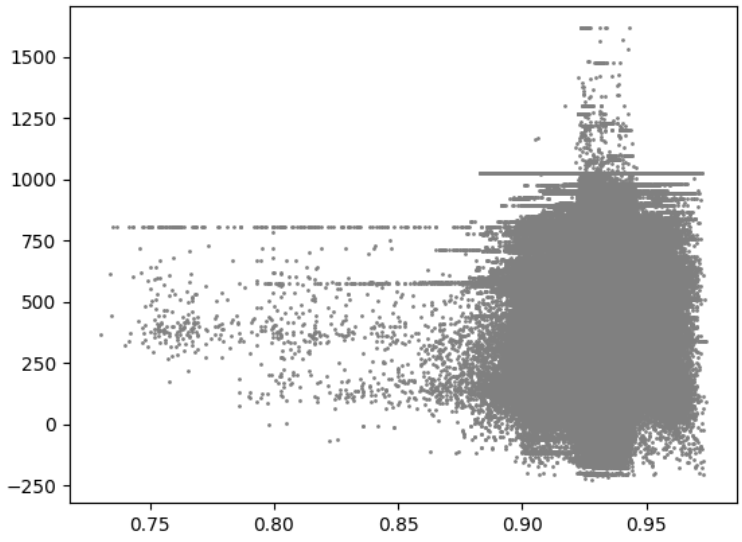


(b) Vectorisation en valeurs relatives

FIGURE 8.20 – Nuages de points des distances cosinus en fonction de la moyenne des CER sur le corpus des mazarinades

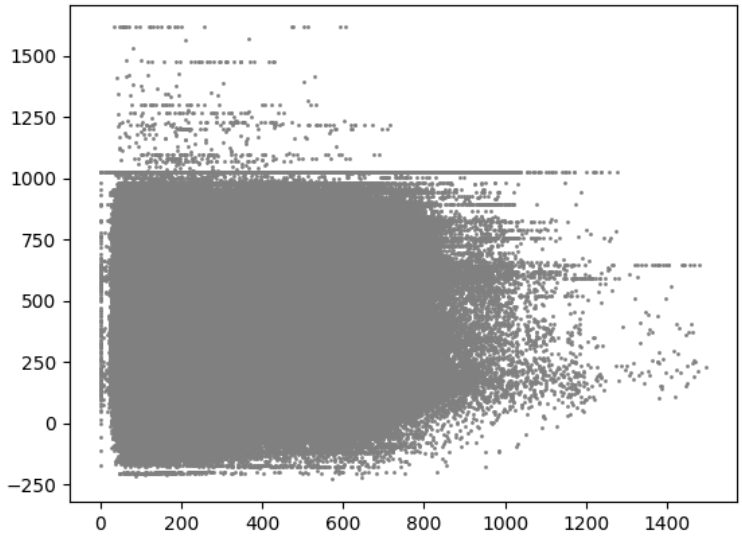


(a) Vectorisation en valeurs absolues

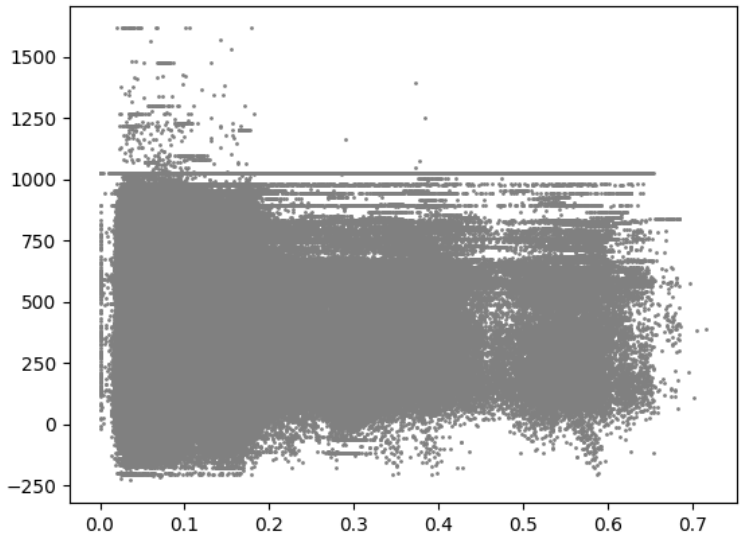


(b) Vectorisation en valeurs relatives

FIGURE 8.21 – Nuages de points des distances Dice en fonction de la moyenne des CER sur le corpus des mazarinades



(a) Vectorisation en valeurs absolues



(b) Vectorisation en valeurs relatives

FIGURE 8.22 – Nuages de points des distances euclidiennes en fonction de la moyenne des CER sur le corpus des mazarinades

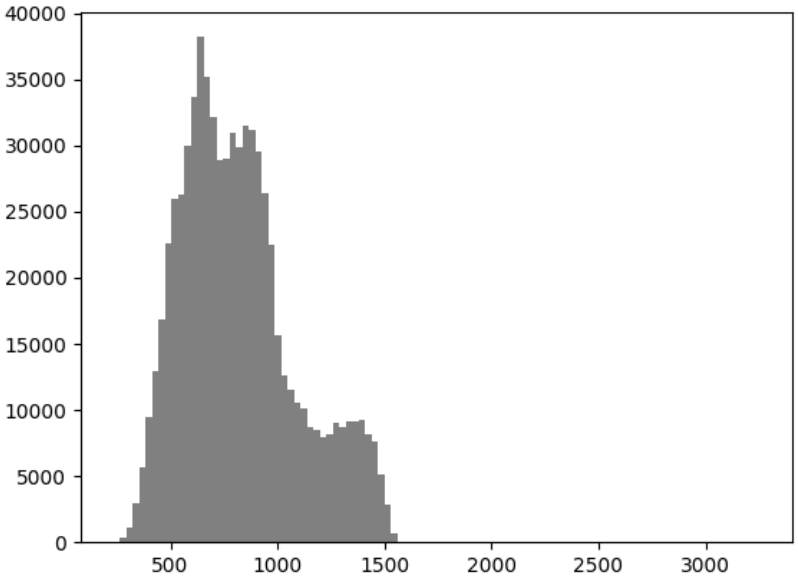


FIGURE 8.23 – Histogramme des scores d'alignement sur les pages similaires

Une fois les scores d'alignement entre pages calculés, il est possible de construire deux types de sorties pour nous aider à mener l'analyse des rapprochements effectués. On peut, pour chaque document du corpus, observer les autres documents qui partagent certaines pages similaires selon un certain score. Nous choisissons d'associer à chaque paire de documents non pas la somme brute des scores d'alignement des pages similaires partagées mais cette somme divisée par le nombre de pages similaires rapprochées. Ceci permet de donner plus de crédit aux rapprochements à fort score d'alignement, plus qu'aux documents à grand nombre de page similaire mais faible score d'alignement. On dispose donc d'une sortie au grain document qui permet d'associer des mazarinades entre elles, encore une fois selon un certain score qui permet d'ordonner ces rapprochements selon leur intensité supposée. En outre, nous gardons aussi la sortie au grain page. Ceci prend forme concrètement dans un tableau HTML où les paires sont ordonnées selon le score de l'alignement et où sont présentés les textes océrisés qui ont servi à l'alignement. Cela permet de se rendre compte directement de la pertinence ou non du rapprochement proposé par la méthode ici décrite.

On observe avec le graphique 8.24 qu'une partie limitée des scores d'alignements ont une valeur supérieure à 1000. Nous choisissons donc de n'observer que les pages ayant un score supérieur à 1000 pour limiter les observations et faciliter l'analyse.

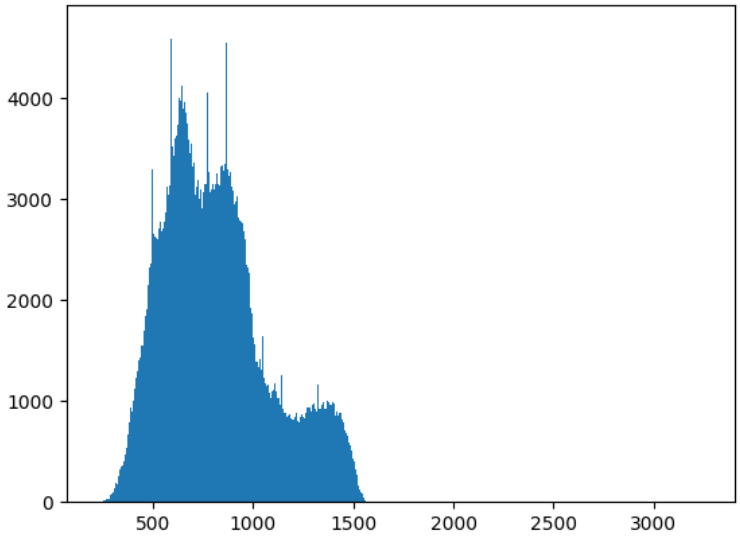


FIGURE 8.24 – Scores des alignements sur le corpus des mazarinades

Généalogie des recueils À partir de la version océrisée et corrigée de la *Bibliographie des mazarinades* de Célestin Moreau, on peut retrouver les entrées comportant le terme *recueil*. Ceci permet de ramener le problème de l'identification automatique des passages à aligner et de l'identification, *in fine*, des recueils éditoriaux, à un problème supervisé. Il s'agirait d'aligner les pages des documents figurant dans la *Bibliographie des mazarinades* avec comme titre *recueil*. Notons que l'on ne trouve pas dans cette ressource d'entrée avec « réédition ». Nous ne chercherons donc que les recueils éditoriaux.

Voici la liste des mazarinades titrées par Moreau avec *recueil* : 3, 19, 27, 54, 56, 68, 112, 126, 546, 567, 675, 748, 780, 783, 788, 830, 831, 1031, 1168, 1241, 1380, 1385, 1452, 1471, 1550, 1636, 1638, 1639, 1646, 1648, 1661, 1740, 1896, 2088, 2093, 2099, 2278, 2355, 2390, 2436, 2447, 2451, 2511, 2538, 2574, 2638, 2728, 2744, 2813, 2847, 2851, 2866, 2913*, 2941, 3034, 3035, 3036, 3037, 3038, 3039, 3040, 3041, 3042, 3043*, 3044, 3045, 3046, 3047, 3048, 3049, 3050, 3051, 3052, 3053, 3054, 3055, 3505, 3634, 3679, 3724, 3789, 3855, 3886, 4007, s3-70, s1-189, lab-78, lab-190, lab-332, soc_41, soc_98.

Nous avons donc, pour toutes les mazarinades, calculé un taux d'alignement (donné par la librairie BioPython) de sorte à obtenir une liste comme[6] :

— Moreau1837-3 Moreau1837-2 6 10029.0 1671.5

— Moreau2045 Moreau2045-2pp.81-84 2 3084.0 1542.0

6. La ressource est structuré comme : premier identifiant, second identifiant, nombre de pages alignées, somme des taux d'alignement

— Moreau1837-3 Moreau3suppl74 1 1447.0 1447.0
— Moreau1837-3 Moreau742 7 10054.0 1436.2857142857142
— Moreau1837-3 Moreau117 1 1427.0 1427.0
— Moreau1837-3 Moreau1837-3 5 7124.0 1424.8
— Moreau1837-3 Moreau1865pp.352-357 86 121809.0 1416.3837209302326
— Moreau528 Moreau1837-2 6 8497.0 1416.1666666666667
— Moreau742 Moreau3suppl74 3 4246.0 1415.3333333333333
— Moreau742 Moreau1837-2 18 25475.0 1415.2777777777778
— Moreau1837-3 Moreau1865-3pp.352-357 93 131619.0 1415.258064516129
— Moreau528 Moreau742 6 8482.0 1413.6666666666667
— Moreau1837-3 Moreau1895-2 4 5643.0 1410.75
— Moreau528 Moreau1837-3 6 8448.0 1408.0
— Moreau742 Moreau1837-3 18 25329.0 1407.1666666666667
— Moreau1837-3 Moreau1865-2pp.352-357 89 125208.0 1406.8314606741574
— Moreau1837-2 Moreau1837-2 5 7032.0 1406.4
— Moreau528 Moreau3suppl74 1 1406.0 1406.0
— Moreau1837-3 Moreau528 4 5618.0 1404.5
— Moreau1865-3pp.352-357 Moreau3suppl74 36 50500.0 1402.7777777777778
— Moreau1837-3 Moreau1865-4pp.352-357 74 103665.0 1400.8783783783783
— Moreau528 Moreau117 1 1400.0 1400.0
— ...

À partir de cela, nous avons filtré les entrées de cette ressource pour aboutir à une version tronquée de cette ressource mais avec uniquement les identifiants des mazarinades qui ont recueil dans leur titre. Ensuite, nous avons récupéré les pages alignées correspondant aux identifiants Moreau en question. Il aura fallu identifier un seuil de taux d'alignement pour trouver un compromis entre le bruit que cela génère (pages qui n'ont rien en commun) et la dureté du filtre. J'ai regardé les alignements proposés par mon code avec des taux arbitraires en observant s'il y avait beaucoup de bruit (pages alignées qui n'ont pas de véritable texte en commun) ou si les pages étaient vraiment alignées. Finalement, on trouve le seuil 1447 pour que les pages soient vraiment alignées.

Néanmoins, ce seuil ne permet pas de retrouver les recueils éditoriaux de façon satisfaisante car uniquement 32 pages sont retrouvées.

8.1.3 Conclusion

Cette recherche s'est conclue rapidement au profit de la recherche exposée dans la section suivante. Effectivement, nous pensons que l'échelle de la page et la vectorisation en unigramme ne permettent pas de retrouver des pages alignées. D'abord, parce que le calcul des alignement est trop couteux à l'échelle de la page (voir introduction de ce chapitre) mais aussi car la vectorisation en unigrammes ne permet pas, contrairement à l'identification du bruit, de conserver la séquencialité des énoncés.

8.2 Vers une étude automatique de la proximité des documents selon les épisodes repérés par Moreau

Une autre piste d'étude est donc de changer d'échelle et de calculer des similarités selon le document (et non plus selon la page). Pour mener une expérience nouvelle que le pur alignement – chose qui n'est peut-être pas à refaire (voir TextPair) – nous choisissons de nous intéresser aux épisodes repérés par Célestin Moreau. Il s'agit de certaines mazarinades qui ont été repérées comme appartenant à des moments forts de la Fronde. Nous disposons d'une ressource au format JSON structurant, pour toutes les années de la Fronde, les épisodes de Moreau en leur associant une date et la liste des mazarinades concernées. Elle est de la forme :


```
"1648":
[
{"Etiquette Moreau": "Avril 1648", "date":"00/04/1648", "liste":
["2305", "935", "940", "2989", "1742", "927", "937", "2294",
"938", "3600", "953", "3307", "2473", "577", "1498", "3251",
"2822", "3736", "3685", "2616", "2270", "1827", "2313", "3345",
"936", "959", "2812", "1297", "536", "1669", "3494", "1585",
"647 bis", "2461", "1542", "1610", "1614", "1564", "3219",
"1644", "1106", "61", "3051", "3050", "2576", "1510", "56",
"3595", "2444", "2636", "3466"]}
]
```


On dispose des épisodes suivants :

— Avril 1648
— Le roi sort de Paris
— Arrêt du parlement contre le cardinal Mazarin
— Le prince de Conty est reconnu généralissime de l'armée du parlement
— Arrivée du duc de Beaufort à Paris
— Lettre du parlement de Paris aux autres parlements du royaumeAvril 1648
— Déclaration du roi qui supprime les charges et offices du parlement
— Arrestation de La Raillère
— Déclaration du roi par laquelle six jours sont donnés aux habitants de Paris pour rentrer dans le devoir
— Prise de Charenton
— Mort de Charles Ier roi d'Angleterre
— Arrestation du chevalier de La Valette
— Réception de l'envoyé de l'archiduc Léopold au parlement
— Départ des députés pour la conférence de Ruel
— Signature du traité de Ruel
— Le parlement accepte le traité de Ruel

— Entrée de l'armée de l'archiduc Léopold en France
— Paix de Saint-Germain
— Arrivée du prince de Condé à Paris
— Voyage de Flandre
— Départ du prince de Condé pour la Bourgogne
— Affaire du jardin de Renard
— Siège de Cambrai
— Arrestation de Morlot
— Entrée du roi dans Paris
— Anniversaire de la naissance du roi
— Coup de pistolet de Guy Joly
— Avant la paix
— Après la paix
— Suite du procès de Beaufort, Gondy et Broussel
— Arrestation des princes
— Départ du roi pour la Normandie
— Départ du roi pour la Bourgogne
— Déclaration du roi contre la duchesse de Longueville
— Départ du roi pour la Guyenne
— Ouvertures de paix par l'archiduc Léopold
— Paix de Bordeaux
— Translation des princes au Havre
— Bataille de Réthel
— Mazarin était revenu la veille de l'armée de Champagne
— Remontrances du parlement pour la liberté des princes
— Mazarin sort de Paris
— Mise en liberté des princes
— Arrivée des princes à Paris
— Serment d'union de l'assemblée de la noblesse
— Retraite du Coadjuteur
— Déclaration du roi pour exclure du conseil les étrangers et les cardinaux
— Le prince de Condé se retire à Saint-Maur
— Le Tellier, Servien et de Lionne sont éloignés de la cour
— Lecture des articles dressés par la Cour contre le prince de Condé
— Déclaration de la majorité du roi
— Le prince de Condé se retire en Guyenne
— Déclaration du roi contre les princes de Condé, de Conty, etc
— Le roi mande au cardinal Mazarin de rentrer en France
— Suite de l'arrêt du 29 décembre 1651
— Le conseiller Bitaut est fait prisonnier par le maréchal d'Hocquincourt
— Traité du duc d'Orléans et du prince de Condé
— Arrivée du cardinal Mazarin à Poitiers
— Affaire d'Angers
— Arrêt du parlement pour le payement des rentes de l'Hôtel de Ville
— Entrée du duc de Lorraine et des Espagnols en France
— Combat de Mont-de-Marsan

— Siège de Miradoux
— On apprend à Paris que le coadjuteur a été nommé cardinal
— Le duc de Nemours passe la Seine à Mantes et vient à Paris
— Prise de Saintes par les troupes royales ; Affaires de la Guyenne
— Mademoiselle part de Paris pour se rendre à Orléans
— Le prince de Condé arrive à l'armée des ducs de Beaufort et de Nemours
— Combat de Bleneau
— La cour part de Gien pour se rapprocher de Paris
— Députation du duc de Rohan, de Chavigny et de Goulas à Saint-Germain
— Retour de Mademoiselle à Paris- Combat devant Étampes
— Prise de Saint-Denys par le prince de Condé
— Siège d'Étampes
— Le duc de Lorraine vient à Paris
— Descente de la châsse de Sainte-Geneviève
— Retraite du duc de Lorraine
— Émeute aux abords du parlement- Guerre des Menardeaux
— Combat du faubourg Saint-Antoine
— Levée du siège de Villeneuve d'Agénois
— Le comte d'Harcourt abandonne son armée
— Incendie de l'Hôtel de Ville
— Le roi déclare aux députés du parlement qu'il consent à éloigner le cardinal Mazarin
— Mort de Mancini
— Lieutenance générale du duc d'Orléans
— Duel des ducs de Beaufort et de Nemours
— Translation du parlement à Pontoise
— Le cardinal Mazarin s'éloigne de la cour
— Arrivée du roi à Compiègne
— Assassinat de l'épinglier
— Premier édit d'amnistie
— Députation du clergé
— Assemblée du Palais-Royal
— Deuxième déclaration d'amnistie
— Députation de six corps de marchands
— Les colonels de la milice s'assemblent pour convenir d'une députation au roi
— Le prince de Condé sort de Paris
— Novembre 1652

8.2.1 Similarité entre documents et cohérence interne des épisodes

Une première voie d'analyse peut être de calculer la similarité entre tous les documents et d'observer, sachant les épisodes, quelle est la moyenne des distances entre les documents desdits épisodes. Pour cela, nous calculons toutes les similarités cosinus entre tous les documents du corpus et réalisons la moyenne

de ces distances pour tous les épisodes. On dispose finalement de la liste suivante :

— Avril 1648 : 0,803590095124685
— Le roi sort de Paris : 0,779413056170111
— Arrêt du parlement contre le cardinal Mazarin : 0,830808845177887
— Le prince de Conty est reconnu généralissime... : 0,839037836076568
— Arrivée du duc de Beaufort à Paris : 0,799500876611549
— Lettre du parlement de Paris aux autres parlements... : 0,821052417981051
— Déclaration du roi qui supprime les charges et... : 0,847211951492704
— Arrestation de La Raillère : 0,818100720712523
— Déclaration du roi par laquelle six jours sont... : 0,83082356584715
— Prise de Charenton : 0,793144442969777
— Mort de Charles Ier roi d'Angleterre : 0,835060591339666
— Arrestation du chevalier de La Valette : 0,812194512954097
— Réception de l'envoyé de l'archiduc Léopold au parlement : 0,822495919745729
— Départ des députés pour la conférence de Ruel : 0,820202266878355
— Signature du traité de Ruel : 0,821316307395687
— Le parlement accepte le traité de Ruel : 0,823953408697001
— Entrée de l'armée de l'archiduc Léopold en France : 0,837591937345247
— Paix de Saint-Germain : 0,808980003544293
— Arrivée du prince de Condé à Paris : 0,793086810016212
— Voyage de Flandre : 0,764085520145328
— Départ du prince de Condé pour la Bourgogne : 0,799515918207598
— Affaire du jardin de Renard : 0,795840005624404
— Siège de Cambrai : 0,813868045819803
— Arrestation de Morlot : 0,833354732917031
— Entrée du roi dans Paris : 0,842554925448273
— Anniversaire de la naissance du roi : 0,801819077813997
— Coup de pistolet de Guy Joly : 0,807006626589155
— Avant la paix : 0,802456711443616
— Après la paix : 0,78942483821809
— Suite du procès de Beaufort, Gondy et Broussel : 0,807236676134984
— Arrestation des princes : 0,820338578987973
— Départ du roi pour la Normandie : 0,831878491939846
— Départ du roi pour la Bourgogne : 0,795566315907613
— Déclaration du roi contre la duchesse de Longueville : 0,824643783531607
— Départ du roi pour la Guyenne : 0,842553040949353
— Ouvertures de paix par l'archiduc Léopold : 0,760157242511271
— Paix de Bordeaux : 0,811173233918711
— Translation des princes au Havre : 0,805492667389227
— Bataille de Réthel : 0,815700089003477
— Mazarin était revenu la veille de l'armée de Champagne : 0,839337977597051
— Remontrances du parlement pour la liberté des princes : 0,836176868614452
— Mazarin sort de Paris : 0,789798930711989
— Mise en liberté des princes : 0,826694843246414
— Arrivée des princes à Paris : 0,836676174120806
— Serment d'union de l'assemblée de la noblesse : 0,822283839055326

— Retraite du Coadjuteur : 0,813638259726037
— Déclaration du roi pour exclure du conseil les étrangers... : 0,834608133824435
— Le prince de Condé se retire à Saint-Maur : 0,814625026811891
— Le Tellier, Servien et de Lionne sont éloignés de la cour : 0,815282617815681
— Lecture des articles dressés par la Cour contre le prince de Condé : 0,78374289828003
— Déclaration de la majorité du roi : 0,825388919693191
— Le prince de Condé se retire en Guyenne : 0,750140813362382
— Déclaration du roi contre les princes de Condé, de Conty, etc. : 0,801053379806668
— Le roi mande au cardinal Mazarin de rentrer en France : 0,819552727006333
— Suite de l'arrêt du 29 décembre 1651 : 0,833975125956277
— Le conseiller Bitaut est fait prisonnier par le maréchal... : 0,816926490482934
— Traité du duc d'Orléans et du prince de Condé : 0,832669304186508
— Arrivée du cardinal Mazarin à Poitiers : 0,818664102126283
— Affaire d'Angers : 0,789725251059252
— Arrêt du parlement pour le payement des rentes de l'Hôtel de Ville : 0,786495212601675
— Entrée du duc de Lorraine et des Espagnols en France : 0,797709665901551
— Combat de Mont-de-Marsan : 0,769008056171471
— Siège de Miradoux : 0,805553926277403
— On apprend à Paris que le coadjuteur a été nommé cardinal : 0,788718998806942
— Le duc de Nemours passe la Seine à Mantes et vient à Paris : 0,824122925426498
— Prise de Saintes par les troupes royales ; Affaires de la Guyenne : 0,823232483884565
— Mademoiselle part de Paris pour se rendre à Orléans : 0,801388204981389
— Le prince de Condé arrive à l'armée des ducs de Beaufort et... : 0,778255792973253
— Combat de Bleneau : 0,82390220039628
— La cour part de Gien pour se rapprocher de Paris : 0,824261473607038
— Députation du duc de Rohan, de Chavigny et de Goulas à... : 0,817053874428141
— Retour de Mademoiselle à Paris- Combat devant Étampes : 0,818512602124936
— Prise de Saint-Denys par le prince de Condé : 0,800453813744007
— Siège d'Étampes : 0,782220253440329
— Le duc de Lorraine vient à Paris : 0,812179170813733
— Descente de la châsse de Sainte-Geneviève : 0,794547260002636
— Retraite du duc de Lorraine : 0,829873271755542
— Émeute aux abords du parlement- Guerre des Menardeaux : 0,809677766185116
— Combat du faubourg Saint-Antoine : 0,76938974431086
— Levée du siège de Villeneuve d'Agénois : 0,741647392125527
— Le comte d'Harcourt abandonne son armée : 0,74385332558815
— Incendie de l'Hôtel de Ville : 0,792430363864501
— Le roi déclare aux députés du parlement qu'il consent à... : 0,774236907627185
— Mort de Mancini : 0,810104431360402
— Lieutenance générale du duc d'Orléans : 0,811104554201402
— Duel des ducs de Beaufort et de Nemours : 0,817392447278068
— Translation du parlement à Pontoise : 0,824779513949587
— Le cardinal Mazarin s'éloigne de la cour : 0,798896352807492
— Arrivée du roi à Compiègne : 0,825520183441626

— Assassinat de l'épinglier : 0,82023949252802
— Premier édit d'amnistie : 0,817704707548381
— Députation du clergé : 0,817978960939869
— Assemblée du Palais-Royal : 0,8297087793834
— Deuxième déclaration d'amnistie : 0,816831543525972
— Députation de six corps de marchands : 0,822826912758896
— Les colonels de la milice s'assemblent pour convenir d'une... : 0,81413462097008
— Le prince de Condé sort de Paris : 0,708235533348832

On observe que les moyennes des similarités entre documents varient entre $0,7$ et $0,8$ ce qui suggère une bonne cohérence interne des épisodes et que, en définitive, les documents des épisodes se ressemblent globalement. Effectivement, ils couvrent les mêmes moments historiques ; il n'est donc pas étonnant, quoique intéressant à démontrer et pertinent dans la démonstration de l'efficacité de la méthode, que les similarités soient importantes.

8.2.2 Représentation en ACP des épisodes

Les documents pouvant être vectorisés au grain caractères, il devient possible de vectoriser les épisodes eux-mêmes en procédant à la somme des vecteurs des documents composant les épisodes. Ceci permettra de réaliser une réduction de dimension par ACP (Analyse en Composantes Principales) pour représenter chaque épisode sur un même plan. Ceci permettra d'observer les épisodes les plus proches les uns des autres et les plus éloignés les uns des autres. Une autre application pourrait être de dater approximativement un nouveau document arrivant dans le corpus car, l'expérience l'a montré, le corpus des mazarinades n'est pas fermé et l'on retrouve encore certains documents qui en relèvent. Il suffira de vectoriser ce nouveau document et de calculer les similarités entre tous les épisodes et de trouver la plus faible pour l'associer à un épisode déjà existant.

Pour analyser ce graphe, il faut commencer par dire que l'individu qui se placerait aux coordonnées $(0,0)$ serait considéré comme un individu virtuel moyen. C'est-à-dire qu'il n'existe pas mais qu'il est, en la sémantique des deux axes, ce qu'il y a de plus « moyen ». Tout ce qui s'en échappe est donc atypique au sens des deux axes. On identifie donc au moins deux groupes : celui des épisodes relativement moyens et proches du point de coordonnées $(0,0)$ et celui des épisodes qui s'en éloignent.

8.3 Conclusion

Ces expériences montrent toute la difficulté de mener une expérience en contexte non supervisé. Plus encore, l'alignement textuel s'est soldé par un échec ; l'impossible calcul de toutes les similarités pour toutes les paires de documents aura mené à un arrêt de cette expérience. Par ailleurs, on se rend compte que le travail avec les épisodes Moreau, s'il devait être développé, est beaucoup plus fécond. Et pour cause, l'introduction de cette ressource consiste,

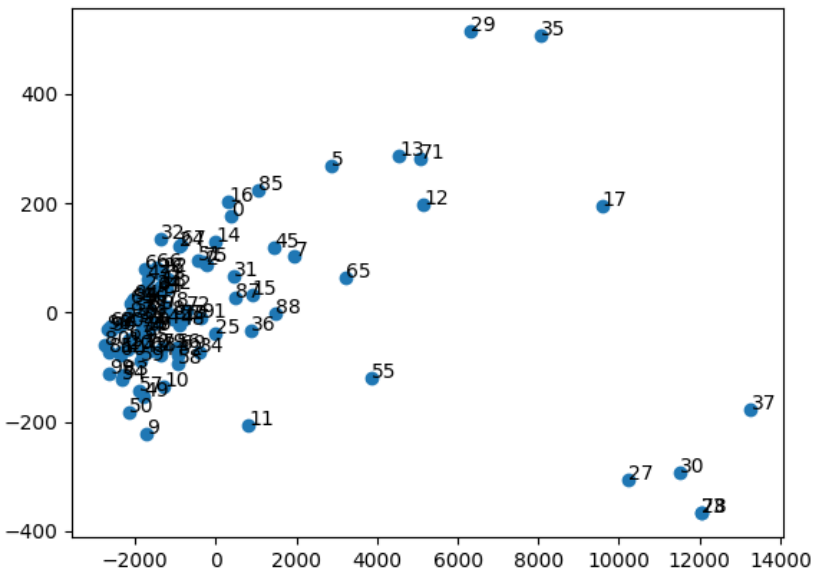


FIGURE 8.25 – Représentation en ACP des vecteurs correspondants aux épisodes Moreau

en fait, à ramener l'étude des similarités dans un context supervisé, car il existe bien une référence à laquelle comparer les résultats.

Chapitre 9

Conclusion

Sommaire

9.1 Contributions et perspectives de cette thèse

Les contributions de cette thèses sont plurielles, et nous tentons ici d'en proposer une synthèse. L'état de l'art aura permis de mettre en relief une typologie des origines et des erreurs d'OCR tout en proposant une vue globale de l'impact de ces erreurs. Les expériences menées en textométrie ont permis de montrer que les données textuelles bruitées sont pertinentes pour l'étude des phénomènes fréquents d'un corpus tout en montrant l'impossible étude des phénomènes rares car, comme les erreurs souvent, ils sont uniques. Il y a donc une confusion computationnelle entre les phénomènes rares et les erreurs. Par ailleurs, cette expérience aura proposé en sous texte une méthodologie pour appréhender ce genre de corpus imparfait : l'échelle de la page permet effectivement un retour facilité au fac-similé numérique ce qui est sans doute le plus important de la démarche. Ceci permet une vérification de l'intuition suggérée par les données océrisées. En outre, l'expérience des similarités textuelles, si elle est globalement un échec, permet de conclure sur l'impossibilité d'étudier des données textuelles bruitées en contexte non supervisé. D'ailleurs, la constitution du modèle de prédiction du taux d'erreurs s'est bien faite sur un autre corpus que celui des mazarinades, corpus qui était transcrit diplomatiquement. Cela a permis de proposer une évaluation au sens strict et telle qu'on l'imagine scientifiquement parlant ; évaluation *supervisée* pour pouvoir calculer des métriques. Néanmoins il convient de nuancer le propos : la conclusion de l'impossible évaluation en contexte non supervisée est peut-être trop hâtive, ou au moins incomplète.

Effectivement, le manque de temps ne nous aura pas permis de soulever les phénomènes concrets mis en jeu par le basculement non supervisé des évaluations. Il s'agit donc de proposer cette piste plutôt en perspectives qu'en conclusion fermée.

Par ailleurs, cette thèse a permis de proposer une chaîne de traitements facile d'utilisation permettant de passer d'un document PDF à un fichier TXT pour les documents du XVIIème siècle. Si cette chaîne de traitements n'est pas une révolution scientifique, elle a le mérite d'exister, et ce pour un public en demande d'outil « clef en main » : les chercheurs en littérature. Effectivement, le recherche en informatique tend à se spécialiser, et le chercheur en informatique, lui, maîtrise de fait toutes les techniques premières. Ce n'est donc pas un enjeu pour lui de réaliser une telle océrisation. En revanche, lorsqu'on est dix-septièmiste, c'est un enjeu de taille, voire le plus souvent un verrou. En outre, cette thèse a permis de proposer un modèle de prédiction du taux d'erreurs sans vérité de terrain. Ceci est un réel enjeu dans les campagnes d'océrisation massive puisque cela permet l'économie de la fastidieuse constitution d'une vérité de terrain. Il devient plus facile, plus rapide et plus économique de comparer des modèles d'OCR pour sélectionner le plus performant. Chose qui, si elle apparaît théoriquement comme une évidence nécessaire, n'est pas toujours réalisée en raison des difficultés précitées. Et c'est le cas des océrisations de Gallica par exemple, qui sont inconsistantes dans le temps (logiciels et métriques d'évaluation variables). Ce travail a aussi permis de montrer la pertinence du T_{con} et du T_{lex} [1] dans l'étude de la quantité d'erreurs d'OCR dans un corpus, le premier étant presque gratuit puisque directement proposé par le logiciel d'OCR (quoiqu'il faille être en mesure de créer un programme capable de parcourir un fichier XML), le second nécessitant une ressource lexicale externe. On peut donc considérer le premier taux comme minimaliste par comparaison au second qu'on peut qualifier de maximaliste. À cet égard, l'entropie conditionnelle de Shannon est aussi un bon indice pour l'estimation du taux d'erreurs et les unigrammes se sont révélés plus informatifs en ce sens. Finalement, un programme d'alignement de séquences a été créé et permet, facilement en donnant en entrée deux textes océrisés, de sortir les passages alignés en même temps qu'en leur donnant un score (d'alignement, donc).

De manière évidente, plusieurs pistes émergent de cette thèse pour continuer les travaux. Indépendemment des conclusions informatiques, il apparaît, en plus d'être nécessaire, urgent de proposer aux dix-septièmistes un outils « clef en main » et *user friendly*, n'exigeant aucune compétence informatique pour être utilisé. Néanmoins, on ne nie pas les compétences informatiques fortes de chercheurs tels que G. Roe, I. Galleron, S. Gabay, J.-B. Camps, etc. – pour se limiter à la France. Nous pouvons dans ce cas parler de l'hétérogénéité des compétences des chercheurs de ce domaine, ce qui ne bat pas en brèche l'urgence d'un « équipement » informatique. Les différentes recherches menées indépendamment les unes des autres pourraient être mises en commun dans un projet unique

1. Le premier est l'agrégation des taux de confiance au caractère fournis par les logiciels d'OCR ; le second correspond à la proportion de mots appartenant à une ressource lexicale connue.

permettant la constitution de ce genre d'outil, lequel serait multimodal. On pourrait suggérer par exemple : l'océrisation rapide d'un document PDF ou d'une image, le prétraitement des numérisations, l'alignement de deux textes et la possibilité de sortir des analyses textométriques sur des phénomènes fréquents si un corpus est renseigné. En plus du côté « pratique » d'un tel outil, il apparaît nécessaire dans un contexte où beaucoup de recherches littéraires se tournent vers le numérique. Effectivement, l'ingénierie du TAL est très couteuse en temps et nécessite de procéder à des compromis entre littéraires et informaticiens. Et cela se fait pour chaque projet de recherche de la sorte. Il y a donc une « non unification » des pratiques en même temps que les efforts ne sont pas concentrés ; l'économie de moyen semble donc, elle aussi, urgente. [2] Sur un autre plan, il apparaît nécessaire de continuer le travail sur les similarités textuelles. Celui-ci s'est avéré particulièrement difficile à mener en pratique mais apparaît aussi fécond dans le questionnement sur l'évaluation et l'élection entre un contexte supervisé ou non supervisé. En effet, il est possible, avec les mazarinades, de proposer une évaluation supervisée (en utilisant la liste des épisodes de Célestin Moreau) en même temps qu'une évaluation non supervisée peut être menée. Un outil de décision sur la qualité de la similarité proposée pourrait être développé, par exemple sous le principe des applications contemporaines de rencontres matrimoniales (comme *Tinder*). Il est effectivement possible, *a minima*, de lire pour procéder à cette évaluation non supervisée. La comparaison de ces deux types d'évaluation permettrait de voir si dans la pratique on arrive à dépasser la nécessaire supervision. Toutefois, une limite de cette perspective est qu'en procédant à la validation des rapprochements de pages similaires, il s'agirait en soi d'une annotation, certe en aval, mais tout de même d'une annotation. Cela reviendrait, peut-être, à ramener une fois de plus le problème entre le carcan de la supervision...

9.2 Un accès restreint aux données issues d'OCR

L'océrisation est vecteur d'erreurs, de bruit, on l'a dit. Même si nous l'avons rappelé à plusieurs reprises, le terme *bruit* cache le phénomène du silence pourtant omniprésent dans les erreurs d'OCR. Sans vérité de terrain donc, on perd définitivement cette information. L'argument selon lequel l'OCR est une bonne solution pour éviter la transcription manuelle et diplomatique de tout un corpus perd alors en crédibilité. Dans tous les cas, il apparaît nécessaire de transcrire une partie du corpus pour évaluer les modèles d'OCR (mais aussi pour permettre l'évaluation de certaines tâches de TAL). Avec la limite que les conclusions tirées sur les performances des modèles ne seront vraies que pour les données transcrites. Or, on l'a dit, nuance est de rigueur en contexte OCR car la quantité d'erreurs est très erratique et d'un corpus à l'autre, mais aussi d'un sous-corpus à un autre ; cette quantité varie extrêmement. On serait donc tenté de transcrire plus encore...

2. Nous pouvons tout de même noter ici l'existence des *consortiums* d'*Humanum* : *Cahier, Corli, Biblissima et Coste* ainsi que le travail d'I. Galleron qui s'attaque à cette lourde tâche d'unification des pratiques et des corpus.

Une question demeure donc : est-il préférable d'investir dans des transcripteurs-annotateurs pour disposer de données textuelles propres, ou bien l'ingénierie des évaluations non supervisées et des transcriptions minimales est-elle plus rentable ? Car c'est bien de cela dont il est question : on érige l'informatique comme un moyen efficient. Mettre en doute ce postulat (ou dictat ?) semble, arrivé à ce point de la réflexion, urgent.

Lorsque les données textuelles proviennent d'OCR, on est en rien certain d'avoir accès à l'information qu'on cherche (c'est-à-dire à l'information contenu dans les ouvrages). Plusieurs entraves s'interposent : la constitution du corpus numérique, l'enrichissement des données, le bruit de l'OCR, le silence (dû principalement aux mauvaises segmentations), mais aussi le problème de ce qu'il est possible ou non de repérer avec de telles données sans référence. On ne lit pas un corpus numérique, tout au plus on le parcourt. L'argument selon lequel on ne peut tout transcrire pour faire des recherches se file aussi pour l'usage des données océrisées : on ne peut pas tout lire, on veut donc utiliser l'informatique pour palier à ce problème, pour tenter de tout embrasser en même temps. L'informatique est donc appelée en renfort pour récupérer les données mais aussi pour les lire. Il est tentant de croire qu'elle le peut ; mais l'absence de référence, le bruit dans les données et la quantité importante des données (d'où un coût computationnel : temps long des calculs ce qui ralentit les recherches) sont autant de freins à l'efficience conférée à cette discipline. En pratique donc, la masse acquise « rapidement » (et cela aussi est relatif fonction des logiciels utilisés) ne permet pas, en tout cas directement, un accès aux données contenues dans les documents. Il est toujours nécessaire d'ajouter encore deux étapes au processus : une étape de filtre automatique (on va chercher à ne présenter que certains résultats qui semblent plus probablement pertinents au prisme de la question posée) et une étape d'analyse manuelle (pour comprendre ce que le logiciel donne en sortie). Dans l'état actuel des choses, les technologies ne sont pas assez avancées (le seront-elles un jour ?) pour être en lien direct avec les données et permettrent de faire i) gagner du temps au chercheur expert et ii) de trouver de nouvelles connaissances. Un exemple concret est celui de l'étude menée sur les mazarinades burlesques : il aura toujours été nécessaire de procéder à un retour interprétatif littéraire (appuyé par des lectures). Le logiciel de textométrie ne propose aucune conclusion, encore moins de conclusion nouvelle non repérée par les chercheurs experts. En substance, il convient de poser la question de la pertinence de l'OCR dans l'objectif du gain de temps ; le coût pour manipuler les données et pour, disons, les décortiquer, contrebalance largement le temps gagné. Toutefois, ce constat est à relativiser selon les questions de recherche. L'existence de métadonnées minimales (comme la date, le genre, etc.) permettrait de rendre suffisantes des données océrisées, par exemple pour comparer les usages du s long ou l'absence de pronom sujet devant les verbes impersonnels.

9.3 De l'évaluation supervisée à l'évaluation non supervisée

Le problème fondamental de ce travail n'est donc pas les erreurs d'OCR mais bien l'absence d'une vérité de terrain totale qui aurait permi de comparer différentes méthodes pour une tâche donnée (comme, avec la similarité textuelle, plusieurs vectorisations). Les méthodes en TAL sont difficilement reproductibles en milieux non supervisés. Preuve en est, pour faire le travail de l'estimation automatique du CER, l'attestation du résultat s'est fondée sur les transcriptions du corpus utilisé. Il y a en TAL une nécessité (au moins intellectuelle) à avoir des données de référence pour pouvoir juger de la qualité des modèles et méthodes proprosées. En milieu non supervisé, on peut par exemple choisir d'observer sur un échantillon des données le comportement du modèle ou de la méthode ; mais c'est précisemment se ramener en milieu supervisé. Dans l'état, nous remarquons soit une impossibilité théorique de franchir cette frontière pour la recherche en informatique ; soit un creux dans la recherche à ce niveau que nous avons tenté de mettre en lumière.

Si l'on a créé un modèle capable d'estimer la qualité des données océrisées en milieu non supervisé, il reste encore une question. Pourquoi cette nécessité d'estimer la qualité des données ainsi acquises ? Une réponse est que l'on veut se ramener en milieu pseudo-supervisé, en faisant comme si on avait des données transcrites grâce auxquelles on calculerait ce taux d'erreur. Une autre réponse est que l'on cherche à justifier par ce taux d'erreur le non fonctionnement des outils informatiques en aval. Si le taux d'erreur est important (ce qui souvent est le cas pour de l'OCR de documents anciens) on a dès lors une justification si les algorithmes ne parviennent pas aux fins qu'ils se sont fixées.

9.4 Le renversement du traitement « tout numérique »

Beaucoup de projets de recherche voient aujourd'hui le jour et proposent d'apporter « le nouveau », c'est-à-dire le numérique, dans la recherche. Si pendant des années la recherche s'est contentée de ne traiter les corpus qu'à travers le prisme de la lecture oculaire (et donc lente donc limitée), il y a la prétention, avec l'informatique, de tout embrasser, d'un seul coup, et faire émerger des connaissances nouvelles et probablement plus générales. On est dans une mode « tout numérique ».

Cette thèse, qui part bien de cette illusion, cherche finalement à poser la question de la pertinence de cela. D'abord, c'est le monde numérique qui fait qu'on a une masse de données conséquente à traiter aujourd'hui ; c'est un problème endogène. Ensuite, l'informatique a ses propres contraintes, ses propres carcans, qui sont différents de la recherche, disons, en littérature. Il faut donc, dans une collaboration littéraires-informaticiens, trouver un moyen de communication. Et cette « traduction » des exigences se fait par des compromis :

ce que l'informaticien ne peut faire avec ses outils est transformé en un autre problème (donc une autre solution) et le chercheur en littérature doit s'en satisfaire. Il y a, semble-t-il, une étape nécessaire de « traduction » entre les exigences de chaque discipline et de l'informatique.

Cette difficulté doit être, cela dit, pensée en terme graduel ou de *continuum* ; cela dépend des questions de recherche qui sont posées au type de données dont il est question dans cette thèse (océrisées et bruitées, donc). Pour des questions purement quantitatives, un corpus renseigné un minimum au plan des métadonnées peut « parler» de lui-même assez rapidement en dépit du bruit et du silence. Donc pour des recherches linguistiques, sociolinguistiques, graphiques, orthographiques ou sur la typographie, c'est bien utile. Pour des questions de recherche plus « sophistiquées », induisant plus d'arrière-plan culturel, l'expert apparaît nécessaire. Mais celui-ci doit aussi être capable de convertir sa manière de penser pour imaginer des questions de recherche adaptées ou adaptables à l'informatique ; ce qui est, en effet, un problème car cela signifie que l'outil limiterait les questions.

Index

Bibliographie

[Abiven, 2019] Abiven, K. (2019). Le moment discursif des barricades d'août 1648 : quelle interprétation des récurrences dans le discours sur l'événement ? *Cahiers de Narratologie*, 35.

[Abiven and Lejeune, 2019] Abiven, K. and Lejeune, G. (2019). Analyse automatique de documents anciens : tirer parti d'un corpus incomplet, hétérogène et bruité. *Recherche d'information, document et web sémantique*, 19(No 1, About Variety in Humanities Big Data).

[Abiven et al., 2021] Abiven, K., Tanguy, J.-B., and Lejeune, G. (2021). Exploiter un corpus de données textuelles sans post-traitement : l'écriture burlesque de la fronde. *Humanités numériques*, (4).

[Afroge et al., 2016] Afroge, S., Ahmed, B., and Mahmud, F. (2016). Optical character recognition using back propagation neural network. In *2016 2nd International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)*, pages 1–4. IEEE.

[Anugrah and Bintoro, 2017] Anugrah, R. and Bintoro, K. B. Y. (2017). Latin letters recognition using optical character recognition to convert printed media into digital format. *Jurnal Elektronika dan Telekomunikasi*, 17(2):56–62.

[Assoucy, 1650] Assoucy, C. C. d. (1650). *L'Ovide en belle humeur de Mr Dassoucy. Enrichy de toutes ses figures burlesques.* Charles de Sercy, Paris.

[Ayres-Bennett, 2004] Ayres-Bennett, W. (2004). *Sociolinguistic variation in seventeenth-century France : Methodology and case studies.* Cambridge University Press.

[Balk and Ploeger, 2009] Balk, H. and Ploeger, L. (2009). Impact : working together to address the challenges involving mass digitization of historical printed text. *OCLC Systems & Services : International digital library perspectives.*

[Bar, 1960] Bar, F. (1960). *Le Genre burlesque en France au XVIIe siècle. Étude de style.* Edition D'Artrey, Paris.

[Barbaresi and Lejeune, 2020] Barbaresi, A. and Lejeune, G. (2020). Out-of-the-box and into the ditch ? multilingual evaluation of generic text extraction tools. In *Actes de 12th Web as Corpus Workshop*, pages 5–13, Marseille, France.

[Beaudouin, 2000] Beaudouin, V. (2000). *Rythme et rime de l'alexandrin classique. Étude empirique des 80 000 vers du théâtre de Corneille et Racine.* Theses, Ecole des Hautes Etudes en Sciences Sociales (EHESS).

[Beaudouin, 2004] Beaudouin, V. (2004). Mètre en règles. *Revue française de linguistique appliquée*, 9(1):119–137.

[Beesley, 1988] Beesley, K. R. (1988). Language identifier : A computer program for automatic natural-language identification of on-line text. In *Proceedings of the 29th annual conference of the American Translators Association*, volume 47, page 54.

[Berg-Kirkpatrick and Klein, 2014] Berg-Kirkpatrick, T. and Klein, D. (2014). Improved typesetting models for historical ocr. In Toutanova, K. and Wu, H., editors, *Actes de 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, volume 2, pages 118–123, Baltimore, Maryland, États-Unis. Association for Computational Linguistics.

[Bermes, 2020] Bermes, E. (2020). *Le numérique en bibliothèque : naissance d'un patrimoine : l'exemple de la Bibliothèque nationale de France (1997-2019)*. PhD thesis, Paris, Ecole nationale des chartes.

[Blanke et al., 2012] Blanke, T., Bryant, M., and Hedges, M. (2012). Open source optical character recognition for historical research. *Journal of Documentation*.

[Bordes, 2022] Bordes, L. (2022). Recueils factices de mazarinades. un singulier exemple du fonds aixois de la bibliothèque méjanes. *Pratiques et formes littéraires. Cahiers du Gadges*, (18).

[Brixtel, 2007] Brixtel, R. (2007). Extraction endogène d'une structure de document pour un alignement multilingue. In *Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles. REncontres jeunes Chercheurs en Informatique pour le Traitement Automatique des Langues (Posters)*, pages 367–376.

[Burgy et al., 2020] Burgy, F., Gerson, S., and Schüpbach, L. (2020). Ex imagine ad litteras : projet d'océrisation de la collection de bry. mathesis, Haute école de gestion de Genève, Genève, Suisse.

[Carrier, 1989] Carrier, H. (1989). *La presse de la Fronde (1648-1653), Les Mazarinades, t. I, La conquête de l'opinion*. Droz, Genève.

[Carrier, 1993] Carrier, H. (1993). Morale et politique devant la guerre civile : les hésitations du duc Claude de Saint-Simon, gouverneur de Blaye, pendant la Fronde des Princes (1650). *Cahiers Saint Simon*, 21(1):35–44.

[Carrier, 1996] Carrier, H. (1996). *Les muses guerrières*. Number 26 in Collection des mélanges de la Bibliothèque de la Sorbonne. Klincksieck, Paris.

[Chen et al., 1998] Chen, S. F., Beeferman, D., and Rosenfeld, R. (1998). Evaluation metrics for language models. In *Actes de DARPA Broadcast News Transcription and Understanding Workshop*, pages 275–280, Lansdowne, Virginia, États-Unis. Carnegie Mellon University.

[Chernoff and Lehmann, 1954] Chernoff, H. and Lehmann, E. L. (1954). The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Annals of Mathematical Statistics*, 25(3):579–586.

[Chiron et al., 2017] Chiron, G., Doucet, A., Coustaty, M., Visani, M., and Moreux, J. (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)*, pages 1–4, Toronto, ON, Canada.

[Clematide et al., 2016] Clematide, S., Furrer, L., and Volk, M. (2016). Crowdsourcing an ocr gold standard for a german and french heritage corpus.

[Cohen et al., 2018] Cohen, K. B., Xia, J., Zweigenbaum, P., Callahan, T. J., Hargraves, O., Goss, F., Ide, N., Névéol, A., Grouin, C., and Hunter, L. E. (2018). Three dimensions of reproducibility in natural language processing. In *Actes de LREC International Conference on Language Resources and Evaluation*, volume 2018, pages 156–165, Miyazaki, Japon . NIH Public Access.

[Cronk, 1987] Cronk, N. (1987). La défense du dialogisme : vers une poétique du burlesque. *Burlesque et formes parodiques. Actes du colloque du Mans, 4-7 décembre 1986*, pages 321–338.

[Debailly, 2012] Debailly, P. (2012). *La Muse indignée. Tome I : La satire en France au xvie siècle.* Classiques Garnier.

[Fischer, 1922] Fischer, R. A. (1922). The mathematical foundations of theoretical statistics. *On Philosophical Transactions of the Royal Society of London*, Series A:309–368.

[Folliard, 2020] Folliard, M. (2020). Le bannissement des muses : Théophile de viau ou l'immanence de l'inspiration. *Littératures classiques*, 102(2):47.

[Franzini et al., 2018] Franzini, G., Kestemont, M., Rotari, G., Jander, M., Ochab, J. K., Franzini, E., Byszuk, J., and Rybicki, J. (2018). Attributing authorship in the noisy digitized correspondence of jacob and wilhelm grimm. *Frontiers in Digital Humanities*, 5:4.

[Gabay, 2019] Gabay, S. (2019). Ocrising 17th french prints. https://editiones.hypotheses.org/1958.

[Gabay, 2020] Gabay, S. (2020). Ocr17 : Gt for 17th french prints.

[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.

[Galand, 2016] Galand, P. (2016). La muse s'amuse : figures insolites de la muse à la renaissance. *Cahiers d'Humanisme et Renaissance*, 130.

[Gale and Church, 1991] Gale, W. A. and Church, K. (1991). Identifying word correspondences in parallel texts. In *Speech and Natural Language : Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991.*

[Ghosh et al., 2016] Ghosh, K., Chakraborty, A., Parui, S. K., and Majumder, P. (2016). Improving information retrieval performance on ocred text in the absence of clean text ground truth. *Information Processing & Management*, 52(5):873–884.

[Giguet et al., 2020] Giguet, E., Lejeune, G., and Tanguy, J.-B. (2020). Daniel@FinTOC'2 Shared Task : Title Detection and Structure Extraction. In *1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation @COLING'2020*, Proceedings of the 28th International Conference on Computational Linguistics (COLING'2020), Barcelone, Spain.

[Grefenstette, 1995] Grefenstette, G. (1995). Comparing two language identification schemes. In *Proceedings of JADT*, volume 95.

[Grinshpun, 2006] Grinshpun, Y. (2006). Fait de langue et fait de style : *Ô* dans les *élégies de Chénier* et dans *Tête d'Or* de Claudel. *L'Information Grammaticale*, 108(1):27–31.

[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 *Actes de Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 1735–1741, Austin, Texas, États-Unis.

[Habert, 2012] Habert, B. (2012). L'archivage pérenne entre us et abus de la mémoire numérique. *Actes des 11e Journées Internationales, Analyse statistique des Données Textuelles (JADT 2012), Liège (Belgique)*, pages 13–15.

[Hamdi et al., 2019] Hamdi, A., Jean-Caurant, A., Sidère, N., Coustaty, M., and Doucet, A. (2019). An Analysis of the Performance of Named Entity Recognition over OCRed Documents. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, volume 24, pages 333–334, Champaign, United States. IEEE.

[Hill and Hengchen, 2019] Hill, M. J. and Hengchen, S. (2019). Quantifying the impact of dirty OCR on historical text analysis : Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities*, 34(4):825–843.

[Hunston, 2008] Hunston, S. (2008). Collection strategies and design decision. In Lüdeling, A. and Kytö, M., editors, *Origin and history of corpus linguistics*. De Gruyter, Berlin.

[Jouhaud, 2009] Jouhaud, C. (2009). *Mazarinades*, volume 28. Editions Aubier.

[Kay et al., 1994] Kay, M., Röscheisen, M., et al. (1994). Text-translation alignment. *Computational linguistics*, 19(1):121–142.

[Kiessling, 2019] Kiessling, B. (2019). Kraken-an universal text recognizer for the humanities. In ADHO, editor, *Actes de Digital Humanities Conference 2019 - DH2019*, Utrecht, Pays-Bas.

[Kissos and Dershowitz, 2016] Kissos, I. and Dershowitz, N. (2016). Ocr error correction using character correction and feature-based word classification. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 198–203. IEEE.

[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.

[Kullback, 1959] Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York.

[La Noue, 1623] La Noue, P. d. (1623). *Le grand dictionnaire des rimes françoises selon l'ordre alphabétique*. Slatkine, Genève.

[Le Cun and Fogelman-Soulié, 1987] Le Cun, Y. and Fogelman-Soulié, F. (1987). Modèles connexionnistes de l'apprentissage. *Intellectica*, 2(1):114–143.

[Lejeune et al., 2012] Lejeune, G., Brixtel, R., Doucet, A., and Lucas, N. (2012). Daniel : Language independent character-based news surveillance. In *International Conference on NLP*, pages 64–75. Springer.

[Maingueneau, 1983] Maingueneau, D. (1983). *Sémantique de la polémique. Discours religieux et ruptures idéologiques au XVIIe siècle*. L'Age d'Homme, Lausanne.

[Mayaffre, 2006] Mayaffre, D. (2006). Philologie et/ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques ? In Rastier, F. and Ballabriga, M., editors, *XXVIIe Colloque d'Albi Langages et Signification*, pages 15–25, Albi, France. Duteil, Carine and Foulquié, Baptiste.

[McEnery and Hardie, 2011] McEnery, T. and Hardie, A. (2011). *Corpus linguistics : Method, theory and practice*. Cambridge University Press.

[Ménage, 1650] Ménage, G. (1650). *Les origines de la langue française*. A. Courbé, Paris.

[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 and Fort, 2019] Millour, A. and Fort, K. (2019). Unsupervised data augmentation for less-resourced languages with no standardized spelling. In *RANLP*, pages 776–784, Varna, Bulgarie .

[Mittendorf and Schäuble, 2000] Mittendorf, E. and Schäuble, P. (2000). Information retrieval can cope with many errors. *Information Retrieval*, 3(3):189–216.

[Moreau, 1851] Moreau, C. (1851). *Bibliographie des mazarinades*, volume 3. J. Renouard et cie.

[Mounier, 2010] Mounier, P. (2010). Manifeste des digital humanities. thatcamp. *Journal des anthropologues. Association française des anthropologues*, (122-123):447–452.

[Mutuvi et al., 2018] Mutuvi, S., Doucet, A., Odeo, M., and Jatowt, A. (2018). Evaluating the impact of ocr errors on topic modeling. In *International Conference on Asia-Pacific Digital Libraries, ICADL 2018*, pages 3–14, Hamilton, Nouvelle-ZélandeNew Zealand. Springer.

[Nagy, 2016] Nagy, G. (2016). Disruptive developments in document recognition. *Pattern Recognition Letters*, 79:106–112.

[Naji et al., 2011] Naji, N., Savoy, J., and Dolamic, L. (2011). Recherche d'information dans un corpus bruité (ocr). In *Actes 8ème Conférence en Recherche d'Information et Applications CORIA'11*, pages 271–286, Avignon, France.

[Nédélec, 2004] Nédélec, C. (2004). *Les États et Empires du burlesque*, volume 51 of *Lumière classique*. H. Champion, Paris.

[Nédelec, 2020] Nédelec, C. (2020). Les muses camuses des burlesques. *Littératures classiques*, N°102(2):81.

[Noille, 2020] Noille, C. (2020). *Comment sortir les textes du musée computationnel ?: Pour un autre scénario de l'édition numérique (la table de montage)*, pages 39–54.

[Pincemin, 1999] Pincemin, B. (1999). Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative. In *Atelier Corpus et TAL : pour une réflexion méthodologique, Conférence TALN*, volume 99, pages 26–36.

[Plas, 2015] Plas, M. (2015). Retours, détours : sur quelques conceptions du bruit : Shannon, von neumann, ashby et cage. *L'Autre Musique Revue*.

[Rastier, 2011] Rastier, F. (2011). *La mesure et le grain : sémantique de corpus*. Number 12 in Lettres numériques. Champion, Paris.

[Rehman and Saba, 2014] Rehman, A. and Saba, T. (2014). Neural networks for document image preprocessing : state of the art. *Artificial Intelligence Review*, 42(2):253–273.

[Saignol, 2021] Saignol, C. (2021). Quel alignement textuel pour l'étude des réécritures théâtrales ? le cas du pédant joué de cyrano de bergerac (1878-1935). In *Réécrire. Doctorales de l'ED III*.

[Schantz, 1982] Schantz, H. F. (1982). *History of OCR, optical character recognition*. Recognition Technologies Users Association.

[Shannon, 1949] Shannon, C. (1949). *La théorie mathématique de la communication*. Cassini, le sel et le fer edition.

[Sinclair, 2003] Sinclair, J. (2003). Corpora for lexicography. In *A practical guide to lexicography*. Van Sterkenberg. Amsterdam : John Benjamins.

[Smith, 2007] Smith, R. (2007). An overview of the tesseract ocr engine. In *Actes de Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633, Parana, Brésil. IEEE.

[Souvay and Pierrel, 2009] Souvay, G. and Pierrel, J.-M. (2009). LGeRM Lemmatisation des mots en Moyen Français. *Traitement Automatique des Langues*, 50(2):21–44.

[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 e-prints*.

[Springmann et al., 2014] Springmann, U., Najock, D., Morgenroth, H., Schmid, H., Gotscharek, A., and Fink, F. (2014). Ocr of historical printings of latin texts : problems, prospects, progress. In *Actes de First International Conference on Digital Access to Textual Cultural Heritage (DATeCH'14)*, pages 71–75, New York, NY, États-Unis. Association for Computing Machinery.

[Stevens, 1961] Stevens, M. E. (1961). *Automatic character recognition : a state-of-the-art report*. Number 112. US Government Printing Office.

[Taghva et al., 1994] Taghva, K., Borsack, J., and Condit, A. (1994). Results of applying probabilistic ir to ocr text. In *Actes de ACM-SIGIR*, pages 202–211, Dublin, IrlandeIreland. Springer.

[Taghva et al., 1996] Taghva, K., Borsack, J., and Condit, A. (1996). Effects of ocr errors on ranking and feedback using the vector space model. *Information Processing & Management*, 32(3):317–327.

[Taghva et al., 2004] Taghva, K., Nartker, T., and Borsack, J. (2004). Information access in the presence of ocr errors. In *Actes de 1st ACM workshop on Hardcopy document processing*, pages 1–8, Washington (DC), États-Unis.

[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, *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). Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL*, pages 205–217, Nancy, France. ATALA.

[Thoreau, 1854] Thoreau, H. D. (1854). *Walden ou la vie dans les bois*. Thoreau, Henry David.

[Traub et al., 2018] Traub, M. C., Samar, T., van Ossenbruggen, J., and Hardman, L. (2018). Impact of crowdsourcing ocr improvements on retrievability bias. In *Actes de the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL'18, pages 29—36, New York, NY, États-Unis. Association for Computing Machinery.

[Traub et al., 2015] Traub, M. C., Van Ossenbruggen, J., and Hardman, L. (2015). Impact analysis of ocr quality on research tasks in digital archives (tpdl 2015). In *International Conference on Theory and Practice of Digital Libraries*, pages 252–263, Poznań, PolognePoland. Springer.

[Ul-Hasan et al., 2016] Ul-Hasan, A., Bukhari, S. S., and Dengel, A. (2016). Ocroract : A sequence learning ocr system trained on isolated characters. In *Actes de 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 174–179, Santorini, Grèce. IEEE.

[van Strien et al., 2020] van Strien, D., Beelen, K., Ardanuy, M. C., Hosseini, K., McGillivray, B., and Colavizza, G. (2020). Assessing the impact of ocr quality on downstream nlp tasks. In *Actes de the 12th International Conference on Agents and Artificial Intelligence (ICAART)*, volume 1, pages 484–496, Valleta, MalteMalta.

[Vobl et al., 2014] Vobl, T., Gotscharek, A., Reffle, U., Ringlstetter, C., and Schulz, K. U. (2014). Pocoto-an open source system for efficient interactive postcorrection of ocred historical texts. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pages 57–61.

[Zhou, 2010] Zhou, Y. (2010). Are your digital documents web friendly ?: Making scanned documents web accessible. *Information technology and Libraries*, 29(3):151–160.