# Analyse automatique de documents anciens : tirer parti d'un corpus incomplet, hétérogène et bruité

## Automatic analysis of old documents: taking advantage of an incomplete, heterogeneous and noisy corpus

Karine Abiven[1], Gaël Lejeune[1]

[1] Sorbonne Université - STIH, 1 rue Victor Cousin, 75230 Paris Cedex
{prenom.nom}@sorbonne-universite.fr

**RÉSUMÉ.** Cet article concerne un ensemble de textes anciens (datant du milieu du 17e siècle), que les spécialistes d'histoire et de littérature ont l'habitude de nommer "corpus des *mazarinades*". Ces quelque 5500 textes offrent une variété de problématiques qui s'inscrivent pleinement dans le domaine des humanités numériques. Nous montrons en premier lieu qu'il ne s'agit pas à proprement parler d'un corpus puisqu'on ne dispose pas, malgré un important travail bibliographique sur le sujet, d'une définition ni d'un recensement rigoureux de cet ensemble. Il s'agit ensuite de voir l'impact de cette définition instable sur le travail des chercheurs qui s'intéressent à ce "corpus", tout en proposant de corriger ces biais grâce à un outillage automatique. Nous montrons que, si le but est d'exploiter le matériau textuel et non de l'interpréter, il est intéressant de s'autoriser à traiter des données brutes (avec un minimum de traitements préparatoires). Enfin, nous exposons un premier cadre d'application sur la sous-partie de cet ensemble actuellement disponible sous forme numérique : la datation de documents. La méthode utilisée se fonde sur une analyse en chaînes de caractères qui permet à la fois de fonctionner sur un corpus partiellement bruité (états de langue divers, scories de l'océrisation…) et sur un corpus hétérogène, comprenant des documents de tailles et surtout de genres très variés. Nous montrons que, dans certains cas, le bruitage du corpus peut être un avantage pour certaines tâches de classification, notamment grâce à l'utilisation de méthodes exploitant des chaînes de caractères. Les approches en caractères permettent en effet de surmonter un certain nombre de difficultés liées à la variété des données disponibles. Aussi ce travail donne-t-il des outils pour extraire des sous-corpus cohérents, pour exploiter des jeux de données issus de la numérisation en économisant le post-traitement, et pour identifier des métadonnées manquantes : trois enjeux essentiels pour ce "corpus" qui reste encore pour une bonne part à divulguer à la communauté dans un format numérique raisonné.

**ABSTRACT.** In this article we try to tackle some problems arising with noisy and heterogeneous data in the domain of digital humanities. We investigate a corpus known as the *mazarinades* corpus which gathers around 5,500 documents in French from the 17th century. First of all, we show that this set of documents is not strictly speaking a corpus since its coverage has not been thoroughly defined. Then, we advocate that it is possible to get interesting results even in the case of such an incomplete, heterogeneous and noisy dataset by strictly limiting the amount of pre-treatments necessary fro processing texts. Finally, we present some results on a case study on document dating where we aim to complete missing metadata in the *mazarinades* corpus. We exploit a method based on character strings analysis which is robust to noisy data and can even take advantage of this noise for improving the quality of the results.

**MOTS-CLÉS.** Documents anciens, Mazarinades, Fouille de Textes, Datation, Corpus, Numérisation.

**KEYWORDS.** Old documents, Mazarinades, Text Mining, Document Dating, Corpus.

## 1. Introduction

L'analyse de textes anciens intéresse de nombreux chercheurs en humanités numériques, chercheurs en informatique (dans les domaines du Traitement Automatique des Langues et de l'Analyse de Documents Numérisés notamment), mais aussi spécialistes de ces textes (philologues, littéraires, historiens…). Un des objets principaux des humanités numériques est de permettre à ces chercheurs d'horizons différents de trouver un terrain commun quant à la manière d'appréhender les objets de recherche d'une part, et quant à la définition des objectifs de recherche de l'autre.

Dans ce cadre, la simple numérisation des données, textuelles ou non, ne peut constituer une fin en soi : l'exploitation, l'enrichissement et la mise en perspective de ces données sont des enjeux plus cruciaux. Les objectifs

scientifiques, de plus en plus ambitieux, des spécialistes de chaque domaine qui sont les « destinataires des données » invitent à questionner la richesse et la pertinence des différents moyens mis à disposition par la science, notamment informatique. Ces chercheurs sont notamment en demande de capacités d'interrogation, de consultation et de navigation élaborées. Mais au delà, l'outil informatique offre la possibilité d'extraire des régularités à différents niveaux et ainsi de mettre en lumière des propriétés de corpus qui seraient très difficiles à déceler « à l'œil nu ».

Les données auxquelles nous nous intéressons dans cette étude sont un ensemble d'écrits datant de la Fronde (milieu du XVII$^{e}$ siècle) et regroupés sous le nom de *mazarinades*. Il s'agit de quelques milliers de documents imprimés sous des formes assez variables, et qui ont trait aux événements contemporains du gouvernement du cardinal Mazarin : si beaucoup de pièces sont critiques à son égard (d'où le nom de *mazarinades* : épopée parodique de Mazarin), il en est aussi qui prennent sa défense, ainsi que bien d'autres qui traitent plus généralement des problèmes du royaume de France à l'époque. La diversité des points de vue est aussi grande que la diversité formelle du corpus.

Notre objectif est de proposer différentes manières d'accéder aux données de ce corpus, que ce soit dans des perspectives de classification non-supervisée (regroupement de documents, détection de tendances, visualisation…) ou de classification supervisée (datation et attribution d'auteur). Ces deux opérations (datation et attribution) sont particulièrement utiles sur un tel corpus de textes polémiques, car leurs métadonnées sont lacunaires : ils sont souvent anonymes (en raison du danger qu'il y aurait à avouer des écrits transgressifs envers le pouvoir) et parfois non datés (car imprimés de manière expéditive en réaction aux événements). Nous nous intéresserons en particulier à l'identification de l'état du grain d'analyse le plus pertinent pour offrir une véritable plus-value au corpus. Il s'agit notamment de définir la meilleure solution opératoire, parmi tous les états possibles du corpus : depuis la version très bruitée à la sortie d'OCR jusqu'au texte modernisé en français contemporain, en passant par la transcription respectant l'état de langue ancien, la graphie, l'orthographe erratique, etc. (qu'on appellera par commodité *transcription "diplomatique"*). Au delà de l'outillage, une des questions centrales qui nous a animés est celle de savoir si un traitement automatique efficace rendait véritablement nécessaire de disposer d'un jeu de données totalement débarrassé de toutes ses aspérités : scories de l'OCR, mots tronqués, états de langue variables, multilinguisme…Autrement dit, il s'agit de savoir si le pré-traitement des données tel qu'il est défini dans une chaîne de traitement traditionnelle de TAL, parfois imparfait et souvent coûteux, n'est pas plus utile pour rassurer l'humain qui en observe le fonctionnement que pour optimiser les tâches accomplies par la machine.

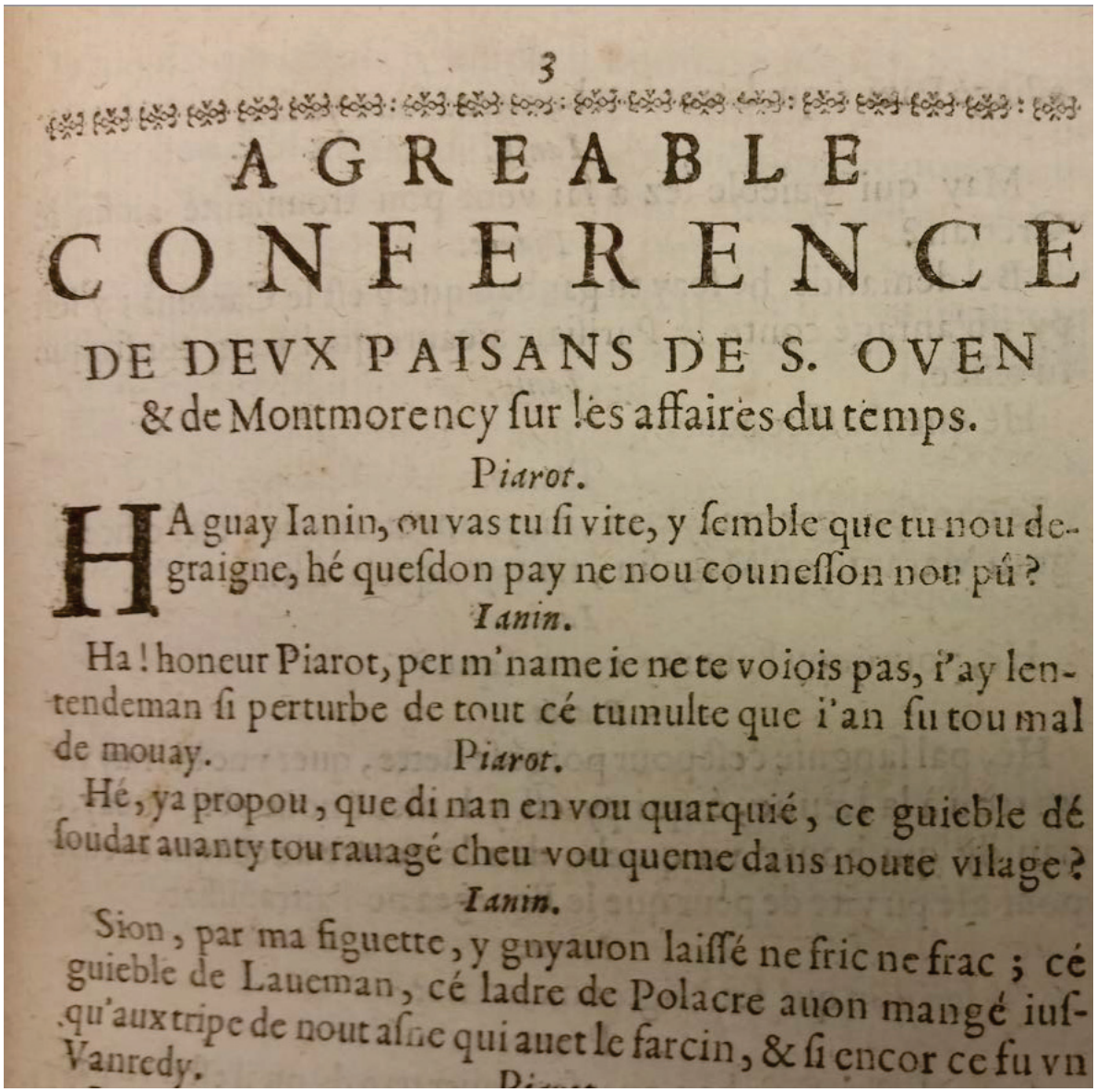
Nous présenterons dans la section 2 notre corpus d'étude, puis, dans la section 3, les différentes approches que nous avons mises en place pour exploiter nos données. Dans la section 4, nous exposerons les premiers résultats que nous avons obtenus en termes de classification et de visualisation. Enfin, nous proposerons quelques conclusions et perspectives de ce travail dans la section 5.

## 2. Le "corpus" des mazarinades : contexte et données

Qu'est-ce donc que les « mazarinades » et s'agit-il à proprement parler d'un corpus ? Dans l'usage critique des littéraires et des historiens, le mot a un sens imprécis : empruntée à un titre de l'époque (la *Mazarinade* attribuée à Scarron, 1651), l'étiquette renvoie d'abord à des écrits à charge contre Mazarin et parus lors de la Fronde (1648-1653). Cette révolte du milieu du XVII$^{e}$ siècle, parfois assimilée à une révolution, est plutôt une révolte parlementaire et nobiliaire qui a gagné la population de Paris et de certaines grandes villes de province contre le pouvoir jugé abusif qu'avait alors Mazarin, bras droit d'Anne d'Autriche pendant la Régence pour minorité de Louis XIV. Par extension dans la tradition, le mot *mazarinade* inclut les réponses à ces pamphlets (toute apologie du ministre), et finalement en vient à dénoter toute pièce imprimée entre 1648 et 1653 ayant trait de près ou de loin aux événements politiques touchant la Fronde (à l'exclusion des livres, dont la production chute, justement aux profit de ces milliers de libelles, c'est-à-dire de petits livres (Jouhaud, 2009, p. 28)).

Se retrouvent donc englobées sous une même appellation des pièces de tailles, de supports matériels, de visées pragmatiques et de genres les plus divers : pamphlets pour ou contre Mazarin, poésies burlesques, chansons dont la circulation fut d'abord orale, lettres fictives. Outre ces écrits relevant de la littérature au sens large, on inclut traditionnellement dans cet ensemble des pièces officielles (actes royaux, arrêts, remontrances de cours souveraines), ainsi que des lettres authentiques et des discours effectivement prononcés puis restitués à l'écrit :

il s'agit alors plutôt de "scansions de l'action" (Jouhaud, 2009, p. 30). On a pu y voir « l'équivalent de tous les registres de la presse moderne, depuis le *Journal officiel* jusqu'au *Canard enchaîné* » (Carrier, 1989, p. 63). D'autres paramètres créent aussi de la diversité : les langues utilisées sont diverses (en majorité le français, mais aussi le latin, l'italien et des versions littéraires de patois, dont on voit un exemple dans la Figure 2.1).



3

AGREABLE
CONFERENCE
DE DEVX PAISANS DE S. OVEN
& de Montmorency ſur les affaires du temps.

*Piarot.*

HA guay Ianin, ou vas tu ſi vite, y ſemble que tu nou degraigne, hé queſdon pay ne nou couneſſon non pû?

*Ianin.*

Ha! honeur Piarot, per m'name ie ne te voiois pas, i'ay lentendeman ſi perturbe de tout cé tumulte que i'an ſu tou mal de mouay.

*Piarot.*

Hé, ya propou, que di nan en vou quarquié, ce guieble dé ſoudar auanty tou rauagé cheu vou queme dans noute vilage?

*Ianin.*

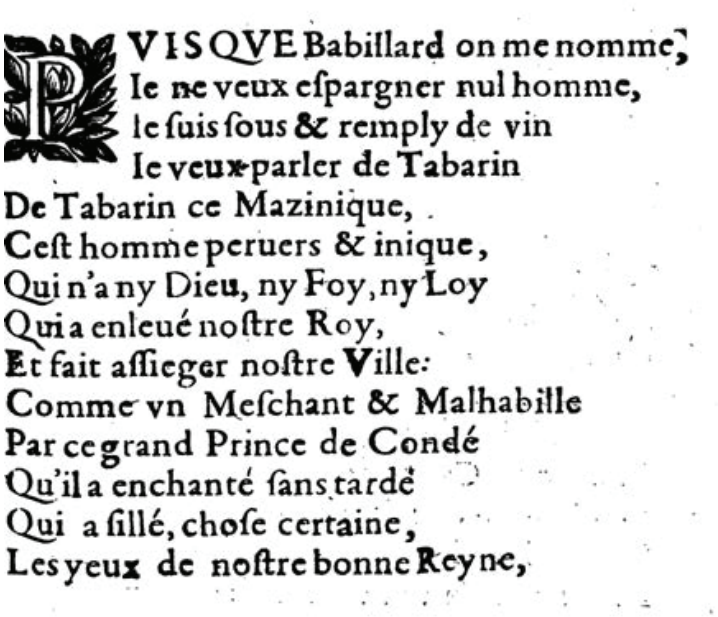
Sion, par ma figuette, y gnyauon laiſſé ne fric ne frac; cé guieble de Laueman, cé ladre de Polacre auon mangé iuſqu'aux tripe de nout aſne qui auet le farcin, & ſi encor ce ſu vn Vanredy.

**FIGURE 2.1.** *Une pièce du corpus : "Agréable conférence de deux paysans de Saint-Ouen et de Montmorency sur les affaires du temps" (dialogue imitant le patois de l'Ile de France, 1649), Réserve de la Bibliothèque Interuniversitaire de la Sorbonne (cote HJR 4= 17 Pièce 16)*

L'état de langue, enfin, n'est pas stable (variantes graphiques, abréviations, orthographe erratique). Ainsi, à la variabilité externe du "corpus" s'ajoutent diverses modalités d'instabilité interne. Il faut y ajouter, dans une perspective de numérisation, l'état inégal de conservation d'imprimés souvent produits dans l'urgence et l'économie de moyens (papier et encre de mauvaise qualité, notamment). Tous ces paramètres induisent une certaine résistance à l'océrisation (voir par exemple la Figure 2.2) ce qui donne une idée de la difficulté d'obtenir de manière automatique une version conforme à la source ou retranscrite en langue moderne (Figure 2.3).

Pour en revenir à la délimitation du corpus, l'ensemble est donc d'ordinaire défini essentiellement par une thématique (les affaires politiques relatives aux événements politiques de la Régence) et la synchronie de ses parties. L'extension du corpus n'est donc pas fixe car ces critères sont sujets à caution. Il est toutefois indispensable de s'appuyer dans un premier temps sur cette définition souple qu'ont proposée les bibliographes et chercheurs des XIX$^{e}$ et XX$^{e}$ siècles, car ils ont fourni des repérages fondamentaux dans ce grand volume de textes. Nous n'aurons donc pas ici pour ambition de discuter les critères de délimitation de ce corpus, même si nous envisageons plus bas des solutions pour le rationaliser grâce à l'outillage numérique.

Ainsi défini approximativement, cet ensemble est estimé à 5500-6000 unités bibliographiques selon l'évalua-



VIS QVE Babillard on me nomme
Ie ne veux eſpargner nul homme,
Ie ſuis ſous & remply de vin
Ie veux parler de Tabarin
De Tabarin ce Mazinique,
Ceſt homme peruers & inique,
Qui n'a ny Dieu, ny Foy, ny Loy
Qui a enleué noſtre Roy,
Et fait aſſieger noſtre Ville:
Comme vn Meſchant & Malhabille
Par ce grand Prince de Condé
Qu'il a enchanté ſans tardé
Qui a ſillé, choſe certaine,
Les yeux de noſtre bonne Reyne,

VIS QVE Babillard on me nomme
JnlËflP le ne veux efpargner nul homme,
WjfflsL le fuis fous & remply de vin
Icvcuxrparlcr deTabarin
DeTabarin ce Mazinique, .
Cefthomrnepcruers & inique,
Qiii n'any Dieu, nyFoy.nyLoy
Quiaenleuénoftre Roy, .
Et fait affieger noftrc Ville :
Commcrvn Mefchant ôC Malhabille
Par ce grand Prince de Condé
Qu'il a enchanté fans tarde* J
Qui a fille, chofe certaine.
Les yeux de noftre bonne Rey ne,

**FIGURE 2.2.** *Exemple de sortie d'océrisation sans post-traitement du Babillard du temps, Paris, N. de la Vigne, 1649*

PVISQVE Babillard on me nomme
Ie ne veux espargner nul homme,
Ie suis saoul, & remply de vin,
Ie veux parler de Tabarin,
De Tabarin ce Mazinique,
Cet homme peruers & inique,
Qui n'a ny Dieu, ny Foy, ny Loy,
Qui à enleué nostre Roy,
Et fait assieger nostre Ville
Comme vn Meschant & Malhabille
Par ce grand Prince de Condé
Qu'il a enchanté sans tardé,
Qui a sillé, chose certaine,
Les yeux de nostre bonne Reyne

PUISQUE Babillard on me nomme
Je ne veux épargner nul homme,
Je suis saoul, et rempli de vin,
Je veux parler de Tabarin,
De Tabarin ce Mazinique,
Cet homme pervers et inique,
Qui n'a ni Dieu, ni Foi, ni Loi,
Qui a enlevé notre Roi,
Et fait assiéger notre Ville
Comme un Méchant et Malhabile
Par ce grand Prince de Condé
Qu'il a enchanté sans tarder,
Qui a sillé, chose certaine,
Les yeux de notre bonne Reine,

**FIGURE 2.3.** *À gauche transcription fidèle (dite "diplomatique") sur* `http://mazarinades.org/`*. À droite orthographe moderne restituée. Les modifications sont marquées en gras.*

tion la plus récente (Haffemayer *et al.*, 2016). Aucune bibliographie exhaustive n'existe à ce jour. Les ressources bibliographiques et catalographiques dont nous disposons sont les suivantes :

**Moreau 1850-1851** (Moreau, 1850) : ce bibliographe du XIX$^{e}$ siècle décompte et décrit plus de 4000 pièces (mais sans se fonder sur la collection la plus riche du monde, alors non cataloguée, de la Bibliothèque Mazarine). Des suppléments complètent sa liste initiale, mais avec des doublons (Moreau, 1862, 1869),

**Socard 1876** (Socard, 1876) étend la délimitation du "corpus" en ajoutant les titres de la riche bibliothèque de Troyes,

**Labadie 1904** (Labadie, 1904) ajoute les titres des mazarinades des fonds bordelais, la Fronde ayant été particulièrement intense et longue à Bordeaux,

**Carrier 1989** (Carrier, 1989) a synthétisé les bibliographies pré-citées et opéré un recensement manuel dans les bibliothèques à travers le monde pour les compléter,

**Haffemayer *et al.* 2016** (Haffemayer *et al.*, 2016) : l'introduction de cet ouvrage collectif rend compte d'un projet en cours mené par les conservateurs de la bibliothèque Mazarine : un catalogue numérique complet des mazarinades, qui devrait être mis en ligne en 2019.

Enfin, l'outil le plus récent qui permet d'exploiter cet ensemble est un corpus numérisé nommé "Projet Mazarinades", issu de la collection de mazarinades de l'Université de Tokyo, mis en ligne par l'équipe des "RIM" ("Recherches Internationales sur les Mazarinades"). Cette collection de 2711 pièces a été numérisée en 2008 et transcrite en 2010-2011 sur un site qui permet l'accès public aux recherches lexicales dans le corpus, à un catalogue des métadonnées incluses dans les notices des textes, ainsi que l'accès au corpus intégral pour les membres inscrits (http://mazarinades.org/, financé par les programmes 20903010, B-22320066, C-26370364 - JSPS KAKENHI). Une convention signée avec les chercheurs qui en sont responsables nous a permis d'exploiter les fichiers sources de ce corpus en ligne, qui sont à l'origine des premiers résultats présentés ici (cf. section 4). Il a été réalisé par une saisie manuelle, ce qui confère à ces données une qualité remarquable pour l'analyse humaine, au-delà de ce qu'on peut espérer obtenir automatiquement ; toutefois, dans le cadre du présent travail, ce procédé ne permet pas de remonter à une version océrisée brute (d'où les tests d'océrisation que nous avons effectués nous-mêmes). En effet, il peut être intéressant d'observer ces différents états intermédiaires, pour l'analyse par l'expert mais aussi pour l'analyse automatique (voir section 3). Par ailleurs, des doublons apparaissent dans les résultats de recherche : il s'avère que ce corpus en ligne contient en fait 1996 pièces uniques, le reste étant des émissions diverses, ou des états différents d'une même édition. Enfin, le corpus est lacunaire et aléatoire, puisque la collection de la bibliothèque de l'Université de Tokyo, toute riche qu'elle soit, contient environ un tiers des pièces existantes. Si cette collection numérisée est pionnière et comble incontestablement un vide important, la constitution du corpus est donc de nature plus opportuniste que raisonnée (McEnery & Hardie, 2011).

L'estimation à 5500-6000 unités textuelles vient de (Carrier, 1989) et regroupe 5200 imprimés et 500 manuscrits environ, modulo 5 % d'erreur selon cet auteur. La marge d'erreur est due au fait qu'un repérage systématique par les catalogues n'était pas possible à son époque et ne l'est toujours pas, même si la situation est en cours d'amélioration en raison des projets de catalogage numérique systématique à la Mazarine déjà évoqués, ainsi qu'à la Bibliothèque Nationale de France (site de l'Arsenal), qui est l'autre collection importante dans le monde. *A fortiori* en 1989, le catalogage était non exhaustif : comme ces libelles sont souvent de brèves pièces, elles ont été reliées depuis le XVII$^{e}$ siècle dans des recueils dits *factices*, c'est-à-dire assemblés par d'anciens possesseurs sans rigueur bibliographique (pour les manuscrits, le catalogage est encore plus incomplet), ce qui renforce le caractère opportuniste de la démarche de constitution de la collection. Le repérage ne peut donc se faire seulement par le titre, mais suppose souvent la fréquentation des ouvrages ; par ailleurs les titres sont parfois trompeurs. En effet, il peut annoncer à première vue un ouvrage sur la Fronde alors qu'il n'en est rien, pour des raisons commerciales ou encore parce que les pratiques d'imitation et de parodie étaient fréquentes. Le travail de Carrier est donc aussi colossal qu'incontournable : ses repérages manuels lui ont permis de découvrir 200 pièces inconnues de ses prédécesseurs bibliographes. Il a également rationalisé les listes de ces derniers, qui comprenaient des doublons notamment en raison de la non-distinction entre éditions et émissions (les réimpressions d'une même édition avec parfois un simple changement de page de titre étaient prises en compte plusieurs fois) ; au contraire, se voyaient regroupées sous un seul numéro des textes distincts (comme les numéros de périodiques, par exemple). S'agissant des manuscrits, Carrier a estimé qu'ils devaient s'élever au dixième des pièces imprimées, soit 500 maximum. Il a surtout établi que la quasi-totalité sont des textes qui ont connu des copies imprimées par la suite ; selon lui, seuls une vingtaine de manuscrits sont des pièces restées sans impression (Carrier, 1989, p. 68). La question se pose de savoir si l'on doit ou non inclure dans le corpus les manuscrits qui, hormis cette poignée de textes non imprimés, sont au plan textuel des doublons par rapport à leur version imprimée.

Les estimations de Carrier posent toutefois plusieurs problèmes. Tout d'abord, il n'a pas pu divulguer de bibliographie exhaustive : les chiffres que nous reprenons ici sont donnés au début d'une étude critique sur le sujet. On ne peut donc s'appuyer sur une liste des titres correspondant aux 5200 imprimés dont parle (Carrier, 1989). La seule bibliographie qu'on possède est celle, incomplète et non dédoublonnée, de Moreau (Moreau, 1850), augmentée de suppléments (Moreau, 1862, 1869). Ensuite, Carrier inclut dans son décompte de 5200 pièces des textes perdus : il recense en effet des pièces « dont nous savons par les contemporains qu'elles furent publiées, mais dont il n'a subsisté aucun exemplaire » (Carrier, 1989, p. 73) ; son but est d'enrichir autant que possible la connaissance sur la production des presses de l'époque (pour comparer la quantité de pièces d'une année sur l'autre par exemple). Mais, outre qu'il ne précise pas le chiffre exact de ces textes virtuels, ces derniers ne sauraient répondre aux critères scientifiques de constitution de corpus, au sens linguistique du terme, lesquels ne doivent inclure que des occurrences attestées. Enfin, Carrier propose plusieurs statistiques dans son ouvrage (nombre de textes par année, par partis politiques, etc.), qu'il a calculées sur un échantillon de mille mazarinades, et non sur l'ensemble (tâche en effet impossible à mener manuellement) : il a choisi 1000 textes jugés représen-

tatifs du corpus. Les statistiques qu'il donne pour l'ensemble sont donc le produit d'une généralisation de tests effectués sur un échantillon relativement faible (20 %) du corpus, dont la représentativité est assez peu étayée (Carrier, 1989, p. 48-49).

Les chiffres mesurant l'étendue du corpus dans cette tradition bibliographique sont présentés dans le tableau 2.1, par ordre chronologique des bibliographes. Il est à noter que parmi les 700 pièces supplémentaires rassemblées par Carrier figurent 200 périodiques non individualisés dans la numérotation Moreau, 200 pièces supplémentaires identifiées à la bibliothèque Mazarine ainsi que 300 pièces en province et à l'étranger. Il évoque également environ 500 manuscrits (non répertoriés) qui ne sont pas comptabilisés dans ces chiffres.

| Auteur | Nombre de Pièces | Commentaires |
|---|---|---|
| Moreau 1850-1869 | 4607 | environ 4200 sans les doubles |
| Socard 1876 | 125 | environ 50 sans les redondances avec Moreau |
| Labadie 1904 | 346 | environ 250 non redondantes |
| Carrier 1989 | 5200 | 4500 des précédents (sans redondance)+700 nouvelles pièces |

***Tableau 2.1.*** *Nombre de pièces recensées dans les principales bibliographies de mazarinades*

L'hétérogénéité du corpus tel que le construit cette tradition est néanmoins à questionner, au regard de la taille, du type de document et du contenu, notamment. S'agissant de la taille, et si l'on définit les mazarinades comme des *libelles* (mot dont le sens premier est "petit livre"), on vise des textes plutôt courts : les mazarinades sont souvent constituées de demi-cahiers de deux feuillets ou de cahiers de quatre feuillets, ce qui donne des brochures de 4, 8 ou 16 pages la plupart du temps. Mais la diversité de taille reste très importante comme nous pouvons le voir dans le tableau 2.2.

| | Total | Moy.(± écart-type) | Min. | Max. |
|---|---|---|---|---|
| # Caractères | 29.838.013 | 14.993 (± 21.3564) | 519 | 382.942 |
| # Tokens | 6.397.988 | 3.215 (± 4.583) | 98 | 81.744 |
| # Tokens Uniques | 1.857.951 | 933 (± 858) | 55 | 12.208 |
| # Phrases | 137.042 | 68 (± 118) | 5 | 2.533 |

***Tableau 2.2.*** *Caractéristiques des 1996 pièces uniques numérisées du corpus "Projet Mazarinades"*

À partir de 1650, les pièces s'allongent et peuvent aller jusqu'à 32 voire 64 pages (Carrier, 1989, p. 86-87). Toutefois, certaines des « mazarinades » les plus connues (le *Mascurat* de G. Naudé, le *Recueil de maximes véritables* de C. Joly) comptent plusieurs centaines de pages. Elles ont certes un rapport étroit avec le reste du corpus : par exemple, le *Mascurat* est une réflexion sur les mazarinades elles mêmes et donne, dès 1649, un premier aperçu bibliographique des libelles parus au début de la Fronde. Mais au plan de la bibliographie matérielle, on est loin du petit livret rédigé et imprimé dans l'urgence. Quant au type documentaire, les documents officiels (actes royaux, arrêts du Parlement, etc.) sont traditionnellement inclus dans cet ensemble. Ces pièces n'ont pourtant rien de pamphlétaire et ne sont pas spécifiques, comme genres, à la Fronde. Mais les instances gouvernementales à l'origine de ces documents sont parties prenantes dans le conflit, et ces actes d'écriture participent donc à la polémique ; en outre, exclure ces textes reviendrait à se priver d'une part importante de l'interdiscours du temps. En effet, le contenu des différents arrêts et déclarations du Parlement ou du roi est amplement glosé dans les pamphlets. Il peut être vu comme raisonnable de garder dans le corpus les pièces officielles ayant trait à la Fronde et d'en exclure ceux qui traitent des affaires courantes du royaume, quoi que cette tâche est coûteuse s'il faut l'accomplir manuellement.

Aussi, pouvoir regrouper de manière semi-automatique les documents selon les genres textuels ou encore à partir de segments de textes communs apparait d'un grand intérêt pour l'expert. Le traitement automatique s'avère précieux pour alléger ces tâches, comme le montre un test effectué pour tenter d'exclure les textes administratifs n'ayant apparemment pas de rapport avec les événements exceptionnels de la Fronde mais traitant des affaires ordinaires. Nous avons pu extraire automatiquement les titres de textes a priori issus des instances officielles (roi et Parlement) : 194 (sur les 1996 pièces uniques numérisées du corpus "Projet Mazarinades"), à partir des mots du titre suivants : *arrest*, *arrests*, *articles*, *ordonnances*, *ordonnance*, *declaration*, *codicille*, *codicile*, *registres*.

Nous avons ensuite effectué manuellement le tri suivant :

— 164 titres correspondraient *a priori* à des textes émis par le roi ou le Parlement en lien avec la Fronde,
— 12 titres relèvent d'affaires *a priori* sans lien direct avec celle-ci (telle traduction d'une ordonnance du roi du Portugal sur ses terres au Brésil, qui se trouve avoir été imprimée dans ces années-là) ; ces 12 textes seraient donc à exclure, après examen rigoureux de leur contenu,
— 18 titres sont à reverser dans le corpus des pièces non officielles, car leur titre comprend les mots ci-dessus sans que ceux-ci en constituent le noyau sémantique (par exemple pour le mot *articles* : *La Relation extraordinaire, contenant le traicté de Mazarin avec le Parlement d'Angleterre. Ensemble les Articles de Composition pour le lieu de sa retraicte dans la Ville de Londres*), ou encore imitent manifestement la structure de titres officiels tout en annonçant un contenu satirique, trahissant un contenu pamphlétaire (par exemple, pour le mot *articles* : *Les Articles des crimes capitaux, dont est accusé le Cardinal Mazarin, & desquels il se doit iustifier*).

Même si cette catégorie de "pièces officielles" demanderait à être mieux définie, ce test vise à montrer les ressources du tri automatique quant à la constitution de corpus, ainsi que la plus-value que l'outillage offre à l'expert, à qui le tri manuel revient *in fine*, mais dans des proportions bien moindres.

Cette opération démontre deux avancées possibles pour la définition de ce corpus grâce au traitement automatique : en premier lieu, elle permet de trier des textes qui devraient être exclus d'un corpus défini comme les "pièces ayant trait aux événements politiques de la Fronde". Ici, on a pu sélectionner automatiquement les 9,7 % de textes potentiellement officiels, dont on a ensuite éliminé manuellement 15,5 % d'items non pertinents. En second lieu, il serait idéalement possible de constituer des sous-corpus qui résoudraient l'impossible délimitation stricte d'un tel volume de textes qui englobe de nombreuses dimensions de l'écrit public (par exemple la question de l'inclusion ou non des pièces officielles dans un corpus essentiellement pamphlétaire) : dans une base de données qui permettrait de sélectionner les métadonnées relatives au type documentaire, on pourrait choisir de sortir les pièces officielles pour constituer un corpus exclusivement pamphlétaire ; si on contraire on cherche à repérer des segments répétés ou à opérer des datations relatives, on pourra constituer un corpus extensif, incluant toutes les pièces relatives à la Fronde. On pourrait aussi par exemple résoudre les difficultés induites par l'hétérogénéité volumétrique, en rendant possible le choix d'un sous-corpus de textes de longueur inférieure ou égale à 16 pages, ce qui représente le gros des "mazarinades" au sens de pamphlets.

Les "mazarinades" de la tradition apparaissent ainsi moins comme un corpus *stricto sensu* que comme une matière dans laquelle on pourra tailler des corpus plus cohérents (signifiants, représentatifs, homogènes), à condition de doter cette masse de textes de métadonnées les plus complètes possible.

## 3. Approche des données bruitées

Le bruitage et l'hétérogénéité des données d'une part et le caractère lacunaire des métadonnées de l'autre apparaissent comme les principaux obstacles au traitement global de cet ensemble de textes. Aussi cherchons-nous en premier lieu à enrichir ces données, sans viser de tâches particulières : il s'agit de rechercher des moyens originaux de traiter les données et partant, de s'affranchir de ces problèmes voire d'en tirer parti. Nous présenterons dans cette section la méthodologie que nous avons mise en place, puis nous exposerons un cas d'application.

### 3.1. *Choisir le bon grain d'analyse : les caractères comme observables*

Nous avons présenté dans la section 2 les principales caractéristiques que nous avons identifiées sur ce corpus : scories d'OCR, multilinguisme et état de langue ou orthographe instables. Ces différents paramètres affectent nécessairement les opérations. Quand nous parlerons par la suite de post-traitement, ce sera au sens de "post-traitement de l'OCR" pour désigner les opérations visant à corriger la sortie de l'OCR pour se rapprocher de l'état originel du document. La notion de pré-traitement sera entendue dans le sens de "pré-traitement pour le TAL" et désignera de manière générale les opérations sur les données permettant de faciliter le travail des outils de TAL placés en aval de la chaîne de traitement. Une approche traditionnelle de TAL consisterait sans doute donc à pré-traiter, à normaliser les données de manière à avoir un état de langue homogène. Cet état de langue permettrait ensuite de tirer parti des outils usuels en TAL, en particulier ceux permettant l'uniformisation des

observables (tokenisation puis lemmatisation) ou leur enrichissement (étiquetage morpho-syntaxique).

Nous prendrons ici le contre-pied de cette approche, pour des raisons à la fois épistémologiques et pragmatiques. D'un point de vue épistémologique, dans la lignée de l'approche *Corpus Integrity* (Dias, 2010), nous souhaitons n'effectuer que les pré-traitements nécessaires à l'utilisation des données et non systématiser l'utilisation de ceux-ci. En effet, dans le cadre de l'analyse de textes en langue ancienne, l'adaptation ou la conception d'analyseurs se révèle à la fois complexe et coûteuse (Guibon *et al.*, 2015). Il ne s'agit pas bien entendu de refuser cette difficulté mais de partir du postulat que tout pré-traitement doit être justifié et doit apporter une plus-value tangible. Si, pour une tâche particulière, le pré-traitement s'avère plus efficace que l'absence de pré-traitement alors nous n'excluons pas d'en tirer parti. D'un point de vue pragmatique, sur les textes anciens issus de l'océrisation, obtenir une retranscription fidèle du texte d'origine exige un travail très important, incluant souvent une supervision par un expert. Ceci peut considérablement repousser le stade où les données deviennent exploitables. En effet, une grande partie de l'effort se concentre alors sur les pré-traitements en amont plutôt que sur les tâches à accomplir en aval. De ce fait, nous proposons de travailler avec des chaînes de caractères comme observables en faisant l'hypothèse qu'il n'est pas obligatoire de disposer d'un grain d'analyse interprétable pour obtenir un résultat utile.

Les approches en caractères ont montré leur robustesse dans un certain nombre de tâches de Traitement Automatique des Langues, depuis des tâches de bas niveau telle que l'identification de langue (Lui & Baldwin, 2012) jusqu'à des tâches plus riches d'extraction d'information (Lejeune *et al.*, 2015) en passant par l'attribution d'auteur où ces approches sont usuelles (Stamatatos, 2009; Sun *et al.*, 2010; Brixtel, 2015).

### 3.2. *Calcul de motifs en caractères*

Les approches en caractères se fondent le plus souvent sur des n-grammes de caractères, avec une volonté de déterminer les valeurs optimales de $n$. Ici nous exploitons des chaînes de caractères un peu plus spécifiques. En effet, l'approche que nous proposons se situe plus généralement à la lisière entre l'algorithmique du texte et la fouille de données. Nous utilisons des motifs (en caractères) fermés et fréquents comme traits pour entraîner un classifieur. Les propriétés de fermeture et de fréquence sont définies de la façon suivante [1]

**Fermeture :** le motif ne peut être étendu vers la gauche ou vers la droite sans diminuer son nombre d'occurrences

**Fréquence :** le motif respecte une borne minimale de nombre d'apparitions (en nombre de textes)

Pour faire la liaison entre les termes utilisés en fouille de texte et ceux utilisés en algorithmique du texte, (Buscaldi *et al.*, 2017) rappellent que les propriétés de fermeture et de fréquence de la fouille correspondent aux propriétés de maximalité et de répétition en algorithmique du texte.

Pour reprendre un exemple classique tiré de (Ukkonen, 2009), si on considère la chaîne de caractères `HATTIVATTIA`, on trouve trois motifs fermés fréquents : `T`, `A` et `ATTI`. `TT` n'est pas fermé car il apparaît systématiquement à chaque occurrence de `ATTI` : son contexte droit est toujours `I` et son contexte gauche `A`. *A contrario*, le motif `T` est fermé car ses contextes gauches et droits varient.

Ces motifs en caractères fermés fréquents constituent une manière condensée de représenter toutes les sous-chaînes de caractères d'un corpus. Pour calculer ces motifs en caractères de manière efficace, en l'occurrence avec une complexité linéaire en la taille des données, nous utilisons ici une implantation en PYTHON [2] de l'algorithme décrit dans (Ukkonen, 2009) qui exploite les tableaux de suffixes augmentés décrits dans (Kärkkäinen *et al.*, 2006).

Du point de vue du TAL, cette technique peut être décrite comme une tokenisation *non-supervisée* en ce sens que les règles de découpage ne sont pas pré-définies mais sont calculées en fonction du corpus donné en entrée. Les tokens obtenus sont des chaînes de caractères mots ou non-mots qui peuvent tout aussi bien être des caractères pris isolément (y compris des caractères non-imprimables), des morphèmes, des groupes de mots ou une combinaison de tout cela (cf. Figure 3.4).

1. Ici le terme s'entend au sens de l'anglais *frequency*, il s'agit donc de fréquence absolue ou d'effectif
2. Disponible sur `https://github.com/gip0/py-rstr-max`

— "iˆ"
— "rdonn"
— "_exploi"
— "_mesme_"
— "_les_lieux_&_en_son_absence_"
— "_aux_Substituts_de_nostredit_Procureur"
— "ontraintes_lb/__solidaires_contre_les_Habitans_des_Parroisses_"

**FIGURE 3.4.** *Exemples de motifs fermés fréquents extraits du document intitulé "ARREST DE LA COVR DES AYDES" (1653, source : corpus "Projet Mazarinades"), les "_" représentent les espaces.*

### 3.3. *Configuration*

De façon traditionnelle dans les approches orientées fouille de données, un défi important est de limiter l'explosion du nombre de motifs, que ce soit pour des raisons calculatoires ou pour des raisons de lisibilité des résultats. Un filtrage classique consiste à appliquer aux motifs deux types de contrainte :

— La contrainte de support par laquelle on définit le nombre minimal ($minsup$) et maximal ($maxsup$) d'objets qui supportent un motif. Ici, cela consiste à définir le nombre minimal et maximal de textes dans lequel un motif apparaît.
— La contrainte de longueur par laquelle on définit la longueur minimale ($minlen$) et maximale ($maxlen$) d'un motif. Ici, il s'agit donc d'une longueur en caractères, sans se limiter aux caractères alphanumériques.

Notre chaîne de traitement fonctionne de la façon suivante :

1. Calcul des motifs fermés fréquents dans tout le corpus de textes (jeu d'apprentissage et jeu de test) ;
2. Filtrage des motifs selon la longueur et/ou le support ;
3. Représentation de chaque texte sous forme d'un vecteur d'effectif des motifs ;
4. Utilisation de l'implantation de SCIKIT du SVM ONE VS REST.

Nous n'avons pas séparé le jeu de données en apprentissage et test du fait de la petite taille du jeu données : moins de 2000 textes et des classes fortement sous-repésentées. Nous avons donc utilisé une validation croisée en 10 strates au moyen de la fonction STRATIFIEDKFOLD de SCIKIT [3]. Concernant le choix du classifieur, nous avons observé que le SVM ONE VS REST était significativement plus performant que des réseaux bayésiens, arbres de décision et forêts d'arbres aléatoires. Ce résultat est conforme aux observations de (Brixtel, 2015) et (Buscaldi *et al.*, 2017) obtenus sur d'autres tâches de classification mais avec le même type de motifs comme descripteurs.

Les motifs que nous avons exploités ne sont pas filtrés sur le support maximal puisque ce paramétrage n'apportait que des modifications marginales des résultats.

— Pas de taille minimale ($minlen = 1$) ;
— Nous avons fait varier la taille maximale ($maxlen$) de 1 à 10 ;
— Nous avons fait varier la contrainte de $minsup$ de 1 à 100 par pas de 10, le $maxsup$ ne bougeant pas.

Nous exploitons un noyau linéaire puisque c'est celui qui a produit les meilleurs résultats avec ce type de traits sur différents types de données et différentes tâches (Brixtel, 2015; Buscaldi *et al.*, 2018). Ceci peut être lié au fait que lorsque le nombre de traits est important, la plus-value du noyau radial sur le noyau linéaire est moins évidente même après paramètrage. Par ailleurs, le noyau linéaire offrait des résultats plus rapides que nos tests avec un noyau radical.

## 4. Expériences de datation automatique

Nous présentons dans cette section les résultats de datation automatique obtenus sur le sous-corpus nommé "Projet Mazarinades". Nous avons choisi de nous attaquer à la tâche de datation car c'est un des premiers problèmes pointés par les bibliographes : 16 % des documents ne mentionnent ni lieu ni date selon des statistiques effectuées sur environ 1000 unités (Carrier, 1991, p. 150), 31 % mentionnant l'un ou l'autre. Les dates ont parfois

3. http://scikit-learn.org/

pu être attribuées *a posteriori* à partir de critères internes au texte (par un des auteurs mentionnés dans le tableau 2.1). En outre, un des enjeux à terme, sur le plan historique, est de pouvoir dater plus finement qu'à l'année, mais plutôt au mois, comme le fait (Carrier, 1989, 1991) afin de situer précisément les textes les uns par rapport aux autres, vu leur masse et la brève période de temps concernée (6 ans).

### 4.1. *Etat de l'art en datation automatique de textes*

La littérature sur le sujet comporte un certain nombre de travaux qui ont pour point commun de travailler au grain mot. (de Jong *et al.*, 2005) ont travaillé sur des documents récents (journaux du début des années 2000) en utilisant des modèles de langue dans lesquels les auteurs associent à chaque mot une probabilité d'occurrence dans une période de temps donnée. L'amélioration du modèle proposée par (Kanhabua & Nørvåg, 2008) se fonde sur un étiquetage morpho-syntaxique et la recherche des lemmes les plus fréquents par période de temps. Mais il s'agit de documents récents, peu bruités pour lesquels les outils d'étiquetage se comportent très bien.

Les éditions 2010 et 2011 du Défi Fouille de Textes (DEFT) proposaient une tâche qui se rapproche de ce que nous traitons ici puisqu'il s'agissait de textes océrisés datant du XIX$^{e}$ siècle. Pour cette tâche, (Garcia-Fernandez *et al.*, 2011), ont proposé cette même idée de modèles de langues temporels en mots. Les auteurs cherchent notamment à détecter les néologismes et les archaïsmes. Ils exploitent également des connaissances externes en utilisant la date de naissance des personnes dont le nom est cité dans les textes. De leur côté, (Raymond & Claveau, 2011) utilisent également une représentation en mots mais remarquent que l'utilisation d'analyseurs morpho-syntaxiques se heurterait au fait que l'entrée est bruitée. Les auteurs proposent une modélisation des erreurs d'OCR : les erreurs d'OCR se traduisent souvent par l'apparition inopinée de ponctuation, et ces erreurs correspondent souvent à des périodes de temps particulières. Ces signes de ponctuation surnuméraires deviennent donc des descripteurs très importants. En effet, nous pouvons observer ce phénomène dans la figure 2.2 (page 4) où divers caractères surnuméraires (dont des signes de ponctuation) apparaissent à l'issue de l'océrisation. Notre approche est donc à comparer à ce dernier travail puisque la recherche des chaînes de caractères nous permet de détecter les scories d'OCR.

### 4.2. *Sous-corpus pour la datation et baselines*

La répartition par date des documents de notre corpus figure dans le tableau 4.3. Nous avons exploité un sous-corpus composé des 1768 pièces qui possèdent une date (88,6% des textes).

| | 1648 | 1649 | 1650 | 1651 | 1652 | Non-daté | Total |
|---|---|---|---|---|---|---|---|
| Nombre | 15 | 1000 | 63 | 103 | 587 | 228 | 1996 |
| Proportion | 0,8% | 50,1% | 3,1% | 5,2% | 29,4% | 11,4% | 100% |

**Tableau 4.3.** *Répartition des 1996 pièces uniques par date (1768 datés et 228 non datés)*

Nous présentons ici des expériences menées sur les documents datés de manière à éprouver la méthode avant de pouvoir traiter ces documents sans date et d'effectuer un travail d'évaluation manuel avec des experts. Nous avons préparé différentes baselines de manière à mesurer la plus-value apportée par les méthodes exploitant les motifs en caractères.

Nous avons conçu des baselines simples opérant au grain mot pour lesquelles nous n'avons pas utilisé de *stop-list*. Ceci pour deux raisons : d'une part car ce filtrage n'a pas amené de plus-value sur la qualité de la classification ni sur le temps de calcul et d'autre part car ceci facilite la comparaison avec les approches en motifs où cette notion de *stop-list* n'existe pas. Ces baselines sont au nombre de 3 et fonctionnent par raffinement successif :

- Baseline 1 (B1) : une approche de sac de mots classique sans utilisation de *stop-list* mais où on élimine les *hapax* (mots d'effectif 1 dans l'ensmeble du corpus)
- Baseline 2 (B2) : qui exploite des n-grammes de mots avec $1 <= n <= 3$, c'est donc l'approche B1 à laquelle on ajoute les 2-grammes de mots et 3-grammes de mots
- Baseline 3 (B3) : une approche en motifs en mots fermés fréquents, c'est donc l'approche B2 moins les motifs redondants

| | Transcription "diplomatique" | | | Orthographe modernisée | | |
|---|---|---|---|---|---|---|
| | MicroF | MacroF | Sim | MicroF | MacroF | Sim |
| B1 : mots | 0,7812 | 0,4335 | 0,825 | 0,8122 | 0,4715 | 0,8645 |
| B2 : n-grammes de mots ($1 <= n <= 3$) | 0,8164 | **0,5865** | 0,8663 | 0,8388 | **0,5974** | 0,899 |
| B3 : motifs en mots ($1 <= len <= 3$) | **0,8294** | 0,5693 | **0,883** | **0,8581** | 0,5725 | **0,9107** |

**Tableau 4.4.** *Résultats de 3 baselines opérant au grain mot pour les deux états de langue : Transcription dite "diplomatique" et orthographe modernisée*

Les résultats de ces baselines figurent dans le tableau 4.4. Nous avons utilisé 3 métriques pour l'évaluation. D'une part, nous avons calculé la Micro F-mesure (F-mesure globale sur le jeu de données) et la Macro F-mesure (moyenne des F-mesure pour chaque classe). Ceci permet de distinguer les méthodes qui se comportent mieux sur les classes peuplées, des méthodes qui obtiennent de bons scores dans les classes faiblement peuplées. Mais, dans le cas de la datation, l'inconvénient d'utiliser une F-mesure classique est bien connu : le caractère binaire des notions de Faux positif/Faux négatif fait que la pénalisation d'un mauvais jugement est la même quel que soit l'écart par rapport à la date réelle. Dans cette configuration toutes les erreurs se valent. Nous avons donc utilisé une mesure de similarité fondée sur une courbe gaussienne proposée par (Grouin *et al.*, 2011). Cette mesure présente l'avantage de pénaliser fortement les écarts de prédiction supérieurs à une année (mesur Sim). La période de temps dans notre corpus étant nettement moins grande (6 ans contre 150 ans) que dans l'article pré-cité, nous avons adapté la pénalité de sorte qu'un écart d'une année constitue un gain de 0,72 (au lieu de 0,97 avec la mesure d'origine), un écart de deux années constitue un gain de 0,28 (contre 0,88) et qu'au delà de 3 ans le gain soit marginal (0,004). Une méthode prédisant systématiquement le centre de la période obtiendrait un score moyen de 0,14 environ.

Nous pouvons observer tout d'abord que l'utilisation des 2-grammes et 3-grammes de caractères dans B2 permet d'obtenir un gain substantiel, en particulier en Macro F-mesure. En effet, B2 est notablement plus efficace sur les classes minoritaires du jeu de données. Si B3 améliore les résultats de B2 sur la Micro F-mesure d'environ un point de pourcentage sur les deux configurations linguistiques, il n'en est pas de même en Macro F-mesure. Il semble qu'en masquant une partie des *features* redondantes contenues dans les n-grammes, les motifs fermés perdent des informations utiles pour les classes minoritaires du corpus. De fait, B3 permet un gain substantiel sur les classes fortement peuplées, ce qui se reflète sur la Micro F-mesure mais au prix d'une perte relativement significative sur la Macro F-mesure. (Brixtel, 2015) a montré, sur une tâche d'attribution d'auteur, que l'élimination de redondance pouvait occasionner une légère perte de performance. Ce point de vue doit probablement être contrebalancé par le fait que les classes de taille réduite sont beaucoup plus facilement influencées par le changement de résultat sur une poignée d'instances.

## 4.3. *Approche fondée sur les caractères*

Nous montrons dans la figure 4.5 les résultats obtenus par la méthode de fouille de motifs en caractères. En abscisse figure la longueur maximale en caractères des motifs utilisés pour la classification (paramètre $maxlen$). Par souci de lisibilité nous n'avons pas fait figurer sur ces courbes les variations sur le paramètre $minlen$. Pour rappel, la transcription dite "diplomatique" restitue aussi fidèlement que possible le texte et l'état de langue originaux, là où l'autre variante opère une modernisation de l'orthographe et une normalisation de la graphie. Notons toutefois que la modernisation qui a été opérée dans le corpus numérisé "Projet Mazarinades" est partielle car automatique, fondée notamment sur la dissimilation des u/v et des i/j, confondus au XVII$^{e}$ siècle : la Figure 2.3 (page 4), où nous avons complété à la main la modernisation de l'extrait, montre l'ampleur de cette tâche si elle est effectuée manuellement. Nous pouvons observer que dans les 4 configurations, la méthode exploitant des motifs en caractères arrive à de bons résultats, y compris en se contentant des motifs de longueur 1 ($maxlen = 1$), puis que les résultats augmentent régulièrement à mesure que des motifs de taille plus grande sont pris en compte. À partir d'une certaine taille, la méthode commence en effet à prendre en compte des indices qui sont plus de l'ordre lexical (cf. Figure 3.4).

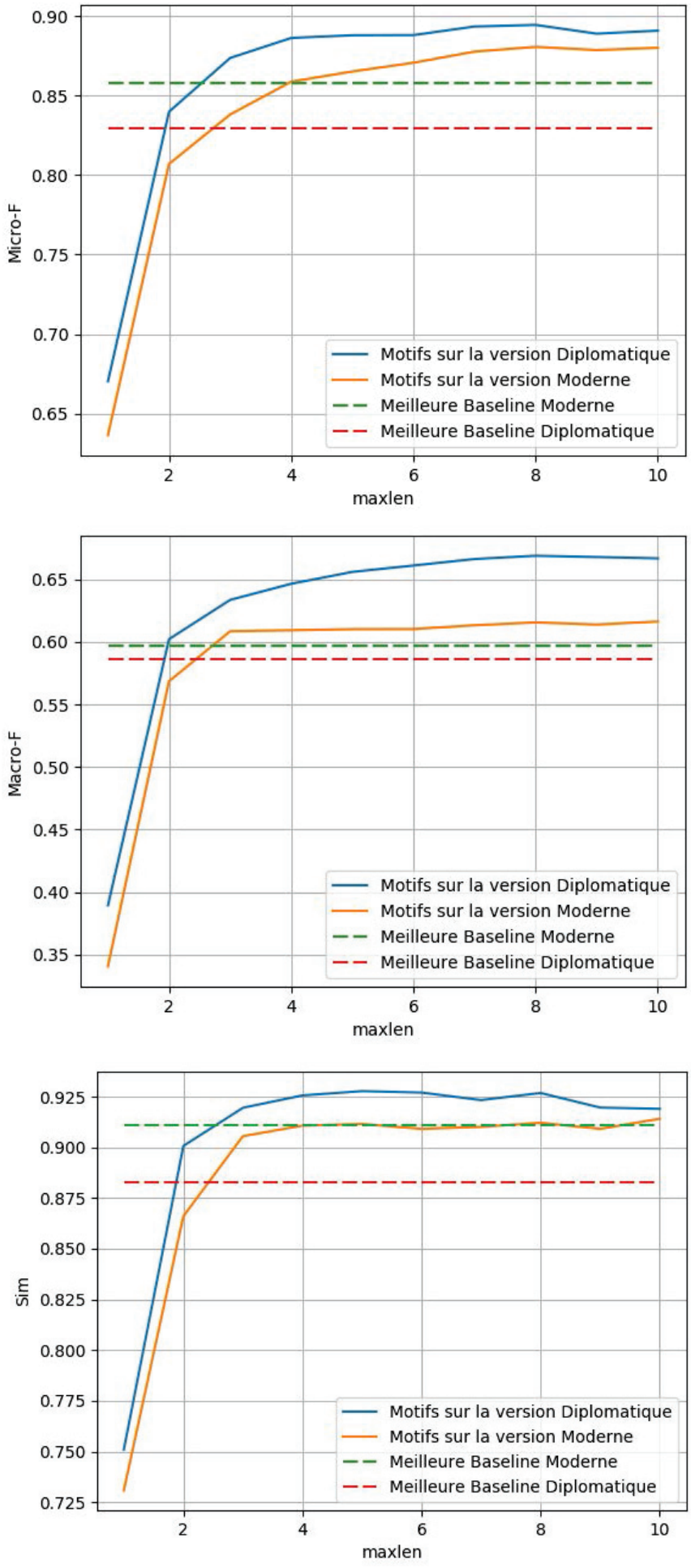



**FIGURE 4.5.** *Résultat des approches fondées sur les caractères en terme de Micro F-mesure, de Macro F-mesure et de similarité de date (Sim) en fonction de la valeur de $maxlen$. En pointillés le score des meilleures baselines pour chaque état de langue.*

De manière générale, la méthode est moins affectée par la variation du matériau au niveau des états de la langue que les baselines qui exploitent le grain mot. Cette méthode fonctionne même légèrement mieux sur les versions "diplomatiques". Ceci semble valider notre hypothèse selon laquelle cette méthode est moins sensible au bruitage des données. Elle semblerait même plus efficace dans tous les cas. Mais, c'est sur la macro-mesure que la plus-value est la plus nette, ce qui amène à nuancer quelque peu le résultat puisque les catégories sont très déséquilibrées, cette mesure étant dès lors moins stable. De façon plus étonnante, c'est sur la mesure de similarité que la plus-value est la moins forte à la fois entre les deux états de langue mais aussi par rapport à la meilleure baseline. Étant donné que cette mesure semble être la plus fiable pour la tâche de datation, il s'avère que seule la variété des états de langue amène une plus-value nette en faveur des motifs en caractères. Il serait sans doute intéressant de voir comment ces méthodes se comporteraient si elles étaient confrontées à des jeux de données où versions "diplomatiques" et versions modernisées seraient mêlées.

## 5. Conclusion et perspectives

Le corpus que nous avons présenté, traditionnellement appelé *mazarinades* par les chercheurs en SHS qui l'exploitent, comprend dans sa définition la plus large 5000 à 6000 documents publiés en France pendant la Fronde (1648-1653). Ce corpus nous semble particulièrement intéressant pour illustrer les collaborations possibles entre différentes spécialités dans le cadre des humanités numériques. Exploité par des chercheurs provenant de divers horizons (historiens, littéraires, etc.), il ne dispose pas à ce jour d'un outillage numérique complet, en raison de sa masse et du caractère lacunaire des métadonnées connues. Un tiers de cet ensemble existe toutefois déjà dans des formats numériques exploitables, auquel l'accès nous a été autorisé pour mener à bien nos premiers tests : le corpus "Projet Mazarinades". Nous ambitionnons de mener conjointement un travail de numérisation qui complèterait cet ensemble et un travail d'exploitation des données. Dans le cadre de cet article, nous avons travaillé sur le jeu de données existant déjà sous forme numérique, d'une part pour interroger la définition des *mazarinades*, d'autre part pour tester une première exploitation de ces données par des techniques de TAL.

Sur le premier point, il est apparu que les techniques de TAL pouvaient assister les experts dans une définition plus sûre du périmètre du corpus lui-même, puisque en l'état, ce que l'on nomme usuellement *mazarinades* n'est pas un corpus à proprement parler mais plutôt l'agrégation d'un travail bibliographique s'étalant sur plusieurs siècles. Grâce à l'automatisation, l'extraction de corpus cohérents à l'intérieur de cette collection disparate est apparue comme désormais envisageable.

Le second point concerne l'exploitation des données, pour laquelle on a testé ici des méthodes de datation automatique des documents, puisque leur date n'est pas toujours connue ou sûre. Si les données dont nous disposons actuellement sont en partie post-traitées, les données futures que nous produirons grâce à la numérisation nécessiteront un post-traitement qui, quel que soit son degré d'automatisation, sera coûteux. Nous avons donc cherché à utiliser une méthode qui soit la plus robuste possible face au bruitage des données, de manière à pouvoir exploiter au fil de l'eau les nouvelles données disponibles. Cette méthode, fondée sur une analyse en chaînes de caractères (mots ou non-mots) a prouvé sa robustesse dans différentes configurations et notamment dans les cas où l'hétérogénéité des données pourrait gêner l'apprentissage. Ces premiers tests interrogent l'équilibre à trouver entre la qualité des données et l'interprétabilité de celles-ci. Pour garantir ces tests, il conviendrait de pouvoir les accompagner d'un score de confiance : charge à l'expert de les confirmer ou infirmer par l'examen des textes (par exemple, s'agissant de la datation, au regard d'indices internes, comme la mention d'événements qui peuvent renvoyer à une chronologie).

Parmi les tâches que nous avons en perspective figurent d'une part la question de l'attribution d'auteur (10 % seulement des auteurs ont signé leurs textes, contre 7 % de pseudonymes et 83 % d'anonymes, selon (Carrier, 1991, p. 77-79)), et d'autre part la détection et le suivi des entités nommées, notamment les noms de lieux comme dans (Moncla *et al.*, 2017), de manière à pouvoir résoudre des tâches d'Extraction d'Information.

## Références

BRIXTEL R. (2015). Maximal repeats enhance substring-based authorship attribution. In *Recent Advances in Natural Language Processing, RANLP 2015, 7-9 September, 2015, Hissar, Bulgaria*, p. 63–71.

BUSCALDI D., GREZKA A. & LEJEUNE G. (2017). Tweetaneuse : Fouille de motifs en caractères et plongement lexical à l'assaut du deft 2017. In *Actes du 13e Défi Fouille de Texte*, p. 65–76, Orléans, France : Association pour le Traitement Automatique des Langues.

BUSCALDI D., LE ROUX J. & LEJEUNE G. (2018). Modèles en caractères pour la détection de polarité dans les tweets. In *Actes du 14e Défi Fouille de Texte*, p. 249–258, Rennes, France : Association pour le Traitement Automatique des Langues.

CARRIER H. (1989). *La Presse de la Fronde (1648-1653) : Les mazarinades, vol. I, La conquête de l'opinion*. Genève : Droz.

CARRIER H. (1991). *La Presse de la Fronde (1648-1653) : Les mazarinades, vol. II, Les Hommes du livre*. Genève : Droz.

DE JONG F., RODE H. & HIEMSTRA D. (2005). Temporal language models for the disclosure of historical text. In *Humanities, computers and cultural heritage : Proceedings of the XVIth International Conference of the Association for History and Computing (AHC 2005)*, p. 161–168 : Koninklijke Nederlandse Academie van Wetenschappen.

DIAS G. (2010). *Information Digestion*. Habilitation à diriger des recherches, Université d'Orléans.

GARCIA-FERNANDEZ A., LIGOZAT A.-L., DINARELLI M. & BERNHARD D. (2011). Méthodes pour l'archéologie linguistique : datation par combinaison d'indices temporels. In *DÉfi Fouille de Textes*, p. 29–40, Montpellier, France.

GROUIN C., FOREST D., PAROUBEK P. & ZWEIGENBAUM P. (2011). Présentation et résultats du défi fouille de textes DEFT2011. In *DEFT 2011 (TALN 2011)*, p. 3–14.

GUIBON G., TELLIER I., PRÉVOST S., CONSTANT M. & GERDES K. (2015). Analyse syntaxique de l'ancien français : quelles propriétés de la langue influent le plus sur la qualité de l'apprentissage ? In *TALN 22*, TALN 22, Actes en ligne, Caen, France.

HAFFEMAYER S., REBOLLAR P. & SORDET Y. (2016). Mazarinades, nouvelles approches. In *Histoire et civilisation du livre, XII, Droz*.

JOUHAUD C. (2009). *Mazarinades. La Fronde des mots*. Paris : Aubier.

KANHABUA N. & NØRVÅG K. (2008). Improving temporal language models for determining time of non-timestamped documents. In B. CHRISTENSEN-DALSGAARD, D. CASTELLI, B. AMMITZBØLL JURIK & J. LIPPINCOTT, Eds., *Research and Advanced Technology for Digital Libraries*, p. 358–370, Berlin, Heidelberg : Springer Berlin Heidelberg.

KÄRKKÄINEN J., SANDERS P. & BURKHARDT S. (2006). Linear work suffix array construction. *Journal of the ACM*, p. 918–936.

LABADIE E. (1904). *Nouveaux Suppléments à la bibliographie des Mazarinades*. Paris : Henri Leclerc.

LEJEUNE G., BRIXTEL R., DOUCET A. & LUCAS N. (2015). Multilingual event extraction for epidemic detection. *Artificial Intelligence in Medicine*. doi : 10.1016/j.artmed.2015.06.005.

LUI M. & BALDWIN T. (2012). langid.py : An off-the-shelf language identification tool. In *In Proceedings of the ACL 2012 System Demonstrations*, p. 25–30.

MCENERY T. & HARDIE A. (2011). *Corpus Linguistics : Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge University Press.

MONCLA L., GAIO M., JOLIVEAU T. & LAY Y.-F. L. (2017). Automated geoparsing of paris street names in 19th century novels. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, GeoHumanities'17, p. 1–8, New York, NY, USA : ACM.

MOREAU C. (1850). *Bibliographie des Mazarinades*. Paris : Jules Renouard.

MOREAU C. (1862). *Supplément à la Bibliographie des mazarinades*, In *le Bulletin du Bibliophile et du bibliothécaire*, p. 786–829. Techener : Paris.

MOREAU C. (1869). *Supplément à la Bibliographie des mazarinades*, In *le Bulletin du Bibliophile et du bibliothécaire*, p. 61–81. Techener : Paris.

RAYMOND C. & CLAVEAU V. (2011). Participation de l'IRISA à DEFT 2011 : expériences avec des approches d'apprentissage supervisé et non-supervisé. In *Challenge DeFT (défi fouille de texte)*, p. 19–27, Montpellier, France.

SOCARD E. (1876). *Supplément à la Bibliographie des Mazarinades*. Paris : H.menu.

STAMATATOS E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538–556.

SUN J., YANG Z., WANG P. & LIU S. (2010). Variable length character n-gram approach for online writeprint identification. In *Multimedia Information Networking and Security (MINES), 2010 International Conference on*, p. 486–490.

UKKONEN E. (2009). Maximal and minimal representations of gapped and non-gapped motifs of a string. *Theoretical Computer Science*, p. 4341–4349.