Des données au corpus

L'exploitation numérique des mazarinades

Karine ABIVEN(1), Gaël LEJEUNE(2)

(1) (2) Sorbonne Université (UR 4509 Sens-Texte-Informatique-Histoire)

1 Le contexte du projet Antonomaz

Le projet Antonomaz (ANalyse auTOmatique et NumérisatiOn des MAZarinades) vise l'exploitation et la visualisation d'un ensemble de documents identifiés depuis longtemps sous le nom de « mazarinades ». Il s'agit de brefs imprimés dont les dimensions sont moindres que celle d'un livre (brochures, ou libelles), qui furent publiés au temps de la Fronde en France, entre 1648 et 1653, et qui sont relatifs de près ou de loin aux mouvements politiques de cette époque. Ces textes intéressent plusieurs communautés : les linguistes, notamment pour la connaissance des variations du français à cette époque ; les historiens, ces textes documentant une révolte majeure de l'histoire de France ; et les littéraires, car c'est un laboratoire formel où à peu près tous les genres de discours disponibles sont politisés, détournés, pastichés.

Le projet Antonomaz recoupe les enjeux classiques du champ des humanités numériques : la préparation et l'exposition d'un corpus et de ses métadonnées, et la volonté de proposer des explorations possibles de ces données et métadonnées. Les objets proposés ici pourraient apparaître un peu à la marge de ce qu'*a priori*, on peut entendre comme « corpus d'auteurs » puisque l'auctorialité pose problème dans ces textes dont les quatre cinquièmes sont anonymes. Ils le sont pour des raisons de contournement de la censure, et de pratique de l'imprimé en temps de crise, qui a permis à des gens de plume de toutes sortes, sans autorité, de produire du discours. On verra toutefois que l'étude des métadonnées, notamment éditoriales, est un des buts du projet, qui entend aussi réfléchir à la notion de corpus.

En effet, si la notion même de *corpus* pose ici problème, c'est parce qu'elle sous-entend *stricto sensu* une certaine homogénéité. Or, tel qu'il a été légué par la tradition, cet

ensemble fait se côtoyer des pièces qui n'ont parfois qu'un point commun matériel (de brefs documents en format *in 4°*) et temporel (imprimés entre 1648 et 1653) : des pièces officielles comme des déclarations royales ou des arrêts du parlement côtoient ainsi des pièces obscènes en vers burlesques, par exemple. Face à cet ensemble quantitativement important et hétérogène, la méthode que nous avons cherché à mettre en place combine encodage classique et travail sur des données bruitées. Nous avons en effet décidé de renoncer, au moins provisoirement, à l'obtention de données textuelles parfaites (dans le sens de manuellement retranscrites et vérifiées) en raison de la masse à considérer. Il s'agit en effet d'environ 5.000 écrits, avec en outre des éditions différentes, soit, sans doute, plusieurs millions d'exemplaires à l'époque, dont encore 1 à 2 millions pourraient être conservés aujourd'hui, selon Jouhaud ([1985] 2009), avec une relativement bonne couverture numérique. L'obtention d'un mode texte parfait pour une telle masse serait très coûteuse, et par ailleurs, ce corpus sous forme retranscrite existe déjà en partie[1]. La volumétrie de l'ensemble peut être plus précisément estimée à environ 71.212 pages sur les 5.085 entrées recensées par les bibliographies du XIX^{e} siècle (à raison de 13,5 pages en moyenne par document) : nous avons calculé que, pour un échantillon d'un peu plus de 100 mazarinades de tailles et langues différentes, 23 minutes sont nécessaires, par page, pour réaliser la transcription avec une structuration XML minimale (13 minutes pour la retranscription elle-même, et 10 minutes pour la relecture). Cette moyenne cache d'importantes disparités selon les pages et ne répond pas à la méthode préconisée de la transcription en double aveugle avec accord inter-annotateurs. Cette méthode rapide représenterait pourtant, ramenée à la taille du corpus, environ 27.000 heures de travail à temps plein pour une personne seule, sans compter l'encodage qui, selon la richesse des traitements, multiplierait « par un facteur cinquante […] le temps de travail, la taille du fichier et la densité informationnelle exploitable » par rapport à un texte brut, selon Gefen (2015 : 5). Nous n'avons pas vérifié cet ordre de grandeur, ayant automatisé un certain nombre de ces post-traitements, mais il va de soi qu'ils augmentent considérablement le temps (et le coût) de production de « bonnes » données.

Aussi nous a-t-il semblé raisonnable non seulement donc de ne pas entreprendre de transcription de l'ensemble des documents disponibles en format image, mais aussi de tailler en amont dans cette masse en isolant des sous-corpus, puisque seuls ces ensembles de documents sont finalement intéressants car homogènes. Ces choix permettent de rejoindre l'épistémologie d'une « libellistique » (au sens de Jouhaud, 2003), c'est-à-dire une étude des libelles comme objets de langage au fonctionnement spécifique, ne se comprenant que dans le contexte serré de leur production, comme « action d'écriture » (Jouhaud, [1989] 2005). Pour de tels écrits publiés en réseaux, liés à l'actualité, souvent pris dans des jeux de questions-réponses, on comprend l'intérêt de constituer des corpus cohérents et de travailler finement les métadonnées pour restituer ces contextes.

Notre démarche est double. Elle passe tout d'abord par la conception de chaînes de traitement pour obtenir des données et les perfectionner peu à peu. Elle consiste

[1] Il s'agit du corpus des RIM ([http://mazarinades.org/], consulté le 20 août 2021), très riche, mais non exhaustif et non téléchargeable ou partageable, ce qui justifie que l'on cherche à compléter la disponibilité en ligne et à créer un corpus répondant aux exigences des principes FAIR (Findability, Accessibility, Interoperability and Reusability).

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

concomitamment à exploiter ces données dans leur forme non préparée, directement issue du processus de reconnaissance automatique de caractères (OCR pour *Optical Character Recognition*). Nous espérons ainsi montrer comment, au plan de la gestion de projet, on peut essayer de faire parler des documents avant d'avoir réussi à tous les préparer sous une forme idéale. En cela, nous cherchons à ne pas imputer systématiquement les insuffisances des résultats de recherches automatiques à la qualité des données : ce sont parfois les méthodes elles-mêmes qui sont insuffisamment robustes à la variation de données non standardisées. On prend ainsi le parti de réfléchir aux observables[2] les plus directement exploitables plutôt que de chercher des observables idéaux. Cela implique, du moins par provision, de considérer les documents en tant que masse de données (puisque, pour un ensemble documentaire ancien, les « mazarinades » représentent une quantité substantielle). Au plan de l'épistémologie des humanités numériques, il convient dès lors de se prémunir du seul attrait du « poids des données[3] » dépourvu de projet interprétatif, et de chercher plutôt, grâce à l'obtention rapide d'une masse suffisante, à reconstruire ces données en corpus, en un « bidouillage artisanal » de données, mais aussi en une tentative de « redéploiement des pratiques interprétatives »[4].

2 Conception d'une chaîne de traitement

Nous avons conçu une chaîne de traitement permettant de passer de manière semi-automatique, avec une intervention humaine réduite aux cas difficiles, de fac-similés numériques à des fichiers XML-TEI : l'accent est mis sur la précision des métadonnées, même si le corps du texte peut rester bruité.

2.1 Glaner des fichiers PDF pour constituer un ensemble documentaire numérique

La première étape consiste à collecter les documents par moissonnage des bibliothèques numériques publiques (*Gallica* et *Mazarinum* notamment) et de Google Livres. Au 7 juillet 2021, date de la rédaction de cette contribution, nous avons récupéré 2.145 PDF (soit environ 29.000 pages) retrouvés grâce à leurs titres et à la numérisation des bibliographies existantes (datant du XIX^{e} siècle), notamment la bibliographie de Célestin Moreau. Cette bibliographie et ses suppléments ont été convertis en tableur interrogeable pour permettre des premières statistiques (nombre de pages, lieu d'impression, etc.[5]). Les métadonnées peuvent ainsi être exploitées en tant que don-

[2] Les observables sont ici définis très simplement comme toute unité susceptible d'être extraite d'un texte (mots, mais aussi caractères, par exemple) pour en permettre l'analyse. Ils ne sauraient être tenus pour le perçu ou le perceptible, ni l'existant, puisqu'ils relèvent d'une construction, et donc de la position de l'observateur ainsi que de la question posée aux données (voir par exemple Culioli 1990 :17). Cette question, fondamentale en linguistique, l'est d'autant plus depuis la mise à disposition de vastes corpus par l'informatique (voir par exemple Bergounioux et Dal, 2016).

[3] Voir la critique de la « technologie sans la science » des exploitations de *big data* (Wieviorka 2006 ; voir aussi Gefen 2015 : 3). Sur la qualification des données en corpus dans les sciences de la culture par opposition à l'usage des *big data*, voir Rastier, 2018.

[4] Ces deux perspectives apparaissent dans un bilan des premières phases des humanités numériques dressé en 2015 par Y. Citton, qui rappelle que certains des « praticiens des humanités numériques revendiquent une pratique relevant d'un bidouillage (*hacking*) pleinement artisanal » mais aussi « impliquant un renouveau et un redéploiement des pratiques interprétatives – tout au contraire d'entraîner leur liquidation dans des flux de données gérées en traitement automatisé » (« Politiques de l'ouverture »).

[5] Voir la description des bibliographies de Moreau et de ses successeurs, ainsi que la mise à disposition de ces ressources structurées et interrogeables, ici : [http://memes.sorbonne-universite.fr/visualisation/Moreau/test.html]. Consulté le 7 juillet 2021.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

nées, les premières qui soient disponibles dans un ensemble documentaire d'une telle ampleur.

La glane de PDF est semi-automatisée grâce à un script fondé sur l'utilisation des API Google et Gallica qui nous permettent d'obtenir, à partir des titres tirés de la bibliographie de Célestin Moreau, des listes d'URL de documents dont le titre est jugé proche. Cette proximité est évaluée *via* une mesure de similarité sur des n-grammes de caractères, mesure qui nous est apparue la plus à même d'être robuste à la variation observée dans les segments comparés (voir Buscaldi *et al.* 2020). La variation dans les titres est due à la coprésence d'orthographes non modernisées dans les catalogues de bibliothèques, et modernisées irrégulièrement par les bibliographes du XIX^{e} siècle.

Enfin, cette collecte s'est accompagnée d'une quête d'homogénéité : des corpus aussi cohérents que possible ont été choisis à l'intérieur de la masse desdites mazarinades pour rendre parlantes les premières applications (un corpus de style burlesque, un autre composé de textes narratifs, un autre de lettres, et un dernier de poésies potentiellement chantées à l'époque). Cette sélection ne se fait pas sans difficultés, car les assignations génériques ne sont pas toujours transparentes ou univoques : le titre *Entretien de Fanchon, Toinon et Nichon, sur l'arrivée de leurs galants, pièce morale, par le sieur Baugion*[6], dont les trois premiers noms propres évoquent l'écriture burlesque ou comique, cache, par exemple, une histoire tragique. La constitution de corpus véritablement cohérents relève en fait d'une étape à part, postérieure, complexe, qui demande un affinage des catégories : nous nous contentons pour le moment de délimiter de grands ensembles (vers et chansons, lettres, récits – ces différents ensembles n'étant pas sans intersections). Ces grands ensembles sont, de notre point de vue en tout cas, utilisables par eux-mêmes en tant que regroupement raisonné de documents PDF ou de fichiers images, la question du mode texte étant traitée de façon séparée, comme nous le verrons à l'étape suivante.

2.2 Concevoir une chaîne de traitement pour obtenir des fichiers en mode texte : la phase d'océrisation

À partir des séries de documents PDF glanés, sont enchaînées les opérations de binarisation, de segmentation, puis de reconnaissance de caractères proprement dite. Cette chaîne de traitement utilise le logiciel Kraken et a été adaptée au projet par Jean-Baptiste Tanguy en intégrant un modèle entraîné sur des données du XVII^{e} siècle (Gabay, Clérice et Reul, 2020). Cette phase, parfois considérée comme relevant de l'ingénierie pure, est importante d'un point de vue épistémologique : parce qu'elle pose d'une part la question de la qualité des données résultantes (et de la mesure de celle-ci) et d'autre part celle de la liaison entre la qualité « observée » ou « ressentie » et l'utilisabilité des données – question qui nous semble trop rarement analysée. Ainsi que nous l'avons évoqué plus haut, il nous semble important de ne pas considérer que l'échec de l'application d'outils numériques (de textométrie, de Traitement Automa-

[6] *Entretien de Fanchon, Toinon et Nichon, sur l'arrivée de leurs galants, pièce morale, par le sieur Baugion*, s. l., 1650, 14 pages. Le choix d'un titre ambigu est concerté, l'auteur conclut ainsi le texte : « Le fiecle eft fi peruers & fi defpraué, que fi i'auois mis vne Intitulation Saincte à ces Remonftrances Chreftiennes, on m'auroit appellé Trouble feftes, Bigot, Mangeur de Chappelets, au lieu qu'en ayant mis vne Burlefque, elle fera debitée beaucoup mieux qu'elle n'eut efté, & i'efpere que par ce moyen le fruict en fera plus grand. »

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

tique des Langues ou encore de *Machine Learning* par exemple) soit systématiquement imputable à la qualité des données textuelles exploitées.

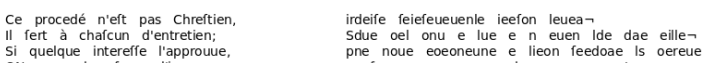


Fig. 1 : Résultats de reconnaissance optique de caractères

Dans la figure 1, on observe deux résultats d'océrisation de qualité très différente. À gauche, un texte pour lequel l'OCR affiche un taux de confiance moyen de 0,98 alors que sur la droite nous voyons un texte pour lequel ce taux tombe à 0,64. Si l'on considère souvent cette mesure comme une bonne manière de quantifier un écart de qualité, le taux de confiance peut être trompeur si on l'interprète comme un pourcentage de données correctes puisque dans le dernier cas le texte est tout simplement illisible. Le taux de lexicalité, pourcentage d'occurrences (*tokens*) figurant dans un lexique de référence, donne en l'espèce un résultat moins engageant (le taux peut être assez faible) mais plus fiable. Toutefois l'analyse textométrique, en dépit de cette hétérogénéité, ne s'est pas avérée sans fruit, comme nous le montrerons plus bas.

À la suite de la glane et de l'océrisation, nous avons mené en parallèle deux opérations d'esprit différent : enrichissement des données d'une part, et exploitation sans nettoyage d'autre part.

2.3 Transcription et passage au format XML-TEI

Nous avons réalisé la transcription d'un petit échantillon (105 textes) qui permet de ré-entraîner le modèle d'OCR (*fine tuning*) et d'avoir un échantillon de test pour mesurer le poids du bruitage de l'OCR dans l'utilisabilité informatique des données. On utilise bien sûr ces données nettoyées dans les fichiers XML.

Ce passage au format XML-TEI a aussi été réalisé pour les documents non-retranscrits. À ce jour, nous avons 447 fichiers XML relus dont 105 pour lesquels la transcription du texte a été corrigée manuellement. Cela représente 2.108.211 occurrences (*tokens*) et 199.734 mots (types, *i. e.* mots considérés comme séquences de lettres), obtenus de manière semi-automatique. Un script génère automatiquement l'en-tête (*header*) pré-rempli pour les métadonnées éditoriales (équipe responsable du travail) et en partie pour les métadonnées textuelles (titres, lieux d'édition et date, exportés automatiquement à partir des données bibliographiques léguées par la tradition, c'est-à-dire Moreau et ses successeurs). Sont ensuite corrigées et abondées ces premières métadonnées : on restitue les liens menant vers les fichiers PDF originels, les liens vers les référentiels en usage (Geonames[7] et ISNI[8]), les métadonnées concernant les lieux où l'imprimé est conservé, les métadonnées et identifiants de la nouvelle « Bibliographie des mazarinades » (conçue depuis 2019 par la Bibliothèque Mazarine[9]) et les annotations spécifiques (vers/prose, commentaire littéraire ou générique éventuel). L'ensemble est encadré par un ODD qui permet d'ajouter des contraintes spécifiques

[7] [https://www.geonames.org/], consulté le 7 juillet 2021.
[8] [https://isni.org/page/search-database/], consulté le 7 juillet 2021.
[9] [https://mazarinades.bibliotheque-mazarine.fr]. Consulté le 7 juillet 2021.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

au projet[10]. Un dépôt Github permet de stocker et de partager ces premières données préparées[11].

Le corps du texte n'est, la plupart du temps, pas entièrement « propre » : la figure 2 montre toutefois que le bruitage est souvent concentré sur la page de titre, qui présente des spécificités typographiques comme des casses (romain et italique) et des tailles de caractères différents, d'ailleurs particulièrement grandes pour l'époque dans les mazarinades (Carrier, 1991 : 202), ce qui pourrait affecter la reconnaissance des caractères, même avec un modèle entraîné sur des données contemporaines de ce corpus. À part ces zones critiques, d'ailleurs facilement identifiables, le corps du texte en lui-même est relativement peu affecté par l'automatisation du processus d'obtention du texte.

La structuration du texte est pour le moment minimale (les paragraphes, strophes et vers sont balisés, ainsi que les retours à la ligne), obtenue semi-automatiquement, avec des vérifications, nécessaires en particulier pour la numérotation des pages. Une structuration par séquences serait d'un grand intérêt, pour des textes versifiés où la polymétrie n'est pas rare, et où des couplets de chansons, des épigrammes et autres formes brèves se mêlent parfois au corps du texte. Ce serait un moyen, certes modeste, de contribuer à une structuration séquentielle des textes qui fait partie des modalités de « lisibilité augmentée » permises par le numérique et proposées par Noille (2020 : 52).

L'intérêt de ce passage en TEI se situe surtout au plan des métadonnées, lesquelles sont un point d'intérêt décisif pour ce corpus. En effet, les libelles sont des écrits qui contournaient souvent le système de la censure *a priori,* en vigueur sous l'Ancien Régime : privilèges ou permissions d'imprimer et de vendre sont peu fréquents, et l'anonymat est presque la règle (83 %[12]). Aussi la recherche d'indices pour retrouver les coordonnées éditoriales est-elle un travail très important, puisqu'en dépit de ces lacunes, ces textes demandent à être situés pour être compris. Sur ce point, on utilise en priorité les métadonnées des catalogues des bibliothèques, et surtout de la Base Bibliographique des Mazarinades de la Bibliothèque Mazarine, évoquée ci-dessus : ce nouvel outil en ligne refond, depuis 2019, toutes les connaissances accumulées depuis deux siècles sur ces textes. Ci-dessous en figure 3, l'exemple de l'en-tête d'un texte intitulé *Plaintes du Carnaval et de la foire Saint-Germain, en vers burlesques*, montre que les dernières recherches bibliographiques (en l'occurrence celles de H. Carrier) sont intégrées dans la balise décrivant la source. Cela permet de renseigner l'auteur très vraisemblable du texte (Laurent de Laffemas), lequel avait été ignoré par la tradition du XIX^{e} siècle, à partir de toutes les recherches ultérieures sur le sujet. De même, la date précise a pu être ajoutée.

[10] [https://github.com/Antonomaz/ODD]. ODD préparé par Alexandre Bartz, ingénieur du projet en 2021.

[11] [https://github.com/Antonomaz/Corpus], consulté le 20 août 2021.

[12] Selon l'estimation de Carrier (1991 : 77-79).

[13] https://github.com/Antonomaz/Corpus/blob/main/Mazarinades/2701-2800/Moreau2794_GALL.XML, consulté le 20 août 2021.

[14] https://github.com/Antonomaz/Corpus/blob/main/Mazarinades/2701-2800/Moreau2794_GALL.XML, consulté le 20 août 2021.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

```
<p>
   <pb n="2" facs="https://gallica.bnf.fr/ark:/12148/bpt6k56119160/f3.item.r=bpt6k56119160"/>
   <ref target="#GALL_bpt6k56119160_2.png">[Bandeau]</ref>
   <title type="sub">4.,. .¬. ------¬--¬¬aaprr „,,,,a, --a,,,,,,,,,,aa¬ ,, ,a -a-,,,,,---
      ,aa S 2 gS5 S A Sle . ue SQaS&amp;amp;QaS&amp;amp;oS&amp;amp; 'SQe8 dS&amp;amp;ee8 ! TE 7 PLAINTE DV
      CARNAVAL, D 1 1 1 1 ; T T T EI E D ET DE LA FOIRE S. GERMAIN EN VERS BVRLESQVES. E T
      E D II EIEc</title>
   <lg>
      <l>
         <lb/>
         <ref target="#GALL_bpt6k56119160_3.png">[Lettrine]</ref>Iença ma petite
         belote,</l>
      <l>
         <lb/> Approche toy mufe falote,</l>
      <l>
         <lb/> Chere Maiftreffe de Ssaron,</l>
      <l>
         <lb/>Qui n'aimes pas l'air fanfaron,</l>
      <l>
         <lb/>Dont fe chantent les funerailles</l>
      <l>
         <lb/>Des Heros morts dans les batailles;</l>
      <l>
         <lb/>Et quoy que nous nous preparions</l>
      <l>
         <lb/>A donner quelques horions</l>
      <l>
         <lb/>Sur les oreilles Polonoifes,</l>
      <l>
         <lb/>Alemandes, Baſques, Françoifes,</l>
      <l>
         <lb/>Et celles generalement</l>
      <l>
         <lb/>Que condamne le Parlement;</l>
```

FIG. 2 : Corps du texte, structuré *a minima*, sans post-correction manuelle des caractères[13]

```
<sourceDesc>
   <bibl>
      <ref target="https://gallica.bnf.fr/ark:/12148/bpt6k56119160"/>
      <author>
         <persName role="auteur_présumé" source="Carrier">
            <forename>L'Abbé</forename>
            <surname>Laffemas, De</surname>
         </persName>
      </author>
      <title source="Moreau">Plaintes du Carnaval et de la foire Saint-Germain, en vers
         burlesques.</title>
      <pubPlace ref="geonames:2988507">Paris</pubPlace>
      <publisher ref="isni:0000000362335140">
         <persName>
            <forename>Claude</forename>
            <surname>Huot</surname>
         </persName>
      </publisher>
      <date when="1649-02" source="Carrier">mi-février 1649</date>
      <extent>
         <measure unit="page" quantity="8"/>
      </extent>
      <note type="format">in-4</note>
      <note type="Moreau_identifier">2794</note>
      <note type="BM_identifier">BM00386</note>
      <ref type="BM_notice"
         target="https://mazarinades.bibliotheque-mazarine.fr/ark:/61562/bm49245"
   </bibl>
   <msDesc>
      <msIdentifier>
         <settlement>Paris</settlement>
         <institution>Bibliothèque nationale de France</institution>
         <repository>Département Littérature et art</repository>
         <idno type="cote">YE-3945</idno>
```

FIG. 3 : Encodage de l'en-tête en vue de l'interopérabilité des documents[14]

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

Plusieurs chantiers de recherche devraient être possibles grâce à cette structuration des métadonnées. Nous essayons, par exemple, d'encoder les ornements présents dans ces imprimés, en espérant identifier ainsi le matériel d'imprimerie utilisé par certains imprimeurs non déclarés (il n'était pas sans risque de se désigner comme l'imprimeur de pièces parfois virulentes). L'idée est de pouvoir faire intervenir les résultats du projet dans un écosystème numérique : en liant de manière pérenne et interopérable les données avec les référentiels en vigueur, et plus spécifiquement avec les bibliothèques numériques, mais aussi, idéalement, en abondant ces ressources catalographiques et bibliographiques à notre tour, si nous arrivions à identifier, par exemple, des officines d'imprimerie.

Mais le coût de ces chantiers d'encodage est important et dépend aussi de l'avancée des numérisations et de la mise à jour des métadonnées des bibliothèques publiques. Ainsi, un dixième seulement du corpus est converti à jour en XML, dont seulement une partie a fait l'objet d'une retranscription fidèle au document d'origine.

3 Expériences d'exploitation de données bruitées

Pour tester l'exploitabilité de ces données, nous avons combiné une double réflexion, d'une part sur les formats nécessaires au requêtage et à l'exposition des résultats et, d'autre part, sur la nature des observables qu'il est utile de confier à la machine. L'idée est ici qu'on peut espérer indexer les données telles qu'elles sont (c'est-à-dire éventuellement non préparées), tout en proposant une sortie qui préserve l'intégrité de la lecture, non pas en mode texte (*a fortiori* s'il est bruité), mais en fac-similé numérique (image du texte en formats PDF, PNG ou TIFF). Noille (2020) rappelle que le mode texte n'est pas une forme de textualité cognitivement viable pour une vraie lecture (p. 48), opposant ainsi « human reading » (p. 51) à « distant reading ». En prenant acte de cette nécessaire distinction, nous essayons de proposer des résultats qui permettent la lecture « humaine », tout en menant une exploration automatique de données : l'intuition est que ce que cherche le lecteur est avant tout un document, et que le mot, la séquence ou le caractère, plus largement tout observable utilisé par la machine, n'est qu'un truchement non pertinent pour l'œil humain.

3.1 Quels observables pour l'indexation des données ?

De plus en plus de travaux s'intéressent à la robustesse des systèmes automatiques, c'est-à-dire à leur capacité à travailler sur des données imparfaites, non normalisées. Ceci intéresse des tâches de différents niveaux, y compris des tâches *a priori* très sensibles à la variation, telles que la reconnaissance d'entités nommées (Ortiz *et al.* 2020). Concernant notre corpus, notre ambition est pour le moment plus mesurée : il s'agit simplement de repérer des similarités entre des portions de documents (phrases, paragraphes, pages…). Le cadre expérimental dans lequel nous nous sommes placés est le suivant : dans quelle mesure le fait de naviguer dans la version bruitée ou dans la version « propre » du corpus affecte les parcours de lecture que l'on pourrait proposer *via* des mesures de similarité classique (distance cosinus sur des vecteurs de mots ou de chaînes de caractères). Afin de répondre à cette question, nous avons exploité la sous-partie de notre corpus qui a été retranscrite manuellement à ce stade du projet :

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

105 documents représentant 1.107 pages au total, que nous appellerons sous-corpus « propre ». Nous avons constitué un autre sous-corpus avec la version océrisée non corrigée de ces différentes pages (sous-corpus bruité). Nous avons regardé si nous pouvions apparier la version « propre » d'une page donnée avec sa version « bruitée ». Le résultat s'est révélé prometteur, puisque dans 97% des cas, nous pouvions apparier ces deux versions, c'est-à-dire que la page la plus similaire à la version « propre » est la version « bruitée » de cette page. Les erreurs proviennent principalement de pages contenant peu de texte, en particulier les pages de titre, dotées d'éléments graphiques que l'on sait problématiques pour les modèles d'OCR. Le reste est constitué de pages dont l'océrisation est si mauvaise qu'elle empêche le déchiffrage humain. Ainsi, sous réserve d'un bruitage raisonnable, il semble que les mesures de similarité ne soient pas particulièrement affectées par le bruitage. Deux conclusions peuvent être tirées à ce stade. D'une part, la contrainte principale est de choisir de vectoriser les textes à l'aide de descripteurs adaptés, en l'occurrence les n-grammes de caractères, puisque ce sont des descripteurs plus robustes que les mots en contexte bruité[15]. D'autre part, cette expérience porte à penser que des données non « préparées » peuvent néanmoins permettre diverses opérations automatiques classiques tels que la fouille de texte, l'utilisation de moteurs de recherche et le *clustering*.

3.2 Un exemple de classification non supervisée (*clustering*)

Cette expérience vise à sélectionner des textes appariés dans un corpus assez conséquent, sans formulation de requête explicite. Il s'est agi de repérer des points communs (sans que l'humain ne préjuge par quels observables) à partir d'un fac-similé entier : l'imprimé *L'Adieu de Mazarin, Burlesque*[16]. Le savoir bibliographique (que l'on a pris soin d'encoder, comme vu plus haut, pour permettre la justification des rapprochements observés) nous apprend qu'il s'agit d'une contrefaçon (sans doute rouennaise) du *Passe-port et l'adieu de Mazarin, en vers burlesques*, publiée en 1649, et écrite, selon toute vraisemblance (quoique non signée), par Paul Scarron[17]. Ce *Passe-port* peut être daté de début mars 1649 : il est assurément postérieur à l'arrêt du 8 janvier prononçant le bannissement de Mazarin, « perturbateur du repos public », hors du royaume (c'est à cet éloignement que renvoie le « passeport » et « l'adieu » donnés au ministre), et à « un blocus de deux mois » mentionné dans le texte. La contrefaçon est probablement datée des semaines qui suivent, comme le confirment les rapprochements ci-dessous.

Le calcul de similarité rapproche ce document d'une dizaine de pièces du corpus disponible des 179 textes en vers burlesques que nous avons océrisés[18]. On peut proposer quelques pistes interprétatives de ces similarités. D'abord le motif de l'« adieu », ce mot, anaphorisé 39 fois dans la première page de texte du document source (voir

[15] En effet, si un mot est mal reconnu par le logiciel de reconnaissance optique de caractères, « Mazorin » au lieu de « Mazarin » par exemple, les n-grammes de caractères qui le composent permettent malgré tout des rapprochements (les deux formes partagent les tri-grammes 'Maz' et 'rin').

[16] [https://gallica.bnf.fr/ark:/12148/bpt6k55764575.r=bpt6k55764575?rk=21459;2.] Pour les métadonnées récemment établies :
[https://mazarinades.bibliotheque-mazarine.fr/cgi-bin/koha/opac-detail.pl?biblionumber=49715].
Pages consultées le 20 août 2021.

[17] Voir Carrier, 2012 : 33-44.

[18] 179 textes ; voir le corpus de mazarinades burlesques déposé ici : [https://github.com/Antonomaz/Corpus_burlesque].

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

Figure 4), génère naturellement des rapprochements de type lexico-stylistiques. Le résultat du calcul de similarité rapproche ce texte de *Le Congé burlesque de l'armée normande*[19]. On repère ainsi un intertexte, qui indique la circulation provinciale du texte de Scarron, puisque ce libelle est paru, à Rouen, sans doute début avril 1649, soit, après la paix de Saint-Germain qui mit aussi fin à la révolte normande et après la contrefaçon rouennaise qu'est l'*Adieu de Mazarin, Burlesque.* En recyclant le motif de l'adieu pour donner congé à la troupe ridicule des frondeurs normands, le texte inverse l'orientation argumentative de son modèle.

Les calculs rapprochent également du document source *La Juliade ou Discours de l'Europe à Monseigneur le duc d'Orléans sur l'esloignement du cardinal Mazarin* […]. On y retrouve une nouvelle fois l'anaphore inspirée de *L'Adieu* de 1649, très structurante pour ce texte : vingt occurrences anaphoriques de « grand » (p. 3), quatorze de « coup » (p. 6), trente de « fourbe » (p. 13-14[20]). *Grand* qualifie ainsi le Duc d'Orléans, réputé avoir chassé Mazarin, ainsi l'auteur d'un remarquable « coup » politique contre le « fourbe » (voir un extrait Figure 5). Le motif satirique et polémique anti-Mazarin est ainsi réinvesti dans une perspective élogieuse, voire encomiastique.

Ces rapprochements permettent donc de constituer très facilement un petit sous-corpus de textes d'inspiration scarronienne – comme beaucoup de vers burlesques de la Fronde, sur lesquels Scarron a imprimé une marque décisive –, puis de chercher à expliquer la circulation de tel imprimé dans un espace de publication, ou de tel motif stylistique. La plupart de ces rapprochements repose manifestement sur le repérage automatique du mot *adieu,* embrayeur d'anaphore. Aussi ce genre de série se retrouverait-elle sans doute facilement avec une table de concordance classique, puisqu'elle repose sur des reprises de mots.

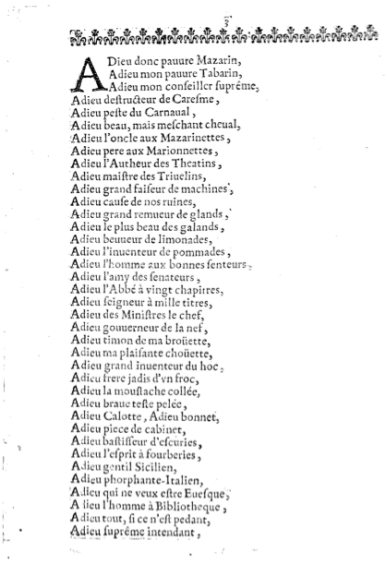
Mais d'autres résultats permettent de regrouper des motifs polémiques non visibles à l'aide d'une recherche textométrique ciblée ; les observables ayant présidé au rapprochement des textes sont d'ailleurs plus opaques à l'œil humain. Si l'on reprend le texte source, on remarque qu'il condense les griefs contre Mazarin, grâce à l'effet-liste procuré par l'anaphore de « adieu » (voir Figure 4) : celui est décrit comme fourbe, « perturbateur du repos public », mais aussi comme introducteur en France de toutes sortes de luxes dispendieux (« Adieu l'inventeur de pommades,/ Adieu l'homme aux bonnes senteurs »), joueur invétéré (« Adieu grand inventeur du hoc », le hoc étant un jeu de cartes), « bastisseur d'écuries » et surtout possesseur d'une bibliothèque richement dotée (la bibliothèque Mazarine), emblème de la « suspecte démesure » d'un ministre réputé ignorant (Sordet 2019 : 466) : « Adieu l'homme à Bibliotheque » (*L'Adieu de Mazarin*, p. 3 ; voir Figure 4). On retrouve ces arguments par la reprise des mêmes anaphores, dans le *Songe burlesque de Polichinel sur le Despart de Jules Mazarin* (« Adieu belle Bibliothèque », p. 7) mais aussi dans une pièce en prose qui ne

[19] *Le Congé burlesque de l'armée* normande, jouxte la copie Imprimée à Roüen, 1649. Il s'agit de la copie d'une édition rouennaise, sans doute de David Ferrand, auteur fameux de la *Muse normande* (voir [https://mazarinum.bibliotheque-mazarine.fr/records/item/14624-redirection], consultée le 20 août 2021). Pour les métadonnées rénovées de la bibliothèque Mazarine : [https://mazarinades.bibliotheque-mazarine.fr/cgi-bin/koha/opac-detail.pl?biblionumber=49031], consultée le 20 août 2021.

[20] À rapprocher de l'anaphore de « bougre » dans la célèbre *Mazarinade* de Scarron de 1651 : « Bougre bougrant, bougre bougré,/Et bougre au suprême degré,/Bougre au poil, et bougre à la plume,/Bougre Sodomisant l'Etat […] » (Scarron, 2021, p. 469).

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

se rapproche pas formellement du texte source : *La Lettre de Remerciment envoyée au cardinal mazarin… avec la harangue de Dame Denise au large Chaperon des Halles*[21]. Cette pièce en langue poissarde reproduit les charges à l'encontre de Mazarin à travers des cuirs, souvent associés à la parlure dite populaire : « vous avez laissé une belle Barbuliteque » (p. 11) […] avec « tant de biaux et gros vomules » (p. 15). Le rapprochement thématique est net, et les déformations phonétiques et graphiques plaisamment associées au langage des harangères (produisant, d'une certaine manière, du bruit, puisqu'elles introduisent de la variation) n'ont pas entravé le calcul de similarité.



ADieu donc pauure Mazarin,
Adieu mon pauure Tabarin,
Adieu mon conſeiller ſuprême,
Adieu deſtructeur de Careſme,
Adieu peſte du Carnaual,
Adieu beau, mais meſchant cheual,
Adieu l'oncle aux Mazarinettes,
Adieu pere aux Marionnettes,
Adieu l'Autheur des Theatins,
Adieu maiſtre des Triuelins,
Adieu grand faiſeur de machines,
Adieu cauſe de nos ruines,
Adieu grand remueur de glands,
Adieu le plus beau des galands,
Adieu beuueur de limonades,
Adieu l'inuenteur de pommades,
Adieu l'homme aux bonnes fenteurs,
Adieu l'amy des fenateurs,
Adieu l'Abbé à vingt chapitres,
Adieu ſeigneur à mille titres,
Adieu des Miniſtres le chef,
Adieu gouuerneur de la nef,
Adieu timon de ma broüette,
Adieu ma plaiſante choüette,
Adieu grand inuenteur du hoc,
Adieu frere jadis d'vn froc,
Adieu la mouſtache collée,
Adieu braue teſte pelée,
Adieu Calotte, Adieu bonnet,
Adieu piece de cabinet,
Adieu baſtiſſeur d'eſcuries,
Adieu l'eſprit à fourberies,
Adieu gentil Sicilien,
Adieu phorphante-Italien,
Adieu qui ne veux eſtre Eueſque,
Adieu l'homme à Bibliotheque,
Adieu tout, ſi ce n'eſt pedant,
Adieu ſuprême intendant,

Fig. 4 : *L'Adieu de Mazarin, en vers burlesques*[22]

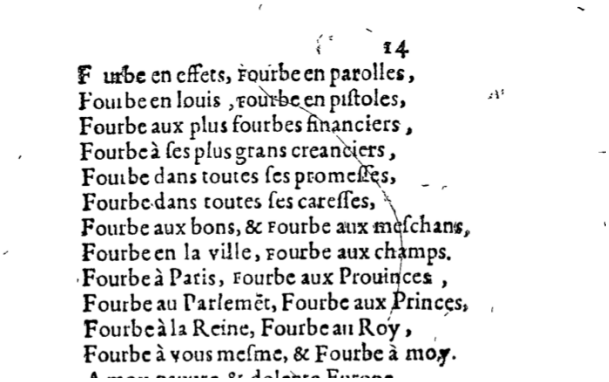
Les opérations de classification non supervisée sont sans doute pertinentes à proportion de la préparation du corpus de travail. Aussi, d'une certaine manière, on a préparé ce qu'on allait trouver : ici, le corpus d'entrée étant les pièces titrées « en vers burlesques », on a construit des sous-corpus de pièces versifiées avec une forte affiliation scarronnienne et des pièces à charge contre Mazarin brassant les mêmes arguments. Enfin, il convient de souligner que le dernier texte envisagé, pour lequel les chaînes de caractères mises en relation ne sont pas immédiatement sémantisables, induit une certaine confiance à accorder aux rapprochements opérés par l'algorithme qui peuvent donner un sentiment de « boîte noire ». La lecture humaine peut rendre

[21] *Lettre de remerciment envoyee au cardinal Mazarin sur la Lettre qu'il a escrite à une Dame de la Cour pour l'accommodement de ses affaires, avec la harangue de dame Denise au large chaperon des Halles, – deputée vers son Everminence* [sic] *par la Communauté de toutes les Beurrieres, de la Ville, Cité & Université de Paris, touchant les bien faits qu'elle a receus d'elle pendant tout le Caresme dernier. Ridendo dicere verum, quis vetat ?,* Paris, 1651.

[22] *Sur l'imprimé à Paris*, chez Claude Huot, ruë Saint Jacques, 1649, p. 3.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

raison du rapprochement, mais non des observables sur lesquels il est fondé (laissant ouverte l'hypothèse d'un rapprochement difficile à justifier du point de vue du sens).


14

Fourbe en effets, Fourbe en parolles,
Fourbe en louis, Fourbe en piſtoles,
Fourbe aux plus fourbes financiers,
Fourbe à ſes plus grans creanciers,
Fourbe dans toutes ſes promeſſes,
Fourbe dans toutes ſes careſſes,
Fourbe aux bons, & Fourbe aux meſchans,
Fourbe en la ville, Fourbe aux champs.
Fourbe à Paris, Fourbe aux Prouinces,
Fourbe au Parlemẽt, Fourbe aux Princes,
Fourbe à la Reine, Fourbe au Roy,
Fourbe à vous meſme, & Fourbe à moy.

Fig. 5 : Extrait de *La Juliade*[23]

4 Conclusion

La publication numérique de plusieurs corpus taillés dans la masse des mazarinades cherche à offrir un accès rénové à des documents historiques et littéraires, à penser cet accès en collaboration avec les institutions patrimoniales, et à proposer des moyens de fouille aux communautés intéressées. Mais, pour reprendre Noille (2020), qui en appelle à une réflexion sur les usages de ce patrimoine numérique, il convient de proposer des « mise(s) en scène des données disponibles » (p. 51) et « des scénarios de consultation intuitifs et productifs » (p. 53), notamment en laissant le mode texte à l'arrière-plan (peu importe alors qu'il soit bruité) pour insister sur la mise en relation dynamique des *fac-simile* numériques. Notons finalement que ces rapprochements entre documents peuvent être assistés, mais non conclus par la machine : sans doute la recherche automatique de similarités permet-elle de suggérer des pistes d'exploration des données, lesquelles sont toujours à affiner manuellement par la connaissance des corpus et des requêtes qu'on cherche à y construire.

[23] Titre complet : *La Juliade ou Discours de l'Europe à Monseigneur le duc d'Orléans sur l'esloignement du cardinal Mazarin & le retour des Princes* (Paris, 1651). Le fragment présenté se situe p. 14.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352