Vers dans les Pamphlets de la Fronde : Analyse Numérique des versifications (1648-1653)

Yuyan QIAN

Table des matières

Introduction

Pour commencer à travailler sur ce corpus, il faut d'abord identifier quelques termes importants qui me permets également de comprendre les tâches à réaliser. En même temps, plusieurs méthodes de lecture sont testées sur ce corpus afin d'avoir une approche relativement simple et efficace. Après avoir un aperçu sur les caractéristiques des données, j'ai commencé à mettre en pratique l'extraction du sous-corpus sur un petit jeu de donnée (test). L'étape suivante servit à phonétiser des vers et à compter les syllabes du vers. Pour ce faire, j'ai combiné deux recettes : phonétiser des vers avec le dictionnaire de référence et l'algorithme de Lenzo. Enfin, j'ai analysé les résultats d'un échantillon (1-100) pour vérifier la précision de phonétisation et les types de vers. Si possible, cette analyse pourrait être généralisée à tout le corpus pour une étude plus approfondie.

Terminologie de versifications

Dans l'élaboration initiale de cette étude, j'ai adopté des définitions élémentaires des termes utilisés afin de pouvoir généraliser la théorie sous-jacente.

En phonologie, une syllabe est comprise comme une séquence de phonèmes qui constitue une unité fondamentale de prononciation. Chaque syllabe est structurée autour d'un élément central, généralement une voyelle, connu sous le nom de noyau syllabique.

Concernant la distinction entre vers et prose, il est important de noter que la poésie est soumise à des contraintes spécifiques de forme et de rythme, qui diffèrent de l'écriture en prose.

Le vers, quant à lui, représente l'unité de base de la poésie. Il s'agit d'une séquence de mots disposée en ligne, qui se termine par un retour à la ligne. La mesure d'un vers est définie par le nombre de syllabes qu'il contient. Pour évaluer correctement la longueur d'un vers, il est donc essentiel de compter les syllabes telles qu'elles sont effectivement prononcées.

1 Lecture et découverte du corpus

Pour me familiariser avec les fichiers du corpus, j'ai analysé les deux différents formats de fichiers : XML et JSON. Deux méthodes sont testées sur XML : lire comme un texte brut (*read()*) et parser l'arbre de XML (*beautifulSoup*). La première méthode est simple, mais elle a inclus trop d'information annotée. Les annotations en français moderne pourraient perturber la phonétisation des vers. Quand même cette fonction de lecture servit à donner un aperçu sur tout le corpus et cela a réussi à extraire les titres des fichiers contenant les vers. La deuxième méthode de parser prend trop de temps bien qu'elle puisse lire tous les méta-données. Lire le format JSON, c'est une intermédiaire entre les deux méthodes précédentes de lecture de XML. En autre, j'ai remarqué qu'un fichier *Moreau3860* a été supprimé du dossier *xml*.

Enfin, une découverte est que dans un seul fichier, il existe des vers et des proses en même temps. Cela sera une difficulté en séparation des vers et des proses.

1.1 Observation de la proportion de documents en vers/prose selon les années

Une partie de ce projet examine la proportion de documents en vers et en prose à travers différentes années, en s'appuyant sur les métadonnées associées à chaque fichier, notamment l'année de publication. Une fonction spécifique a été développée pour quantifier le nombre de documents en vers et en prose par an. En analysant la répartition annuelle de ces documents, on constate une prédominance notable des documents en prose par rapport à ceux en vers, ce qui reflète les tendances générales observées dans le corpus. Il est estimé que les textes en vers représentent environ 34% du total, avec une majorité d'octosyllabes. La figure générée par cette fonction montre une répartition des

documents moins de 25% en vers pour chaque année parce que l'on a pris en compte seulement les documents annotés seulement en vers. Il est important de souligner que, même parmi les documents classifiés comme prose, un certain nombre contiennent des passages en vers. Cette nuance est cruciale pour une compréhension complète de la composition du corpus et souligne la diversité textuelle au sein de ce dernier.

Plus précisément, au sein du corpus des Mazarinades, on observe une diminution progressive de la proportion de documents en vers entre 1648 et 1656. Cette analyse inclut également les documents pour lesquels l'année de publication reste inconnue.

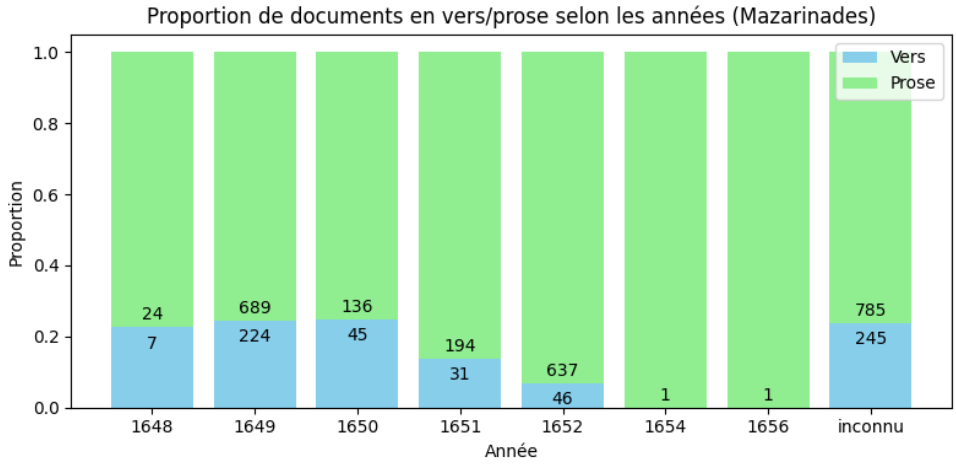

Figure 1 – Proportion de documents en vers/prose selon les années

2 Extraction du sous-corpus

Pour simplifier le travail, j'ai choisi de travailler sur un sous-corpus (1-100). Les fichiers annotés seulement en vers sont extraits dans un dossier *sous_corpus_test.json* avec leurs chemins relatifs. En même temps, j'ai testé sur un seul fichier toutes les fonctions définies pour analyser les types de vers. Après avoir vérifié que les fonctions sont correctes, j'ai appliqué les fonctions sur tous les fichiers du sous-corpus (20 fichiers en vers).

3 Phonétisation : Déjà fait... mais pas satisfaisant

La méthodologie pour analyser la versification repose essentiellement sur la phonétisation des vers. Le dictionnaire GLÀFF a déjà accompli une partie de ce travail en fournissant la phonétisation pour un grand nombre de mots. Toutefois, certains termes ne sont pas couverts par GLÀFF, que ce soit en raison de l'absence de prononciation dans le dictionnaire ou à cause de la spécificité du lexique de l'ancien français. Face à cette lacune, il est nécessaire de trouver une méthode complémentaire capable de phonétiser les mots non reconnus par GLÀFF. Voici les outils que j'ai testés, mais qui n'ont pas donné de résultats satisfaisants :

— Outil en ligne *Syllaber* : Bien que les résultats soient assez convaincants, son utilisation en ligne n'est pas adaptée au traitement d'un grand corpus de données.résultat assez satisfaisant, mais en ligne et pas pratique pour un grand corpus

— Script avec *Espeak* : Les résultats obtenus ne sont pas suffisamment précis pour nos besoins (trop de mots non reconnus)

— Comptage des lettres de voyelles : Cette méthode a engendré de nombreuses erreurs et manque de flexibilité pour traiter les mots inconnus. Les règles sont trop nombreuses à définir.

— *Plint*[1] : L'adaptation des codes de cet outil, qui vérifie la validité d'un poème en fonction des contraintes de métrique, de rime et de type de rime, a malheureusement échoué.

Il est donc impératif de poursuivre la recherche d'un outil ou d'une méthode plus efficace qui pourrait compléter GLÀFF dans la phonétisation des mots en ancien français pour une analyse de versification précise et fiable.

4 Phonétisation avec l'Algorithme de Lenzo (Text-to-Phone)

Après avoir exploré diverses méthodes de phonétisation, j'ai opté pour l'algorithme de Lenzo en tant que solution privilégiée pour la phonétisation de tokens en ancien français non reconnus par GLÀFF. Dans la première ébauche du rapport, j'avais déterminé plusieurs objectifs clés, notamment affiner la précision de la phonétisation, envisager l'adoption ou la création d'un phonétiseur adapté à l'ancien français, et finalement, intégrer des modèles d'apprentissage automatique pour la classification. L'intégration de l'algorithme de Lenzo adresse ces préoccupations en exploitant l'apprentissage non paramétrique, notamment l'utilisation d'arbres de décision, pour la phonétisation des mots en ancien français.

4.1 Principe de l'algorithme de Lenzo

Le package Perl nommé *Lenzo (t2p)*[2] offre une méthode pour établir des règles de correspondance graphème-phonème basées sur des dictionnaires de prononciation. Cette approche génère des règles de transformation de lettres en sons pour la prononciation des mots, en se fondant sur un ensemble d'exemples de prononciation.

J'ai sélectionné *t2p* pour la phonétisation des textes en ancien français, car il offre une flexibilité et une puissance supérieures à celles des algorithmes basés sur le décompte de règles prédéfinies. L'avantage principal de *t2p* réside dans sa capacité à généraliser au-delà de son ensemble d'apprentissage initial, permettant la phonétisation de mots inédits pour le programme. Cette fonctionnalité est particulièrement pertinente pour traiter des mots

1. Voir http ://plint.a3nm.net/en/about
2. Voir http ://www.cs.cmu.edu/afs/cs.cmu.edu/user/lenzo/html/areas/t2p/

en ancien français qui ne figurent pas dans le vocabulaire connu, un aspect crucial pour ce projet.

4.2 Préparation du dictionnaire de phonèmes

t2p s'appuie sur un dictionnaire de prononciation, comme le CMU Pronouncing Dictionary, pour construire des arbres de décision modélisant les mots. Pour l'adapter au français, j'ai constitué un dictionnaire de prononciation spécifique au français, s'inspirant également de GLÀFF. Étant donné l'ampleur du dictionnaire GLÀFF, j'ai décidé d'exclure les formes les moins courantes et les plus fréquentes dans le corpus Frantext du 20e siècle[3], afin de maintenir une taille gérable pour le dictionnaire de prononciation et d'optimiser la précision des règles de phonétisation. L'utilisation intégrale des formes de GLÀFF aurait abouti à un dictionnaire excessivement volumineux pour *t2p*. Par ailleurs, les formes extrêmement fréquentes dans *Frantext 20e* n'apportaient pas de précision supplémentaire aux règles de prononciation. Suite à cette sélection, le dictionnaire a été réduit à 5818 formes.

Les prononciations, qu'elles soient issues du dictionnaire de prononciation ou générées par l'algorithme, sont représentées en phonèmes *SAMPA*. Pour structurer et aligner le dictionnaire, j'ai développé un script nommé 2.0_creerLex0.ipynb, qui transforme les formes et génère un fichier texte (Lenzo/lex0.txt). Cette étape d'alignement a permis de valider et de conserver 4440 mots dans le dictionnaire de Lenzo.

4.3 Intégration de l'algorithme de Lenzo dans le pipeline de phonétisation

```
11    ILluſtre    Senateur, Hero  Icomparable,
      [[i,lystR], s@nat_9R, _@Ro, ikO~_paRabl_]

11    Qu'on       ne   m'aceuſe     pas  te    croyant      Adorable
      [k____jO~, [n@], m___as_2z_, [pA], [t@], [kRwa,jA~], [a,dO,Rabl]]

11    Si     i'eſeue    dta   gloire  vn   ſi    fameux   Autel
      [[si], i___@z___, dta, [glwaR], v_, [si], [fa,m2], [o,tEl]]

12    Quel    honnbur   dans   ce   lieux   ne   doit-on   point   te    rendre
      [[kEl], _On_byR, [dA~], [s@], [lj2], [n@], dw_a_jO~, [pwE~], [t@], [RA~dR]]

12    Dans    les   ſiecles   paſſez  vit-on   iamais  mortel,
      [[dA~], [le], [sjEkl], [pa,se], vi__jO~, jam_E_, [mOR,tEl]]'
```

Figure 2 – Exemple de phonétisation d'un vers avec GLÀFF et Lenzo

Dans cet exemple, les deux méthodes de phonétisation, GLÀFF et Lenzo, sont combinées pour obtenir une phonétisation du vers. Les phonèmes en bleu foncé entre crochets

3. *Frantext 20e* est un corpus de 30 millions de mots composé de romans contemporains (20e siècle) extraits de la base de données Frantext

sont ceux qu'on trouve dans la GLÀFF, tandis que les phonèmes en rouge sont ceux générés par t2p. Il y a quand même des erreurs dans les transcriptions de lenzo, par exemple, le mot *vit-on* est transcrit en [vijO] au lieu de [vitO]. Dans la partie des voyelles, Lenzo fonctionne d'une manière satisfaisante.

5 Analyse des résultats d'un échantillon (1-100)

Pour illustrer des versifications, j'ai choisi un échantillon (1-100) pour analyser les résultats de phonétisation. Les résultats sont présentés dans les figures suivantes. Elle montre clairement la distribution de la moyenne de syllabes et du nombre de vers dans un sous-ensemble des Mazarinades. Avec des barres indiquant la moyenne de syllabes par vers et une ligne pour le nombre de vers par fichier, on observe une variation de la longueur des vers, suggérant une prédominance de vers de taille moyenne, probablement des octosyllabes, conformément aux estimations des critiques littéraires. Cependant, la présence d'alexandrins semble moins fréquente dans cet échantillon, car il n'y a pas de pics significatifs correspondant à une moyenne proche de 12 syllabes. (environ 5% des vers sont des alexandrins dans cet échantillon).

La taille moyenne des vers se situe principalement entre sept et huit syllabes, et la distribution est relativement uniforme. Cette observation souligne deux lacunes dans les études antérieures : les limitations inhérentes au corpus et les imperfections de l'algorithme de phonétisation de Lenzo. Dans le corpus, une confusion fréquente entre les lettres 'v' et 'u' a été notée. De plus, les titres et les parties colophoniques pourraient influencer négativement la phonétisation des vers, tandis que l'inclusion de chiffres latins pour la pagination interfère avec l'analyse métrique.

La phonétisation selon Lenzo n'est pas exempte d'erreurs, notamment avec des mots non reconnus par le système. Une solution envisageable est l'enrichissement du dictionnaire de prononciation pour construire un arbre de décision plus robuste. L'ajout de mots fonctionnels tels que *le, la, les, ces, ce, des, etc.*, a conduit à une nette amélioration de la détection du schwa post-consonantique /l/. Cela démontre que le développement de l'arbre de décision est susceptible d'affiner le décompte syllabique.

Le nombre de vers par fichier varie considérablement, ce qui peut refléter des différences dans la longueur des textes. Pour établir avec précision la proportion d'alexandrins ou la présence de poésies à strophes, il serait nécessaire d'effectuer une analyse plus approfondie, en examinant les motifs de syllabes et en utilisant des informations typographiques qui pourraient révéler des structures strophiques. Cette figure constitue donc un point de départ pour une enquête plus détaillée sur la nature poétique des Mazarinades.

Une grande lacune dans mon analyse est l'absence de strophe dans les fichiers. En effet, la moyenne de syllabes par vers est une mesure globale qui ne permet pas de distinguer les strophes. Si l'on lit les vers par strophe, les types de vers seront plus représentatives. Il n'est pas facile de tirer les strophes à partir le corpus d'origine. Cela pourrait être une piste pour une étude plus approfondie.

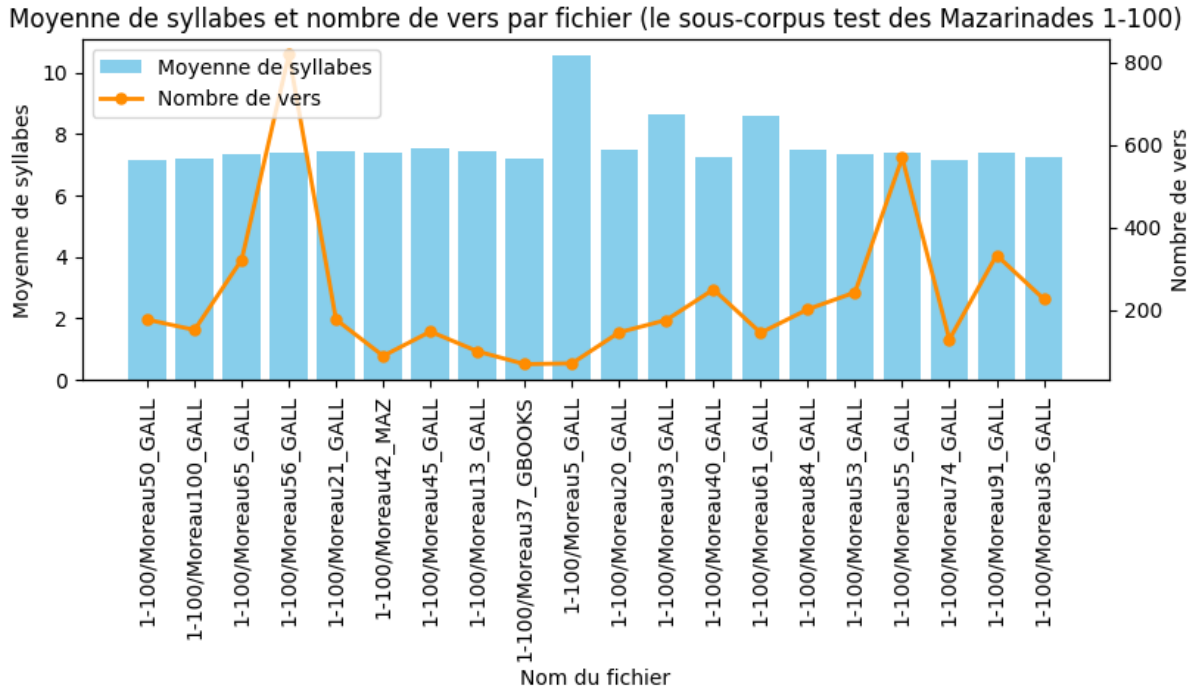

Figure 3 – La moyenne de syllabes par vers selon les fichiers du sous-corpus (1-100)

6 Conclusion

En conclusion, ce rapport a discuté les nuances complexes de la phonétisation dans le cadre de la versification en ancien français, en se concentrant sur l'analyse métrique des Mazarinades. La méthodologie adoptée a mis en lumière les défis inhérents à la phonétisation de mots non reconnus par le dictionnaire de référence et la recherche d'une méthode complémentaire a été au cœur de nos efforts.

J'ai examiné divers outils et méthodes, allant de l'utilisation de services en ligne à des scripts automatisés, en passant par des approches heuristiques basées sur le comptage de voyelles. Malgré leurs contributions, aucun de ces outils n'a fourni de solution pleinement satisfaisante, soulignant ainsi la complexité de phonétiser avec précision un corpus de texte historique.

L'algorithme de Lenzo, bien qu'imparfait, s'est révélé être un pas significatif vers une meilleure compréhension de la phonétique de l'ancien français. L'ajout ciblé de lexiques au dictionnaire de prononciation et l'amélioration de l'algorithme ont permis d'améliorer la reconnaissance des phonèmes, en particulier le schwa, qui est crucial dans la métrique poétique.

L'analyse quantitative des vers a révélé des tendances intéressantes dans la distribution de la longueur des syllabes et a identifié des lacunes dans les études précédentes, notamment concernant la confusion des lettres et les perturbations induites par les éléments non-phonétiques du texte.

Ces découvertes ouvrent la voie à des recherches futures, notamment le développement de méthodes de phonétisation plus raffinées et l'expansion des dictionnaires pour englober une plus grande variété de lexiques en ancien français. L'objectif ultime est de parvenir à une analyse métrique plus précise et détaillée qui non seulement éclaire la pratique poétique de la période de la Fronde, mais enrichit également notre compréhension de l'évolution de la langue française.

Ce rapport, par conséquent, non seulement rend compte des progrès accomplis, mais met également en lumière les défis persistants et les opportunités pour une enquête approfondie, soulignant l'importance continue de la phonologie historique dans les études littéraires.