

28 MAI 2018

AnTOnoMAZ



ANalyse AuTOMatique et NumérisatiOn des MAZarinades

Karine Abiven, Gaël Lejeune

EA 4509-STIH & UMR 8599 CELLF



Document confidentiel –
ne peut être reproduit ni diffusé
sans l'accord préalable
de Sorbonne Université.

Présentation du corpus

6000 écrits (publiés en 5 ans : 1648-1653, la Fronde)

- Données de bibliographie matérielle

Imprimés à 95 %

Manuscrits encore en partie non catalogués

10 pages en moyenne

Imprimés dans l'urgence (papier de mauvaise qualité, typographie inégale)

- Métadonnées

80% des pièces anonymes (et 45 % où l'éditeur est inconnu)

Datation douteuse ou inconnue dans 11 % des cas minimum

- Etat de la numérisation

3000 pièces environ sans version numérique en mode texte

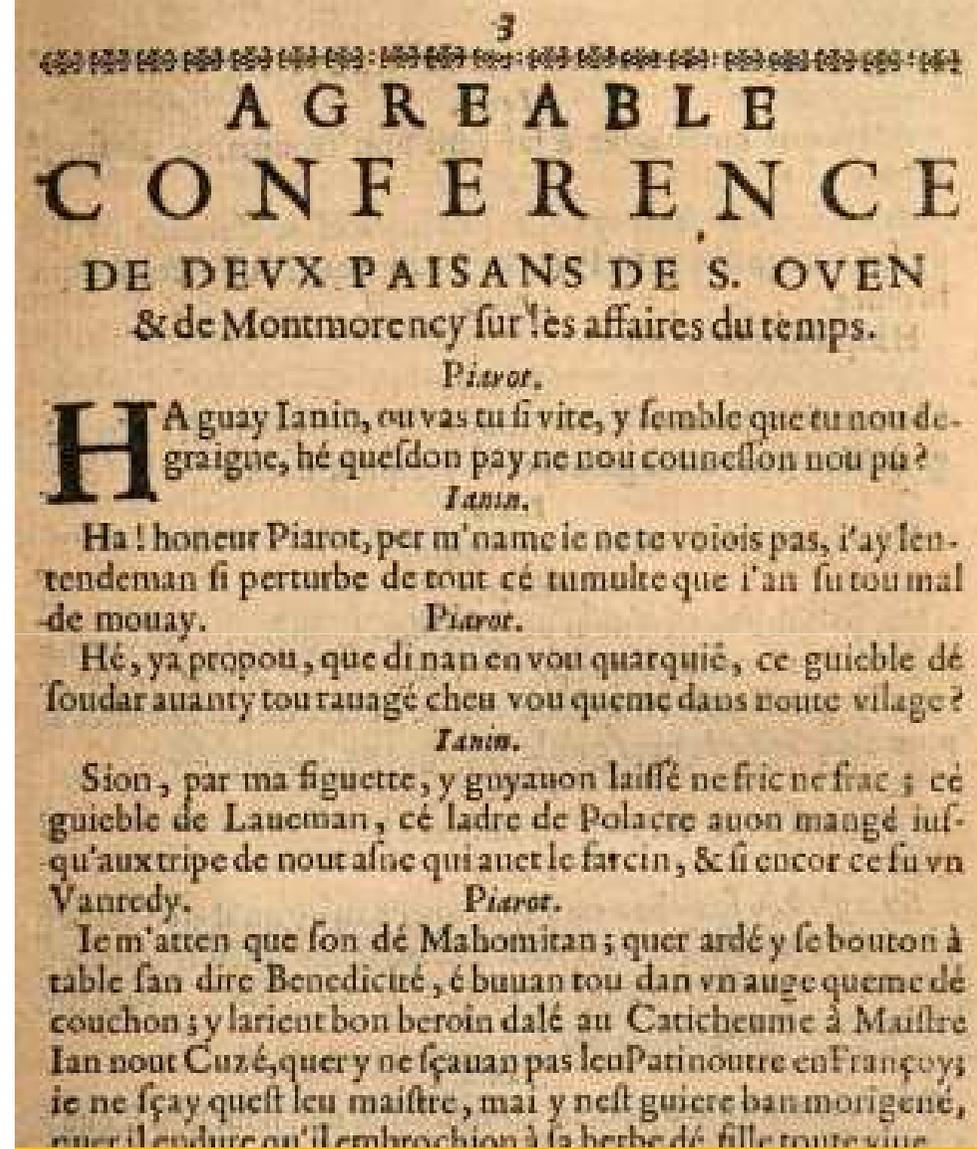
Un corpus bruité

Etats de conservation des imprimés
d'origine : résistance à l'océrisation

Etat de langue : variantes graphiques,
abréviations, orthographe

Langues diverses : français, latin, italien,
patois

Variabilités externes : genres de discours
(très divers, titres peu transparents), formes
(vers/ prose), etc.



Un corpus bruité (II)

 VIS QVE Babillard on me nomme,
le ne veux espargner nul homme,
le fuis sous & remply de vin
le veu~~x~~ 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 fans tardé
Qui a fillé, chose certaine,
Les yeux de nostre bonne Reyne,

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

Un corpus bruité (III)

PVISQVE Babillard on me nomme
Je ne veux espargner nul homme,
Je suis saoul, & remply de vin,
Je 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 Malhabille
Par ce grand Prince de Condé
Qu'il a enchanté sans tarder,
Qui a sillé, chose certaine,
Les yeux de notre bonne Reine,

Un corpus en partie numérisé

Données de l'université de Tokyo : 2711 pièces numérisées

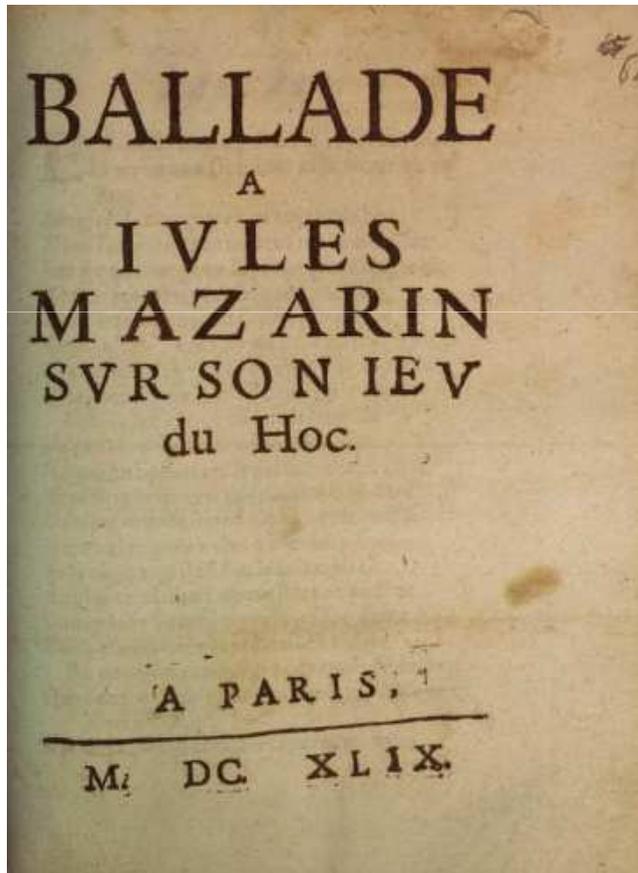
- Inconvénients techniques :

**Doublons dans les résultats de recherche : 1990 pièces uniques
Pas de version océrisée brute**

- Inconvénients philologiques et matériels :

**Numérisation sur microfilm (mauvaise qualité)
Corpus lacunaire (collection incomplète)
Métadonnées lacunaires**

Le problème de l'anonymat



Recherches internationales sur les **Mazarinades**

RECHERCHER

ACCUEIL > RECHERCHE > AFFICHAGE D'UNE OCCURRENCE EN CONTEXTE

Mazarinade n° B_13_63a

Image de la page

Anonyme [1649], *BALLADE DV MAZARIN GRAND IOVEVR DE HOC.*, françaisRéférence RIM : MU_506. Cote locale : B_13_63a.

BALLADE
DV
MAZARIN
GRAND IOVEVR
DE
HOC.

A PARIS,
Chez IEAN BRVNET, ruë neuue saint
Louys, au Canon Royal, proche le Palais.

M. DC. XLIX.

1. PRÉSENTATION & DROITS
2. MODE D'EMPLOI
3. GUIDE D'ÉDITION
4. GÉOLOCALISATION des Mazarinades (fonds, collections, catalogues, etc.)
5. LISTE DES IMPRIMEURS
6. LISTE DES AUTEURS CONNUS
7. LISTE DES GENRES & CATÉGORIES
8. AUTRES DOCUMENTS

ACCÈS PUBLIC AU CORPUS

RECHERCHE
CATALOGUE
CHARIOT
LIENS UTILES

Attributions par des indices philologiques

« Je vous envoie mes cinq balades, que S. A. R. a voulu faire imprimer, et par le premier ordinaire je vous enverray de quoy vous divertir »
(Lettre de Jacques Carpentier de Marigny à Pierre Lenet, 25/07/1652)

52. Balades (*sic*) servant à l'histoire. [568.]

L'exemplaire sur lequel ma note a été faite, était incomplet. Il faut quatre ballades qui sont les quatre premières du n° 570. Malgré la différence des titres, la ballade à *Jules Mazarin sur son jeu de hoc* et celle *sur la naissance de la Fronde* ne font qu'une seule et même pièce.

De cette double rectification il résulte que les *Balades servant à l'Histoire*, publiées d'abord en 1651 [568], ont eu en 1652 une seconde édition réellement *revue et augmentée* [570], et que la *Balade du Mazarin grand joueur de hoc* est de Marigny.

Célestin Moreau, *Bibliographie des Mazarinades*, 1851-1853

Cryptonymes, pseudonymes, Attributions douteuses

Anonyme, *SVITTE ET VNZIESME ARRIVÉE DV COVRIER FRANÇOIS, APPORTANT TOVTES LES Nouvelles de ce qui s'est passé depuis sa dixième arriuée iusqu'à present*, 1649.

D. B. [signé] / **Cyrano de Bergerac, Savinien de [?]**, *LE CONSEILLER FIDELE*, Jean Brunet, 1649.

M. E. G. E. N. R. S., *LA VERITÉ DV ROYALISTE PRESENTÉE AV ROY, PAR M. E. G. E. N. R. S. President. A SA MAIESTÉ. écoute Lecteur*, 1652.

Sandricourt, ? de [?], *LE VISAGE DE LA COVR, ET LA CONTENANCE DES GRANDS, AVEC LEVR CENSVRE. ET Le Dialogue du Roy, & du Duc d'Anjou, Auec la Mamman. EN PROVERBES*, 1652.

Le problème des faux et de la notion même d'auctorialité

Gondi, Jean-François Paul / cardinal de Retz [faux], *HARANGVE FAITE AV ROY PAR MONSEIGNEVR LE CARDINAL DE RETZ, Faite à Compiegne le onziesme Septembre 1652. , 1652.*

Cf **Gondi, Jean-François Paul / cardinal de Retz,** *LA VERITABLE HARANGVE FAITE AV ROY, PAR MONSEIGNEVR LE CARDINAL DE RETZ, POVR LVY DEMANDER la Paix, & son retour à Paris, au nom du Clergé, & accompagné de tous ses Deputez. Prononcé à Compiegne le 12. Septembre 1652., 1652.*

De Guénégaud [signé] ; Du Tillet [signé] (conseiller secrétaire du roi, greffier en chef au parlement de Paris), *SECOND ARREST DV CONSEIL DV ROY, tenu à Pontoise le 23. Iuillet, Portant Cassassion de l'Arrest du Parlement de Paris, des 19. & 20. de Iuillet 1652, 1652 [?].*

Anonyme ; Du Tillet [signé] [faux], *LES IVSTES PLAINTES DE LA CROSSE ET DE LA MITRE DV COAIVTEVR DE PARIS, 1652.*

Données pour la datation automatique

1990 textes avec une distribution très hétérogène :

- 1648 : 1 %
- 1649 : 51 %
- 1650 : 3 %
- 1651 : 5 %
- 1652 : 29 %
- N.D. : 11 %

	Total	Moy.	Ecart-type	Min.	Max.
Caractères	29.838.013	14.993	+21.364	519	382.942
Tokens	6.397.988	3.215	+4.583	98	81.744
Phrases	137.042	68	+118	5	2.533
Vocabulaire	1.857.951	933	+858	55	12.208

Objectifs et Méthode pour la datation

Limiter les pré-traitements :

- corpus integrity
- ne pas écraser les observables

Comparer les différents états :

- langue(s) d'époque ou français moderne
- avec et sans balises

Analyse au grain caractère : robustesse au bruit et aux variations orthographiques

Apprentissage Automatique : SVM linéaire (Un contre le Reste)

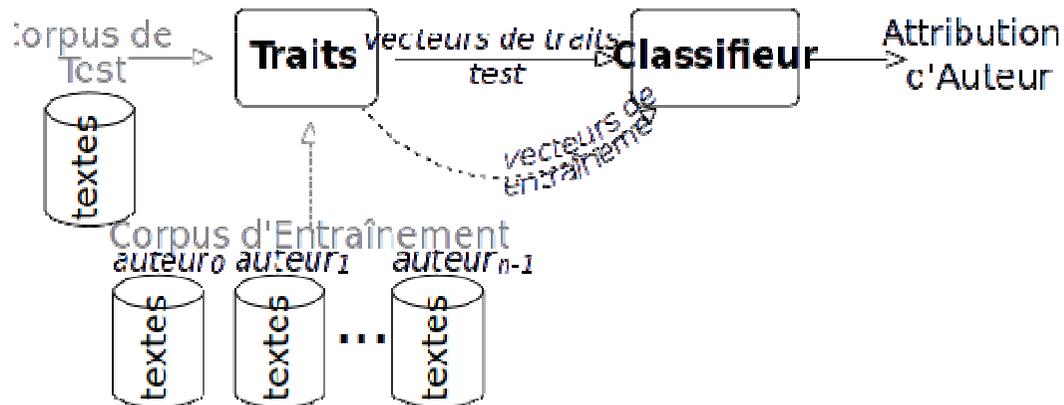
Quelques précisions sur la méthode

Calcul des chaînes de caractères répétées maximales (Ukkonen, Kärkkäinen, Brixtel)

- équivalent à des motifs fermés fréquents (effectif ≥ 2)
- calcul en temps linéaire

Adapté aux corpus bruités et à la configuration multi-classe :

- Corpus Multilingue (Lejeune et al. 2015)
- Attribution d'auteur (Brixtel et al. 2015)
- Données hétérogènes (Lejeune et Zhu 2018)



Premiers Résultats

80 % de micro F-mesure, 54,9 en macro F-mesure

Intervalle de temps trop petit pour affiner la métrique

Les caractères mieux que les mots

Les motifs moins bien que les n-grammes

L'intervalle 3-5 le plus prometteur

Difficultés et perspectives

Équilibre entre la qualité et l'interprétabilité ?

Donner un score de confiance ?

Travailler directement sur la sortie d'OCR ?

Et l'attribution d'auteur ?



MERCI

