28 MAI 2018

AnTOnoMAZ

ANalyse AuTOmatique et NumérisatiOn des MAZarinades

Karine Abiven, Gaël Lejeune

EA 4509-STIH

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

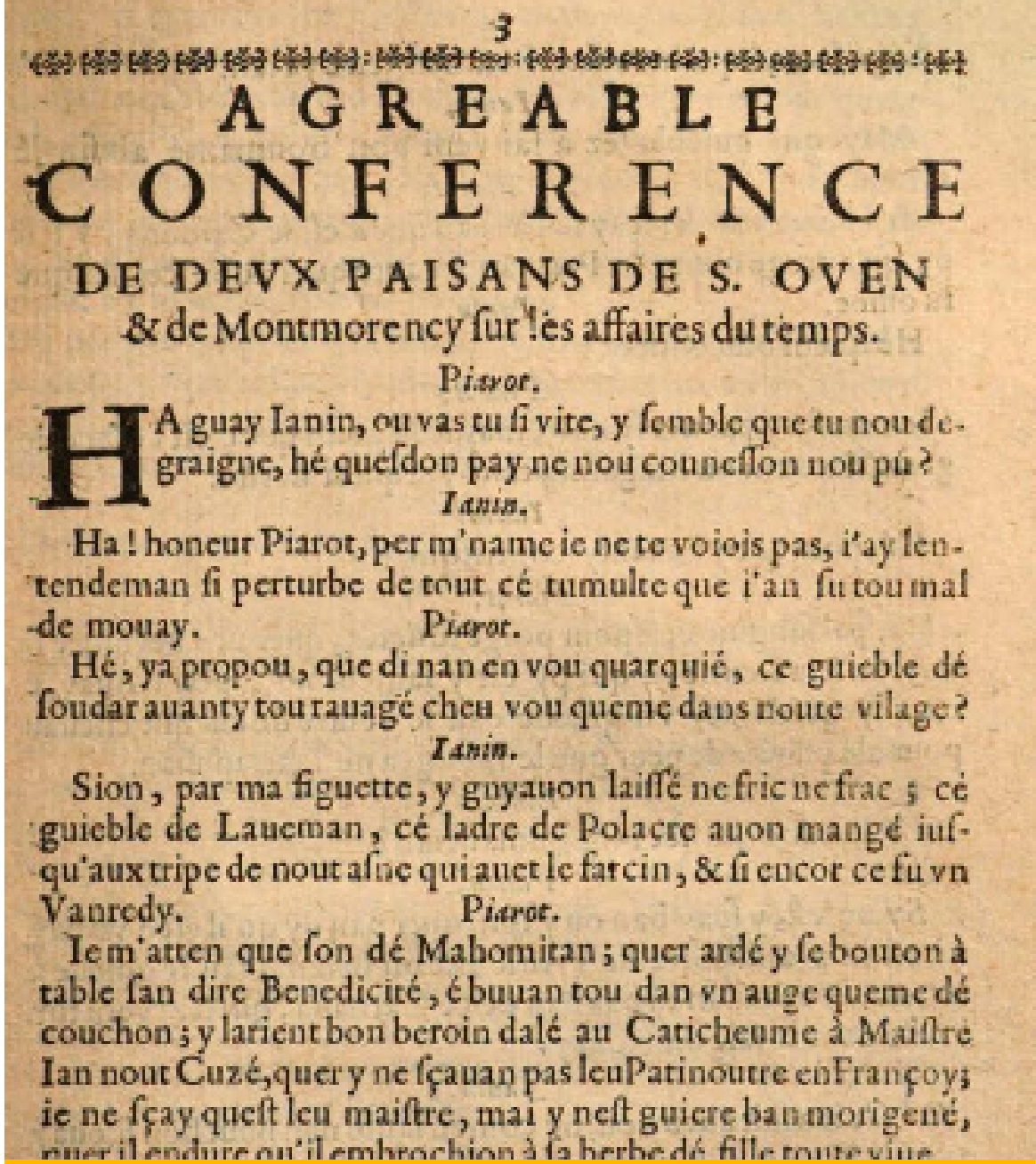
Un corpus bruité

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.


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 counession nou pú?

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?

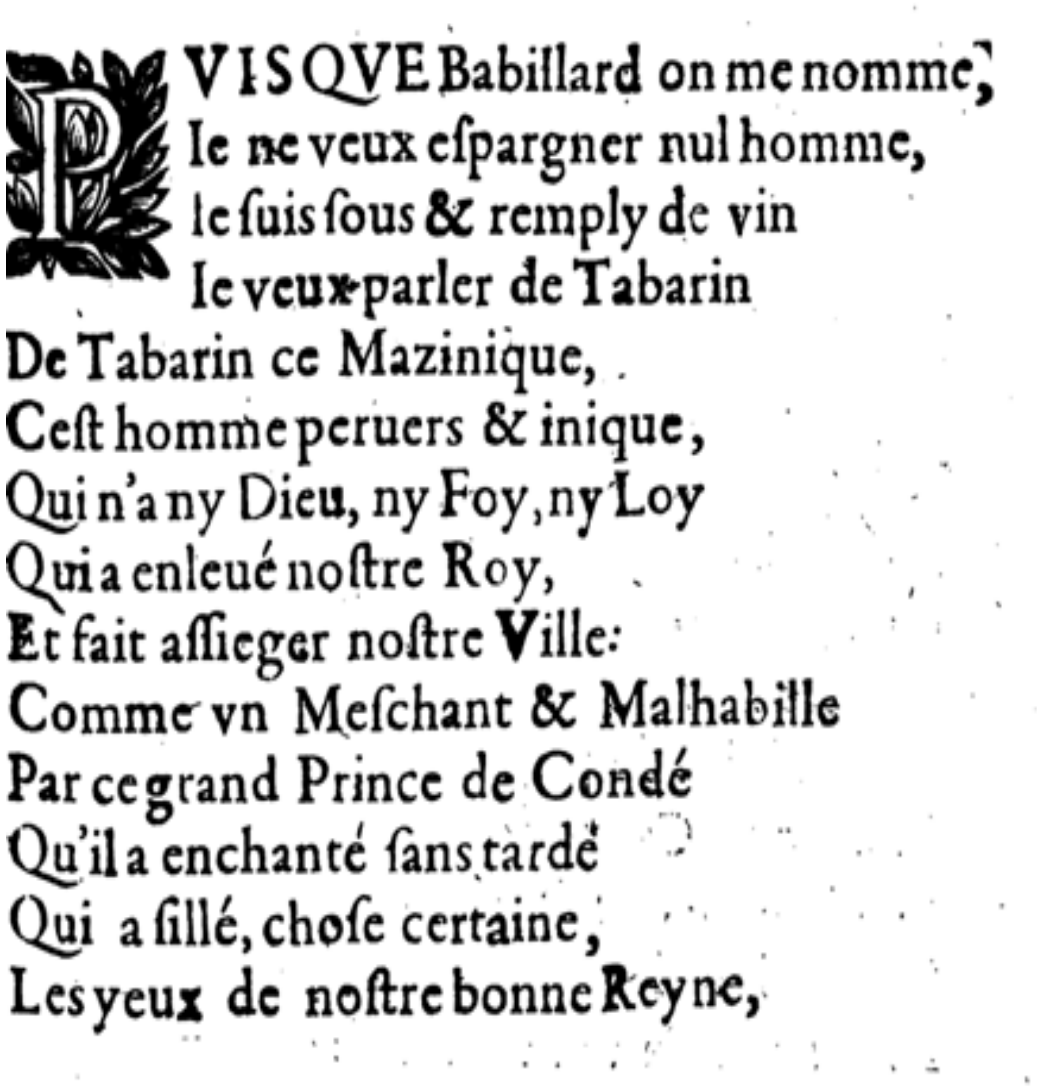
Ianin.

Sion, par ma figuette, y guyauon 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.

Piarot.

Ie m'atten que son dé Mahomitan; quer ardé y se bouton à table san dire Benedicité, é buuan tou dan vn auge queme dé couchon; y larient bon beroin dalé au Catcheume à Maistre Ian nout Cuzé, quer y ne sçauan pas leu Patinoutre en Françoy; ie ne sçay quest leu maistre, mai y nest guiere ban morigené, quer il endure qu'il embrochion à sa berbe dé fille toute viue

Un corpus bruité (II)



PVIS 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 fillé, choſe certaine,
Les yeux de noſtre bonne Reyne,

Le Babillard du temps en vers burlesques, *Paris, N. de la Vigne, 1649. http://mazarinades.org/*

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,

TEST OCÉRISATION

Un corpus bruité (III)

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,

Transcription en mode texte
http://mazarinades.org/

PUISQUE Babillard on me nomme
Je ne veux **é**pargner nul homme,
Je suis saoul, et rempli de vin,
Je veux parler de Tabarin,
De Tabarin ce Mazinique,
Cet homme per**v**ers et inique,
Qui n'a n**i** Dieu, ni Foi, ni Loi,
Qui **a** enlevé notre Ro**i**,
Et fait assi**é**ger n**o**tre Ville
Comme un Méchant et Malhabi**l**e
Par ce grand Prince de Condé
Qu'il a enchanté sans tard**er**,
Qui a sillé, chose certaine,
Les yeux de notre bonne Re**i**ne,

Modernisation orthographique à partir de celle de http://mazarinades.org/

TRANSCRIPTIONS

Un corpus en partie numérisé

LETTRES SORBONNE UNIVERSITÉ

Données du « projet Mazarinades » (équipe des RIM)

http://mazarinades.org/

JSPS KAKENHI. Les programmes de recherche soutenus dans le cadre de ce projet reçoivent les codes d'identification: 20903010 (en 2008), B-22320066 (en 2010-2013), C-26370364 (en 2014-2016).

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**

Attributions par des indices philologiques

LETTRES SORBONNE UNIVERSITÉ

« Je vous envoye mes cinq balades, que S. A. R. a voulu faire imprimer, et par le premier ordinaire je vous enveray 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

LETTRES SORBONNE UNIVERSITÉ

Anonyme, *SVITTE ET VNZIESME ARRIVÉE DV COVRIER FRANÇOIS, APPORTANT TOVTES LES Nouvelles de ce qui s'est passé depuis sa dixiéme 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

LETTRES SORBONNE UNIVERSITÉ

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

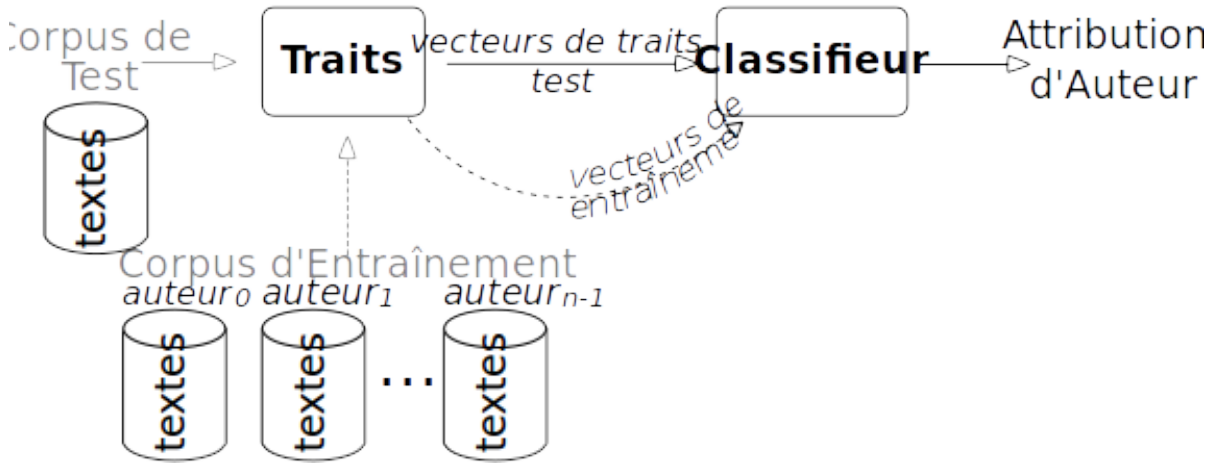
Analyse au grain caractère : robustesse au bruit et aux variations orthographiques

Apprentissage Automatique : SVM linéaire (Un contre le Reste)

Quelques précisions sur la méthode

Calcul des chaînes de caractères répétées maximales (Ukonnen, Kärkkainen, 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

LETTRES SORBONNE UNIVERSITÉ

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

LETTRES SORBONNE UNIVERSITÉ

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

LETTRES SORBONNE UNIVERSITÉ