Le Projet ANTONOMAZ (ANalyse auTOmatique et NumérisatiOn des MAZarinades) exploite un corpus constitué de 6000 écrits ayant pour objet les affaires politiques de la régence du cardinal Mazarin (traditionnellement appelés «Mazarinades»). Ces textes souvent brefs (10 pages en moyenne) ont été édités sous forme imprimée pour la plupart, mais une partie est manuscrite. Ce corpus est d'un grand intérêt pour la connaissance de la période (parfois considérée comme une première « révolution » française).

Ce corpus est encore sous-exploité en raison de son accès difficile : outre sa masse textuelle importante qui résiste à l'analyse manuelle, ce sont des textes qui demeurent opaques (leur auteur est inconnu dans plus de 80 % des cas) et dont la datation est pour partie incertaine.

D'un autre côté, ces documents présentent aussi un certain nombre de difficultés pour les méthodes classiques de Traitement Automatique des langues. Le corpus est bruité : les imprimés d'origine ne sont pas d'une égale qualité ni dans des états de conservation homogènes, et de ce fait résistent à l'océrisation ; la transcription est d'autant plus complexe pour les pièces manuscrites. Par ailleurs, l'état de langue (français de la première moitié du XVIIe siècle) constitue un verrou scientifique important : les variantes graphiques des lemmes sont nombreuses, les abréviations souvent utilisées et l'orthographe erratique. Ce dernier point est principalement dû au fait que les pièces ont souvent été imprimées dans l'urgence, dans un contexte de crise politique et d'explosion de la presse écrite (plusieurs dizaines de pièces pouvaient être composées par semaine). Il convient d'ajouter qu'un certain nombre de pièces sont écrites en latin ou mêlent le latin au français, et De nombreuses autres composées en divers patois. Enfin l'hétérogénéité des métadonnées est accentuée par des variabilités externes : genres (plus de 60 genres de discours différents utilisés, sous des titres souvent déceptifs), forme (composition en vers aussi bien qu'en prose), etc. Un autre obstacle interprétatif de ce corpus pour les experts réside dans le caractère lacunaire des métadonnées : l'anonymat est fréquent (ce qui est classique dans le cas de pamphlets), et la datation parfois douteuse. C'est sur ces deux points (attribution d'auteur d'une part et datation d'autre part) que l'analyse automatique nous a paru féconde.

Nous présenterons les expériences menées sur ce corpus sur ces deux plans. L'approche que nous développons se fonde sur une analyse en chaînes de caractères et une exploitation des macro-structures textuelles. Cette approche nous a permis d'obtenir plusieurs résultats intéressants en exploitant le caractère "bruité" des données plutôt qu'en cherchant à le corriger. En d'autres termes, nous nous sommes efforcés d'adapter nos outils d'analyse plutôt que de chercher à adapter les données. Nous montrons que l'exploitation des chaînes de caractères, que ce soit au moyen d'apprentissage classique ou d'apprentissage profond, est particulièrement bien adaptée aux données bruitées et/ou hétérogènes. De plus, ces méthodes s'avèrent en mesure de généraliser efficacement assez rapidement ce qui permet d'exploiter des techniques d'apprentissage profond sur des données de taille modeste. Nous proposerons également quelques perspectives, notamment sur la détection et le profilage des erreurs de numérisation.