Des données au corpus : l'exploitation numérique des mazarinades

Karine Abiven*[1] and Gael Lejeune*[1]

[1]STIH EA 4509 – Université Paris-Sorbonne - Paris IV – France

Résumé

Le projet Antonomaz (" ANalyse auTOmatique et NumérisatiOn des MAZarinades[1] ") se propose d'explorer un ensemble de brefs imprimés parus en France lors de la Fronde (1648-1653) – entre 5000 et 5500 unités, environ 68 000 pages. Utilisés par des communautés multiples (linguistiques, historiens, littéraires), ces écrits posent pourtant des problèmes d'accès spécifiques : comme ces petits livrets étaient peu couteux, et produits en masse, ils ne sont pas rares, et les exemplaires sont nombreux, y compris numériques (d'Europeana à Google Livre en passant par les bibliothèque numériques Gallica ou Mazarinum, que le présent projet contribue à abonder en fac-similés numériques de haute qualité). Mais pour que ces textes soient exploitables, les tâches demeurent nombreuses, à des niveaux disciplinaires divers :
* au plan de la linguistique de corpus (il reste à les constituer en corpus cohérents) ;

* de la bibliographie matérielle et de la philologie numérique (documentation de métadonnées floues pour des pamphlets souvent anonymes, et imprimés à la va-vite) ;

* du TAL (élaboration de méthodes efficaces d'océrisation appropriées pour ces documents anciens ; classification, supervisée ou non, qui permet par exemple d'obtenir des cluster de textes ainsi mieux compréhensibles par de nouveaux contextes) ;

* des humanités numériques (par exemple, imaginer des visualisations originales qui puissent permettre de mettre en réseaux des textes d'actualité, et donc intrinsèquement dépendants de leur contexte).

Nous espérons obtenir différents résultats :

* abonder en métadonnées nouvelles la Base Bibliographique de la Bibliothèque Mazarine (datation des documents, atelier d'imprimerie d'origine, parti politique d'origine) ;

* construire et partager des chaines de traitements pour diverses tâches (océrisation de PDF anciens, annotation automatique, repérage d'entités nommées, etc.) ;

* fournir aux diverses communautés intéressées une base de données requêtable et téléchargeable.

https://cahier.hypotheses.org/antonomaz.

*Intervenant

Mots-Clés: corpus, mazarinades, pamphlets, contextes, TAL, données, métadonnées, classification, philologie numérique