Enrichissement d'un corpus multiséculaire de théâtre

Aaron Boussidan[1], Philippe Gambette[1], Adrien Roumégous[1]

Le corpus réuni par Paul Fièvre au format XML-TEI sur le site theatre-classique.fr a été utilisé dans plusieurs études en humanités numériques ou en traitement automatique des langues (Karsdorp et al. 2015 ; Douguet, 2018), parfois après un reformatage pour le rendre plus standard et lui appliquer divers outils de visualisation (Glorieux 2016 ; Fischer et al. 2019). Nous nous intéressons à l'enrichissement de ce corpus à partir d'autres collections de pièces de théâtre constituées dans le cadre de travaux étudiants (Canu & Carpentier 2021), de projets de recherche comme le corpus Malherbe du projet Anamètre (Renault 2018), de relectures collaboratives comme sur Wikisource ou individuelles comme celles de Michel Capus sur le site théâtre-documentation.com.

Cet enrichissement présente à la fois des enjeux quantitatifs et des enjeux qualitatifs de représentativité. Nous évaluons ces enjeux sur la base du corpus constitué par Céline Fournial pour étudier de façon systématique les sources du théâtre classique français (Fournial 2019). *Hyperpièces*, disponible sur celinefournial.github.io/hyperpieces, est composé de 104 comédies, 216 tragi-comédies et 273 tragédies, sur la période 1550-1650. Le corpus *French DraCor*, dérivé de celui de theatre-classique.fr couvre environ un quart de ce corpus, avec une surreprésentation des comédies : près d'un tiers des comédies référencées dans Hyperpièces sont contenues dans *DraCor*. Quant aux corpus de théâtre-documentation (TD), Bibliothèque dramatique (BD) et Wikisource (WS), ils couvrent respectivement environ 17%, 10% et 1% du corpus *Hyperpièces*. Par ailleurs, les corpus TD et BD apportent chacun une vingtaine de pièces absentes de *DraCor*. Le corpus WS a davantage d'intérêt sur les siècles plus récents.

Se posent alors diverses questions, liées à des problématiques techniques (détection des doublons entre les différents corpus et conversions de format nécessaires pour aboutir à celui du corpus de référence) ou à l'uniformité éditoriale du corpus, par exemple pour les didascalies (Galleron, 2021). En particulier, pour les pièces des siècles les plus anciens, la question de la normalisation de la langue a été gérée différemment selon les sources.

Nous proposons une chaine de traitement informatique visant à uniformiser le format des pièces de théâtre collectées, en adoptant les choix effectués au sein du projet DraCor et en extrayant des données sources les informations requises pour compléter les métadonnées et structurer correctement le texte des pièces. Elle est constituée par un ensemble de scripts Python mis à disposition sous licence libre sur github.com/AaronFive/StageHyperpieces. Un outil de normalisation automatique fondé sur des règles (Bawden et al. 2022) nous permet d'estimer l'état de langue des pièces ajoutées et en proposer une caractérisation, en quantifiant le nombre de règles appliquées.

Nous montrons enfin comment ces questions de normalisation peuvent être contournées en utilisant ces corpus pour des analyses qui s'intéressent à la structure des pièces plutôt qu'à leur contenu textuel. Plusieurs objets mathématiques peuvent en effet être utilisés pour représenter la structure d'une pièce de théâtre, notamment des réseaux de personnages (Lotker 2021), matrices de co-présences (Marcus 1970, Brainer & Neufeldt 1974, Douguet 2015) ou encore modèles de mots paramétrés (Boussidan 2021).

[1] LIGM, Université Gustave Eiffel, CNRS

Références

- Bawden, Rachel ; Poinhos, Jonathan ; Kogkitsidou, Eleni ; Gambette, Philippe ; Sagot, Benoît ; Gabay, Simon. « Automatic Normalisation of Early Modern French. » *LREC 2022 - 13th Language Resources and Evaluation Conference*, 2022.
- Brainerd, Barron ; Neufeldt, Victoria. « On Marcus' methods for the analysis of the strategy of a play. » *Poetics*, 3(2) :31–74, 1974.
- Boussidan, Aaron. *Modélisation de pièces de théâtre*, mémoire de master 2, Université Gustave Eiffel, 2021.
- Canu, Amélie ; Carpentier, Claire. « La Bibliothèque dramatique. L'édition numérique d'un corpus de pièces de théâtre du XVII^{e} siècle. » *Dix ans avec CAHIER : des corpus d'auteurs pour les humanités à leur exploitation numérique*, 2021.
- Douguet, Marc. *La composition dramatique : La liaison des scènes dans le théâtre français du XVII^{e} siècle*, thèse de doctorat en Langues et littératures françaises, Université Paris 8, 2015.
- Douguet, Marc. « Les hémistiches répétés. » *JADT'18* : 215, 2018.
- Fischer, Frank ; Börner, Ingo ; Göbel, Mathias ; Hechtl, Angelika ; Kittel, Christopher ; Milling, Carsten ; Trilcke, Peer. « Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. » *Proceedings of DH2019*.
- Fournial, Céline. *Imitation et création dans le « théâtre moderne » (1550-1650) : la question des cycles d'inspiration*, thèse de doctorat en littérature et civilisation française, Sorbonne Université, 2019.
- Galleron, Ioana. « Pour un balisage sémantique des textes de théâtre : le cas des didascalies. » *Sens public*, 2021.
- Glorieux, Frédéric. « Dramagraphie 0.2. » Carnet Hypothèses *J'attends des résultats*.
- Karsdorp, Folgert ; Kestemont, Mike ; Schöch, Christof ; van den Bosch, Antal. « The love equation: Computational modeling of romantic relationships in French classical drama. » *6th Workshop on Computational Models of Narrative (CMN'15)*, 2015.
- Lotker, Zvi. *Analyzing Narratives in Social Networks*, Springer Cham, 2021.
- Marcus, Solomon. *Mathematical poetics*. Bucharest, The Publishing House of the SRR Academy (1970).
- Renault, Richard. « Corpus Malherbe : corpus de textes versifiés du XVII^{e} au début du XX^{e}. » *Journée d'étude CORLI : Traitements et standardisation des corpus multimodaux et web 2.0*, 2018.