
Modélisation des textes littéraires : entre temps limité et désir d'exhaustivité (retour sur deux projets d'éditions numériques)

Sarah Orsini*^{1,2}

¹Histoire et Sources des Mondes antiques – École Normale Supérieure - Lyon, Université Lumière - Lyon 2, Université Jean Moulin - Lyon III, Université Jean Monnet [Saint-Etienne], Centre National de la Recherche Scientifique – France

²École normale supérieure - Lyon – École normale supérieure - Lyon (ENS Lyon) – France

Résumé

Le tournant numérique a profondément reconfiguré les modalités d'analyse et d'édition des textes. Il a également transformé la façon de représenter le savoir et de le rendre accessible, ce qui induit une mutation importante des pratiques de lecture. Cela ouvre des perspectives d'une grande créativité : il nous appartient de réinventer nos méthodes de travail pour produire de nouvelles formes de savoirs. Cela a également induit la redéfinition de concepts-clés de la littérature tels que le texte, le livre ou le document, en phase avec le développement de la nouvelle philologie (Driscoll, 2010).

De fait, un apport précieux de la nouvelle philologie est l'attention portée à des " faits " sur le document qui dépassent le texte prêt-à-lire et qui sont regroupées en diverses " dimensions " variant en fonction de la lecture qu'on en fait : linguistique, sémantique, paléographique, littéraire, génétique, culturelle, etc ... (Sperberg-McQueen, 2009 in Pierazzo, 2014). L'édition du document est donc une modélisation (McCarty, 2005) qui résulte de choix éditoriaux. Or, face à la grande créativité offerte par le tournant numérique, comment réaliser ces choix de façon satisfaisante ? Quelles dimensions privilégier et quel degré de précision viser ? Où s'arrêter pour que le projet soit terminable ?

Cette transformation des pratiques éditoriales, aussi créative soit-elle, n'est pas toujours confortable. En effet, nos méthodes héritées de la Renaissance et du XIXe siècle continuent de faire école : nous sommes toujours formé.e.s aux méthodes des éditions critiques sur support papier, et ces méthodes permettent toujours de réaliser de fréquentes publications. De fait, comme le souligne Elena Pierazzo, si tous les livres suivent plus ou moins la même méthode, en raison des limites physiques de la page, les éditions numériques sont toutes différentes, car elles font toujours l'objet d'expérimentations méthodologiques et ne sont pas limitées en termes d'espace. L'espace presque illimité permet également de multiplier les lectures possibles du document, mettant chacune en lumière une de ses " dimensions ". Face à cette multitude d'opportunités, il faut alors choisir les intentions de l'édition, " ce qu'elle fait ", pour aboutir à un résultat.

Aujourd'hui, des méthodes d'édition numérique ont été établies selon des centres d'intérêts principaux, souvent hérités des méthodes philologiques des siècles précédents (éditions philologiques

*Intervenant

avec apparat, reconnaissance des entités nommées, indexation, analyse linguistique automatisée, alignement de traduction), avec des guides de l'encodage à la visualisation. Mais d'une part chaque projet d'édition associe à sa manière ces méthodes, rendant nécessaire l'établissement d'outils de visualisation sur-mesure. D'autre part dans certains domaines il n'existe pas de protocole fixe, ce qui offre une grande liberté, mais induit aussi d'importants défis techniques. Par exemple, dans les éditions génétiques, malgré une volonté générale d'analyser et de donner à lire le texte dans sa chronologie, l'échelle de l'analyse et les éléments analysés varient en fonction des textes et des projets (Grésillon, 1994, Pierazzo, 2014).

Nous avons mené cette réflexion méthodologique dans le cadre du doctorat, dont une des missions consistait à réaliser un prototype d'édition génétique : un outil de lecture chronologique des brouillons du poème néolatin *Crepereia Tryphaena* de Giovanni Pascoli (1892). L'intérêt d'une telle édition est d'analyser le processus d'écriture, et c'était particulièrement intéressant dans le cas de la composition d'un poème néolatin qui supposait un travail de reconstitution de la langue latine, des contenus et des références intertextuelles. Notre démarche était exploratoire : il s'agissait de tester le plus grand nombre d'axes de lecture possible à appliquer à un corpus réduit de brouillons :

- encodage en XML-TEI de la chronologie de l'écriture pour chaque brouillon, de l'échelle de la page à l'échelle de la lettre
- typologie des gestes d'écriture (analyse des types de modification et des causes)
- analyse métrique
- parallélisation des éléments sémantiquement proches d'un brouillon à l'autre (échelle du mot, du groupe de mots, de la strophe)
- encodage des références intertextuelles explicites et implicites.

L'encodage chronologique a abouti à un outil de lecture chronologique des brouillons au moyen de boutons permettant d'afficher le texte d'une campagne à l'autre (inspiré du *Proust Prototype*, André, Pierazzo, 2013, il ne repose toutefois pas sur une transcription intégrée au fac-similé). La typologie des gestes d'écriture a donné lieu à des tableaux statistiques, ainsi qu'à des commentaires au fil de la lecture chronologique. Cela a nécessité l'élaboration de visualisations sur-mesure, qui sont encore aujourd'hui à l'état de prototype et dont l'aspect peut sembler peu attirant puisqu'il a été réalisé sans aucune aide en développement et design. Or, l'apport de ces deux disciplines permettrait de proposer des solutions pour représenter le suivi d'un passage sélectionné d'un brouillon à l'autre (vision synoptique), ou représenter le travail métrique (lacunes comblées progressivement au moyen de tests plus ou moins satisfaisants), autant d'informations qui ne sont pour l'instant interrogeables qu'en XPath ou XQuery pour étayer le commentaire génétique.

Outre les difficultés techniques supposées par ces expérimentations, la principale limite rencontrée est celle du temps. En effet, un tel encodage n'est pas automatisable. Il est seulement possible de réaliser grossièrement la parallélisation des éléments proches avec un traitement automatique des langues (lemmatisation, co-occurrences et requêtes XQuery). Ainsi, pour traiter un corpus entier avec une telle précision, il faudrait être très nombreux ou y consacrer des dizaines d'années. Mais est-il souhaitable d'étendre une telle démarche à l'ensemble du corpus latin de l'auteur, voire du corpus italien ? Faut-il préférer une transcription de davantage de textes mais d'une moindre précision ? Est-ce que le lecteur en apprendra davantage sur l'écriture pascolienne si l'on traite tout le corpus de cette façon que si l'on choisit quelques dossiers représentatifs ? Et jusqu'où peut-on assister le lecteur dans sa confrontation avec le brouillon ? Pour comprendre profondément le geste d'écriture, une fréquentation assidue des brouillons reste nécessaire, et n'est que facilitée par l'édition. De fait, certains aspects du processus d'écriture restent inexplicables, soit parce que les raisons d'un geste ne sont pas explicites, soit parce que les mots nous manquent pour décrire ce qui relève d'une expérience sensible (c'est pourquoi il est bien plus facile d'explicitier les mises au point

des derniers brouillons que les premières étapes, notamment la création de nouveaux contenus). L'autre limite rencontrée est celle de la subjectivité de telles analyses : même dans le cas où le déchiffrement serait aisé, la datation relative des éléments du brouillon relève de l'interprétation de l'éditeur. Tout en permettant au lecteur d'accéder à une transcription plus facile à lire et dynamique, elle déforme nécessairement le document en le modélisant.

À ces enjeux herméneutiques s'ajoutent les problématiques de la visibilité : quelle visibilité peut avoir un prototype (expérimentation d'une exhaustivité dans la précision de l'encodage) par rapport à l'édition génétique d'un corpus complet (exhaustivité dans la transcription des textes) ? Comment situer une telle production dans la foison des projets numériques ?

Nous pourrions penser que de telles problématiques sont caractéristiques de la génétique textuelle. Cependant, elles concernent également des projets d'édition de textes classiques. Le projet AgroCCol pour lequel nous travaillons désormais consiste à établir une édition numérique de type encyclopédique des textes agronomiques latins et grecs. Pour l'instant, pour des raisons de temps, nous avons sélectionné des textes sur la culture des céréales et des légumineuses, afin de pouvoir réaliser un encodage correspondant à une multiplicité d'axes de lecture :

- indexation de termes techniques pour constituer un dictionnaire sur l'agriculture antique
- constitution d'un thésaurus des noms de plantes et d'outils
- identification des entités nommées
- encodage thématique

Certes, davantage de tâches sont automatisables entièrement (notamment le relevé des occurrences du dictionnaire technique) ou partiellement (analyse thématique), nous travaillons en équipe et bénéficions d'aides techniques importantes. Cependant, le choix d'une analyse qualitative (même si assistée par ordinateur) nous a poussé à restreindre notre corpus, et l'échéance du projet nous contraint parfois à arrêter la précision de nos analyses thématiques à ce qui a été établi. Cela pose à nouveau des problèmes de subjectivité (sur quels critères les textes ont-ils été choisis ? sur quels critères des thèmes sont-ils associés à des mots du texte ? l'encodage thématique est-il constant d'un texte à l'autre ?), de visibilité (quelle sera la place d'un tel corpus par rapport à des corpus anciens plus larges sans élucidation sémantique ou littéraire tels que *Perseus* ? En tant que lecteur, comment trouver une telle publication ?). Se pose également la question du rapport au lecteur : cette sélection satisfera-t-elle sa curiosité, même si elle ne traite pas de l'ensemble de l'agriculture ?

Ainsi, la facilitation de l'accès aux documents, aux textes ou aux données de la recherche augmente les possibilités de traiter des volumes plus grands et d'aspirer à une exhaustivité en matière de corpus (œuvre intégrale, mouvement, associations), tandis que la TEI nous pousse à viser une exhaustivité en termes de précision d'encodage. Mais le souhait de continuer à mener des analyses de précision correspond à un temps long de la recherche et ne peut pas être subordonné uniquement à la demande ou à des échéances courtes. Dans ce contexte, la démarche des données ouvertes et réutilisables est tout à fait stimulante. Cependant, elle ne peut aboutir que si l'on organise des synergies entre les publications numériques, soit en préférant travailler à la suite d'un projet préexistant, soit en construisant ensemble des répartitions entre divers projets, sans pour autant les uniformiser. Des projets de plateformes telles que *Biblissima*, ou *Fonte Gaia* pour les textes italiens pourraient constituer une première étape pour sortir de cette tendance à l'isolement des projets numériques les uns par rapport aux autres.

—

Cette présentation peut être adaptée au format de communication courte si nécessaire.

Bibliographie indicative

TEI Consortium, eds. *TEI P5 : Guidelines for Electronic Text Encoding and Interchange*. [Version 4.1.0]. [Last updated on 19-08-2020]. TEI Consortium. URL : <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html> [Consulté le 20/09/19].

André Julie et Elena Pierazzo, " Le codage en TEI des brouillons de Proust: vers l'édition numérique ", *Genesis. Manuscripts – Recherche – Invention*, no 36, Sigales, 2013, p. 155-161. DOI : <https://doi-org.acces.bibliothequediderot.fr/10.4000/genesis.1159> [consulté le 16/11/2020]

CASENAVE Joana, " Le positionnement éditorial dans l'édition critique numérique ", *Digital Studies/Le champ numérique*, vol. 9, no 1, 2019. DOI : <http://doi.org/10.16995/dscn.348> [consulté le 15 janvier 2020].

Driscoll Matthew J., " The Words on the Page: Thoughts on Philology, Old and New ", dans Judy Quinn et Emily Lethbridge (éd.), *Creating the Medieval Saga: Versions, Variability and Editorial Interpretations of Old Norse Saga Literature*, Denmark, University Press of Southern Denmark, 2010, p. 85-102.

Grésillon Almuth, *Éléments de critique génétique: lire les manuscrits modernes*, Paris, Presses Universitaires de France [1994], 2016.

McCarty Willard, *Humanities computing*, Basingstoke, Palgrave Macmillan, 2005.

Orsini Sarah, *Les Carmina de Giovanni Pascoli: édition traduite et commentée d'une sélection de poèmes latins et édition numérique d'une sélection de brouillons*, Université Lyon 2, Università Roma Tre, 2019.

URL de l'édition génétique du poème *Crepereia Tryphaena* : [:https://github.com/SarahOrsini/Edition_genetique.L](https://github.com/SarahOrsini/Edition_genetique.L) [consulté le 16/11/2020]

Pierazzo Elena, *Digital Scholarly Editing: Theories, Models and Methods.*, , sans lieu, 2014. Existe en format papier: édition Ashgate, 2015.

Sperberg-McQueen C. Michael, " How to Teach your Edition How to Swim ", *Literary and Linguistic Computing*, vol. 24/1, 2009, p. 27-52.

Mots-Clés: XML, TEI, génétique textuelle, édition critique, modélisation, visualisation, données ouvertes