McDUFF, Lydia 20294887
SAVARD-ARSENEAULT, Adrien 20155684

Présentation du corpus

Travail présenté au
Pr Dominic FOREST

dans le cadre du cours
SCI6203 – Intelligence artificielle et données textuelles

École de bibliothéconomie et des sciences de l'information
Université de Montréal
10 novembre 2025

Critères généraux

Le corpus s'intitule « Littérature critique anglophone autour du *Roman de Silence* ». Il est constitué de documents rassemblés par Adrien Savard-Arseneault depuis l'automne 2020, dont plusieurs qui ont été ajoutés en été 2023, puis en automne 2025. Il est en cours de conversion et d'annotation par Lydia McDuff et Adrien Savard-Arseneault. Ces documents ont été publiés au cours des cinquante dernières années, le plus vieux datant de 1985 et le plus récent, de juillet 2025. Le corpus peut être cité de la façon suivante :

Savard-Arseneault, A. et McDuff, L. (2025). *Littérature critique anglophone autour du* Roman de Silence [corpus]. Université de Montréal. `https://github.com/pitoftar/sci6203a25/tree/main/corpus`

Constitué de 90 documents scientifiques (articles, chapitres de livres, thèses et mémoires), le corpus a une taille totale de 1 032 363 mots. Les documents comptent généralement entre 6 000 et 12 000 mots, avec une moyenne de 11 470 mots par document. La médiane est de 8 463,5 mots.

57 documents sont accessibles en ligne grâce à une licence institutionnelle, 20 documents sont en accès ouvert en ligne (dont 3 avec un embargo), 8 documents sont disponibles à la Bibliothèque des lettres et sciences humaines de l'Université de Montréal, 4 documents sont sous licence CC BY-NC-ND 4.0 et 1 sous licence CC BY-SA 4.0. La majorité des documents constituant le corpus ne sont donc pas complètement libres de droits.

Critères technologiques

Le corpus a été constitué en fouillant les bases de données pertinentes[1] au sujet à l'aide de la requête `"roman de silence" OR "romance of silence"`. L'outil de recherche ne permettait pas toujours les requêtes combinées : dans ces cas, les deux termes de recherche ont été indiqués séparément. Lorsque cela était possible, des filtres ont également été appliqués à la recherche afin de ne retenir que les documents en langue anglaise. Les textes ont été individuellement survolés afin de s'assurer de leur pertinence avec le sujet avant d'être ajoutés au corpus.

[1] Les bases de données interrogées sont les suivantes : JSTOR, Project Muse, ProQuest, Web of Science, Google Scholar, ainsi que le catalogue des documents disponibles dans les bibliothèques de l'Université de Montréal *via* WorldCat.

En raison de sa nature, la majorité du corpus (83%, soit 75 documents sur 90) a été récupérée au format PDF. De ce nombre, un peu moins de la moitié des documents sont nativement numériques alors que 39 sont issus de la numérisation. Le texte est alors mis en forme à l'aide de procédés de reconnaissance optique des caractères (ROC). Le reste du corpus (15 documents) a été récupéré au format HTML ou texte brut (`.txt`). Nous expérimentons actuellement avec plusieurs manières pour extraire le texte des PDF afin de les baliser dans un format plus structuré et exploitable (HTML ou XML).

Certaines données du corpus (langue, longueur des textes, provenance, année, nom de l'autrice[2]) ont été annotés manuellement par les membres de l'équipe à l'aide du logiciel Zotero. Les annotations de cette classe ont été traduites et structurées au format JSON.

Critères informationnels

Tous les textes composant le corpus sont liés aux disciplines de la littérature, de la philologie ou des études médiévales. Certains intègrent aussi des approches d'études féministes ou de codicologie. Ils partagent tous un sujet, soit l'étude du *Roman de Silence*, un texte médiéval du XII^{e} siècle. Puisque la tâche que nous envisageons réaliser est l'analyse linguistique de la désignation du personnage principal, prénommé Silence, dans le corpus critique, il était nécessaire de rassembler le plus grand nombre de textes possibles traitant de ce sujet.

Critères linguistiques

Le corpus est constitué de 27 chapitres de livres, une retranscription d'une communication de colloque, 54 articles de revues scientifiques évalués par les pairs et 8 thèses ou mémoires. Le registre de langue est soutenu, et comprend quelques termes spécialisés. Tous les textes sont rédigés en langue anglaise. Certains d'entre eux comprennent toutefois des citations issues du texte original, en ancien français. Celles-ci sont généralement indiquées entre guillemets (“”) ou placées dans des blocs de citation à l'écart du texte.

[2]Le féminin a été choisi pour alléger le texte et proposer une alternative au masculin comme genre neutre.

Difficultés rencontrées et commentaires

Puisqu'à notre connaissance aucun corpus sur la littérature critique anglophone autour du *Roman de Silence* n'avait été constitué auparavant, nous avons dû le faire nous-mêmes. Cela dit, le repérage de documents pertinents n'a pas été un défi en soi : les difficultés rencontrées résident plutôt dans l'annotation et la transformation des documents.

Afin de pouvoir comparer la désignation de Silence dans les textes critiques avec les genres des autrices, nous avons annoté chaque document dans Zotero en ce sens. Pour ce faire, certains des documents comprenaient une biographie de l'autrice, mais pour la grande majorité nous avons dû rechercher le nom de l'autrice sur un moteur de recherche[3]. Cette partie de la recherche, bien que nécessaire, s'est révélée chronophage.

Par ailleurs, tel que mentionné dans les critères technologiques, la majorité des documents n'étant disponibles que sous format PDF. Nous devrons donc trouver une façon efficace de convertir leur texte sous un format davantage approprié à la fouille de texte et qui, nous l'espérons, ne nécessitera peu de correction manuelle de notre part.

[3]Par souci de rigueur et de respect envers les autrices, nous ne voulions pas présumer leur genre selon leur prénom.