Pour une détection automatique de l'espace textuel des personnages romanesques

Jean Barré, Pedro Cabrera Ramírez, Frédérique Mélanie, Ioanna Galleron

Pour une détection automatique de l'espace textuel des personnages romanesques

Jean Barré
ENS - PSL - Translitterae
Lattice UMR 8094
jean.barre@ens.psl.eu

Pedro Cabrera Ramírez
ENS - PSL
Lattice UMR 8094
pedro.cabrera.ramirez@ens.psl.eu

Frédérique Mélanie
CNRS
Lattice UMR 8094
frederique.melanie@ens.psl.eu

Ioanna Galleron
Sorbonne Nouvelle
Lattice UMR 8094
ioana.galleron@sorbonne-nouvelle.fr

Résumé

Cette communication présente une approche théorique pour l'étude des personnages de roman grâce à leur détection automatique. Nous introduisons d'abord la notion d'espace textuel des personnages comme base théorique pour analyser la résolution des chaînes de coréférence dans la fiction romanesque, puis nous présentons un cas d'étude : celui de deux personnages secondaires de *Manon Lescaut*, le lieutenant de police et Tiberge.

1 Introduction

De combien de personnages sont constitués les romans ? Malgré son apparente naïveté, cette question reste encore un problème ouvert dans les études littéraires. En effet, pour effectuer ce type d'analyse, les spécialistes recourent à l'annotation manuelle Lavocat (2020), très coûteuse en temps. Ces méthodes ne permettent donc pas pour le moment d'élargir la focale et de proposer une lecture distante Moretti (2000), seule à même d'étudier les personnages et leur rôle dans la structure du récit de manière empirique. En effet, une étude de la démographie des personnages à plus grande échelle requiert nécessairement une approche computationnelle.

La détection automatique de personnages est une problématique de recherche active dans le domaine des études littéraires computationnelles. Les développements récents ont permis d'élargir la compréhension des personnages au-delà de la reconnaissance d'entités nommées :

— La détection de l'agentivité, soit la capacité d'un personnage à entreprendre des actions intentionnelles et autonomes au sein du récit, rend possible la reconnaissance des entités qui jouent un rôle actif, même s'ils ne sont pas les protagonistes principaux tels que « le cocher » ou « la grenouille » (Vala et al., 2015; Karsdorp et al., 2015).

— La résolution de la coréférence, soit le fait pour plusieurs éléments linguistiques de renvoyer à une même entité, rend possible de mieux distinguer les personnages secondaires dans des textes littéraires par rapport à d'autres personnages mineurs, moins significatifs dans le développement de l'intrigue (Bamman et al., 2014).

Dans la continuité des récentes améliorations permises par l'utilisation de grands modèles de langue pour la détection des personnages (Bamman et al., 2019), une équipe du Lattice (CNRS – École Normale Supérieure – Sorbonne Nouvelle) a développé fr-BookNLP, un algorithme de Traitement Automatique des Langues (TAL) entraîné sur des textes littéraires. Ces derniers ont été annotés pour la détection de personnages de roman avec la résolution de leur chaîne de coréférence. Cette dernière fait référence au processus d'identification des cas où différents mots (ou signifiants) font référence à un même personnage (ou signifié) dans un récit. Il s'agit de détecter et de relier toutes les mentions d'un personnage spécifique tout au long du récit, même si elles sont exprimées à l'aide de noms, de pronoms ou de descriptions différents.

Par exemple, deux chaînes de coréférence sont présentes dans cette phrase

> ***Mon* frère** *m'*embrassa très tendrement
> mais **il** ne *me* parla point.

La première renvoie au narrateur, la seconde à son frère.

L'utilisation d'un tel outil influence nécessairement la notion de personnage, puisqu'il requiert

de la formaliser et de la saisir en un ensemble de caractéristiques finies. Nous introduisons la notion *d'espace textuel des personnages*, qui permet d'opérationnaliser l'analyse de la présence d'un personnage et de son évolution au fil du récit.

2 Un espace textuel des personnages

2.1 Conceptualisation

Nous envisageons le personnage d'un récit dans sa dimension textuelle, car une approche computationnelle du personnage repose nécessairement sur la présence linguistique de ce dernier et notre capacité à détecter cette dernière. Il nous a semblé pertinent d'approfondir les travaux de Jahan et Finlayson (2019) en abordant ce problème par la chaîne de coréférence. Ainsi, nous considérons l'espace textuel du personnage comme l'ensemble de ses mentions (les mots ou ensemble de mots du texte faisant référence à ce personnage). Si l'on considère le texte comme ligne continue, l'espace textuel associé à un personnage correspondrais à la distribution géographique de toutes les mentions d'un personnage au fil du récit.

Notre approche se fonde sur le modèle fr-BookNLP qui a été conçu pour récupérer l'information de la coréférence spécifiquement dans des textes littéraires. Il représente un personnage comme étant un ensemble d'éléments linguistiques liés entre eux. Les sorties du modèles vont nous permettre de mesurer l'espace textuel des personnages comme une manière de mesurer leurs différents rôles - continus et significatifs pour les personnages principaux et plus ponctuels et incidents pour les autres.

2.2 French BookNLP

French BookNLP est une adaptation pour le français des travaux menés pour l'anglais par Bamman et al. (2014, 2019, 2020). Le projet se fonde d'une part sur un corpus annoté en coréférence dans le cadre du projet ANR Democrat Landragin (2021), et d'autre part sur le typage des chaînes de coréférence en un ensemble fini d'entités (`PER`, `LOC`, `ORG`, `FAC`, `GPE`, `VEH`). Un corpus de 20 textes littéraires (les 10 000 premiers tokens) a été annoté avec ces étiquettes et est disponible en différents formats : BRAT, SACR (Oberle, 2018), CONLL [1].

La chaîne de traitement a permis la conception de quatre modèles indépendants pour (i) la détection des entités nommées, (ii) des évènements littéraires Grunspan et al. (2022), (iii) des citations et (iv) de la coréférence de ces entités. C'est ce dernier qui est, à notre sens, un des plus fertiles pour l'analyse littéraire. Nous réduisons l'analyse aux chaînes de type PER, puisque la difficulté de la résolution de la coréférence est principalement liée aux chaînes longues, c'est à dire aux personnages de romans.

Le tableau 1 propose un aperçu des performances de fr-BookNLP. Pour la détection de l'ensemble des mentions coréférant à un personnage, le f1 score du modèle performe à 0.87, ce qui est assez éloigné de l'état de l'art en TAL pour la reconnaissance d'entités nommées, mais assez proches des résultats de Bamman (2021) [2] pour la reconnaissance des personnages sur un corpus littéraire en anglais. Le MUC est une mesure d'évaluation couramment utilisée pour les tâches de résolution des coréférences. Elle évalue la capacité du modèle à regrouper correctement les mentions qui se réfèrent à la même entité.

TABLEAU 1 – Évaluation de fr-BookNLP

		precision	rappel	F_1
PER		0.85	0.89	0.87
Coreference	MUC	85,06	85,10	85,08

Il est important de noter que fr-BookNLP est un modèle en construction et que ces résultats ne sont que temporaires. Une des difficultés du modèle réside dans la duplication de chaînes très longues, ce qui implique que le modèle récupère plusieurs chaînes de coréférence pour le personnage principal. L'étude des personnages principaux n'est pas encore pertinente à ce stade de développement, et l'on se focalisera dans la suite de cet article sur l'étude des personnages secondaires.

2.3 Approche du texte

Manon Lescaut (dans sa version normalisée proposée par Wikisource) nous a paru être un cas intéressant pour deux raisons. D'une part, la diversité des personnages est importante tout au long du récit, avec une grande variabilité des dénominations des différents personnages. D'autre part, une des spécificités de *Manon Lescaut* réside dans le fait que de nombreux personnages réapparaissent à différent moments du récit, ce qui complexifie la tâche

1. L'ensemble des fichiers et des scripts du modèles sont disponibles sur le GitHub du projet (Lattice, 2023).

2. https ://github.com/booknlp/booknlp

pour notre modèle. Ce dernier a été entraîné sur des textes du XIX^{e}siècle, mais nous faisons l'hypothèse que les problèmes posés par un état de langue antérieur seront assez marginaux pour ne pas affecter notre expérience.

Le processus d'annotation automatique opérationnalise la récupération de l'espace textuel de chaque personnage du roman, modélisant sa présence narrative dans le texte grâce au repérage des différentes occurrences renvoyant à cette entité. En effet, comme on peut le voir dans la figure 1 avec une visualisation dans SACR, les syntagmes nominaux (substantifs et noms propres) ou pronominaux sont annotés, y compris les expansions. Les syntagmes renvoyant à un même référent sont reliés pour former une chaîne de coréférence, et le type PER est attribuée à chaque élément intégré.

Les apparitions discontinues du personnage sont ainsi fusionnées dans une seule chaîne de coréférence qui, dans les cas des personnages importants, est distribuée sur l'ensemble du texte.

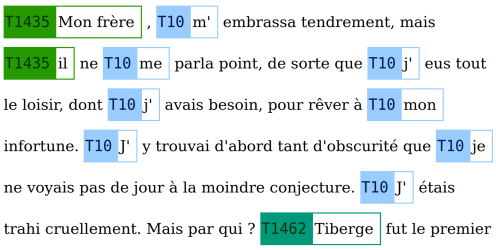

FIGURE 1 – Visualisation de la coréférence dans SACR, chaque couleur est un personnage, chaque nombre son rang d'apparition

3 Intermittence des personnages secondaires

Une fois le texte de *Manon Lescaut* annoté automatiquement par le modèle, nous découpons le récit à l'aide d'une fenêtre roulante de 1 000 tokens afin de récupérer le signal d'apparition des différents personnages. Nous focalisons notre attention sur deux personnages secondaires que le modèle repère à des moments précis du récit.

La figure 2 montre les occurrences de la coréférence de *Tiberge* le long de la fenêtre glissante. La particularité de ce personnage est sa présence discontinue au fil du roman – c'est donc un personnage secondaire. Il est particulièrement actif à trois moments du récit, que le modèle détecte bien ici. Ils correspondent au début du récit, quand Tiberge est au séminaire avec le chevalier, aux occasions

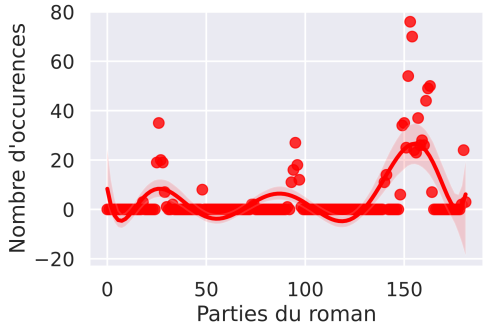

FIGURE 2 – Les points montrent les occurrences de la coréférence de Tiberge au fil du roman, et la courbe montre son signal d'apparition grâce à une régression non-linéaire.

où le chevalier lui demande du secours financier, et à sa poursuite du chevalier à la Nouvelle Orléans. Ce signal n'aurait pas été retrouvé avec seulement le nom propre de Tiberge puisque ce dernier ne compte que 43 occurrences contre 714 concernant l'ensemble de ses mentions.

La figure 3 décrit l'apparition de *M . le lieutenant général de police* le long de la fenêtre glissante. Ce personnage est très intéressant puisqu'il n'a qu'une présence limitée et circonscrite à deux moments précis : une première fois lors de l'enfermement du chevalier à Saint-Lazare et de Manon à la Salpêtrière, puis une seconde fois au moment du conflit avec M. de Gxxx. Mxxx. et son fils. Son importance est pourtant clé, puisque son arrivée dans l'action correspond à deux nœuds de l'intrigue qui déterminent le destin des deux personnages principaux. Une nouvelle fois, le nombre de mentions relatifs au lieutenant de police équivaut à 314 occurrences, contre seulement 19 pour le mot *lieutenant*. Ainsi, la récupération de la coréférence permet de rendre l'analyse de la présence du personnage plus fine.

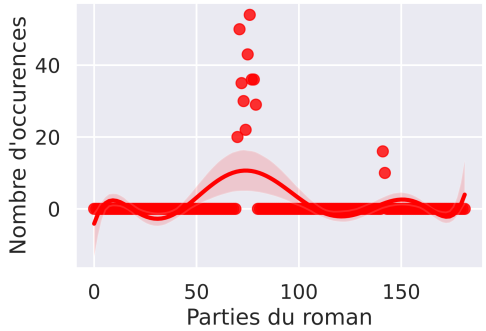

FIGURE 3 – Les points montrent les occurrences de la coréférence du lieutenant de police au fil du roman, et la courbe montre son signal d'apparition grâce à une régression non-linéaire.

4 Conclusion

La notion d'espace textuel des personnages nous permet de mieux saisir la distribution de leur présence au fil du récit, et la résolution de la coréférence permet d'opérationnaliser cette notion. Grâce à ce cadre théorique et technique, il nous est possible appréhender le rôle des personnages secondaires dans les romans, et pourrait nous permettre à terme d'améliorer notre compréhension de la structure narrative. L'intermittence des apparitions des personnages découpe en effet le récit en scènes distinctes et rend compte du fractionnement de la présence effective des personnages.

La méthode présentée ici est dépendante des sorties du modèle de résolution de coréférence fr-BookNLP, avec les erreurs qu'elles peuvent comporter. Cependant, on a vu qu'il était dans l'intérêt des études littéraires computationnelles de s'approprier des outils du TAL avancés afin de mener à bien des analyses originales et apporter un nouveau point de vue sur une notion fondamentale de la théorie littéraire.

Ces premiers travaux ouvrent la porte à de nombreuses pistes de recherches. Il faudrait tout d'abord ré-entraîner le modèle fr-BookNLP sur des romans d'autres époques ou genres littéraires, afin d'évaluer sa performance et sa capacité à s'adapter à des contextes différents. On pourrait également examiner les relations entre les courbes de présence des personnages et des événements clés dans le récit, afin d'explorer les interactions entre les personnages et leur impact sur le récit. Une analyse à grande échelle de l'espace du personnage permettrait de lier des variations significatives avec les catégories définies par le schéma actanciel de Greimas Greimas (1966).

Remerciements

Le doctorat de Jean Barré est financé par l'EUR Translitteræ (programme « Investissements d'avenir » ANR-10- IDEX-0001-02 PSL* et ANR-17-EURE-0025). Cette recherche a également bénéficié du soutien du Réseau international de recherche Cyclades du CNRS (Corpora et Linguistique Computationnelle pour les Humanités Numériques).

Bibliographie

David Bamman. 2021. BookNLP.

David Bamman, Olivia Lewke, et Anya Mansoor. 2020. An annotated dataset of coreference in english literature.

David Bamman, Sejal Popat, et Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North*, pages 2138–2144. Association for Computational Linguistics.

David Bamman, Ted Underwood, et Noah A. Smith. 2014. A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 370–379. Association for Computational Linguistics.

Algirdas Julien Greimas. 1966. *Sémantique Structurale : Recherche de Méthode*. Larousse, Paris.

Claude Grunspan, Frédérique Mélanie-Becquet, Jean Barré, Laurette Chardon, Ioana Galleron, Marco Naguib, Clément Plancq, Olga Seminck, et Thierry Poibeau. 2022. Event annotation for literary corpora analysis. *DH2022*.

Labiba Jahan et Mark Finlayson. 2019. Character identification refined: A proposal. In *Proceedings of the First Workshop on Narrative Understanding*, pages 12–18. Association for Computational Linguistics.

Folgert Karsdorp, Marten van der Meulen, Theo Meder, et Antal van den Bosch. 2015. Animacy detection in stories. In *Workshop on Computational Models of Narrative*.

Frédéric Landragin. 2021. Le corpus DEMOCRAT et son exploitation. présentation. *Langages*, 224(4) :11–24.

Lattice. 2023. fr-litbank.

Françoise Lavocat. 2020. L'étude des populations fictives comme objet et le « style démographique » comme nouveau concept narratologique. *Atelier de théorie littéraire*.

Franco Moretti. 2000. Conjectures on world literature. *New Left Review*, 1.

Bruno Oberle. 2018. SACR: A Drag-and-Drop Based Tool for Coreference Annotation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Hardik Vala, David Jurgens, Andrew Piper, et Derek Ruths. 2015. Mr. bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 769–774, Lisbon, Portugal. Association for Computational Linguistics.