De l'annotation intégrale à l'analyse des réseaux de personnages : un modèle pour la REN dans les textes littéraires en français

Motasem Alrahabi*— Arthur Amalvy— Vincent Labatut**— Perrine Maurel***

* *Sorbonne Université, ObTIC*
** *Avignon Université, Laboratoire Informatique d'Avignon – UPR 4128*

RÉSUMÉ. *Les corpus textuels sont au cœur des humanités numériques, mais leur exploitation à grande échelle reste limitée par l'absence d'outils adaptés, notamment pour la reconnaissance d'entités nommées. Pour y remédier, nous présentons un corpus de sept romans français du* XIX*e siècle, annotés intégralement. Cette ressource permet d'entraîner et d'évaluer des outils robustes aux spécificités littéraires. Nous proposons un modèle de REN basé sur CamemBERT, performant sur ce type de textes, et montrons son intérêt via l'extraction de réseaux de personnages, ouvrant de nouvelles perspectives sur les dynamiques narratives. Les ressources sont librement accessibles en ligne.*

MOTS-CLÉS : *reconnaissance d'entités nommées, corpus littéraires annotés, réseaux de personnages, apprentissage supervisé.*

TITLE. *From complete annotation to character network analysis: A NER model for literary texts in French*

ABSTRACT. *Textual corpora are central to Digital Humanities, but their large-scale use remains limited by the lack of suitable tools, particularly for Named Entity Recognition. To address this, we present a fully annotated corpus of seven 19th-century French novels. This resource enables the training and evaluation of tools robust to literary specificities. We propose a CamemBERT-based NER model, effective on such texts, and demonstrate its relevance through character network extraction, opening new perspectives on narrative dynamics. All resources are freely available online.*

KEYWORDS: *Named entity recognition, annotated literary corpora, Network analysis, supervised learning.*

1. Introduction

La constitution de corpus textuels dans le domaine des humanités numériques (HN) est en plein essor (Ehrmann *et al.*, 2023). À mesure que ces ressources gagnent en volume et en complexité, leur exploration nécessite des instruments d'analyse performants, conçus pour accomplir une variété de tâches en traitement automatique des langues (TAL). Parmi celles-ci, la reconnaissance d'entités nommées (REN), associée à des tâches comme la résolution de coréférence, joue un rôle clé dans l'interprétation et la valorisation des corpus dans différents contextes. En littérature, des problématiques fondamentales liées aux personnages, aux lieux et aux dimensions temporelles occupent une place centrale et ouvrent la voie à de nombreuses applications : identification des noms de lieux (Berragan *et al.*, 2023), analyse des réseaux de personnages (Labatut et Bost, 2019), recherche documentaire guidée par entités (Guo *et al.*, 2009), enrichissement sémantique par lien aux bases de connaissance (Labusch et Neudecker, 2022), ou encore reconstruction de biographies à partir de mentions dispersées dans les corpus (Fokkens *et al.*, 2018). Il faut cependant noter que les corpus textuels issus des HN pourraient être davantage valorisés grâce à des outils d'analyse adaptés (Egloff et Picca, 2020). Ceux-ci, étant généralement issus de l'application de méthodes d'apprentissage supervisé, souffrent d'un défaut de données annotées. Si certaines tâches bénéficient déjà de corpus ou de modèles adaptés pour le français — comme le corpus Democrat pour la résolution de coréférences (Landragin, 2021), ou celui de Durandard *et al.* (2023) pour la détection du discours direct — d'autres, en revanche, restent moins couvertes. C'est notamment le cas de la REN et de la résolution d'alias, lacunes que nous nous proposons de combler dans cet article.

L'une des principales difficultés de la REN réside dans la variabilité des performances des modèles selon la langue, l'époque ou le domaine des textes analysés (Liu *et al.*, 2021), une difficulté accentuée dans les corpus littéraires et historiques, qui se caractérisent par une grande diversité de types de texte, un bruit important lié à l'OCR, des variations diachroniques des conventions d'écriture, et un manque de ressources annotées adaptées. Les meilleures approches existantes (Ehrmann *et al.*, 2023 ; Yamada *et al.*, 2020) obtiennent d'excellents résultats sur des documents contemporains courts, comme les textes journalistiques ou encyclopédiques, qui constituent l'essentiel de leurs données d'apprentissage. Cependant, cette efficacité diminue sensiblement dans des contextes plus complexes, tels que les textes littéraires, notamment narratifs, qui sont marqués par une grande richesse et diversité stylistique (Silva et Moro, 2024). Cette difficulté est accentuée par la variation diachronique, un aspect souvent négligé qui exige des outils capables de traiter les particularités des textes historiques (Ehrmann *et al.*, 2023). Par ailleurs, les défis sont encore plus importants pour les langues peu documentées ou éloignées de l'anglais, pour lesquelles les ressources et les outils sont largement insuffisants. Pour surmonter ces obstacles, il est crucial de concevoir des corpus et des modèles adaptés, permettant ainsi d'améliorer les performances des systèmes REN tout en répondant à des besoins plus spécifiques et diversifiés en TAL.

Notre étude, qui vise à combler ce manque, apporte une contribution majeure en proposant un nouveau corpus de référence (ou *gold standard*) en français, annoté en entités nommées (EN), ainsi qu'un nouveau modèle de REN spécifiquement adapté aux documents littéraires. Notre corpus se compose de romans français du XIX^{e} siècle, annotés dans leur intégralité, sans échantillonnage. Nous adaptons un guide d'annotation existant pour créer ces ressources, ce qui nous permet d'entraîner un modèle de REN et d'évaluer la pertinence des architectures neuronales sur ce type de documents. En termes d'application, nous démontrons l'intérêt de ce modèle en l'utilisant pour extraire des réseaux de personnages de romans (section 4.3). Pour les besoins de cette analyse ainsi que pour remédier au manque de jeux de données disponibles, nous ajoutons à notre corpus une couche d'annotation ciblant la tâche de résolution d'alias, c.-à-d. le fait d'identifier les différentes expressions faisant référence à un même personnage.

L'article est organisé en quatre parties. Nous faisons tout d'abord le point sur les principales contributions concernant la REN dans le domaine du TAL (section 2). Nous présentons ensuite notre corpus annoté en EN, en décrivant la méthode suivie et les difficultés rencontrées (section 3). La partie suivante porte sur l'entraînement de notre modèle de REN et sur son application à l'extraction de réseaux de personnages (section 4). Nous concluons avec une discussion des apports du modèle proposé, mais aussi de ses limites (section 5), et en suggérant des pistes d'amélioration pour les travaux futurs.

2. REN et résolution d'alias : un bref état de l'art

Dans cette section, nous abordons dans un premier temps les méthodes existantes conçues pour reconnaître les EN dans des textes littéraires (section 2.1), avant de fournir une brève description des corpus disponibles pour entraîner et pour évaluer des méthodes automatiques à effectuer cette tâche (section 2.2). Nous traitons également du cas de la résolution d'alias dans la section 2.3.

2.1. *Approches et méthodes de REN*

La REN dans les textes littéraires est complexifiée par un certain nombre de facteurs spécifiques qui les distinguent suffisamment d'autres domaines pour avoir un effet mesurable sur les performances. On peut notamment mentionner la diversité linguistique et stylistique (archaïsmes, néologismes, figures de style), les références implicites (périphrases, titres, surnoms), les contextes narratifs complexes (variations selon les narrateurs ou les dialogues), l'ambiguïté (noms communs utilisés comme noms propres), ainsi que des entités enrichies ou spécifiques influencées par le multilinguisme et les références culturelles (références mythologiques, objets symboliques, entités fictives) (van Dalen-Oskam *et al.*, 2014 ; Dekker *et al.*, 2019 ; Labatut et Bost, 2019).

Plusieurs méthodes ont donc été développées pour la REN dans les corpus littéraires, allant des approches à base de règles et d'expressions régulières aux modèles d'apprentissage automatique supervisés et non supervisés. Les méthodes à base de règles, qui reposent sur des répertoires géographiques (*gazetteers*) et sur des règles spécifiques au domaine, permettent de reconnaître les entités (Moncla *et al.*, 2017), mais leur manque de généralisation constitue une limite importante. Les approches d'apprentissage automatique traditionnel ont établi des bases solides pour la REN (Kogkitsidou et Gambette, 2020). Plus récemment, les méthodes d'apprentissage profond, qui intègrent les plongements lexicaux contextuels et l'apprentissage par transfert, surpassent nettement les autres (Sprugnoli, 2018), mais elles nécessitent en contrepartie de grands corpus annotés pour leur entraînement. L'arrivée des *Transformers* préentraînés, notamment avec des modèles comme BERT et GPT, a marqué une avancée majeure, offrant une meilleure prise en compte du contexte et des spécificités syntaxiques des textes littéraires (Devlin *et al.*, 2019). En français, des modèles comme CamemBERT (Martin *et al.*, 2020) et FlauBERT (Le *et al.*, 2020) sont adaptés aux particularités de la langue, tandis qu'en anglais, des modèles préentraînés tels que BERT et RoBERTa (Liu *et al.*, 2019) sont souvent affinés sur des corpus littéraires.

L'évolution de ces méthodes, des règles aux modèles *Transformers* préentraînés, illustre des avancées majeures, mais les textes littéraires continuent à poser des défis à surmonter tels que la longueur des documents (Amalvy *et al.*, 2023), la difficulté à reconnaître certains noms particuliers (Dekker *et al.*, 2019), la richesse stylistique (Silva et Moro, 2024), le manque de données ou encore l'ambiguïté de certaines entités. Pour répondre à ces enjeux, Ehrmann *et al.* (2023) proposent des stratégies telles que l'apprentissage par transfert, l'apprentissage actif, l'utilisation de modèles de langage historiques, ainsi que la normalisation et l'entraînement sur des données adaptées au contexte.

2.2. *Corpus annotés en REN*

Un petit nombre de travaux s'intéressent à l'annotation de données en REN dans les textes littéraires. Bamman *et al.* (2019) introduisent LitBank, reposant sur des extraits de 100 œuvres du domaine public en anglais (environ 211 000 tokens au total), et couvrant six catégories d'entité définies par ACE 2005 (Linguistic Data Consortium, 2005). De même, Dekker *et al.* (2019) ont créé OWTO, une ressource basée sur 40 chapitres extraits d'autant de romans en anglais (300 phrases par roman), et axée sur les personnages. Dans Frontini *et al.* (2020), des extraits de 400 mots ont été annotés dans chacune des neuf langues du corpus ELTeC (European Literary Text Collection) (Schöch *et al.*, 2021). Les étiquettes comprenaient des catégories telles que les gentilés, les professions, les œuvres d'art, les personnes, les lieux, les événements et les organisations, mais elles ont été simplifiées par la suite pour éviter les annotations imbriquées et les chevauchements (Frontini *et al.*, 2020). Alrahabi *et al.* (2024) ont proposé un corpus annoté d'extraits de trois romans français du XIX^{e} siècle,

couvrant trois catégories d'entité (personnage, lieu et divers) avec une granularité fine pour chaque catégorie. Enfin, le corpus Travel Writings (Sprugnoli, 2018) couvre 38 récits de voyage en anglais (de 1850 à 1940) avec 100 000 tokens annotés, et se concentre sur les entités géographiques.

En complément des corpus annotés manuellement, certains projets utilisent des annotations automatiques, souvent sur des volumes plus importants. Par exemple, BDCamões (Grilo *et al.*, 2020) est un vaste corpus portugais OCRisé (4 millions de mots), couvrant 14 genres littéraires sur cinq siècles, avec des annotations automatiques pour les lieux, organisations, œuvres, événements et entités diverses. En français, GeoNER (Kogkitsidou et Gambette, 2020) se concentre sur trois textes littéraires des XVI^{e} et XVII^{e} siècles, explorant l'impact de la normalisation historique sur la REN géographique.

Du point de vue méthodologique, il est à noter que les guides d'annotation proposés concernent toujours des textes courts ou limités. Jørgensen *et al.* (2020) forment leur guide à partir de l'annotation d'un corpus d'environ 600 000 tokens (soit moitié moins que notre corpus), composé principalement de textes journalistiques, mais aussi de rapports gouvernementaux et de blogs. Le guide ACE08 (Li, 2008) concerne deux corpus linguistiques faits d'extraits de blog, d'actualité, de conversation à la radio ou au téléphone, qui sont des formats courts par nature. Ainsi, le corpus anglophone est composé de six sous-corpus d'une longueur moyenne de 44 166 mots, et le corpus arabophone est composé de trois sous-corpus d'une longueur moyenne de 35 000 mots.

Ces choix méthodologiques illustrent une tendance générale : malgré la diversité des travaux existants, une limitation commune réside dans l'échantillonnage partiel des œuvres littéraires, où seules des portions — parfois très courtes — sont annotées. Cela restreint leur exploitation en termes d'évaluation, car les modèles existants peuvent généralement prendre en entrée l'entièreté de ces extraits. Cependant, les limites de taille de contexte des modèles ou le manque de ressources de calcul mènent souvent à l'application d'une fenêtre glissante quand il s'agit de longs documents. Cette pratique est préjudiciable à la performance de la tâche de REN car les modèles n'ont pas toujours à disposition suffisamment de contexte à l'inférence (Amalvy *et al.*, 2023), ce que ne permettent pas de mesurer des corpus composés de courts documents. Afin de surmonter cette limitation, nous constituons un corpus de romans français entièrement annotés en EN, particulièrement utile pour l'analyse des textes littéraires longs.

2.3. *Résolution d'alias*

La tâche de résolution d'alias consiste à extraire, pour un personnage, tous les alias qui y font référence dans le texte. Dans cet article, nous considérons comme alias d'un personnage toute mention qui serait annotée lors de la phase d'annotation de REN : nous excluons donc les mentions génériques comme les pronoms. Cette

tâche peut être considérée comme une version de la résolution de coréférences limitée aux personnages. Dans le cas de romans complets, cette tâche est considérée comme très difficile, et les approches neuronales récentes se heurtent au coût computationnel du traitement de textes aussi longs (Guo *et al.*, 2023 ; Gupta *et al.*, 2024). La résolution d'alias hérite de ces difficultés, et plusieurs approches se basent donc plutôt sur un ensemble de règles (Vala *et al.*, 2015 ; Cuesta-Lazaro *et al.*, 2022), utilisant une première passe de résolution de coréférences comme une entrée du système. Il s'agit d'une tâche moins étudiée que la REN, et plusieurs auteurs ne s'y intéressent que pour résoudre des problèmes de plus haut niveau, comme l'attribution du locuteur (Cuesta-Lazaro *et al.*, 2022) ou la classification de personnages (Bamman *et al.*, 2014).

La résolution d'alias constitue donc une tâche de bas niveau importante qui précède l'exploitation de textes littéraires lorsqu'on s'intéresse à leurs personnages. Au vu de son importance, nous enrichissons les annotations de notre corpus de REN en relevant également les alias des personnages, ce qui constitue une première pour des romans complets en français. Ces annotations permettront de développer et d'évaluer des systèmes de résolution d'alias à une échelle réaliste.

3. Constitution du corpus

Cette section décrit les méthodes que nous mettons en œuvre pour élaborer notre corpus : sélection des textes (section 3.1), règles appliquées (section 3.2) et déroulement du processus d'annotation (section 3.3). Elle fournit également une brève évaluation du corpus annoté (section 3.4) et revient sur les difficultés rencontrées (section 3.5).

3.1. *Romans sélectionnés*

Notre étude porte principalement sur les romans français du XIX^{e} siècle, un genre qui s'est imposé comme modèle en raison de sa structure narrative distincte, articulée autour des figures de l'auteur, du narrateur et des personnages. Ces romans, caractérisés par une narration souvent au passé et à la troisième personne, conjuguent récit et description. Cette alternance permet aux auteurs de faire avancer l'intrigue tout en marquant des pauses pour installer le décor, ancrant ainsi l'action dans un environnement qui reflète la société de l'époque. Un autre avantage concret de ce choix est que les œuvres produites à cette période sont libres de droits, ce qui facilite le partage de romans entiers sans soulever de problème juridique.

Notre corpus se compose de sept romans, sélectionnés pour leur richesse en personnages et en intrigues, ainsi que pour leur représentativité des divers courants littéraires. Ces œuvres suivent les trois grandes étapes du processus de canonisation décrites par Evans (2000) : elles sont sélectionnées par les critiques pour leur mérite esthétique, validées par les enseignants et universitaires pour l'enseignement,

puis pérennisées par les éditeurs à travers des rééditions successives. Les textes sélectionnés sont listés et décrits dans le tableau 1.

Roman	**Auteur**	**Publication**	**Tokens**
Les Trois Mousquetaires	Alexandre Dumas	1849	294 989
Le Rouge et le Noir	Stendhal	1854	216 445
Eugénie Grandet	Honoré de Balzac	1855	80 659
Germinal	Émile Zola	1885	220 273
Bel-Ami	Guy de Maupassant	1901	138 156
Notre-Dame de Paris	Victor Hugo	1904	221 351
Madame Bovary	Gustave Flaubert	1910	148 861

TABLEAU 1. *Liste des romans constituant notre corpus avec, pour chacun, l'auteur, l'année de publication, et le nombre de tokens*

Aucun prétraitement ni normalisation préalable n'a été appliqué aux documents collectés. Totalisant 1 320 734 tokens, l'ensemble des œuvres provient du site Wikisource[1]. Les URL correspondantes sont répertoriées en annexe (tableau 9).

3.2. *Règles d'annotation*

Le travail décrit par Sims et Bamman (2020) illustre la complexité accrue de l'annotation des EN dans les longs documents, en montrant que l'attribution de locuteurs et la résolution de coréférences posent des défis similaires. La fréquence élevée des mentions et la complexité narrative renforcent la nécessité de modèles adaptés pour les textes littéraires, avec des performances élevées pour garantir la cohérence dans ces textes. Ce constat est confirmé dans d'autres études telles que (Amalvy *et al.*, 2023 ; Tay *et al.*, 2021 ; Labatut et Bost, 2019). La dernière mentionne que la prose littéraire, plus complexe que la prose journalistique sur laquelle la plupart des modèles sont habituellement entraînés, affecte la performance des méthodes génériques pour des tâches NLP variées, comme la modélisation de l'intrigue ou la détection de personnages. Les auteurs soulignent également que les œuvres de fiction possèdent des caractéristiques spécifiques rendant inefficaces certains outils standards de traitement, ce qui renforce la nécessité de méthodes adaptées pour garantir la cohérence et la précision dans l'annotation de textes littéraires longs.

Pour l'annotation de notre corpus en EN, nous choisissons d'utiliser un guide récent, proposé dans le cadre du projet Universal NER[2] (Mayhew *et al.*, 2024). Nous l'avons sélectionné en raison de de son caractère très généraliste, afin de nous permettre de faire évoluer progressivement nos directives d'annotation et de les

1. https://fr.wikisource.org/wiki/Wikisource:Accueil
2. https://www.universalner.org/guidelines/

adapter aux besoins spécifiques de notre corpus. Universal NER, qui est lui-même basé sur le projet Universal Dependencies (UD)[3], propose un guide d'annotation multilingue et un ensemble de données standardisées, visant à établir des bases cohérentes pour la REN dans une dizaine de langues typologiquement variées. Dans sa version actuelle, il couvre notamment l'anglais, le chinois, le russe, le danois et le serbe, mais il n'inclut pas encore le français. Selon le guide d'Universal NER, tous les noms de personnes (réelles ou fictives), d'organisations et de lieux doivent être annotés, à condition que le mot soit ou contienne un nom propre et qu'il ait une référence unique pointant vers une entité spécifique et qui soit constante dans le temps. Le jeu de balises comprend trois types principaux d'entité : les *Personnes* (PERS), les *Lieux* (LOC) et les *Organisations* (ORG), ainsi qu'une catégorie supplémentaire, *Autre* (OTHER) pour les entités jugées pertinentes mais ne correspondant pas à ces types principaux. Nous adaptons ces catégories et ces directives à notre corpus, comme détaillé ci-après.

3.3. *Déroulement de l'annotation*

L'équipe, composée d'un coordinateur de projet et de trois annotateurs, a mené une campagne d'annotation en trois phases reposant sur la plateforme libre d'étiquetage de données Label Studio[4].

Première phase : annotation exploratoire. Nous annotons un échantillon selon le guide, chaque passage étant traité par deux annotateurs. Nous classons dans les catégories PERS, LOC et ORG tout ce qui est explicitement défini et sans ambiguïté. Par exemple, dans le cas de PERS, cela inclut les humains, les animaux (« *— Qu'est-ce que cette Djali ? — C'est la chèvre.* »[5]), les figures mythologiques ou religieuses (« *le tendon d'Achille* », « *Dieu* »), les figures personnifiées (« *Le suisse, alors, se tenait sur le seuil, au milieu du portail à gauche, au-dessous de la Marianne dansant* ») et les figures fictives (« *les accents divins du désespoir de Caroline dans le Matrimonio segreto le firent fondre en larmes* »).

Parmi les premières consignes d'annotation que nous avons jugées importantes d'ajouter au guide, citons :

– annoter la mention continue la plus englobante d'une EN (« *la compagnie des gardes de M. Essart* » [ORG][6]) ;

3. https://universaldependencies.org/
4. https://github.com/HumanSignal/label-studio
5. Dans cet exemple comme dans l'ensemble de cet article, le passage souligné désigne l'extrait pertinent dans son contexte au sein de l'œuvre d'origine.
6. Certaines catégories, comme les dates ou les monnaies, ne sont pas prises en compte dans notre guide. Si la mention la plus englobante appartient à l'une de ces catégories (p. ex. « *soie Louis XVI* »), elle n'est pas annotée. On examine alors la portion la plus restreinte pour déterminer si une annotation est nécessaire.

– ne pas annoter les EN imbriquées (p. ex. « *M. Essart* » dans l'exemple précédent) ;

– annoter les descripteurs qui accompagnent les EN (p. ex. « *la rue Boursault* », « *le prince de Guerche* »).

Deuxième phase : concertation des annotateurs. Cette étape permet à l'équipe d'analyser les difficultés et les désaccords de la première phase, et d'adapter le guide à nos besoins. Parmi ces ajustements, citons notamment :

– annoter les appellations et les titres nobiliaires, car le statut social des différents personnages est souvent central au récit (p. ex. « *Sa Majesté le roi Charles X* [PERS] ») ;

– annoter les déterminants et les prépositions (p. ex. « *le roi* ») car ils permettent parfois de préciser le genre de la personne[7] ;

– nous ajustons l'étiquette (OTHER), définie dans le guide d'Universal NER pour inclure des catégories comme les nationalités et les langues, afin de mieux répondre à nos besoins spécifiques. Nous l'utilisons pour annoter des EN associées aux trois catégories principales, mais qui sont *indéterminées* ou *ambiguës*. Pour nous, la notion d'*indéterminé* concerne les entités qui renvoient à une référence spécifique dans le roman, mais qui, une fois sorties de ce contexte, ne renvoient plus à une personne, à un lieu ou à une organisation précise[8]. Cette catégorie inclut les noms communs représentatifs (« *Tout le monde s'inclina vers le Patron* », « *la bohémienne* », « *les environs de Paris* », « *Français contre Français* », etc.). La notion d'*ambigu*, quant à elle, regroupe les entités que le contexte ne permet pas de classer clairement dans l'une de nos trois catégories. Dans l'exemple « *être responsable par-devant Notre-Dame la Critique* », le contexte manque de précisions pour trancher, et la catégorisation de *Notre-Dame* reste ambiguë : bien qu'il s'agisse de la cathédrale, elle est ici personnifiée.

Troisième phase : annotation finale. Le guide d'annotation, désormais adapté, est appliqué à l'ensemble du corpus. Comme dans la première étape, chaque phrase est traitée par deux annotateurs pour garantir la qualité et la cohérence.

Le tableau 2 présente le nombre total d'entités annotées par roman, ainsi que par catégorie, réalisées par les trois annotateurs. Le corpus produit est publiquement accessible en ligne[9].

7. En français, les particules définies (*le, la, les, l'*) indiquent le genre, le nombre et parfois l'élision, ce qui les rend cruciales pour l'analyse, contrairement à l'anglais où *the* est invariable.
8. L'un des avantages de notre démarche est que les annotateurs disposent d'une vision globale du contexte du roman, contrairement à certains travaux où seuls des échantillons sont annotés.
9. https://github.com/obtic-sorbonne/7-romans

Roman	Nombre de mentions d'entité					Tokens par
	PERS	LOC	ORG	OTHER	**Total**	**mention d'entité**
Les Trois Mousquetaires	8 583	962	45	255	9 845	1,90
Le Rouge et le Noir	4 351	806	39	78	5 274	1,89
Eugénie Grandet	1 389	187	13	34	1 623	1,65
Germinal	3 439	838	195	40	4 512	1,43
Bel-Ami	1 773	400	25	46	2 244	1,77
Notre-Dame de Paris	3 631	1 409	17	94	5 151	2,01
Madame Bovary	2 004	403	11	70	2 488	1,54
Total pour le corpus	25 170	5 005	345	617	31 137	1,80

TABLEAU 2. *Nombre de mentions d'entité par roman et par type, et moyenne du nombre de tokens qui les composent*

3.4. *Accord inter-annotateur*

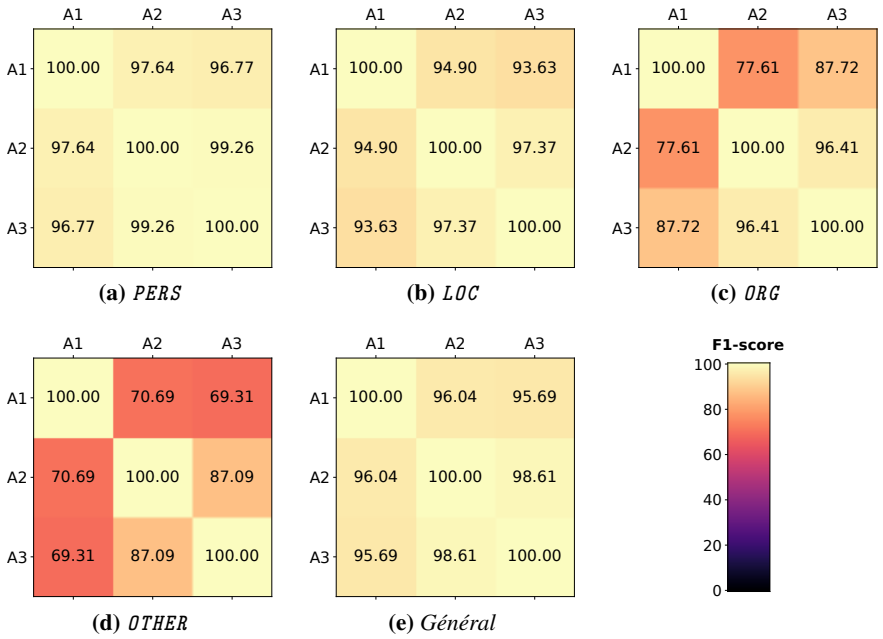
Nous avons fait appel à trois annotateurs aux profils très diversifiés. A1 correspond à un annotateur débutant, au profil mixte entre littérature et informatique. A2 correspond à un annotateur semi-expérimenté, au profil littéraire, qui avait déjà participé à des campagnes d'annotation à plusieurs mains. A3 correspond à un annotateur débutant, au profil informatique.

Nous mesurons l'accord inter-annotateur en calculant le F1-score au niveau des entités, et le coefficient Kappa de Cohen (Cohen, 1960) au niveau des tokens. Nous calculons Kappa en nous limitant aux tokens identifiés comme appartenant à une entité par au moins un annotateur, ignorant ceux qui ne sont annotés par personne. Cette méthode permet d'éviter de surestimer l'accord, en excluant les tokens qui n'ont aucune probabilité d'appartenir à une entité, ces derniers étant majoritaires (Brandsen *et al.*, 2020).

Classe	F1-score	Kappa
PERS	97,89	93,71
LOC	95,30	86,45
ORG	87,25	75,57
OTHER	75,70	47,70
Moyenne	89,04	75,86
Global	96,78	93,35

TABLEAU 3. *Accord inter-annotateur par type d'entité. Les deux dernières lignes indiquent respectivement la moyenne par classe et le score obtenu en considérant toutes les entités sans distinction de classe.*

En considérant toutes les entités sans distinction de classe, nous obtenons un F1-score de 96,78 et un Kappa de Cohen de 93,35, ce qui confirme la cohérence de nos annotations. Mais ces scores agrégés peuvent masquer des disparités, aussi nous les décomposons selon deux dimensions : le type d'entité et l'annotateur. Le tableau 3 détaille l'accord inter-annotateur par classe d'entité : bien que l'accord global soit excellent, la classe `OTHER` semble poser davantage de difficultés, comme nous le détaillons dans la section suivante. La figure 1 donne le détail de l'accord par paire d'annotateurs, et montre que les scores plus faibles pour `ORG` et `OTHER` sont dus à un désaccord plus prononcé du premier annotateur.



(a) *PERS* **(b)** *LOC* **(c)** *ORG*

(d) *OTHER* **(e)** *Général*

FIGURE 1. *Accord inter-annotateur exprimé en F1-score, pour chaque paire d'annotateurs et type d'EN*

3.5. *Difficultés rencontrées*

L'annotation des EN présente plusieurs défis, tenant à la diversité des catégories, des styles littéraires et des contextes d'utilisation. Nous les abordons en harmonisant les annotations et en distinguant deux types principaux de désaccord.

Le premier type regroupe les désaccords faciles à résoudre, qui sont dus à des oublis, à des incertitudes ou à des erreurs d'interprétation. Voici quelques exemples :

– dans la phrase « *Je fais le Sénat au Salut, et, de temps en temps, des chroniques littéraires pour la Planète* », l'entité *La Planète* a été annotée à tort par un annotateur comme `LOC` au lieu d'`ORG` ;

– « *On les nommait à la Chambre la bande à Walter* » : après concertation, nous optons pour `ORG` plutôt que `LOC` ;

– dans la phrase « *Sera à la Saint-Jean gelée,* », le nom propre *Saint-Jean* pourrait porter à confusion, car il s'agit, avec le déterminant, d'un évènement empruntant le nom d'une personne par antonomase : nous annotons la mention la plus restreinte (*Saint-Jean*) en `PERS`.

Le second type concerne les désaccords de fond, plus difficiles à résoudre, car complexes et porteurs d'une certaine ambiguïté. Ils relèvent fréquemment de la catégorie `OTHER`, créée et ajustée lors de la deuxième phase. Voici quelques exemples :

– dans l'énoncé : « *Elle s'empara d'un prie-Dieu et s'agenouilla.* », l'entité nommée est relevée par tous les annotateurs, mais sa catégorisation suscite un débat. En effet, *prie-Dieu* désigne un objet, mais une lecture littérale du terme évoque également une référence au personnage divin (*Dieu*), soulignant l'usage de l'objet pour prier ce dernier ;

– les notions « indéterminé » et « ambigu » de la catégorie `OTHER` peuvent varier en fonction de l'interprétation de l'annotateur, comme dans le cas de « *l'affaire Morel* », qui pourrait être considérée comme `OTHER` ou comme `PERS`, auquel cas seul le mot *Morel* serait annoté ;

– l'inclusion ou l'exclusion de certains syntagmes dans les EN, comme *Français* ou *huguenots* dans l'exemple qui suit, soulève des questionnements, la capitalisation reflétant souvent des conventions typographiques plutôt qu'une intention emphatique de l'auteur : « *où Français devaient combattre contre Français [...] En effet, le sac de La Rochelle, l'assassinat de trois ou quatre mille huguenots [...]* » ;

– les variations stylistiques entre romans et auteurs ont profondément influencé l'évolution du guide d'annotation. Par exemple, Victor Hugo et Alexandre Dumas se distinguent par un usage particulièrement fréquent de noms communs pour désigner certains personnages, contrairement à d'autres auteurs du corpus qui privilégient généralement les noms propres. Ainsi, dans *Notre-Dame de Paris*, une approche d'annotation basée uniquement sur les noms propres ne couvre que 28 % des mentions d'*Esmeralda*, tandis que les désignations par « *l'égyptienne* » et « *la bohémienne* » représentent respectivement 45 % et 26 % de ces mentions.

Ces quelques cas illustrent la complexité de l'annotation des EN dans les textes littéraires, nécessitant une approche flexible et adaptée aux enjeux spécifiques de chaque contexte.

4. Modèle et application proposés pour la REN

Nous mettons à profit notre corpus en entraînant un modèle de REN sur celui-ci (section 4.1). Afin de mobiliser nos annotations dans l'entraînement du modèle, la majorité des désaccords de fond furent levés afin de ne présenter qu'une seule et même annotation au modèle : pour cela, nous avons privilégié l'une des interprétations au profit de l'autre en fonction du consensus atteint au sein de notre équipe. En ce qui concerne la seule exception que nous avons jugée irrésolvable, le modèle a reçu l'une des deux annotations de façon aléatoire. Compte tenu du manque de jeux de données littéraires en français, ce modèle entend avancer l'état de l'art en permettant une bonne performance d'inférence sur ce type de texte (section 4.2). Nous illustrons son utilité sur une tâche d'extraction de réseaux de personnages (section 4.3).

4.1. *Méthode d'entraînement*

Il existe principalement deux modèles encodeurs préentraînés à base de Transformers pour le français : CamemBERT (Martin *et al.*, 2020) et FlauBERT (Le *et al.*, 2020). Leurs architectures, basées sur BERT (Devlin *et al.*, 2019), sont très proches, de même que leurs performances sur différentes tâches.

Nous nous basons sur le modèle CamemBERT-base, de 110 millions de paramètres. Nous ajoutons à celui-ci une couche neuronale de classification, et formalisons la tâche de REN comme un problème de classification de tokens en nous basant sur le format BIO (Ramshaw et Marcus, 1995). Nous effectuons un ajustement fin de nos modèles sur notre corpus entier pendant un maximum de dix cycles d'apprentissage avec arrêt prématuré, avec un taux d'apprentissage de $1 \cdot 10^{-5}$. En pratique, notre stratégie d'arrêt prématuré revient à entraîner le modèle pendant environ trois cycles d'apprentissage. Pour traiter le déséquilibre entre les classes, nous pondérons notre fonction de coût en divisant le nombre d'exemples de la classe ayant le plus d'exemples (`PERS`) par le nombre d'exemples de chaque classe. Pendant l'entraînement et l'inférence, le modèle prédit les entités présentes dans chaque paragraphe sans contexte supplémentaire. Nous distribuons notre modèle librement en ligne[10].

4.2. *Évaluation du modèle*

Afin d'évaluer la performance de notre modèle, nous procédons par validation croisée en 7 blocs : nous produisons 7 modèles différents, chacun entraîné sur un ensemble unique de 5 livres, avec un jeu de développement d'un livre, et évalué sur le dernier livre restant (nous nommons cette configuration *romans complets*). Pour donner une idée de la meilleure performance possible du modèle, nous réalisons

10. http://huggingface.co/compnet-renard/camembert-base-literary-NER-v2

une expérience supplémentaire où nous utilisons un jeu de test correspondant à un septième de chaque roman, un jeu de développement correspondant à un autre septième, et un jeu d'entraînement correspondant à cinq septièmes de chaque roman (configuration *romans échantillonnés*). Cette expérience mesure la performance du modèle lorsque celui-ci a observé des entités du jeu de test au moment de l'apprentissage.

Nous utilisons la librairie `seqeval`[11] dans son mode par défaut, et rapportons nos scores en termes de précision, rappel et F1-score. Le tableau 4 recense les résultats que nous obtenons sur les différents romans, tandis que le tableau 5 détaille les résultats par types d'entité. Nous ne comparons pas notre modèle avec d'autres modèles entraînés sur des textes de domaines différents, car les différences de modalité d'annotation entre corpus les pénaliseraient injustement.

Roman	**F1**	**Précision**	**Rappel**
Les Trois Mousquetaires	71,15	68,49	74,02
Le Rouge et le Noir	88,97	85,23	93,04
Eugénie Grandet	88,56	85,58	91,77
Germinal	89,94	87,54	92,48
Bel-Ami	87,13	82,47	92,34
Notre-Dame de Paris	75,70	73,48	78,04
Madame Bovary	88,25	84,02	92,92
Romans complets	81,21	78,19	84,48
Romans échantillonnés	91,53	88,12	95,21

TABLEAU 4. *Performance de notre modèle de REN évalué par validation croisée. Les résultats par roman sont obtenus dans la configuration* romans complets*.*

Dans l'ensemble, nos résultats sont cohérents avec ceux obtenus précédemment par CamemBERT sur d'autres corpus de domaines différents en français, comme sur le jeu de données journalistique French Treebank (Martin *et al.*, 2020). Nous observons une forte différence entre les configurations *romans échantillonnés* et *romans complets* : le fait de donner accès au modèle à des entités du jeu de test à l'entraînement augmente sensiblement la performance (plus de 10 points de F1-score). Deux romans posent plus de difficultés : *Notre-Dame de Paris* et *Les Trois Mousquetaires*. Dans le cas des *Trois Mousquetaires*, nous analysons manuellement les résultats et observons que le modèle a de grandes difficultés à reconnaître le personnage principal *d'Artagnan*. Ce résultat rejoint les observations de Dekker *et al.* (2019), qui notent que les mentions de personnages qui comportent des caractères spéciaux (en l'occurrence, une apostrophe) sont plus difficiles à détecter. Pour ce qui est de *Notre-Dame de Paris*, les erreurs de précision comme de rappel se concentrent sur les descriptions définies, qui ne contiennent pas de nom propre (*« le roi »*, *« le cardinal »*, etc.). Du côté des résultats par classe, si on observe de bonnes

11. https://github.com/chakki-works/seqeval

Classe	F1	Précision	Rappel
PERS	83,51	81,89	85,20
LOC	81,80	78,69	85,17
ORG	55,74	42,39	81,34
OTHER	34,08	25,15	52,86

TABLEAU 5. *Performance de notre modèle de REN sur les différentes classes*

performances sur les classes PERS et LOC, les classes ORG et OTHER sont très difficiles à détecter. Cela peut s'expliquer par la difficulté inhérente du problème (l'accord inter-annotateur étant plus faible pour ces deux classes comme indiqué dans le tableau 3, mais aussi par le déséquilibre des classes : notre corpus comporte presque 73 fois plus d'entités PERS que d'entités LOC.

Il est à noter que ces résultats sous-estiment probablement les capacités du modèle final que nous mettons à disposition, car celui-ci est entraîné sur l'entièreté des sept romans du corpus.

4.3. *Application à l'extraction de réseaux*

Nous nous tournons maintenant vers une application de notre modèle de REN au domaine de la littérature, à travers la tâche d'extraction de réseaux de personnages (Labatut et Bost, 2019). Par conséquent, elle se concentre sur les EN de type PERS.

Les réseaux de personnages sont des graphes où les sommets représentent des personnages et les arêtes traduisent les relations entre eux. Ils sont particulièrement utiles pour de nombreuses applications. D'une part, ils offrent une visualisation claire des relations entre les personnages d'une œuvre littéraire, permettant ainsi de comprendre intuitivement la dynamique des interactions. D'autre part, ils constituent un outil précieux pour l'analyse littéraire, en apportant une nouvelle perspective sur la structure narrative et les relations interpersonnelles. Enfin, ces réseaux fournissent une modélisation d'un texte narratif long sous la forme compacte d'un graphe d'interactions. Ce type de modèle peut être exploité pour traiter automatiquement différentes tâches, telles que la classification de genre littéraire (Hettinger *et al.*, 2015 ; Holanda *et al.*, 2019), la segmentation d'histoires (Min et Park, 2016) ou l'alignement narratif (Amalvy *et al.*, 2024a).

Extraction des réseaux. L'outil Renard[12] (Amalvy *et al.*, 2024b) est un pipeline modulaire sous forme de librairie Python permettant d'extraire des réseaux de personnages (statiques ou dynamiques) à partir de textes narratifs, et d'analyser

12. https://github.com/CompNet/Renard/

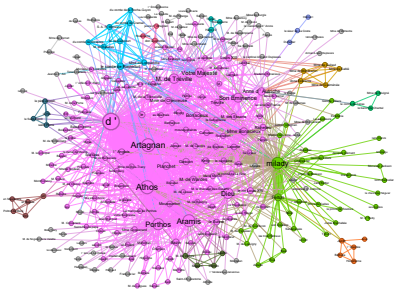
les relations entre personnages, offrant une visualisation de l'évolution des réseaux sociaux au fil du récit. Pour extraire ces réseaux, la librairie résout séquentiellement différentes tâches de TAL. La résolution de la tâche de REN permet d'abord de détecter les mentions des différents personnages dans le texte : il s'agit des entités `PERS` extraites par le système. Puis, la tâche de *résolution d'alias* permet de rassembler les mentions distinctes faisant référence à un même personnage. Finalement, la librairie extrait les interactions entre les personnages pour déterminer leurs relations. Ces relations peuvent être de différents types : cooccurrences, conversations, rencontres, etc. Compte tenu des annotations à notre disposition, nous nous concentrons sur l'extraction de réseaux de cooccurrences : nous considérons que deux personnages ont une interaction s'ils apparaissent à proximité l'un de l'autre dans le texte, ce qui constitue une approximation de leurs interactions réelles. Dans cette expérience, nous fixons arbitrairement le seuil de proximité à 32 tokens pour tous les romans. À plus long terme, un travail devra être mené pour proposer une méthode plus fondée visant à estimer ce paramètre. Afin de mesurer l'intérêt de notre modèle, nous proposons de l'utiliser lors de la phase de REN de `Renard`, puis de mesurer la qualité des réseaux extraits par rapport à des réseaux de référence.

Pour obtenir ces réseaux de référence, nous utilisons nos annotations en REN pour la classe `PERS` afin d'obtenir des mentions de référence. De plus, nous annotons manuellement les sept romans à notre disposition en résolution d'alias. Pour ce faire, nous nous inspirons du guide d'annotation du corpus littéraire *Novelties* (Amalvy et Labatut, 2024). Un annotateur (l'un des auteurs de cet article) a adapté les annotations existantes de ce corpus qui portent originellement sur les versions anglaises de nos romans. Nous notons que cette phase d'annotation nous a en outre permis de corriger de rares erreurs restantes dans nos annotations en REN, lorsque les alias à annoter semblaient incorrects. Nous indiquons le nombre de personnages ainsi annotés pour chaque roman dans le tableau 6.

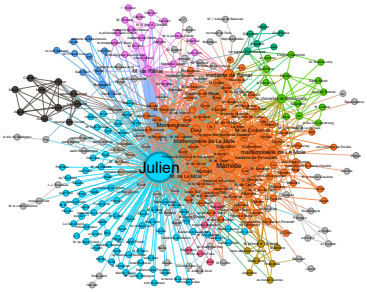
Roman	**Personnages**
Les Trois Mousquetaires	213
Le Rouge et le Noir	318
Eugénie Grandet	107
Germinal	102
Bel-Ami	150
Notre-Dame de Paris	536
Madame Bovary	175

TABLEAU 6. *Nombre total de personnages distincts annotés pour chaque roman*

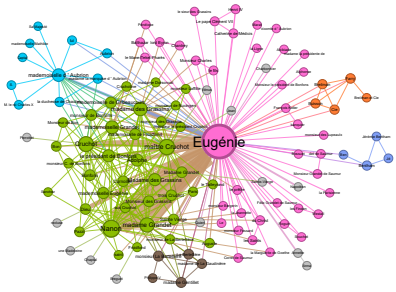
Les réseaux extraits en utilisant notre modèle de REN sont présentés dans la figure 2. Comme pour notre expérience d'évaluation de REN, nous utilisons le modèle dans un schéma de validation croisée pour éviter de surestimer la qualité des réseaux extraits. Le processus étant complètement automatique, on voit apparaître certaines erreurs. Par exemple, dans *Les Trois Mousquetaires*, la présence d'un



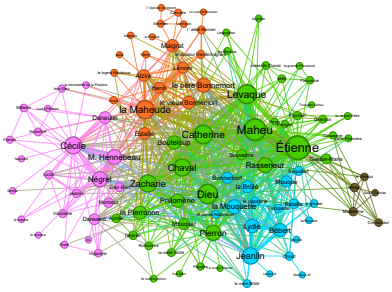
(a) *Les Trois Mousquetaires*



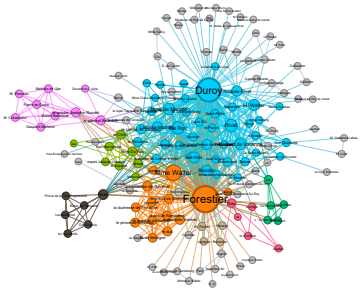
(b) *Le Rouge et le Noir*



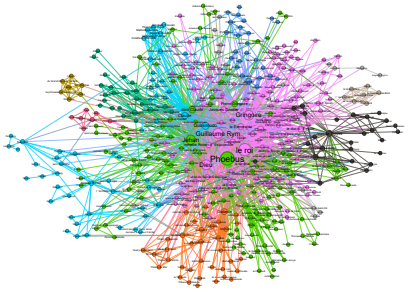
(c) *Eugénie Grandet*



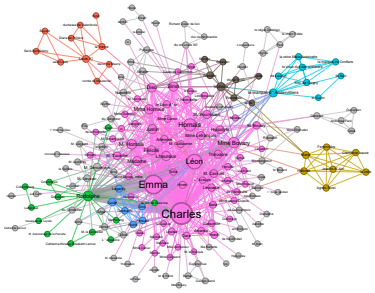
(d) *Germinal*



(e) *Bel-Ami*



(f) *Notre-Dame de Paris*



(g) *Madame Bovary*

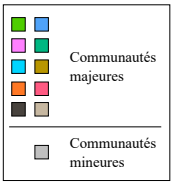


FIGURE 2. *Réseaux extraits par* `Renard` *en utilisant notre modèle de REN sur les différents romans du corpus. Le degré des sommets est indiqué par leur taille, tandis que l'épaisseur des arêtes représente leur poids. La couleur des sommets correspond à leur communauté (gris pour les communautés de trois sommets ou moins).*

sommet incorrect appelé *d'*, ou bien le *Cardinal de Richelieu* représenté par plusieurs sommets (*Son Éminence*, *Richelieu*). Nous revenons sur ce point plus loin, en proposant une évaluation quantitative de ces erreurs. La taille des sommets indique leur degré, leur couleur correspond à leur communauté, et l'épaisseur des arêtes représente leur poids. Les communautés constituent une partition de l'ensemble des sommets V. Nous les identifions au moyen de l'algorithme Girvan–Newman (Girvan et Newman, 2002), une méthode standard consistant à retirer itérativement les arêtes de plus grande intermédiarité. Dans le cadre de l'analyse de réseaux de personnages, les communautés sont souvent interprétées comme des sous-intrigues (Labatut et Bost, 2019). Visuellement, il est possible d'identifier plusieurs types de réseaux dans la figure 2. Certains romans sont clairement construits autour de leur protagoniste, qui centralise de nombreuses relations : c'est notamment le cas de *Le Rouge et le Noir* (Julien Sorel) et d'*Eugénie Grandet* (elle-même). D'autres sont au contraire structurés autour d'une opposition, comme *Les Trois Mousquetaires* (les mousquetaires contre Milady) et *Bel-Ami* (Duroy et Forestier). D'autres encore ont une nature plus chorale : *Germinal* et *Notre Dame de Paris* contiennent de très nombreux personnages, avec des interactions réparties de façon beaucoup plus uniforme. Certains réseaux mettent aussi en évidence l'importance de certaines relations dans l'histoire : le groupe des mousquetaires est très dense dans *Les Trois Mousquetaires*, le lien entre Julien et Mme de Rênal domine dans *Le Rouge et le Noir*, Eugénie et Nanon sont elles aussi fortement connectées dans *Eugénie Grandet*, de même qu'Emma et Charles dans *Madame Bovary*. Ces différentes observations sont cohérentes avec la lecture du roman, et indiquent intuitivement une bonne qualité des réseaux extraits.

Méthode d'évaluation. Afin de mesurer la qualité des réseaux extraits, nous utilisons des métriques qui permettent de mesurer la fidélité du graphe prédit $G_p = (V_p, E_p)$ en termes de sommets et d'arêtes par rapport à un graphe de référence $G_r = (V_r, E_r)$. Pour les sommets, nous utilisons la Précision, le Rappel et le F1 de sommet (Vala *et al.*, 2015) :

$$Pre_V = \max_{f_V} \frac{\sum_{u \in V_p} 1 - \frac{|u - f_V(u)|}{|u|}}{|V_p|} \quad [1]$$

$$Rec_V = \max_{g_V} \frac{\sum_{v \in V_r} [g_V(v) \cap v \neq v_\varnothing]}{|V_r|}, \quad [2]$$

où :

– chaque sommet de V_p et V_r représente l'ensemble des alias d'un personnage ;

– f_V est une fonction associant un sommet de G_p à un sommet de G_r ou au sommet nul $v_\varnothing$ si le personnage prédit n'a pas d'équivalent dans les personnages de référence ;

– similairement, g_V est une fonction associant un sommet de G_r à un sommet de G_p ou au sommet nul ;

– l'expression $[g_N(v) \cap v \neq v_\varnothing]$ vaut 1 si la condition interne est vraie, 0 sinon.

Le F1 de sommet est la moyenne harmonique des deux mesures Pre_V et Rec_V.

Pour ce qui est des arêtes, nous utilisons la Précision, le Rappel et le F1 d'arête (Amalvy *et al.*, 2025) :

$$Pre_E = \max_{f_E} \frac{\left|\{f_E(e) : e \in E_p\} \cap E_r\right|}{|E_p|} \quad [3]$$

$$Rec_E = \max_{g_E} \frac{\left|E_p \cap \{g_E(e) : e \in E_r\}\right|}{|E_r|}, \quad [4]$$

où f_E est une fonction associant une arête de E_p à une arête de E_r ou à l'arête nulle $e_\varnothing$ si une arête n'a pas d'équivalent dans les arêtes de référence, et *vice versa* pour g_E. Enfin, nous utilisons également les versions pondérées de ces mesures d'arête :

$$WPre_E = \max_{f_E} \frac{\sum_{e \in E_p} 1 - \left|w(f_E(e)) - w(e)\right|}{|E_p|} \quad [5]$$

$$WRec_E = \max_{g_E} \frac{\sum_{e \in E_r} 1 - \left|w(e) - w(g_E(e))\right|}{|E_r|}, \quad [6]$$

où $w(e)$ est le poids de l'arête e, normalisé en divisant par le poids maximal du graphe considéré. Ces versions pondérées permettent de prendre en compte la fidélité des poids des arêtes extraites. Par définition, elles ne peuvent atteindre que des scores inférieurs ou égaux à ceux de leurs homologues non pondérées.

Résultats. Nous indiquons la qualité des sommets extraits dans le tableau 7, et la qualité des arêtes dans le tableau 8. La qualité des sommets semble acceptable (de 52,46 à 64,08 $F1_V$), mais les métriques d'arête sont relativement basses (de 20,13 à 48,18 $F1_E$). Nous montrons dans l'annexe B qu'une partie importante des erreurs est due à l'algorithme de résolution d'alias plutôt qu'à la REN. La forte influence de cet algorithme explique, au moins en partie, que nous ne notions pas de corrélation entre la performance de la REN sur un roman et la qualité du réseau extrait de celui-ci. Si *Les Trois Mousquetaires* et *Notre-Dame de Paris* présentaient plus de difficultés en termes de détection d'entités, cela ne se ressent pas fortement sur nos mesures.

De manière générale, les résultats sont un peu en dessous de ceux reportés par Amalvy *et al.* (2025) sur le corpus Litbank (Bamman *et al.*, 2019 ; Bamman *et al.*, 2020), ce qui peut s'expliquer par la différence de longueur entre les textes étudiés. En effet, Litbank ne contient que des extraits d'environ 2 000 tokens, quand les romans de notre corpus peuvent être jusqu'à 150 fois plus longs. Cette longueur implique de plus nombreuses entités à détecter et d'alias à résoudre, ce qui multiplie les risques d'erreur, lors de l'étape de REN comme celle de résolution d'alias.

Il est à noter que, dans cette étude, nous ne prenons en compte que les mentions de personnages annotées dans notre étape de REN. Cela exclut des mentions plus génériques, comme certaines descriptions définies ou les pronoms, qui seraient détectables en utilisant un algorithme de résolution de coréférences. Ces mentions

Roman	$F1_V$	Pre_V	Rec_V
Les Trois Mousquetaires	59,10	47,08	79,34
Le Rouge et le Noir	63,52	57,83	70,44
Eugénie Grandet	52,46	43,37	66,36
Germinal	60,35	47,96	81,37
Bel-Ami	57,25	45,45	77,33
Notre-Dame de Paris	61,81	55,36	69,96
Madame Bovary	64,08	52,47	82,29

TABLEAU 7. *Qualité des sommets des réseaux extraits par un pipeline* `Renard` *utilisant notre modèle de REN*

peuvent avoir une forte influence sur les liens du réseau, comme montré par Amalvy *et al.* (2025).

Roman	$F1_E$	Pre_E	Rec_E	$WF1_E$	$WPre_E$	$WRec_E$
Les Trois Mousquetaires	34,12	27,27	45,56	33,48	26,76	44,72
Le Rouge et le Noir	26,38	25,39	27,45	26,20	25,22	27,26
Eugénie Grandet	20,13	19,69	20,58	19,04	18,63	19,47
Germinal	48,18	43,73	53,65	46,99	42,64	52,32
Bel-Ami	23,91	19,72	30,37	23,52	19,40	29,87
Notre-Dame de Paris	23,45	20,42	27,55	23,24	20,23	27,30
Madame Bovary	29,38	24,79	36,06	29,13	24,58	35,75

TABLEAU 8. *Qualité des arêtes des réseaux extraits par un pipeline* `Renard` *utilisant notre modèle de REN*

Ce travail sur les réseaux de personnages n'est qu'une étude préliminaire, qui appelle plusieurs approfondissements. Il s'agira tout d'abord de mener une analyse descriptive plus poussée, permettant une comparaison avec la littérature existante. Par ailleurs, des expérimentations seront nécessaires afin d'évaluer l'impact des erreurs présentes dans le réseau sur une tâche en aval (Labatut et Bost, 2019), telles que la recommandation ou l'identification des personnages principaux.

5. Conclusion et perspectives

Dans cet article, nous avons présenté un nouveau corpus en français constitué de romans annotés en entités nommées. Notre évaluation a révélé un très bon accord inter-annotateur, ce qui démontre la cohérence de notre processus d'annotation. À la différence de la plupart des jeux de données disponibles en ligne, cette annotation a été menée sur *l'intégralité* de ces romans, et non pas sur des extraits, ce qui en fait

une ressource très précieuse, voire unique, pour le développement et l'évaluation de méthodes de REN sur ce type de textes longs.

Afin d'illustrer cela, nous avons également entraîné un modèle de langage spécifiquement destiné à la REN dans les romans en français. Bien que performant, notre modèle a montré ses limites pour certains romans spécifiques, soulignant les défis posés par les variations stylistiques et narratives, et le besoin de pousser ce travail plus loin. Enfin, nous avons proposé une application de la REN dans les romans, prenant la forme d'une tâche d'extraction de réseaux de personnages. Celle-ci a nécessité de notre part l'ajout d'une nouvelle couche d'annotation à notre corpus, ciblant la tâche de résolution d'alias. Celle-ci est potentiellement plus complexe sur des livres complets, et aucun corpus existant ne proposait jusqu'ici d'annotations de ce type, qui plus est en français. Notre corpus, le code source de nos expériences ainsi que notre modèle sont mis à disposition de la communauté sous licence libre.

Nous identifions trois perspectives directes à notre travail. La première est d'ajuster le guide d'annotation afin d'améliorer encore l'accord inter-annotateur et l'homogénéité des EN. Le guide d'Universal NER a été adapté dans le cadre de notre étude pour intégrer les spécificités linguistiques du français, et pour traiter des cas complexes d'étiquetage des EN (notamment dans la catégorie `OTHER`). Une collaboration avec ce projet pourrait permettre d'affiner ce guide en y intégrant ces particularités tout en préservant sa cohérence multilingue. La seconde piste porte sur la création d'un corpus type *silver* de grande ampleur, qui permettrait de venir enrichir les données disponibles, et potentiellement d'améliorer ainsi l'apprentissage et donc les performances de notre modèle de REN. Une troisième perspective, à plus long terme, est d'appliquer notre méthode d'extraction de réseaux de personnages à un corpus beaucoup plus vaste. Ce serait l'occasion, d'une part, d'analyser les propriétés statistiques des réseaux construits à partir d'un échantillon statistiquement représentatif ; et, d'autre part, de mobiliser ces réseaux pour des tâches de prédiction, telles que la classification ou la régression appliquées aux œuvres littéraires.

6. Bibliographie

Alrahabi M., Brando C., Frontini F., Guide d'annotation manuelle d'entités nommées dans des corpus littéraires, Technical report, Labex OBVIL - L'Observatoire de la vie littéraire, 2024.

Amalvy A., Janickyj M., Mannion S., MacCarron P., Labatut V., « Interconnected Kingdoms : Comparing 'A Song of Ice and Fire' Adaptations Across Media Using Complex Networks », *Social Network Analysis and Mining*, vol. 14, p. 199, 2024a.

Amalvy A., Labatut V., Annotation Guidelines for Corpus Novelties : Part 2 — Alias Resolution, Technical report, Avignon Université, 2024.

Amalvy A., Labatut V., Dufour R., « The Role of Global and Local Context in Named Entity Recognition », *ACL*, p. 714-722, 2023.

Amalvy A., Labatut V., Dufour R., « Renard : A Modular Pipeline for Extracting Character Networks from Narrative Texts », *Journal of Open Source Software*, vol. 9, p. 6574, 2024b.

Amalvy A., Labatut V., Dufour R., « The Role of Natural Language Processing Tasks in Automatic Literary Character Network Construction », *31st International Conference on Computational Linguistics*, p. 8462-8473, 2025.

Bamman D., Lewke O., Mansoor A., « An Annotated Dataset of Coreference in English Literature », *12th Language Resources and Evaluation Conference*, p. 44-54, 2020.

Bamman D., Popat S., Shen S., « An annotated dataset of literary entities », *NAACL*, p. 2138-2144, 2019.

Bamman D., Underwood T., Smith N. A., « A Bayesian Mixed Effects Model of Literary Character », *ACL*, vol. 1, p. 370-379, 2014.

Berragan C., Singleton A., Calafiore A., Morley J., « Transformer based named entity recognition for place name extraction from unstructured text », *International Journal of Geographical Information Science*, vol. 37, nº 4, p. 747-766, 2023.

Brandsen A., Verberne S., Wansleeben M., Lambers K., « Creating a Dataset for Named Entity Recognition in the Archaeology Domain », *Twelfth Language Resources and Evaluation Conference*, p. 4573-4577, 2020.

Cohen J., « A Coefficient of Agreement for Nominal Scales », *Educational and Psychological Measurement*, vol. 20, p. 37-46, 1960.

Cuesta-Lazaro C., Prasad A., Wood T., « What does the sea say to the shore ? A BERT based DST style approach for speaker to dialogue attribution in novels », *ACL*, 2022.

Dekker N., Kuhn T., van Erp M., « Evaluating named entity recognition tools for extracting social networks from novels », *PeerJ Computer Science*, vol. 5, p. e189, 2019.

Devlin J., Chang M.-W., Lee K., Toutanova K., « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding », *NAACL*, p. 4171-4186, 2019.

Durandard N., Tran V. A., Michel G., Epure E., « Automatic Annotation of Direct Speech in Written French Narratives », *ACL*, vol. 1, p. 7129-7147, 2023.

Egloff M., Picca D., « WeDH - a Friendly Tool for Building Literary Corpora Enriched with Encyclopedic Metadata », *LREC*, p. 813-816, 2020.

Ehrmann M., Hamdi A., Linhares Pontes E., Romanello M., Doucet A., « Named Entity Recognition and Classification in Historical Documents : A Survey », *ACM Computing Surveys*, vol. 56, nº 2, p. 1-47, 2023.

Evans A. B., « Jules Verne and the French Literary Canon », *Jules Verne : Narratives of Modernity*, Liverpool University Press, chapter 2, p. 11-34, 2000.

Fokkens A., ter Braake S., Sluijter R., Arthur P. L., Wandl-Vogt E. (eds), *Proceedings of the Second Conference on Biographical Data in a Digital World 2017*, vol. 2119 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018.

Frontini F., Brando C., Byszuk J., Galleron I., Santos D., Stanković R., « Named Entity Recognition for Distant Reading in ELTeC », *CLARIN Annual Conference*, p. 37-41, 2020.

Girvan M., Newman M. E. J., « Community structure in social and biological networks », *Proceedings of the National Academy of Sciences*, vol. 99, nº 12, p. 7821-7826, 2002.

Grilo S., Bolrinha M., Silva J., Vaz R., Branco A., « The BDCamões Collection of Portuguese Literary Documents : A Research Resource for Digital Humanities and Language Technology », *12th Language Resources and Evaluation Conference*, p. 849-854, 2020.

Guo J., Xu G. X., Cheng X., « Named entity recognition in query », p. 267-274, 2009.

Guo Q., Hu X., Zhang Y., Qiu X., Zhang Z., « Dual Cache for Long Document Neural Coreference Resolution », *ACL*, p. 15272-15285, 2023.

Gupta T., Hatzel H. O., Biemann C., « Coreference in Long Documents using Hierarchical Entity Merging », *LaTeCH-CLfL*, p. 11-17, 2024.

Hettinger L., Becker M., Reger I., Jannidis F., Hotho A., « Genre Classification on German Novels », *DEXA*, p. 249-253, 2015.

Holanda A. J., Matias M., Ferreira S. M. S. P., Benevides G. M. L., Kinouchi O., « Character Networks and Book Genre Classification », *International Journal of Modern Physics*, 2019.

Jørgensen F., Aasmoe T., Ruud Husevåg A.-S., Øvrelid L., Velldal E., « NorNE : Annotating Named Entities for Norwegian », *12th Language Resources and Evaluation Conference*, p. 4547-4556, 2020.

Kogkitsidou E., Gambette P., « Normalisation of 16th and 17th Century Texts in French and Geographical Named Entity Recognition », *4th ACM SIGSPATIAL Workshop on Geospatial Humanities*, p. 28-34, 2020.

Labatut V., Bost X., « Extraction and Analysis of Fictional Character Networks : A Survey », *ACM Computing Surveys*, vol. 52, n° 5, p. 89, 2019.

Labusch K., Neudecker C., « Entity Linking for Cultural Heritage Collections with Wikidata », *DH 2022 : Digital Humanities Conference*, 2022.

Landragin F., « Le corpus DEMOCRAT et son exploitation. Présentation », *Langages*, vol. 224, n° 4, p. 11-24, 2021.

Le H., Vial L., Frej J., Segonne V., Coavoux M., Lecouteux B., Allauzen A., Crabbé B., Besacier L., Schwab D., « FlauBERT : Unsupervised Language Model Pre-training for French », *12th Language Resources and Evaluation Conference*, p. 2479-2490, 2020.

Li F., Automatic Content Extraction 2008 Evaluation Plan (ACE08), Technical report, Linguistic Data Consortium, 2008.

Linguistic Data Consortium, ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, Technical report, Linguistic Data Consortium, 2005.

Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V., « RoBERTa : A Robustly Optimized BERT Pretraining Approach », *arXiv*, 2019.

Liu Z., Xu Y., Yu T., Dai W., Ji Z., Cahyawijaya S., Madotto A., Fung P., « CrossNER : Evaluating Cross-Domain Named Entity Recognition », *AAAI Conference on Artificial Intelligence*, p. 13452-13460, 2021.

Martin L., Muller B., Ortiz Suárez P. J., Dupont Y., Romary L., de la Clergerie E., Seddah D., Sagot B., « CamemBERT : a Tasty French Language Model », *ACL*, p. 7203-7219, 2020.

Mayhew S., Blevins T., Liu S., Šuppa M., Gonen H., Imperial J. M., Karlsson B. F., Lin P., Ljubešić N., Miranda L. J., Plank B., Riabi A., Pinter Y., « Universal NER : A Gold-Standard Multilingual Named Entity Recognition Benchmark », *NAACL*, 2024.

Min S., Park J., « Network Science and Narratives : Basic Model and Application to Victor Hugo's Les Misérables », *Studies in Computational Intelligence*, vol. 644, p. 257-265, 2016.

Moncla L., Gaio M., Joliveau T., Le Lay Y.-F., « Automated Geoparsing of Paris Street Names in 19th Century Novels », *1st ACM Workshop on Geospatial Humanities*, p. 1-8, 2017.

Ramshaw L., Marcus M., « Text Chunking using Transformation-Based Learning », *3rd Workshop on Very Large Corpora*, 1995.

Schöch C., Patraș R., Erjavec T., Santos D., « Creating the European Literary Text Collection : Challenges and Perspectives », *Modern Languages Open*, vol. 1, n° 25, p. 1-19, 2021.

Silva M. O., Moro M. M., « PPORTAL_ner : An Annotated Corpus of Portuguese Literary Entities », *Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, p. 12927-12937, 2024.

Sims M., Bamman D., « Measuring Information Propagation in Literary Social Networks », *EMNLP*, p. 642-652, 2020.

Sprugnoli R., « Arretium or Arezzo ? A Neural Approach to the Identification of Place Names in Historical Texts », *CLiC-it*, vol. 2253 of *CEUR Workshop Proceedings*, p. 26, 2018.

Tay Y., Dehghani M., Abnar S., Shen Y., Bahri D., Pham P., Rao J., Yang L., Ruder S., Metzler D., « Long Range Arena : A Benchmark for Efficient Transformers », *ICLR*, 2021.

Vala H., Jurgens D., Piper A., Ruths D., « 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 », *EMNLP*, p. 769-774, 2015.

van Dalen-Oskam K., de Does J., Marx M., Sijaranamual I., Depuydt K., Verheij B., Geirnaert V., « Named entity recognition and resolution for literary studies », *Computational Linguistics in the Netherlands Journal*, vol. 4, p. 121-136, 2014.

Yamada I., Asai A., Shindo H., Takeda H., Matsumoto Y., « LUKE : Deep Contextualized Entity Representations with Entity-aware Self-attention », *EMNLP*, p. 6442-6454, 2020.

Annexes

A. Précisions sur le corpus

Le tableau 9 indique les URL de chaque roman du corpus proposé. Il s'agit des versions brutes qui ont été utilisées lors du processus d'annotation. La figure 3 montre la distribution des EN pour chaque roman et classe d'entité.

Roman	**URL Wikisource** – `https://fr.wikisource.org/wiki/...`
Les Trois Mousquetaires	`.../Les_Trois_Mousquetaires/Texte_entier`
Le Rouge et le Noir	`.../Le_Rouge_et_le_Noir/Texte_entier`
Eugénie Grandet	`.../Eugénie_Grandet`
Germinal	`.../Germinal/Texte_entier`
Bel-Ami	`.../Bel-Ami/Édition_Ollendorff,_1901/Texte_entier`
Notre-Dame de Paris	`.../Notre-Dame_de_Paris/Texte_entier`
Madame Bovary	`.../Madame_Bovary/Texte_entier`

TABLEAU 9. *URL des versions brutes des romans constituant notre corpus*

B. Impact de la REN sur la qualité des réseaux

Afin de vérifier l'impact de la REN sur la qualité des réseaux de personnages, nous réalisons une extraction en injectant à l'entrée du pipeline `Renard` les annotations en

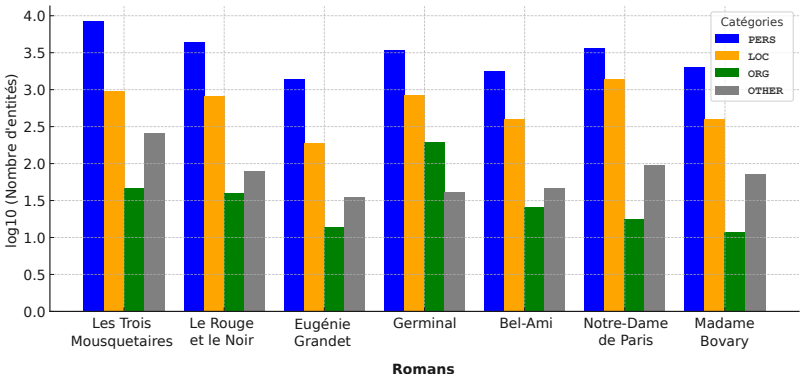


FIGURE 3. *Distribution des EN par roman et par type*

REN de notre jeu de données. Cela nous permet d'observer l'influence de l'algorithme de résolution d'alias sur les résultats, et de donner une borne haute de la performance de l'algorithme d'extraction lorsque l'étape de REN est parfaitement exécutée. Le tableau 10 donne les résultats de cette expérience en termes de mesures de sommets, tandis que le tableau 11 donne les mesures en termes d'arêtes.

Roman	$Prec_V$	Rec_V	$F1_V$
Les Trois Mousquetaires	68,76	92,02	78,71
Le Rouge et le Noir	73,55	83,96	78,41
Eugénie Grandet	66,60	84,11	74,34
Germinal	80,04	85,29	82,58
Bel-Ami	71,44	88,67	79,13
Notre-Dame de Paris	68,04	76,68	72,10
Madame Bovary	71,52	88,00	78,91

TABLEAU 10. *Qualité des sommets des réseaux extraits par un pipeline* `Renard` *en utilisant nos annotations en REN*

Si la REN influence fortement la qualité des sommets extraits (jusqu'à 16,79 $F1_V$ pour *Germinal*), son impact sur les arêtes reste plus limité (8,24 $F1_E$ pour *Madame Bovary*). L'algorithme de résolution d'alias apparaît ainsi comme le principal facteur affectant la qualité, encore imparfaite, des réseaux extraits.

Roman	Pre_E	Rec_E	$F1_E$	$WPre_E$	$WRec_E$	$WF1_E$
Les Trois Mousquetaires	26,56	34,32	29,94	26,32	34,01	29,67
Le Rouge et le Noir	23,39	24,52	23,94	23,31	24,44	23,86
Eugénie Grandet	28,66	30,23	29,42	27,28	28,77	28,00
Germinal	60,04	52,10	55,79	58,44	50,72	54,31
Bel-Ami	27,95	31,89	29,79	27,50	31,37	29,31
Notre-Dame de Paris	21,25	27,00	23,79	20,81	26,45	23,29
Madame Bovary	44,71	48,01	46,30	44,16	47,42	45,73

TABLEAU 11. *Qualité des arêtes des réseaux extraits par un pipeline* `Renard` *en utilisant nos annotations de REN*