Guide d'annotation manuelle d'entités nommées dans des corpus littéraires

Version 2 (1 mars 2024)

Motasem Alrahabi, ObTIC, Sorbonne Université
Carmen Brando, EHESS
Francesca Frontini, CNR-ILC Pisa

Contribution:

Perrine Maurel, Caroline Parfait, Léa Heid, Ada Desideri, Margarite Bordry, Arthur Provenier, Yoann Dupont, James Gawley, Romain Jalabert, Camille Koskas (Sorbonne Université)

Dépôt Github: https://github.com/obtic-sorbonne/NER-Annotation-Guideline

Introduction

Notre étude s'inscrit dans une initiative des humanités numériques littéraires qui vise à étiqueter des entités nommées (EN) dans des corpus littéraires à l'aide des outils du Traitement automatique des langues (TAL). Notre objectif est de mettre au point un modèle d'apprentissage automatique pour les textes fictionnels en prose, en particulier pour les romans du XIXe siècle rattachés aux courants réalistes et naturalistes, qui font figure de modèle dans l'histoire du genre. Il est d'abord nécessaire de procéder à l'annotation manuelle d'EN d'un sous-ensemble des romans pour d'une part, établir un corpus d'entraînement pour créer les modèles, et d'autre part, disposer d'un corpus de validation pour vérifier les performances de ces modèles. Tenant compte du fait que l'élaboration d'un guide d'annotation est un prérequis à toute entreprise de ce genre, le présent document vise donc à fournir les définitions et les consignes pour constituer une annotation de référence. Celle-ci donnera lieu à un balisage manuel des extraits du roman qui sera à la fois homogène et guidé par nos besoins d'analyse.

Le corpus concerné englobe actuellement 3 romans[1]:

- le premier tiers de chacun des six chapitres du roman *Le Ventre de Paris* d'Emile Zola[2],
- un cinquième de chacun des quatorze chapitres du roman *Nana* d'Emile Zola où chaque extrait est composé de paragraphes choisis et listés de manière aléatoire,
- Les premières moitiés des chapitres un et cinq pour chacune de deux parties de *Bel Ami* de Guy de Maupassant.

[1] Les exemples dans ce guide sont extraits du chap. 3 du *Ventre de Paris* et du chap. 9 de *Nana.*
[2] Cette initiative prend la suite du travail décrit ici :
https://github.com/DHNamedEntities/19thCenturyFrenchNovels

Définition d'une entité nommée

Les EN (Entités nommées) sont des types particuliers d'unités lexicales (groupes de formes) qui font référence à une entité du monde concret dans certains domaines spécifiques (humains, sociaux, politiques, économiques ou géographiques).

Dans cette étude, nous considérons que les EN renvoient aux noms propres (NP) ainsi qu'aux descriptions définies (DD)[3] contenant un NP.

Nous allons principalement nous intéresser aux catégories suivantes : **Personnage** et **Lieu**. Une catégorie **Misc** est également disponible pour annoter tout autre type d'EN, comme les organisations, les dates, les mesures, les démonymes ainsi que les noms des produits alimentaires (fromages, vins...) et d'oeuvres, etc.

Le choix de baliser les personnages est motivé par le besoin d'analyser les différentes dénominations servant à les désigner et leur évolution tout au long d'un roman. De même, le balisage des lieux nous permettra d'étudier la dimension géographique du roman, présentant l'intérêt de pouvoir localiser ces lieux sur une carte.

Les personnages et lieux peuvent être dotés d'une ou plusieurs caractéristiques selon la manière à travers laquelle ils sont désignés dans le texte. Ce choix multiple est motivé par le besoin d'attribuer plusieurs caractéristiques à la fois à une EN.

Remarques générales

Sur le balisage d'EN. Lors de l'annotation d'une EN, il faut sélectionner l'ensemble des unités lexicales concernées, autrement dit, il est important de correctement choisir les frontières de l'EN. On doit aussi annoter dans un texte toutes les occurrences d'une EN. Par exemple, on annote la première fois qu'on rencontre *Paris* et si plus loin il y a une deuxième occurrence *Paris*, il faut l'annoter aussi.

Sur la catégorisation d'EN. De manière générale, il faut savoir que le choix de la catégorie dépend uniquement du contexte dans lequel elle est employée, il faut donc tenir compte de l'intention du locuteur et de ce qu'il a voulu désigner.

Sur le cas des EN discontinues et imbriquées. Dans cette première version du guide, nous éviterons les annotations imbriquées et discontinues. Par conséquent, nous prendrons en compte l'expression la plus large possible quand elle est définie et non discontinue. Exemples:

Le roi d'*Italie* (Description définie) --> annoter juste l'Italie en tant que lieu.

Tout près de la rue touristique de *Mouffetard* --> annoter juste Mouffetard comme lieu.

Sur le cas des personnages et lieux évoqués. Les interprétations que l'on pourrait tirer de ces annotations manuelles s'appliquent sur les données annotées dans le cadre de projet, et il est a priori difficile de les généraliser sur d'autres textes.

[3] Une DD est un groupe nominal sous la forme déterminant + nom, comme par exemple, "le Chat noir" et "le 2e président de la République". L'idée de considérer les DD comme une nouvelle catégorie linguistique d'EN est longuement discutée par Ehrmann (2008) et reprise dans le guide d'annotation de la campagne d'évaluation ETAPE (la suite d'ESTER).

Catégories

1) PERSONNAGE :

Il s'agit des personnes ainsi que des animaux qui mènent ou participent à une action quelconque, même éphémère, dans l'histoire. Ces personnages peuvent également être évoqués sans participer à l'action (ex : *Dieu* dans une expression, personnages mythologiques, etc.). Ils sont souvent fictifs mais peuvent être réels quand il est question de personnages historiques célèbres par exemple.

Les personnages sont souvent désignés par leur nom, prénom, surnom, alias ou initiale, accompagnés parfois de la particule désignant le titre de civilité. Ex : *M. Verlaque*, *Gavard*, *Nana*, *madame Belles-fesses, Monsieur L., etc.*

On inclut dans l'annotation des personnages les noms communs qui accompagnent les noms propres (*président, ministre*…), les déterminants (*le, la, les, des*...), les articles définis contractés (*aux, des*…) et les civilités (*Madame, M., mademoiselle*...). On exclut cependant les démonstratifs et les possessifs, sauf dans des cas rares comme *Sa majesté le roi de France, Son Altesse*...

Pour la catégorie Personnage, il existe deux attributs détaillés ci-dessous qui sont optionnels. Autrement dit, lors de l'annotation d'un personnage, il est possible de spécifier l'un, l'autre, les deux ou aucun :

1.1) Rôle: désigne un métier, un rang administratif, un grade militaire, un ordre religieux, un rôle familial ou amical. Ex : *le cousin Florent*, *la mère Méhudin*, *le cousin Quenu, l'ami Claude Lantier*. Il ne faut pas annoter la civilité comme un rôle (l'inclure simplement dans l'annotation). En outre, les épithètes servant à caractériser un personnage ne doivent pas être considérés en tant que rôles, par exemple, dans "*la vieille Méhudin*", l'EN comprend uniquement le NP et l'attribut rôle ne doit pas être coché.

1.2) PersEvoqué: les personnages dotés de cet attribut correspondent aux cas suivants :

- une expression contenant un nom propre qui n'est pas utilisé pour se référer à quelqu'un, par exemple, nom de *Dieu* !
- les personnes célèbres ou figures mystiques sont parfois évoqués dans le texte, comme, par exemple, *Beethoven* dans la phrase prise de Numa Roumestan : *Voilà trente ans que ... cette belle romance de Beethoven*, …
- dans le cas de mises en abyme[4], c'est-à-dire de procédés où l'auteur met en scène une œuvre au sein de l'œuvre (pièce de théâtre, roman), les personnages cités doivent aussi être annotés en tant que PersEvoqué. Ça sera le cas pour "Jupiter", "Iris" et "Ganymède" dans l'exemple issu du chapitre 1 de Nana : *Le premier acte de la Blonde Vénus se passait dans l'Olympe, un Olympe de carton, avec des nuées pour coulisses et le trône de Jupiter à droite. C'étaient d'abord Iris et Ganymède, aidés d'une troupe de serviteurs célestes, qui chantaient un chœur en disposant les sièges des dieux pour le conseil.* À noter que par le contexte, *Blonde*

[4] https://fr.wikipedia.org/wiki/Mise_en_abyme

Vénus fait référence à une pièce de théâtre qui est jouée dans l'action, l'EN est donc taguée en tant que MISC.

L'utilisation de cet attribut peut s'avérer utile pour les analyses des textes a posteriori mais il paraît difficile d'en tenir compte pour la création d'un modèle NER. Nous avons cependant choisi de le garder pour ouvrir les possibilités d'analyse.

Il sera souvent nécessaire d'avoir une connaissance confirmée des personnages du roman pour prendre en compte toutes les manières possibles de désigner un personnage au sein de l'annotation, ce qui ne sera pas le cas de tous les annotateurs. Il peut s'avérer utile de consulter des ressources externes existantes listant la liste des personnages.

Consignes d'annotation pour les personnages:

Ce tableau contient des exemples à annoter en Personnage (l'EN est en italique):

id	Exemples	Définition
1	*Florent* *Nana* *Balthazar*	Les noms propres (NP) pour désigner les personnes et animaux.
2	« Ohé, *nana* ! »	Les noms et surnoms de personnages quand ils ne sont pas capitalisés.
3	*la Faloise*	Les NP comprenant un article défini dans leur nom
4	*les Méhudins* *une Gradelle* *un Méhudin*	En présence de déterminants (définis ou indéfinis, contractés ou non, on exclut les démonstratifs), on les intègre à l'EN. Cela permet de distinguer les groupes et le genre des personnages.
5	*madame François* *M. Gradelle*	Le NP accompagné de la particule ou titre désignant la civilité (*madame ou monsieur* y compris les différentes abréviations *Mme, M.*, etc) avec l'objectif de différencier le genre des personnages. Les civilités ne sont pas marqués en tant que rôle.
6	*madame Belles-Fesses* *la belle Normande, la Normande* *la belle Lisa* *Son Altesse*	Des surnoms comme *madame Belles-Fesses* qui peuvent être utilisés pour désigner un personnage. Selon le cas, il peut avoir besoin d'inclure le possessif ainsi que l'épithète dans l'annotation. Cela est laissé au jugement de l'annotateur.
7	En l'appelant : sa *"Juliette"*	Par le contexte, il peut être explicitement dit qu'il est question d'un nom par l'usage de double guillemets

8	*la mère Chantemesse* *le comte Muffat* *les filles Méhudin* *du marquis de Chouard* *cousin Florent* *ami Florent*	Les descriptions définies (DD)[5] contenant un NP et des rôles (familiaux, amicaux, métiers, ...). On inclut aussi la contraction (ex : du, des) le cas échéant. On annote en personnage avec l'attribut Role coché.
9	nom de *Dieu !* *Beethoven* *Roi Dagobert*	Parfois présents dans des expressions, on les annote en tant que personnages évoqués

Ce tableau contient des exemples à annoter *partiellement* en Personnage (l'EN est en italique):

id	Exemples	Définition
10	La mère de *Louise* l'ami de *Claude Lantier* anciens ministres de *Louis Philippe* la fille aînée *des Macquart*	Ne pas annoter les descriptions de personnes en relation à une autre personne ou d'autres personnes (membres de famille, amis,...), on annote seulement les NP de personnages qui sont référencés. Quand il s'agit de personnes célèbres avec certitude, on coche l'attribut persEvoq.
11	de la Blonde *Vénus*	En présence d'un épithète capitalisé qui accompagne un NP, on annote tout. En présence d'épithètes, on n'inclut pas les déterminants pour éviter une annotation imbriquée innécessaire.
12	mon *Florent* son *frère Florent* mon *Mimi*	On évite d'inclure les possessifs (ex : mon, sa) dans l'EN. Il se peut que des rôles soient utilisés avec le NP et dans ce cas, il faut cocher "role".
13	La cadette, *Claire* cousine germaine, *Augustine Landois* son cousin, *M. Hector de la Faloise* ses fils, *Henri* et *Charles*	En présence d'un nom de personnage suivi d'une virgule et d'une description du rôle, ce dernier n'est pas annoté, seul le NP l'est.

Ce tableau contient des exemples à ne pas annoter en Personnage:

id	Exemples	Définition

[5] Une DD est un groupe nominal sous la forme déterminant + nom, comme par exemple, "le Chat noir" et "le 2e président de la République". L'idée de considérer les DD comme une nouvelle catégorie linguistique d'EN est longuement discutée par Ehrmann (2008) et reprise dans le guide d'annotation de la campagne d'évaluation ETAPE (la suite d'ESTER).

14	le marchand de vin[6] le jeune homme	On n'annote pas les désignations génériques sous forme de groupes nominaux contenant aucun NP
15	... en Nourrice normande …, en Fermière …, en Postillon de Lonjumeau	
16	"Allons-y, mon bonhomme !"	
17	son cousin madame	Les rôles, familiaux ou honorifiques, ne contenant aucun NP

Exemples d'annotation avec contexte (EN en italique):

- Pers sans attribut

  Dès le lendemain, *monsieur Verlaque* commença à mettre le nouvel inspecteur au courant de la besogne.

  … jugeait l'intérieur *des Quenu-Gradelle* trop endormi.

- Avec l'attribut Role:

  La mère Méhudin, comme on la nommait, était longtemps...

  … et *le comte Muffat*, qui vinrent saluer silencieusement …

- Avec l'attribut PersEvoque :

  nom de *Dieu* ***!***

2) LIEU :

Ce sont des unités administratives (régions, départements...), constructions humaines, lieux géographiques naturels, etc. Ces lieux peuvent être réels ou fictifs. Les EN de lieux concernent également les lieux où se déroule une action ou un lieu évoqué.

On inclut dans l'annotation des lieux les noms communs quand ils sont accompagnés de noms propres (*rue, place*…) et les déterminants (*le, la, les, des*...). On exclut cependant les possessifs.

Les attributs de la catégorie Lieu sont décrits ci-dessous.

2.1) Lieu géographique naturel: Il s'agit par exemple: - des fleuves, rivières, lac… - des montagnes, massifs… - des éléments du système solaire (*le soleil, Mars*…).

2.2) Région administrative: Il s'agit d'un mot ou groupe de mots pouvant parfois être employé pour référer à la fois à une région, son peuple et son gouvernement. Ces régions dites «

[6] Bamman et al. (2019) propose d'inclure l'ensemble de DD en tant qu'entité littéraire, nommée ou non-nommée, dans le contexte des textes de fiction.

administratives » sont délimitées par l'homme à l'aide de frontières imaginaires (*la ville de Marseille*, *le cinquième arrondissement, le département de la Moselle, le continent Indien…*) par opposition aux régions naturelles.

2.3) Construction humaine: bâtiments et aux autres constructions humaines fonctionnelles permanentes (*la mairie de Lyon, l'église Saint Jacques…*) ou axes de circulation (odonymes: *l'avenue des Martyres, la place de la Concorde...*).

2.4) LieuEvoqué : indique si un lieu est clairement évoqué et n'est pas un lieu où se déroule une action dans le roman, par exemple, *Hollande* et *Angleterre* dans : *Les poissons blancs de Hollande et d'Angleterre encombraient aussi le marché*. C'est également le cas pour : le roi d'*Italie,* ici Italie est annoté en tant que lieu évoqué.

Attention, quand on dit lieu où se déroule l'action, ceci doit être interprété au sens large et même indirect; par exemple dans :

il connaissait, rue Cuvier, près du Jardin des Plantes, une dame veuve, dont le mari avait eu la direction des postes à Plassans,

"une sous-préfecture du Midi", "Plassans", "Midi" sont des lieux évoqués.

C'est également le cas dans l'exemple suivant :

Il leur apprit qu'il était rentré en France, grâce aux papiers d'un pauvre diable, mort entre ses bras de la fièvre jaune, à Surinam où "Surinam" fait partie de cet univers fictif, même si c'est de manière marginale ; par contre *des recettes du Midi*, "Midi" represente un lieu évoqué.

Par contre, quand le lieu fait clairement partie du nom propre d'événement ou organisation, toute l'expression sera annoté comme MISC (*la guerre de Crimée*).

Tout comme les personnages évoqués, il est généralement nécessaire d'avoir une connaissance des lieux dans un roman pour avoir des annotations précises.

Consignes d'annotation pour les lieux:

Ce tableau contient des exemples à annoter en Lieu (l'EN est en italique):

id	Exemples	Définition
18	*Le Havre* *Les Halles* *au Vigan*	Les NP comprenant un article défini dans leur nom
19	*des Halles* *aux Halles* *la Seine* *l'arc de Triomphe* *au pont de Neuilly* *la rue Pirouette* *l'église Saint-Eustache* *le Luxembourg*	En présence de déterminants (définis ou indéfinis, contractés ou non), on les intègre à l'EN.

21	*la rue Serpente* *la ville de Paris* *le quartier des Halles*	Le terme générique du lieu compris dans l'EN comme par exemple, rue et ville car il rend compte de sa fonction et de sa nature. On inclut également les déterminants. Pour les rues, ajouter l'attribut Construction Humaine
22	*les côtes de la Guyane* *le nord de l'Europe* *les rues de Paris*	Les lieux imprécis (avec les déterminants)
23	*du fort de Bicêtre* *les Halles centrales*	On annote les noms de lieux historiques qui ont existé et qui ont été détruits ou qui ont changé de fonction, de forme ou de localisation, c'est le cas par exemple de l'ancienne configuration des halles à Paris.

Ce tableau contient des exemples à annoter partiellement en Lieu (l'EN est en italique):

id	Exemples	Définition
24	chez *monsieur Lebigre* chambre de *Florent* chez *les Muffats* la maison de *Florent*	Ne pas annoter les lieux nommés en rapport avec un personnage, néanmoins on annote le personnage qui y est référé dans la catégorie personne.
25	la préfecture de *la Seine* une sous-préfecture de *Midi* des filles du *Nord* les pêches du *Midi* des filles de *Provence*	Si l'expression la plus large désigne une catégorie de type MISC, opter pour une expression moins large qui dénote un lieu bien défini. Il faut aussi activer l'attribut LieuEvoq pour l'ensemble des exemples proposés.
26	les nouvelles *Halles* la grande *Thérèse* les grandes *Halles* la vieille *Méhudin* le petit *Quenu*	Ne pas inclure dans l'annotation les épithètes qui ne font pas partie du nom vernaculaire de lieu.
27	le roi d'*Italie* le roi de *Prusse*	Seulement annoter les noms de lieux évoqués dans les titres de personnes. A cocher aussi l'attribut lieuEvoque.
28	le mur de *Sainte Eustache* la pointe de *Saint-Eustache* le cadran lumineux de *Saint-Eustache* les toits *des Halles* la chapelle de la *Vierge* la grande chambre de *la rue Royer-Collard* des pavillons de la boucherie et de *la Vallée*	Ne pas annoter les parties mais la construction humaine référencée en tant que lieu.

Ce tableau contient des exemples à ne pas annoter en Lieu:

id	Exemples	Définition
29	je suis allé à la préfecture hier la préfecture acceptait pavillons église poissonnerie	Ne pas annoter un lieu désigné seulement par un nom commun et ne contenant aucun nom propre.

Exemples d'annotation avec contexte (EN en italique):

- Avec l'attribut Nature :

  Il se souvenait des côtes de la Guyane, des beaux temps de la traversée.

- Avec l'attribut DecoupageAdmin :

  la direction des postes à Plassans, ...

- Avec l'attribut ContructionHumaine :

  Placé à l'encoignure droite de la rue Pirouette, sur la rue Rambuteau, flanqué de quatre petits…

  C'était un grand garçon osseux, soigneusement rasé, avec un nez maigre et des lèvres minces, qui demeurait rue Vavin, derrière le Luxembourg.

  …de cet immense développement des Halles, qui lui donnait, ...au milieu des rues étranglées de Paris,...

- Avec l'attribut LieuEvoque :

  L'arrivage des écrevisses d'Allemagne, en boîtes et en paniers, était très fort ce matin-là.

  Il y a deux contrôles, disait-il, celui de la préfecture de la Seine et celui de la préfecture de police.

  Elle ne se rassit pas, elle marchait, fiévreuse, allant de la glace de la cheminée à un miroir de Venise

3) MISC (divers):

Toute autre catégorie, particulièrement les organisations, les démonymes, les unités monétaires, les dates, les véhicules ainsi que les noms de fromages, de vins et d'oeuvres. Autrement dit, il s'agit de catégoriser les entités, appartenant à toute autre catégorie, qui sont désignées par un nom propre capitalisé (le trait de la majuscule est fort dans le contexte littéraire.

Consignes d'annotation pour les Misc:

Ce tableau contient des exemples à annoter en MISC (l'EN est en italique):

id	Exemples	Définition

30	*au Dépôt de la préfecture de police* *Corps législatif* *l'administration de la Ville* *des Chambres* *l'Ecole de droit* *Etat* *l'Etat* *l'Empire*	Les noms d'organisations avec les déterminants
31	*les Port-Salut* *ananas Pompadour*	Les fromages, les fruits et légumes, etc. aussi avec les déterminants
32	*le Parisien* *Provençaux* *aux Parisiens*	Les démonymes avec les déterminants. Ne pas les prendre en compte s'ils sont en minuscule.
33	*les Anglaises*	Les démonymes dans le rôle d'une catégorie Misc (les fruits, dans l'exemple).
34	*du Léoville*	Le vin…
35	*le Paris mondain*	
36	*la Révolution*	
37	*l'écurie Vandeuvres, l'armée d'Orient*	Ici on considère que toute l'expression est un nom propre d'institution, même si les noms *écurie* et *armée* sont en minuscule.
38	*la guerre de Crimée,*	Ici on considère que toute l'expression est un nom propre d'événement, il s'agit en effet du nom de l'événement d'après un référentiel.
39	*la frégate* *le Canada*	Les moyens de transport en lien avec les actions sont annotés en Lieu; ajouter l'attribut constructionHumanie

Ce tableau contient des exemples à annoter *partiellement* en MISC (l'EN est en italique):

id	Exemples	Définition
40	un *Murillo* les *Montreuil* rougissants un *Gutenberg* pensif	On annote en Misc.

41	à la *Daumont*	Dans les expressions de type à la manière de, on annote le NP sans le déterminant en Misc
42	ministre des *Finances*	

Quelques références bibliographiques

Bamman, D., Popat, S., & Shen, S. (2019). An Annotated Dataset of Literary Entities. In NAACL.

Ehrmann, M. (2008). Les Entités Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation. Informatique et langage [cs.CL], Paris Diderot University, Français.

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., & Gravier, G. (2005). The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News. In Proceedings of the 9th European Conference on Speech Communication and Technology (InterSpeech'05), Lisbon, Portugal, p. 1149-1152. Guide d'annotation disponible ici: URL.

Poibeau, T. (2005). Sur le statut référentiel des entités nommées. In Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles, Articles longs, pages 171–180, Dourdan, France. ATALA.

Rosset, S., Grouin, C., & Zweigenbaum, P. (2011). Entités nommées structurées : guide d'annotation Quaero. Technical Report 2011-04, LIMSI-CNRS.