

Ce que BookNLP fait aux textes

Panorama du projet

Frédérique Mélanie-Becquet & Jean Barré

15 avril 2023

Lattice : ENS-PSL-CNRS

Une équipe

- Lattice
 - Thierry Poibeau*, Frédérique Mélanie-Becquet*
 - Jean Barré*, Pedro Cabrera*, Ioana Galleron*, Claude Grunspan,
Olga Seminck, Ana Duron-Tejedor*
 - Clément Plancq, Marco Naguib, Martial Pastor*
 - Frédéric Landragin
- Collaborations
 - Laurette Chardon - CRISCO (Université de Caen Normandie)
 - Motasem Alrahabi, Johanna Cordova, Ada Desideri - ObTIC (SCAI –
Sorbonne Université)

BookNLP

L'annotation

Multilingual BookNLP

French BookNLP

Le corpus

L'annotation

Le modèle

BookNLP en application

Potentiels pour l'analyse littéraire

Limites actuelles du modèle

Une application concrète

Un projet de D. Bamman (Berkeley, 2014-2016)

But : annotater des romans...

- Entités annotées
- Coréférence
- Evénements
- Prises de parole (dialogue, ...)

<https://github.com/dbamman/litbank>

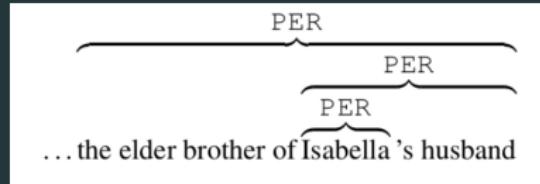
BookNLP : Les entités

The entity annotation layer of LitBank covers six of the ACE 2005 categories in text:

- People (PER): *Tom Sawyer, her daughter*
- Facilities (FAC): *the house, the kitchen*
- Geo-political entities (GPE): *London, the village*
- Locations (LOC): *the forest, the river*
- Vehicles (VEH): *the ship, the car*
- Organizations (ORG): *the army, the Church*

À noter :

- annotation de personnages (vs entité nommée PERS)
- annotation imbriquée



BookNLP : La coréférence

Coreference	One may as well begin with [Helen] _x 's letters to [[her] _x sister] _y
-------------	---

BookNLP : multilingual project

D. Bamman (Berkeley, 2020)

Créer des corpus annotés avec le même schéma d'annotation

Deux idées principales

- Intégrer les réseaux de neurones (et plus généralement les modèles de langage)
- Etendre le modèle à 4 langues (en plus de l'anglais) : japonais, russe, allemand, espagnol

Le corpus



- 18 romans français
 - bloc : 10 000 premiers tokens
 - date : XIX^e
- 184 000 tokens - 14 208 entités annotées
- Annotation en Co-référence

Date	Author	Title
1830	Honoré de Balzac	Sarrasine
1836	Théophile Gautier	La morte amoureuse
1841	George Sand	Pauline
1856	Victor Cousin	Madame de Hautefort
1863	Théophile Gautier	Le capitaine Fracasse
1873	Émile Zola	Le ventre de Paris
1881	Gustave Flaubert	Bouvard et Pécuchet
1882-1883	Guy de Maupassant	Mademoiselle Fifi, nouveaux contes (1)
1882-1883	Guy de Maupassant	Mademoiselle Fifi, nouveaux contes (2)
1882-1883	Guy de Maupassant	Mademoiselle Fifi, nouveaux contes (3)
1901	Lucie Achard	Rosalie de Constant, sa famille et ses amis
1903	Laure Conan	Élisabeth Seton
1904-1912	Romain Rolland	Jean-Christophe (1)
1904-1912	Romain Rolland	Jean-Christophe (2)
1917	Adèle Bourgeois	Némoville
1923	Raymond Radiguet	Le diable au corps
1926	Marguerite Audoux	De la ville au moulin
1937	Marguerite Audoux	Douce Lumière

Le corpus

Les mentions



Les chaines
=> Relie un ensemble de maillons

The screenshot shows a software interface for managing chains. On the left, there's a sidebar with 'Annotation' and 'Type de schéma' set to 'CHAINE'. Below this is a list of chains with small icons next to them. The main area displays a text excerpt:

Deux hommes parurent.
L'un venait de la Bastille, l'autre du jardin des Plantes. Le plus grand, vêtu de toile, marchait le chapeau en arrière, le gilet déboutonné et sa cravate à la main. Le plus petit, dont le corps disparaissait dans une redingote marron, baissa la tête sous une casquette à visière pointue.

Quand ils furent arrivés au milieu du boulevard, ils s'assirent, à la même minute, sur le même banc.

Pour s'essayer le front, ils retirèrent leurs coiffures, que

default

Console BOUARDDETPECUCHET - CHAINE 318 unités

NB MAILLO 318 REF Bouard et Pécuchet

<https://txm.gitpages.huma-num.fr/textometrie/>

Le corpus

[#5] Au delà T8 du canal , entre T13 les maisons T13 que séparent T14 des chantiers , le grand ciel pur se découpaient en plaques d'outremer, et sous la réverbération T90 du soleil , les façades blanches, les toits d'ardoises. T28 les quais de granit éblouissaient. Une rumeur confuse montait au loin dans l'atmosphère tiède ; et tout semblait engourdi par le désœuvrement du dimanche et la tristesse des jours d'été. T254 Deux hommes parurent. T534 L'un venait de T594 la Bastille , T791 l'autre T881 du Jardin des Plantes .T534 Le plus grand , vêtu de toile, marchait le chapeau en arrière, le gilet déboutonné et T534 sa cravate à la main.T791 Le plus petit , T791 dont le corps disparaissait dans une redingote marron, baissait la tête sous une casquette à visière pointue.

[#6] Quand T254 ils furent arrivés T882 au milieu T2 du boulevard , T254 ils s'assirent, à T883 la même minute , sur le même banc.

[#7] Pour s'essuyer le front, T254 ils retirèrent T254 leurs coiffures, que T254 chacun posa près de T254 soi ; et T791 le petit homme aperçut, écrit dans le chapeau de T791 son voisin : Bouvard ; pendant que T534 celui-ci distinguait aisément dans la casquette T791 du particulier en redingote le mot : Pécuchet.

Chains and Links — Mozilla Fi...
Débuter avec Firefox file:///home/frederique/Documents/BookN 50% ★
https://github.com/us... Ninja_Flask CNRS Autres marque-pages
about:blank
T16
T20
T254 Deux hommes
ils
ils
ils
leurs
chacun
soi
nous
notre
nos
ils
leurs
Ils
ils
eux
ils
Leurs
Bouvard et Pécuchet
ils
Leurs
Ils
C'

<https://boberle.com/projects/coreference-annotation-with-sacr/>

L'annotation : Entités

Quelles entités faut-il annoter ?

Reprise des catégories de Bamman ?

- Personnages
 - PERS, no-Pers (non humains)
- Lieux :
 - GPE, Loc, Fac
- Véhicules
 - VEH
- Indications temporelles, fêtes, etc.
 - Time

L'annotation : Entités

Campagne d'annotation : janvier – août 2021

- annotés en triple aveugle
 - 3 textes : Pauline, Le Capitaine Fracasse, Le Ventre de Paris
- discussions / adjudication
- annotation dans TXM (le reste du corpus)

Corpus de 184 000 mots, 14 208 entités annotées

L'annotation : Entités

Fréquents problèmes de « limite »

- Pers : Dieu, autres personnages non humains (Zeus, animaux qui parlent...); « un nouveau visage apparu en ville », « la foule », « la moitié de la ville » (collectifs, ensembles imprécis), « on »
- GPE / Loc / Fac : la lande (la Lande), la route de Bressuire
- Catégorie Fac en elle-même : annotation minimale : lieu de vie, pièce entière, etc. Mais héros caché dans un placard ?
- VEH : animaux ?

Question de la robustesse

L'annotation : Corréférence

Y'a-t-il ou non corréférence ?

Débuter avec Firefox file:///home/frederique/Doc... 67% ★ Autres marque-pages

jour T254 vint T254 leur dire qu'on T1003 lui avait parlé d' T1246 un domaine ,
à T1253 Chavignolles , entre T1256 Caen et T1262 Falaise . T1246 Cela consistait
en T1266 une ferme de trente-huit hectares , avec T1274 une manière de
château et T1283 un jardin en plein rapport . T254 Ils se transportèrent
dans T1285 le Calvados et T254 ils furent enthousiasmés. Seulement, tant
de T1266 la ferme que de T1274 la maison T1266 l'une ne serait pas vendue
sans T1274 l'autre), on exigeait cent quarante-trois mille francs. T534 Bouvard n'en
donnait que cent vingt mille.

T791 Pécuchet combattit T534 son entêtement, T534 le T791 pria de céder,
enfin T791 déclara qu' T791 il complèterait le surplus. C'était toute T791 sa fortune,
provenant du patrimoine de T1286 sa mère et de T791 ses économies.

Edit Mode | FFAC

about:blank

T1262

Falaise
Falaise
Falaise

T1266

une ferme de trente-huit hectare
la ferme
l'une
leur ferme
Leur exploitation
la ferme
la ferme

T1274

une manière de château
la maison
l'autre
La maison
sa
sa
Elle

T1283

un jardin en plein rapport

L'annotation : Citations

Les citations :

- citation (les mots prononcés, pensés ?, rapportés ?, ...)
- entité (par qui ?)
- introducteur (modalité émise)

```
199 lines (199 sloc) | 8.12 KB
Raw Blame ⚙️ 🗑️

1 T4 Citation 1510 1569 - Mon Dieu, oui, on pourrait prendre le mien à mon bureau !
2 T5 Citation 1571 1606 - C'est comme moi, je suis employé.
3 T1 Citation 1414 1422;1431 1508 - Tiens, nous avons eu la même idée, celle d'inscrire notre nom dans nos couvre-chefs.
4 T6 Introduceur 1423 1426 dit
5 T7 Entite 1427 1429 il
6 R1 Source Arg1:T7 Arg2:T1
7 T2 Citation 2479 2517 - Comme on serait bien à la campagne !
8 T3 Introduceur 2468 2475 échappa
9 T8 Entite 2464 2467 lui
10 R2 Source Arg1:T8 Arg2:T2
11 T9 Citation 3994 4014;4028 4045 - Moi, je suis veuf, et sans enfants !
12 T10 Introduceur 4015 4018 dit
13 T11 Entite 4019 4026 Bouvard
```

Le modèle

- Personnages (et non entités nommées)
- Coréférence
- Événements
- Prises de parole (dialogue, ...)



Le modèle

precision	rappel	F_1
86.01	83,13	85,42

Entités :

precision	rappel	F_1
51.32	70,73	61,02

Évènements :

precision	rappel	F_1
91.95	90,74	91,34

Prises de parole :

Le modèle

		precision	rappel	F_1
Mentions		90,65	90,08	90,37
Corréférence :	<i>MUC</i>	85,06	85,10	85,08
	<i>B³</i>	82,66	56,49	67,11
	<i>CEAF_e</i>	28,50	91,89	43,50
	<i>BLANC</i>	85,81	62,99	69,22
	<i>LEA</i>	64,73	62,47	63,58

Application en études littéraires computationnelles

- Enfin un outil du TAL adapté aux textes littéraires ?
- Quelles réponses ? Quels questionnements ?
- Nouvelles perspectives sur l'histoire littéraire - La lecture distante

[#21] Un bruit de ferrailles sonna sur le pavé dans un tourbillon de poussière :

c'étaient T936 trois calèches de remise T936 qui s'en allaient vers T937 Bercy ,
promenant T939 une mariée avec T939 son bouquet, T940 des bourgeois en
cravate blanche , T941 des dames enfouies jusqu'aux aisselles
dans T941 leur jupon, T943 deux ou trois petites filles , T944 un collégien . La vue
de cette noce amena T254 Bouvard et Pécuchet à parler T946 des
femmes , T946 qu' T254 ils déclarèrent frivoles, acariâtres, têtues. Malgré
cela, T946 elles étaient souvent meilleures que T950 les hommes ; d'autres
fois T946 elles étaient pires. Bref, il valait mieux vivre sans T946 elles ;
aussi T791 Pécuchet était resté célibataire.

T936

trois calèches de remise
qui

T937

Bercy

T939

une mariée
son

T940

des bourgeois en cravate blanche

T941

des dames
leur

T943

deux ou trois petites filles

T944

un collégien

T946

des femmes
qu'
elles
elles
elles

T950

les hommes

Chronotope du récit

- Lier les entités PER, LOC, TIME entre elles

Co-référence - Espace Personnage

- Vision du personnage dans la globalité de ses mentions
- Récupération de tous les adjectifs qui qualifient un personnage
- Récupération de tous les verbes d'action d'un personnage

Comparaison Homme - Machine :

	Humain	BookNLP
Nombre de PER	108	315
Nombre de mention par PER	37.5	26.2
Nombre de mention total	7938	8260

- Incohérence de chaîne - Problème de métrique ?

à T312 cette fille ; mais T312 il voudrait être sans cesse auprès d T312 elle : ce
est bien juste qu T312 il paye pour l'incommodité. — Voyons donc, T312 lui dis

<Frédéric-9127>	133
<Frédéric-7372>	131
<un gaillard de trente ans-5800>	120
<Frédéric-6791>	120
<Arnoux-1890>	118
<Martinon-1766>	114
<Deslauriers-1707>	113
<Mme Arnoux-145>	106

- Résolution de la coréférence sur 512 tokens

Étude la caractérisation

Quels sont les enjeux dans la représentation du genre dans la fiction ?

- Évaluez l'importance du genre dans les différences de description des personnages.
- Temps d'écran des personnages et chaîne de coréférence
- Facteur particulier du genre des écrivain.e.s ?

Un résultat saisissant

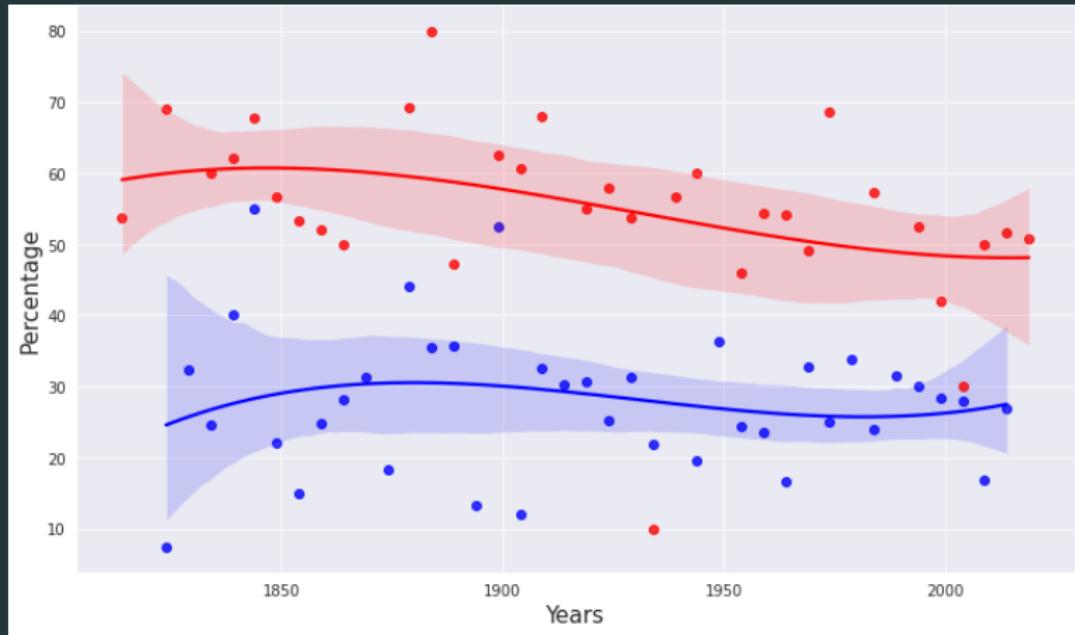


Figure 1: Proportion de la caractérisation des femmes par des auteurs et des autrices, moyenne tous les cinq ans

Conclusion

- Il reste encore du travail
- Mise en ligne de BookNLP - bientôt
- Peut apporter beaucoup aux études littéraires - Changement d'échelle, tendances sur des siècles de littérature
- Automatiser des réseaux de personnages - calcul de densité, de centralité
- Agentivité des personnages
- Étude des chronotopes associés à des genres littéraires / à des auteurs particuliers

Bonus

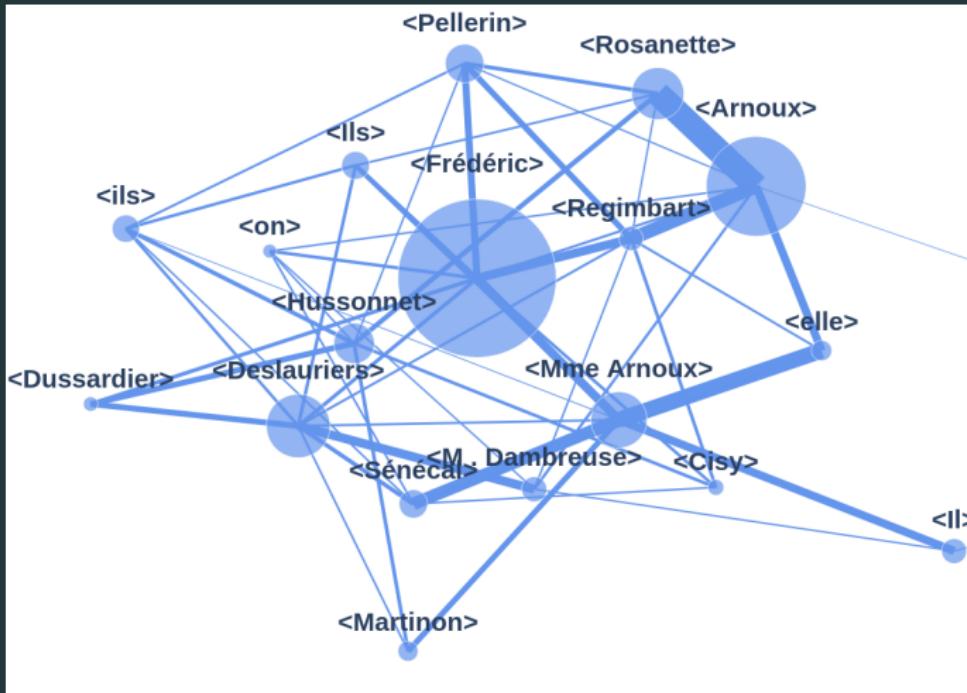


Figure 2 : Réseau de Personnages dans l'Éducation Sentimentale

Questions ?

Projet à Berkeley :

- <https://github.com/dbamman/litbank>
- <https://github.com/booknlp/booknlp>

Partie française développée au Lattice :

- <https://www.lattice.cnrs.fr/projets/booknlp/>
- <https://github.com/lattice-8094/fr-litbank/>