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

BookNLP

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

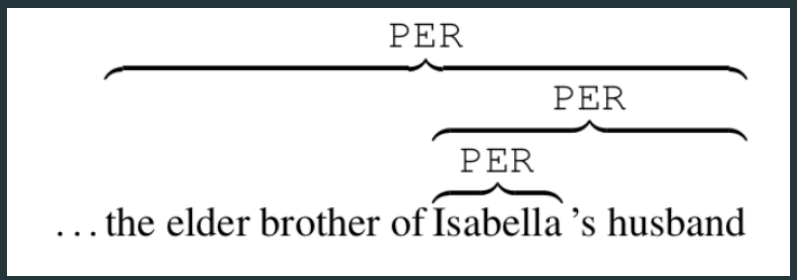
BookNLP : Les entités

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 $[\text{Helen}]_x$'s letters to $[[\text{her}]_x \text{ sister}]_y$

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

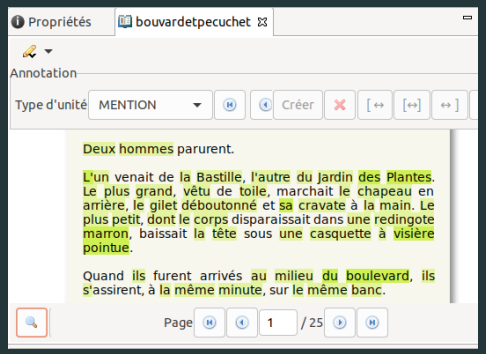
Date	Author	Title
1830	Honoré de Balzac	Sarrasine
1836	Théophile Gautier	La morte amoureuse
1841	George Sand	Pauline
1856	Victor Cousin	Madame de Hautefort
1863	Théophile Gautier	Le capitaine Fracasse
1873	Émile Zola	Le ventre de Paris
1881	Gustave Flaubert	Bouvard et Pécuchet
1882-1883	Guy de Maupassant	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

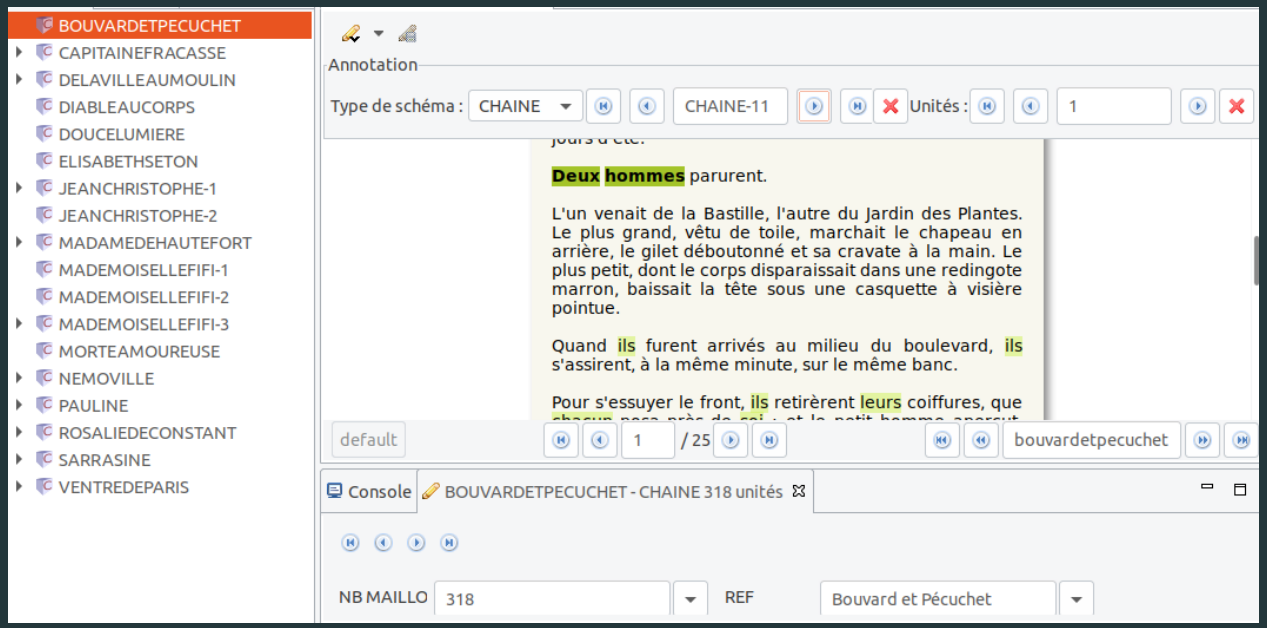
Le corpus

Les chaines

=> Relie un ensemble de maillons

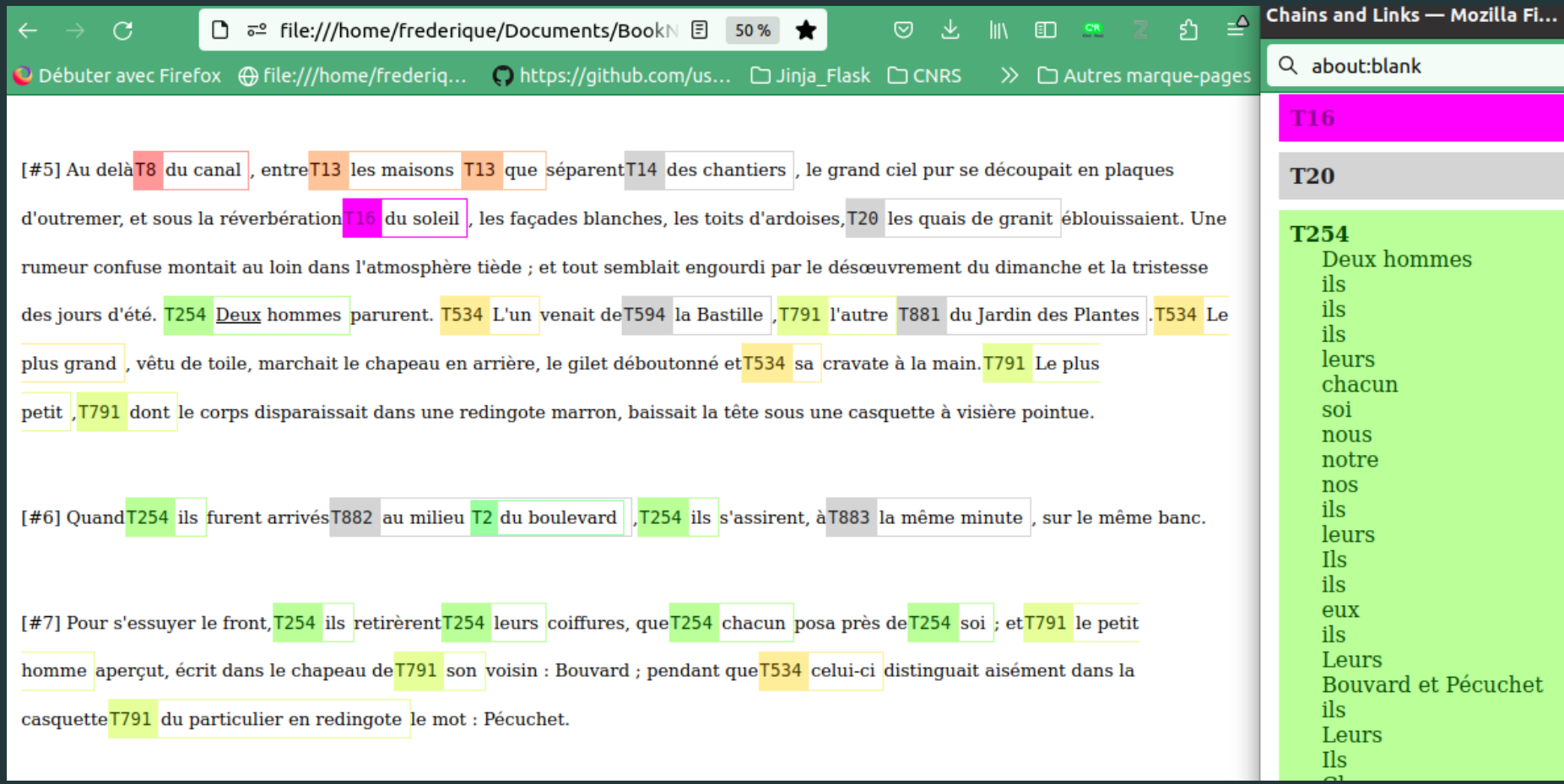
Les mentions





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

Le corpus



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

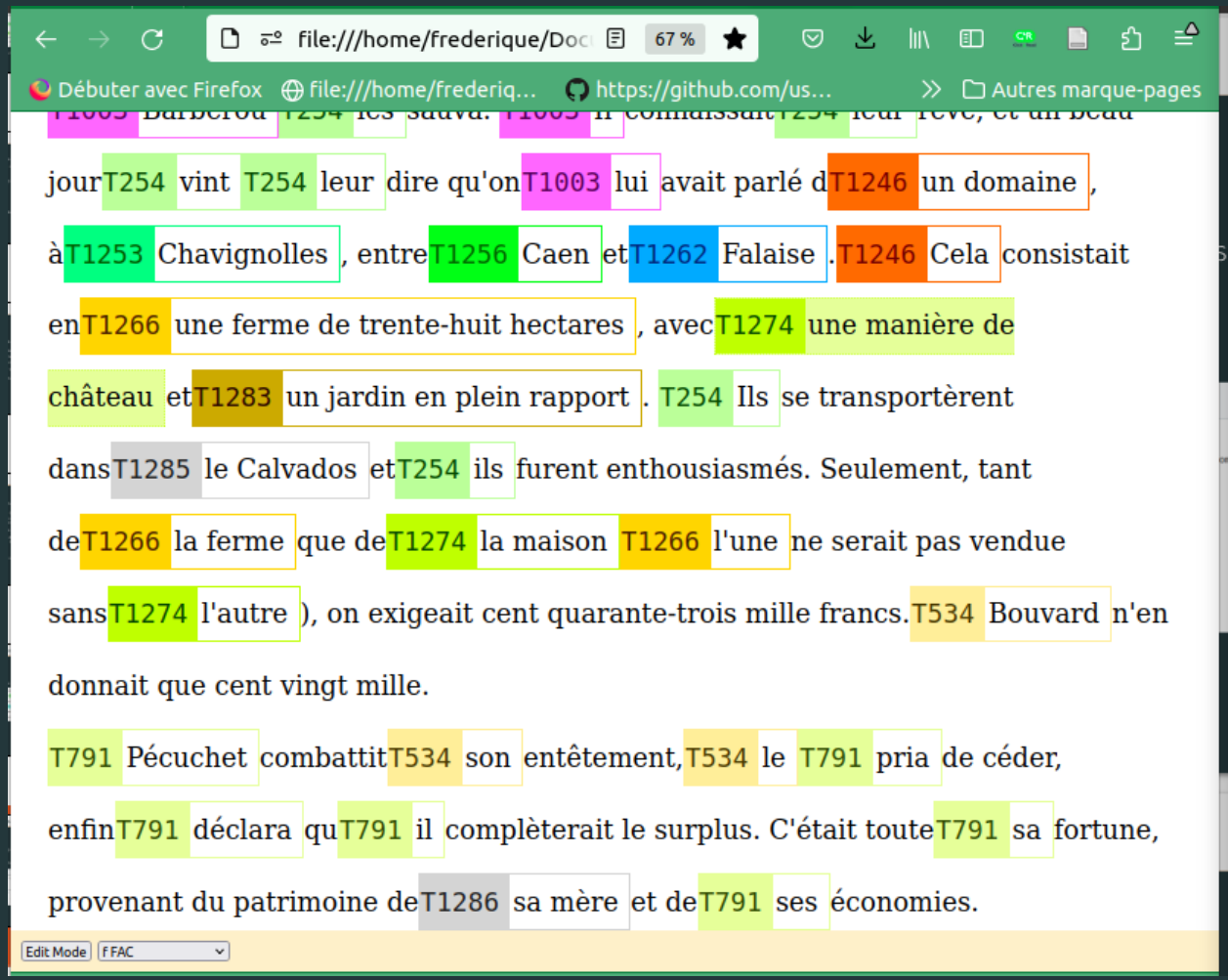
Fréquents problèmes de « limite »

- 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 : Coréférence

Y’a-t-il ou non coréférence ?



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

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

Le modèle

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

Le modèle

Entités :

precision	rappel	F_1
86.01	83,13	85,42

Évènements :

precision	rappel	F_1
51.32	70,73	61,02

Prises de parole :

precision	rappel	F_1
91.95	90,74	91,34

Le modèle

Coréférence :

		precision	rappel	F_1
Mentions		90,65	90,08	90,37
Coreference	*MUC*	85,06	85,10	85,08
	B^3	82,66	56,49	67,11
	$CEAF_e$	28,50	91,89	43,50
	BLANC	85,81	62,99	69,22
	LEA	64,73	62,47	63,58

BookNLP en application

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

Ce que BookNLP fait au texte

[#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.

Ce que BookNLP fait au texte

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

Potentiels pour l'analyse littéraire

Chronotope du récit

- Lier les entités PER, LOC, TIME entre elles

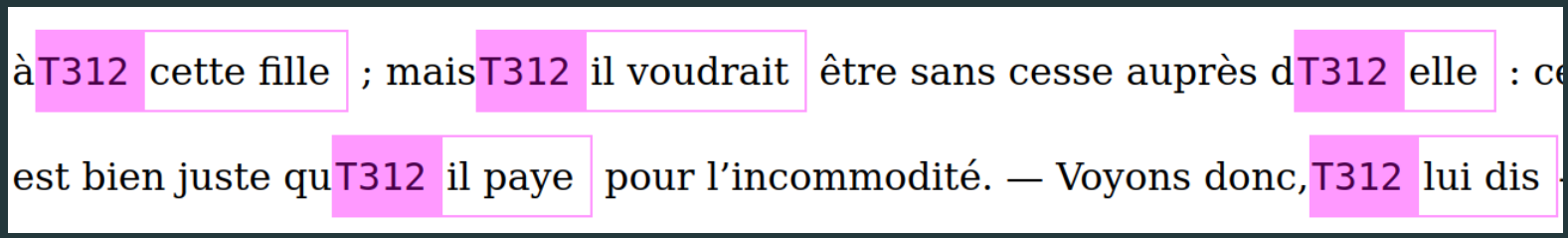
Co-référence - Espace Personnage

- 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 ?



Limites temporaires du modèle

- Résolution de la coréférence sur 512 tokens

<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

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

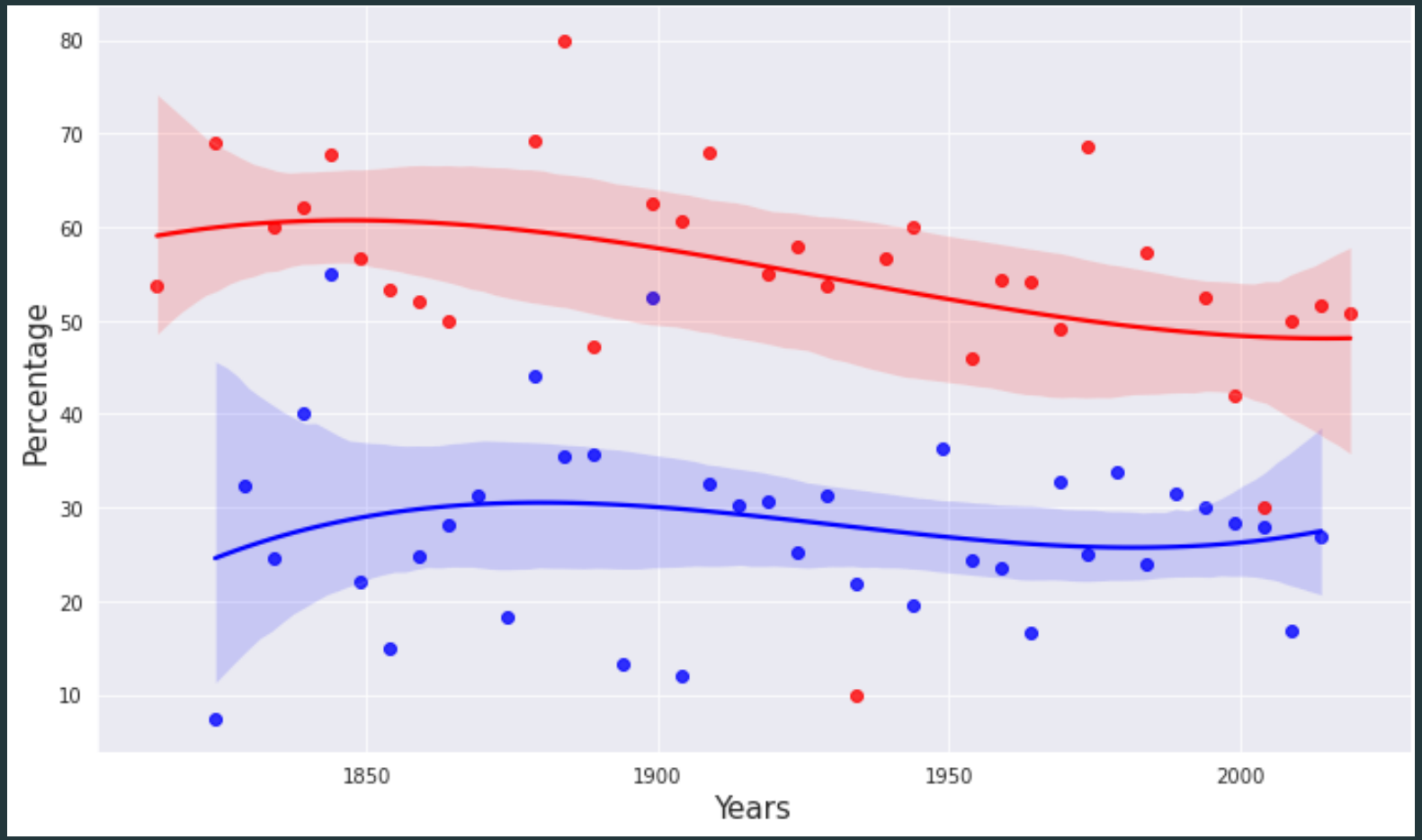


Figure 1 : Proportion de la caractérisation des femmes par des auteurs et des autrices, moyenne tous les cinq ans

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

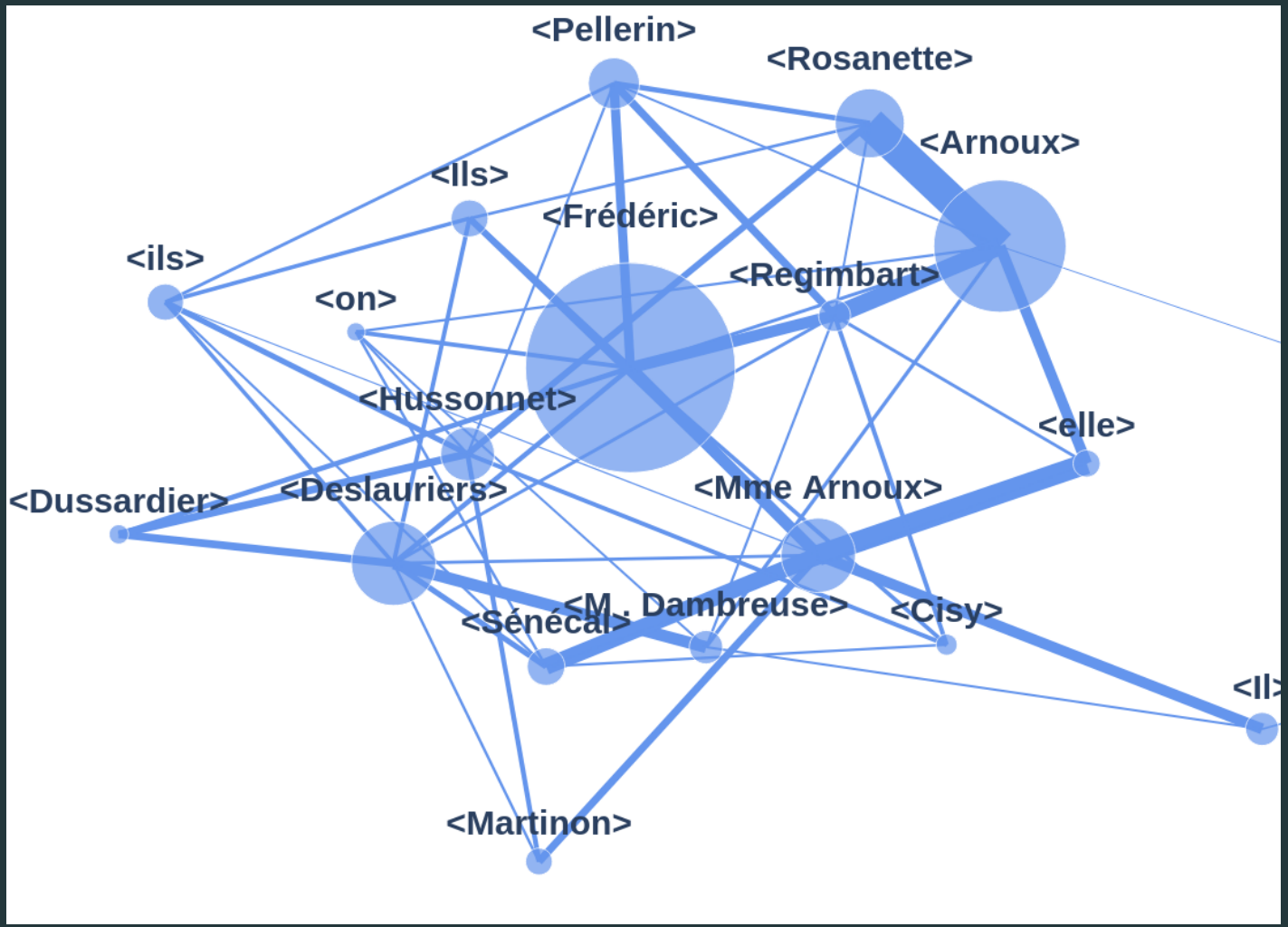


Figure 2 : Réseau de Personnages dans l'Éducation Sentimentale

Questions ?

Liens utiles - BookNLP

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/*