# French BookNLP (LitBank)

Thierry Poibeau

Projet mené en collaboration avec Jean Barré, Laurette Chardon (U. Caen), Ioana Galleron, Claude Grunspan, Frédérique Mélanie, Marco Naguib, Clément Plancq, Olga Seminck

Présentation Sorbonne Université, 19/10/2023

PR[AI]RIE
PaRis Artificial Intelligence Research InstitutE

ENS | PSL

cnrs

# L’ère du big data (1/3)

## How Not to Read a Million Books

**by Tanya Clement, Sara Steger, John Unsworth, Kirsten Uszkalo**

**October, 2008**

[Figure 1] First of all, where does
Million Books Project, which beg
at Carnegie Mellon University; th
million books (“less than 1% of al
Egypt. The “million book” goal w
most notably Google Print (now k
Fair in October 2004, and which h
number equal to all the titles in W
books for years, but these massive
million books?”—a question first
whatever you do, you don't read th

**D-Lib Magazine**
**March 2006**

**Volume 12 Number 3**

**ISSN 1082-9873**

## What Do You Do with a Million Books?

Gregory Crane
Tufts University
<gregory.crane@tufts.edu>

### Introduction

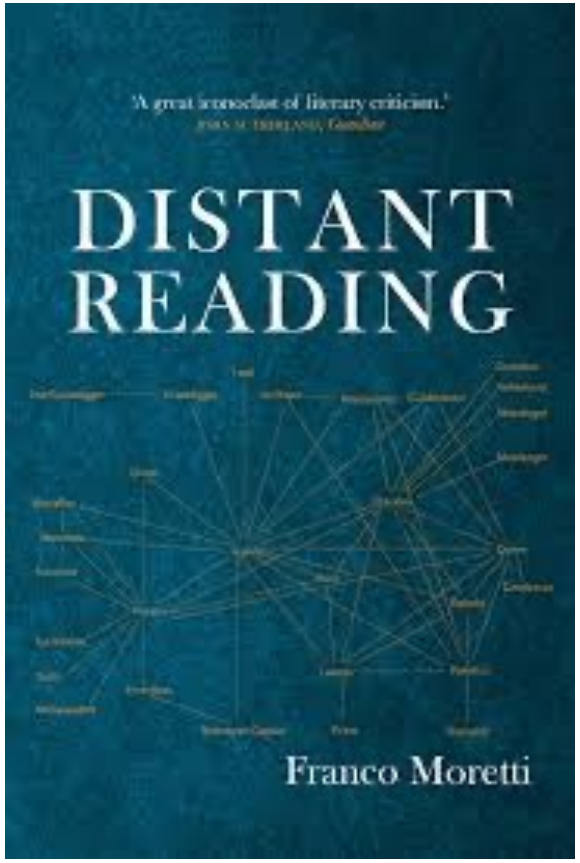
The Greek historian Herodotus has the Athenian sage Solon estimate the lifetime of a human being at c. 26,250 days (Herodotus, *The Histories*, 1.32). If we could read a book on each of those days, it would take almost forty lifetimes to work through every volume in a single million book library. The continuous tradition of written European literature that began with the *Iliad* and *Odyssey* in the eighth century BCE is itself little more than a million days old. While libraries that contain more than one million items are not unusual, print libraries never possessed a million books of use to any one reader. The great libraries that took shape in the nineteenth and twentieth centuries were meta-structures, whose catalogues and finding aids allowed readers to create their own customized collections, building on the fixed classification schemes and disciplinary structures that took shape in the nineteenth century.

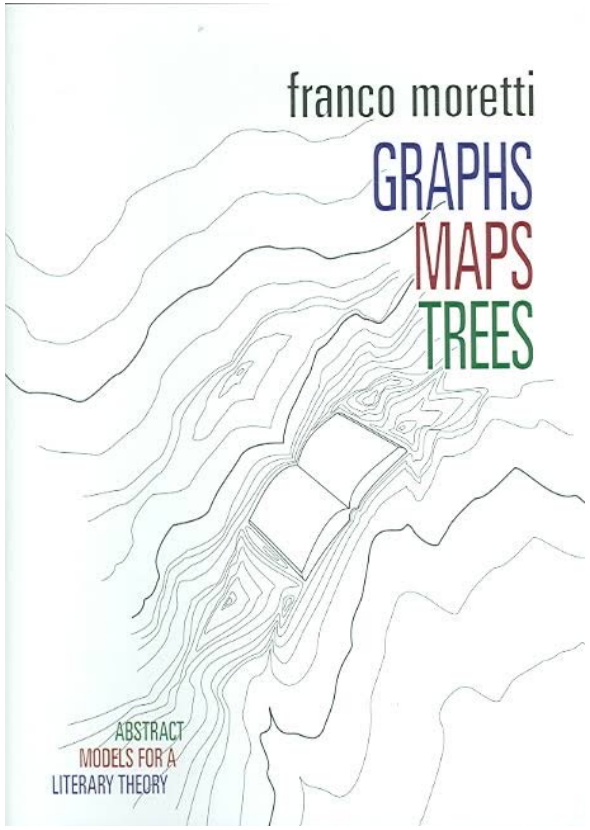
# L’ère du big data (2/3)

Qualitative researchers arrive at the medialab bringing rich data and longing to explore them. Their problem is that qualitative data cannot be easily fed into network analysis tools. Quantitative data can have many different forms (from a video recording to the very memory of the researcher), but they are often stored in a textual format (i.e. interviews transcriptions, field notes or archive documents...). The question therefore becomes: how can texts be explored quail-quantitatively? Or, more pragmatically, how can texts be turned into networks?

« Once Upon a Text: an ANT Tale in Text Analysis »
Tommaso Venturini and Daniele Guido (médialab Sciences Po)

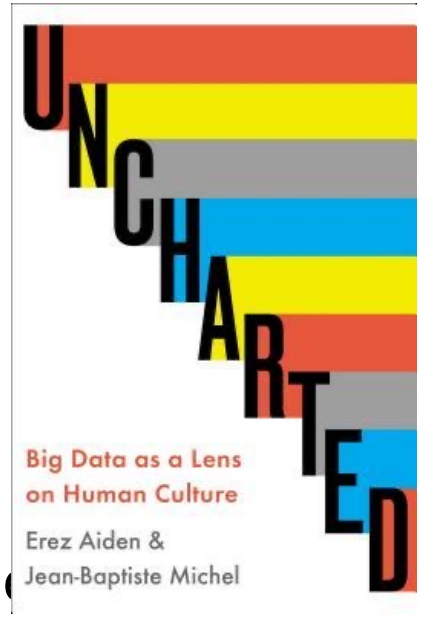
# L’ère du big data (3/3)



# BookNLP : motivations

- Projet lancé initialement par Bamman (Berkeley) en 2014
- Motivations
  - Disponibilité de corpus importants de littérature en anglais (Hathi trust)
  - Nouvelles perspectives d'analyse (cf. Google Books, Google n-grams) – Culturomics
- Mais manque d'outils dédiés
  - Analyse des personnages <> reconnaissance des entités nommé
  - Analyse de la coréférence, analyse des événements, etc.



# A Bayesian Mixed Effects Model of Literary Character

**David Bamman**
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
dbamman@cs.cmu.edu

**Ted Underwood**
Department of English
University of Illinois
Urbana, IL 61801, USA
tunder@illinois.edu

**Noah A. Smith**
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
nasmith@cs.cmu.edu

## Abstract

We consider the problem of automatically inferring latent character types in a collection of 15,099 English novels published between 1700 and 1899. Unlike prior work in which character types are assumed responsible for probabilistically generating *all* text associated with a character, we introduce a model that employs multiple effects to account for the influence of extra-linguistic information (such as author). In an empirical evaluation, we find that this method leads to improved agreement with the preregistered judgments of a literary scholar, complementing the results of alternative models.

learning: entity types learned in this way will be increasingly similar the more similar the domain, author, and other extra-linguistic effects are between them.[1] While in many cases the topically similar types learned under this assumption may be desirable, we explore here the alternative, in which entity types are learned in a way that controls for such effects. In introducing a model based on different assumptions, we provide a method that complements past work and provides researchers with more flexible tools to infer different kinds of character types.

We focus here on the literary domain, exploring a large collection of 15,099 English novels published in the 18th and 19th centuries. By accounting for the influence of individual authors while inferring latent character types, we are able to learn personas that cut across different authors more ef-

ACL, 2014

# Multilingual BookNLP

- Nouveau projet dans la continuité de BookNLP obtenu par D. Bamman en 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 ) pour améliorer les performances de l'analyseur
  - Etendre le modèle à  4 langues (en plus de l'anglais) : japonais, russe, allemand, espagnol + proposer une méthode pour d'autres langues

  - Le Lattice s'est proposé de développer la branche pour le français

# BookNLP : en pratique

- Il faut des corpus !
  - Democrat fournit une bonne base
  - Taille similaire à BookNLP/LitBank (2014) pour l'anglais
  - Annotations en chaînes de coréférence
- Travail d'annotation important
  - Sélection des annotations pertinentes pour BookNLP (par rapport à Democrat)
  - Ajour d'annotations requises par BookNLP
- Travail d'ingénierie important
  - Choix et entraînement de modèles

# Corpus

# Corpus

- 22 extraits de romans français du 19e et début 20e siècle
  - Libres de droits
  - Diffusés librement et gratuitement
  - Fichiers du corpus Democrat : disponibles sur Ortolang
  - Chaque fichier = une vingtaine de pages de roman

- https://github.com/lattice-8094/fr-litbank

# Corpus (cf. Github)

| Date | Author | Title | annot |
|---|---|---|---|
| 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 | |

# Annotation

# Campagne d’annotation

- Janvier – Août 2021
- Quelques textes annotés en triple aveugle, puis discussions / adjudication
- Le reste du corpus -> annotation simple avec révision rapide
- Annotation avec TXM

- Corpus de 184 000 mots, 14 208 entités annotées

# Entités

- Principalement centré sur la notion de personnage et de lieu
- On reprend les catégories de Bamman de manière assez fidèle

- Personnages / pers. non humains (chez Bamman une seule catégorie)
- Lieux : GPE, Loc, Fac
- Véhicules
- Indications temporelles, fêtes, etc.
- Chez Bamman, en plus (2014) : supersense (en fait, familles sémantiques, animaux, vêtements…)

1 PROLOGUE THE STORMING OF SERINGAPATAM ( 1799 ) Extracted from a Family Paper I address these lines -- written in India -- to my relatives in England .

2 My object is to explain the motive which has induced me to refuse the right hand of friendship to my cousin , John Herncastle .

3 The reserve which I have hitherto maintained in this matter has been misinterpreted by members of my family whose good opinion I can not consent to forfeit .

4 I request them to suspend their decision until they have read my narrative .

5 And I declare , on my word of honour , that what I am now about to write is , strictly and literally , the truth .

6 The private difference between my cousin and me took its rise in a great public event in which we were both concerned -- the storming of Seringapatam , under General Baird , on the 4th of May , 1799 .

7 In order that the circumstances may be clearly understood , I must revert for a moment to the period before the assault , and to the stories current in our camp of the treasure in jewels and gold stored up in the Palace of Seringapatam .

8 II One of the wildest of these stories related to a Yellow Diamond -- a famous gem in the native annals of India .

9 The earliest known traditions describe the stone as having been set in the forehead of the four-handed Indian god who typifies the Moon .

10 Partly from its peculiar colour , partly from a superstition which represented it as feeling the influence of the deity whom it adorned , and growing and lessening in lustre with the waxing and waning of the moon , it first gained the name by which it continues to be known in India to this day -- the name of THE MOONSTONE .

11 A similar superstition was once prevalent , as I have heard , in ancient Greece and Rome ; not applying , however ( as in India ) , to a diamond devoted to the service of a god , but to a semi-transparent stone of the inferior order of gems , supposed to be affected by the lunar influences -- the moon , in this latter case also , giving the name by which the stone is still known to collectors in our own time .

12 The adventures of the Yellow Diamond begin with the eleventh century of the Christian era .

13 At that date , the Mohammedan conqueror , Mahmoud of Ghizni , crossed India ; seized on the holy city of Somnauth ; and stripped of its treasures the famous temple , which had stood for centuries -- the shrine of Hindoo pilgrimage , and the wonder of the Eastern world .

Extrait de LitBank (visualisation avec Brat, sur le serveur du labo)

# Entités : problèmes d'annotation

- 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 ? (cf. Bamman « tout peut être un lieu »)
  - VEH : animaux ?
- Question de la robustesse

— C'est toi, la *Goualeuse* (1) — dit l'homme en blouse — tu vas me payer l'*eau d'aff* (2), ou je te fais danser sans violons!

— Je n'ai pas d'argent — répondit la femme en tremblant; car cet homme inspirait une grande terreur dans le quartier.

— Si ta *filoche* est *à jeun* (3), l'*ogresse* du tapis-franc te fera crédit sur ta bonne mine.

— Mon Dieu..... je lui dois déjà le loyer des vêtements que je porte.....

— Ah! tu raisonnes? — s'écria le Chourineur; et il donna dans l'ombre et au hasard un si violent coup de poing à cette malheureuse, qu'elle poussa un cri de douleur aigu.

Eugène Sue, *Les Mystères de Paris* (Gallica)

# Coréférence

- Democrat fournit une base solide pour le travail
- Typage des entités non typées dans Democrat
  - On n’annote pas les noms abstraits, etc
  - On ne retient que les chaînes correspondant à un type faisant partie du modèle BookNLP
- Vérification manuelle d’annotations Democrat
  - Peu d’erreurs dans l’ensemble -> on ne corrige pas (ou minimalement) Democrat

# Events chez Bamman

- Annotation parfois surprenante chez Bamman
- La tête peut être un nom (fog, rain)
- Conventions d'annotation :
  - Pas d'annotation si le V est lié à un modal (*it could rain*)
  - Pas d'annotation en cas de négation
  - Pas d'annotation si le V n'est pas dans le récit principal

- On essaie d'adapter l'annotation (élargissement) et les normes BookNLP mais annotation plus « large »

1 CHAPTER I In Chancery London .
2 Michaelmas term lately over , and the Lord Chancellor sitting in Lincoln 's Inn Hall .
3 Implacable November weather .
4 As much mud in the streets as if the waters had but newly retired from the face of the earth , and it would not be wonderful to meet a Megalosaurus , forty feet long or so , waddling like an elephantine lizard up Holborn Hill .
5 Smoke lowering down from chimney-pots , making a soft black drizzle , with flakes of soot in it as big as full-grown snowflakes -- gone into mourning , one might imagine , for the death of the sun .
6 Dogs , undistinguishable in mire .
7 Horses , scarcely better ; splashed to their very blinkers .
8 Foot passengers , jostling one another 's umbrellas in a general infection of ill temper , and losing their foot-hold at street-corners , where tens of thousands of other foot passengers have been slipping and sliding since the day broke ( if this day ever broke ) , adding new deposits to the crust upon crust of mud , sticking at those points tenaciously to the pavement , and accumulating at compound interest .
9 Fog everywhere .
10 Fog up the river , where it flows among green aits and meadows ; fog down the river , where it rolls defiled among the tiers of shipping and the waterside pollutions of a great ( and dirty ) city .
11 Fog on the Essex marshes , fog on the Kentish heights .
12 Fog creeping into the cabooses of collier-brigs ; fog lying out on the yards and hovering in the rigging of great ships ; fog drooping on the gunwales of barges and small boats .
13 Fog in the eyes and throats of ancient Greenwich pensioners , wheezing by the firesides of their wards ; fog in the stem and bowl of the afternoon pipe of the wrathful skipper , down

# Events dans French BookNLP

- On garde certaines conventions
  - La tête peut être un nom (pluie, brouillard)
- On annote tous les V conjugués se rapportant à des événements
  - On annote même s'il y a un modal ou une négation
  - On prévoit des scripts pour supprimer automatiquement les V avec négation et/ou modaux
- Est-ce une annotation utile ? Réutilisable ?
- Faut-il continuer à annoter les événements?

# Citations / Niveaux de discours

- Repérer les dialogues et les prises de parole (discours direct / discours indirect), rattachement personnage – prise de parole
- Niveaux de récit
  - Descriptions
  - Pensées, rêves
  - Flash-back, digressions
- Travail en collaboration avec l'équipe Obtic/Scai (Motasem Alrahabi, Glenn Roe, Carmen Brando)

# Repository Github

- Tout est en ligne
  - https://github.com/lattice-8094/fr-litbank

# Apprentissage

# Mise au point d’annotateurs dédiés

- Apprentissage à partir de l’annotation manuelle
- Format BIO multi-niveau
- Apprentissage profond
  - Version française de BERT (Camembert, plus stable que Flaubert)
  - Spécialisation de la couche de sortie (fine tuning) pour l’annotation des entités et des événements (séquences)
  - Approche similaire pour les prises de parole
  - Algo plus spécialisé pour la coréférence (cf. CROC, thèse de Loïc Grobol) : relations, et pas simplement séquences

# Entités

| precision | rappel | $F_1$ |
|---|---|---|
| 86.01 | 83,13 | 85,42 |

**Table 2 :** Test result FR-LitBank

- Résultats proches de ceux de Bamman (2014), plus de données permettrait sans doute une amélioration (cf. Bamman 2022, 5 fois plus de données)

# 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 |

**Table 3 :** Test result FR-LitBank

- Relative incohérence des résultats : mesurez non adaptées à des textes longs

# Evénements & Prises de parole

Evénements

| precision | rappel | $F_1$ |
|---|---|---|
| 51.32 | 70,73 | 61,02 |

**Table 4 :** Test result FR-LitBank

Prises de parole

| precision | rappel | $F_1$ |
|---|---|---|
| 91.95 | 90,74 | 91,34 |

**Table 5 :** Test result FR-LitBank

# Commentaires

- Résultats globalement bons malgré grande variation des exemples et volume d’annotation limité
- Tâches encore proche du TAL traditionnel
  - Entité vs personnage ?
  - Intérêt de la notion d’événement ?
- Quelles études mener à partir de là ?

# Romanciers et romancières du XIXème siècle : une étude automatique du genre sur le corpus GIRLS

Marco Naguib [1] , Marine Delaborde [1] , Blandine Andrault [1] , Anaïs Bekolo [1] , Olga Seminck [1] Détails

1 Lattice - Lattice - Langues, Textes, Traitements informatiques, Cognition - UMR 8094

**Résumé** : Cette étude porte sur les différences entre les romans français du XIXe siècle écrits par des hommes et ceux écrits par des femmes en trois étapes. Premièrement, nous observons que ces textes peuvent être distingués par apprentissage supervisé selon ce critère. Un modèle simple a un score de 99% d'exactitude sur cette tâche si d'autres œuvres de la même personne figurent dans le jeu d'entraînement, et de 72% d'exactitude sinon. Cette différence s'explique par le fait que le langage de l'individu est plus distinctif qu'un éventuel style propre au genre. Deuxièmement, notre étude textométrique met au jour des stéréotypes de genre chez les hommes et les femmes. Troisièmement, nous présentons un modèle de coréférence entraîné sur des textes littéraires pour étudier le genre des personnages. Nous montrons ainsi que les personnages féminins sont plus nombreux chez les femmes, et prennent généralement une place plus proéminente que chez les hommes.

**Mots-clés** : Genre | littérature | apprentissage automatique | modèle tranformer | coreference | textometrie | calcul de spécificités.

# Conclusion

# Conclusion & Perspectives

- Outils en partie adaptés au domaine littéraire
- Intérêt pour des études littéraires (cf. Naguib et al;, 2022)

- Qu’est-ce qu’un personnage ?
- Dériver des réseaux de personnages et caractériser les œuvres sur cette base ?
- Identifier des structures narratives propres
  - Narration / description / dialogues
  - Repérage de scènes (Darmstadt)
  - Structures de récits (flashbacks, structures imbriqués, récit dans le récit…)

- Merci pour votre attention !