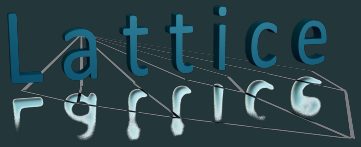


Operationalizing Canonicity

A Quantitative Study of French 19th and 20th Century Literature

Jean Barré, Thierry Poibeau, Jean-Baptiste Camps

June 19, 2023

Lattice Lab : ENS-PSL-CNRS

1. Introduction
2. Corpus
3. Determining canonical factors
4. Methods
    - Textual features
    - Statistical Modeling
5. Results
    - Canonicity at the novel scale
    - Canonicity at the author scale
6. Canonical selectivity in an author's production
7. Conclusion

Introduction

"How is the selection of works and names destined for immortality made?". (Lanson, 1895)[1]

Canon formation in the sociocultural field :

- “Selective tradition” (Pollock, 1999)[2]
- The canon “embodies literary legitimacy itself”(Casanova, 2008)[3]
- Canonization as a sociological process (Bourdieu, 1992)[4]
- Is there some textual evidence to these research ?

Computational literary studies & the literary canon

- Distant reading. (Moretti, 2000)[5]
- Literary studies not really familiar with the main structuring lines of literary history ? (Underwood, 2019)[6]

Main research questions :

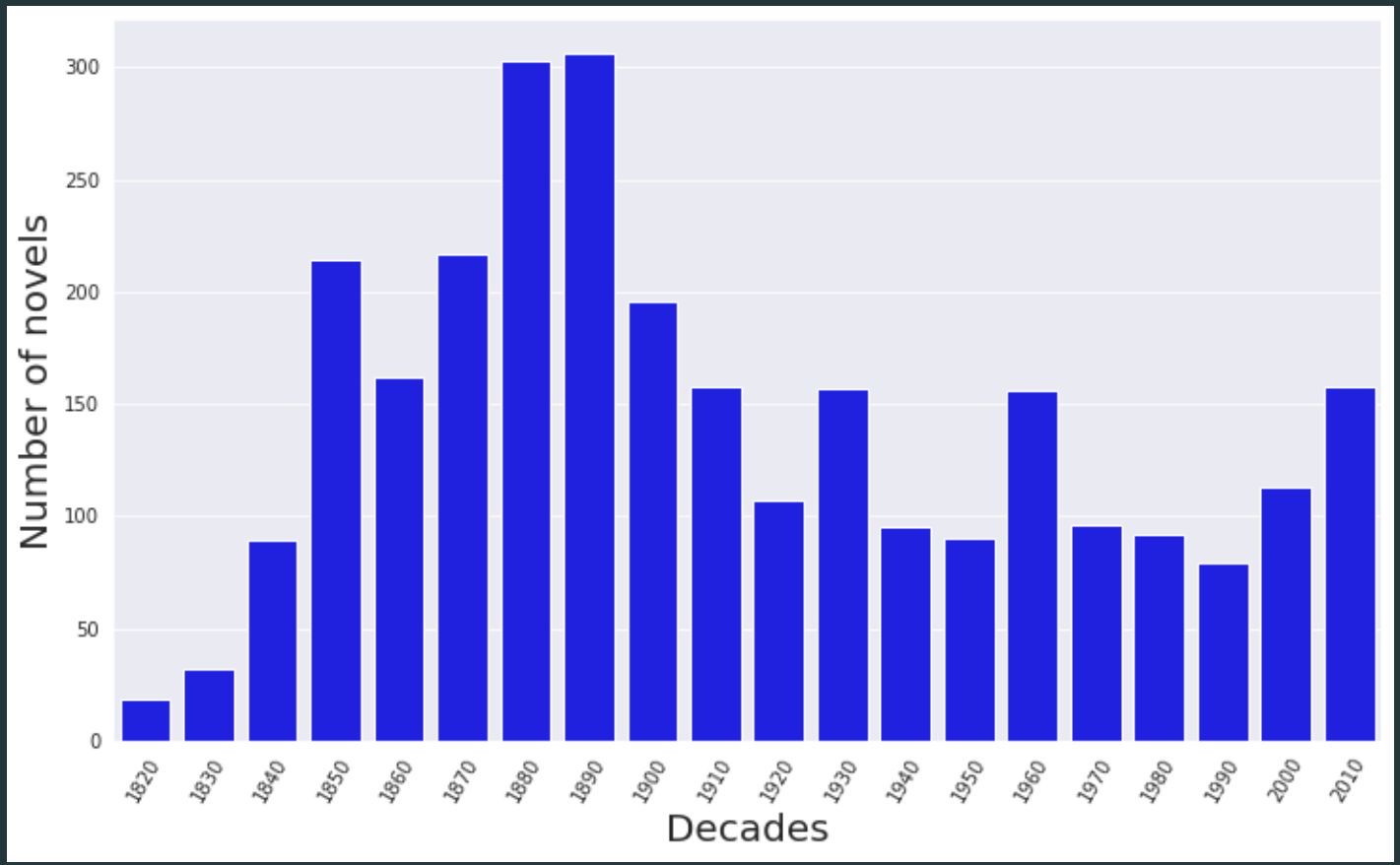
- Can we operationalize canonization process ?
- Can we link canonicity with some textual dynamics ?
- Textual properties as a causal phenomenon ? Or as a product of the canonization process ?

Corpus

Corpus

Corpus Chapitres (Leblond, 2022)[7] :

- 2960 novels
- 79.301 mean number of tokens per novel



Determining canonical factors

Determining canonical factors

The French literary canon through its contemporary reception

- The school canon

Determining canonical factors

The French literary canon through its contemporary reception

- The school canon
- The academic canon

Determining canonical factors

The French literary canon through its contemporary reception

- The school canon
- The academic canon
- The canon of the agrégation

Determining canonical factors

The French literary canon through its contemporary reception

- The school canon
- The academic canon
- The canon of the agrégation
- The canon of publishers

Determining canonical factors

The French literary canon through its contemporary reception

- The school canon
- The academic canon
- The canon of the agrégation
- The canon of publishers
- The canon of criticism

Determining canonical factors

The French literary canon through its contemporary reception

- The school canon
- The academic canon
- The canon of the agrégation
- The canon of publishers
- The canon of criticism
- The political canon

Our French Literary Canon

Twofold granularity

- novel scale - 306 items (10% of the corpus)
- author scale - 1173 items (40%)
- School institution as a canon maker ?

Cosine Similarity Heatmap

- Strong similarity between the school based factors
- literary award list clearly different

Methods

Textual features

Bag-of-words

- 1000 Most Frequent n-grams
- Focus on function words -> unconscious way of writing (stylometry) -> unconscious markers of the canon selection ?

Statistical Modeling

Binary classification

- SVM - Scikit-Learn (Pedregosa, 2011) [8]
- GroupKFold - Cross Validation pipeline dealing with idiolectal bias
- Baselines
- Metrics (Balanced accuracy, f1 score, ...)

Results

Balanced accuracy : 0.708 Baseline : 0.496

	precision	recall	f1-score	support
canon	0.728	0.668	0.697	306
non_canon	0.691	0.748	0.719	304
full dataset				610

Table 1: Results of the evaluation of the model, novel scale

Canonicity at the novel scale

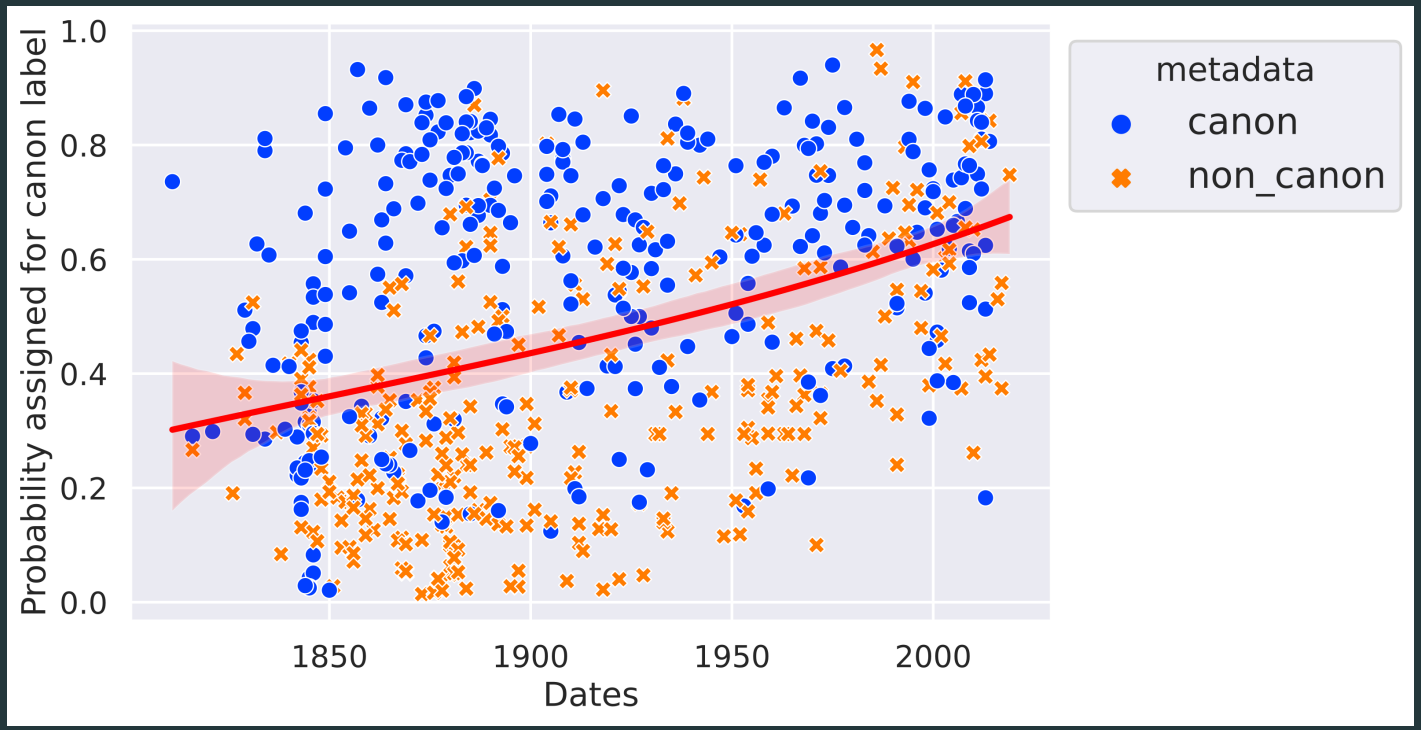


Figure 1: Predicted probability to be canonical, novel scale

Canonicity at the novel scale

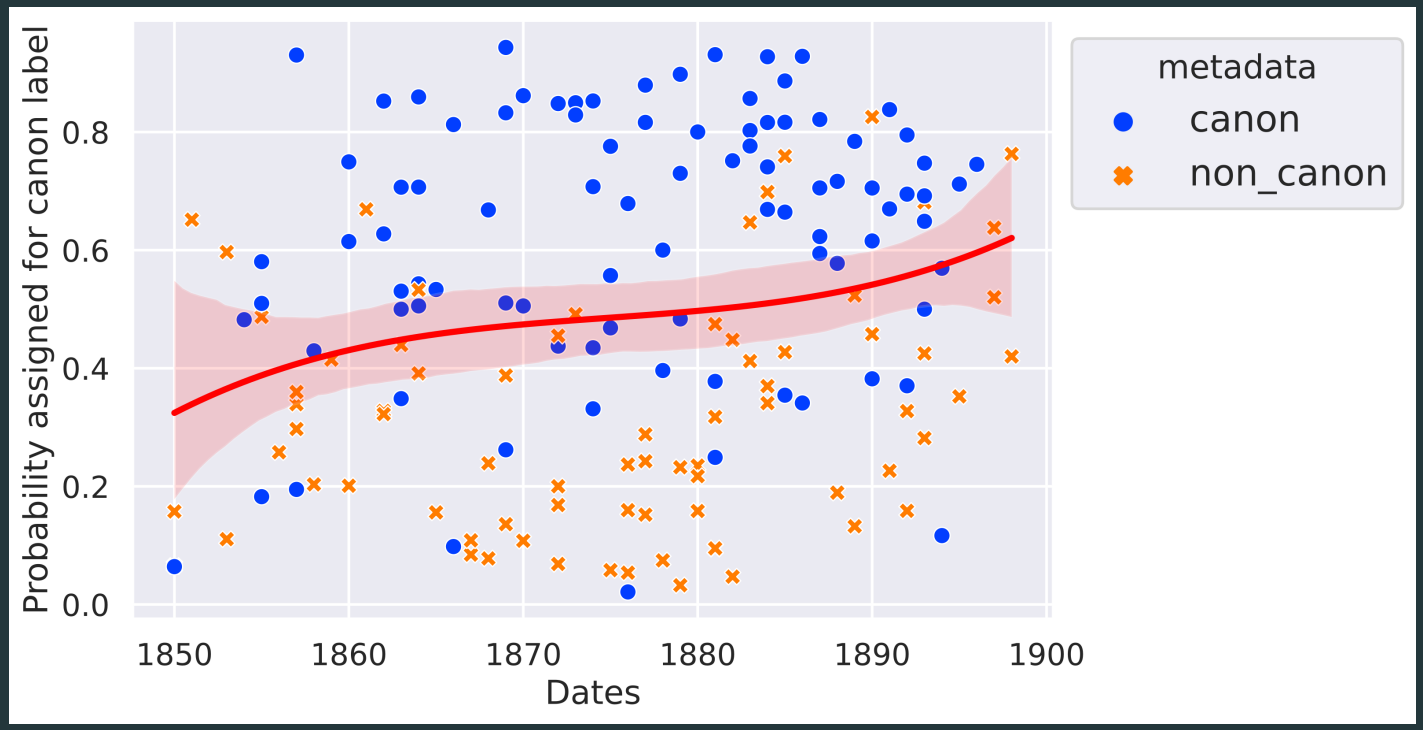


Figure 2: Predicted probability to be canonical, 1850-1900, novel scale

Balanced accuracy : 0.741 Baseline : 0.516

	precision	recall	f1-score	support
canon	0.721	0.645	0.681	1173
non_canon	0.782	0.836	0.808	1787
full dataset				2960

Table 2: Results of the evaluation of the model, author scale

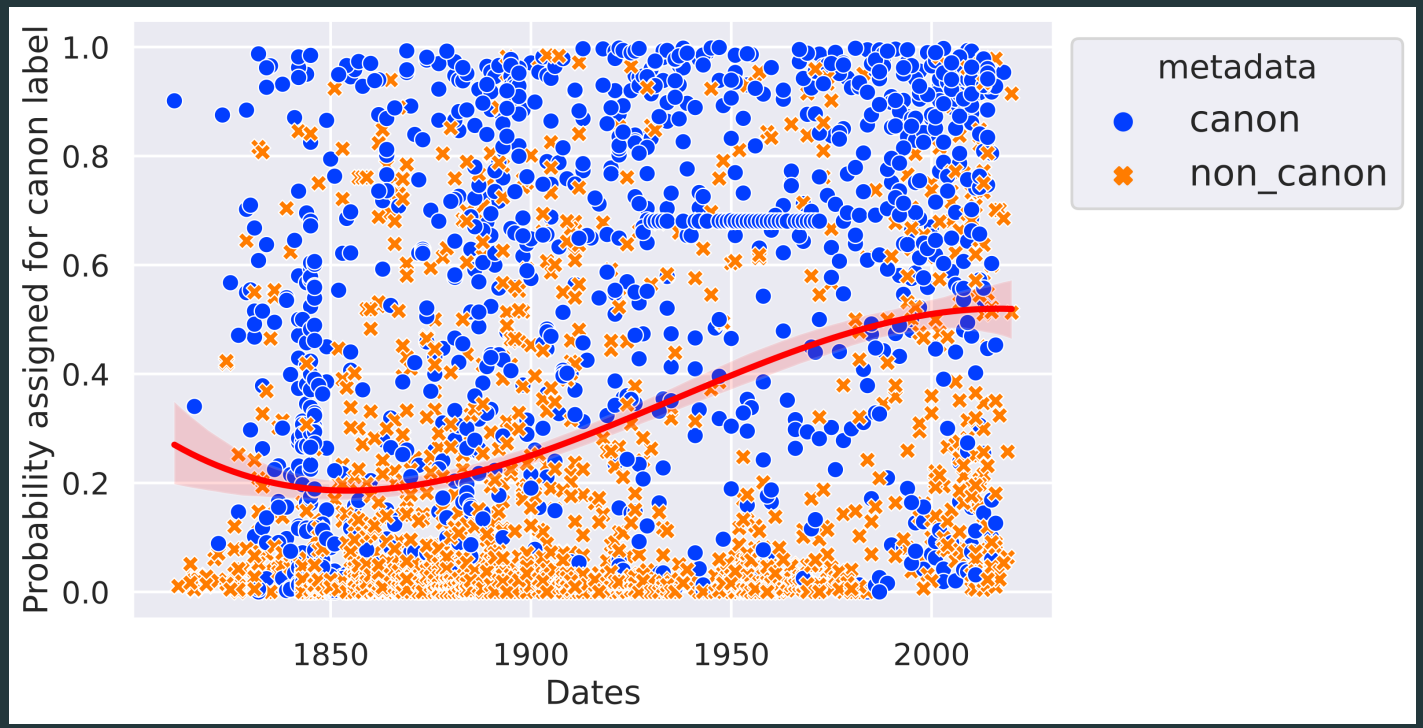


Figure 3: Predicted probability to be canonical, author scale

Canonicity at the author scale

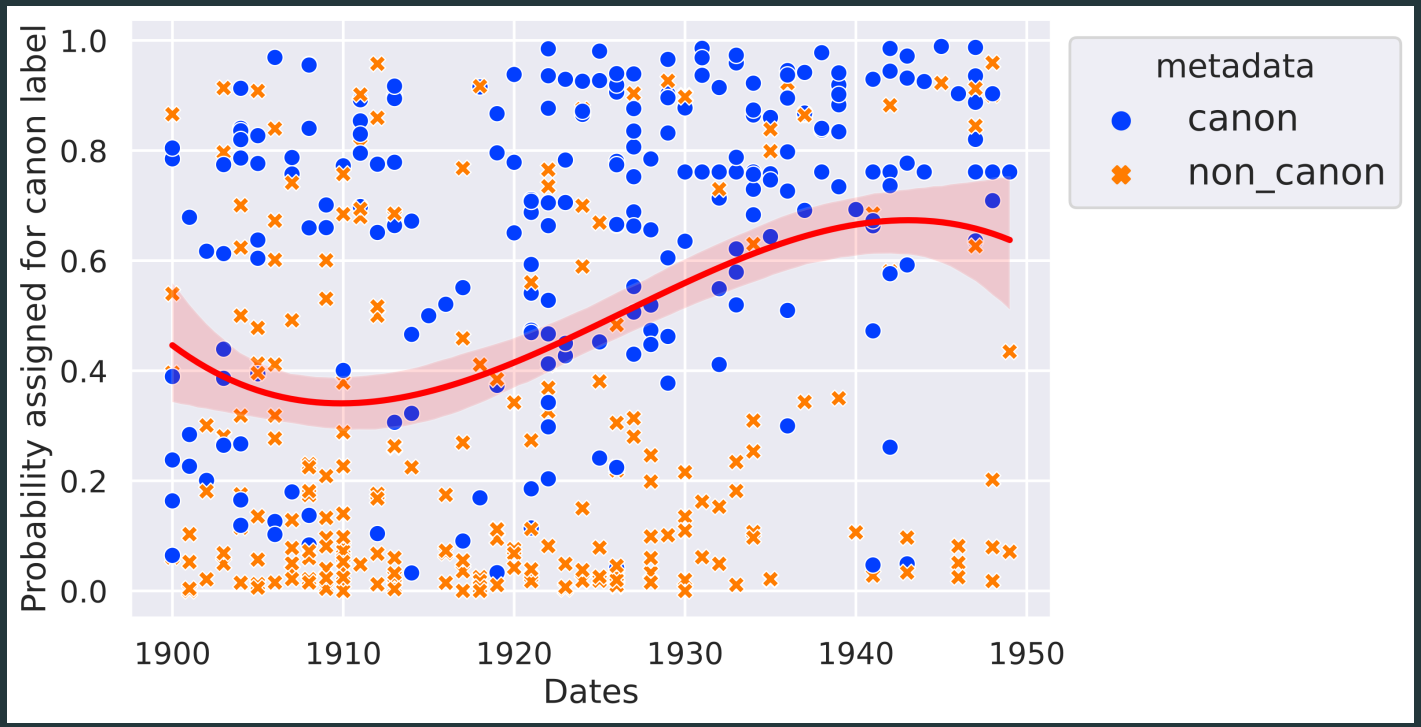
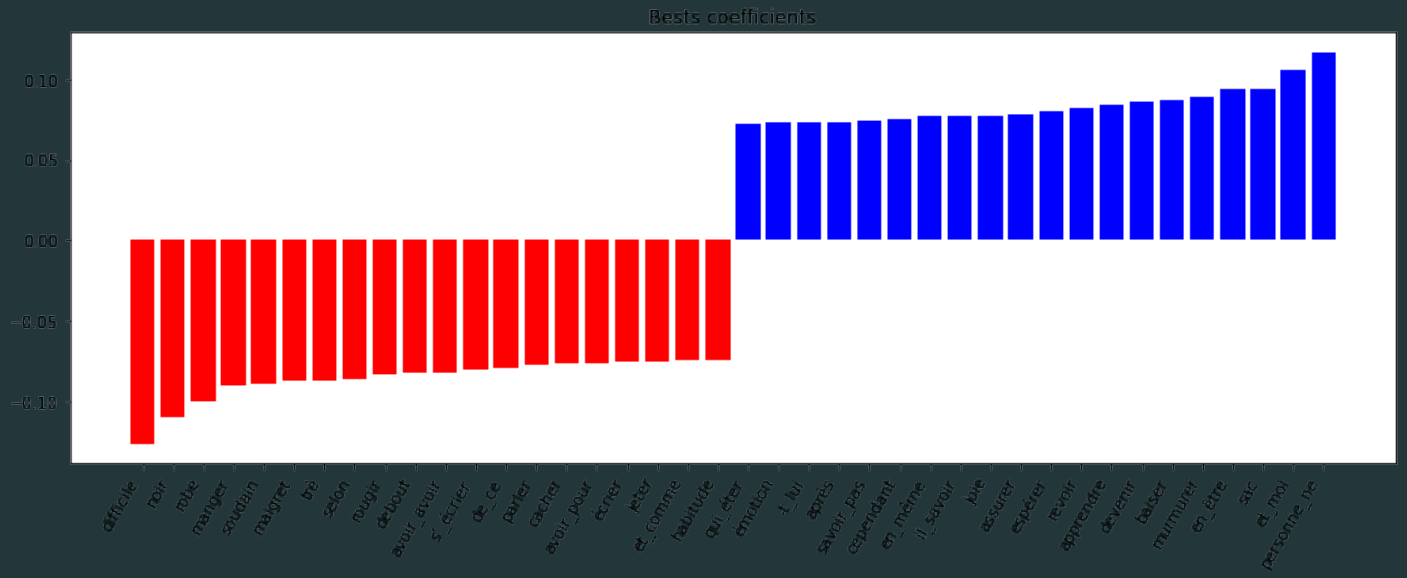


Figure 4: Predicted probability to be canonical, author scale

Model's discriminant coefficient



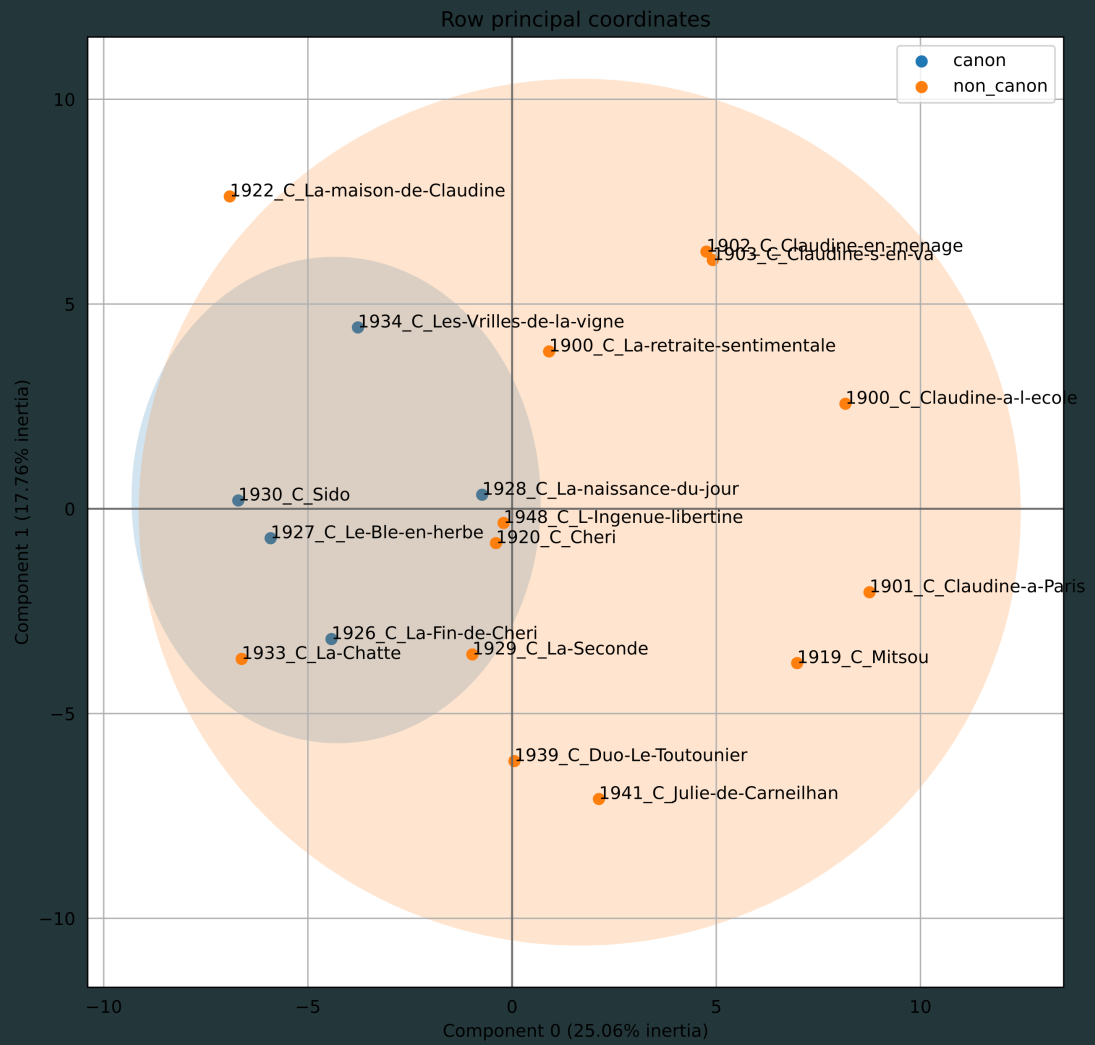
Model coefficients insights

- Complexity of the sentence (auxiliaries, conjunctions, substantive nouns)
- A more colloquial register for non-canonized novels

Canonical selectivity in an author’s production

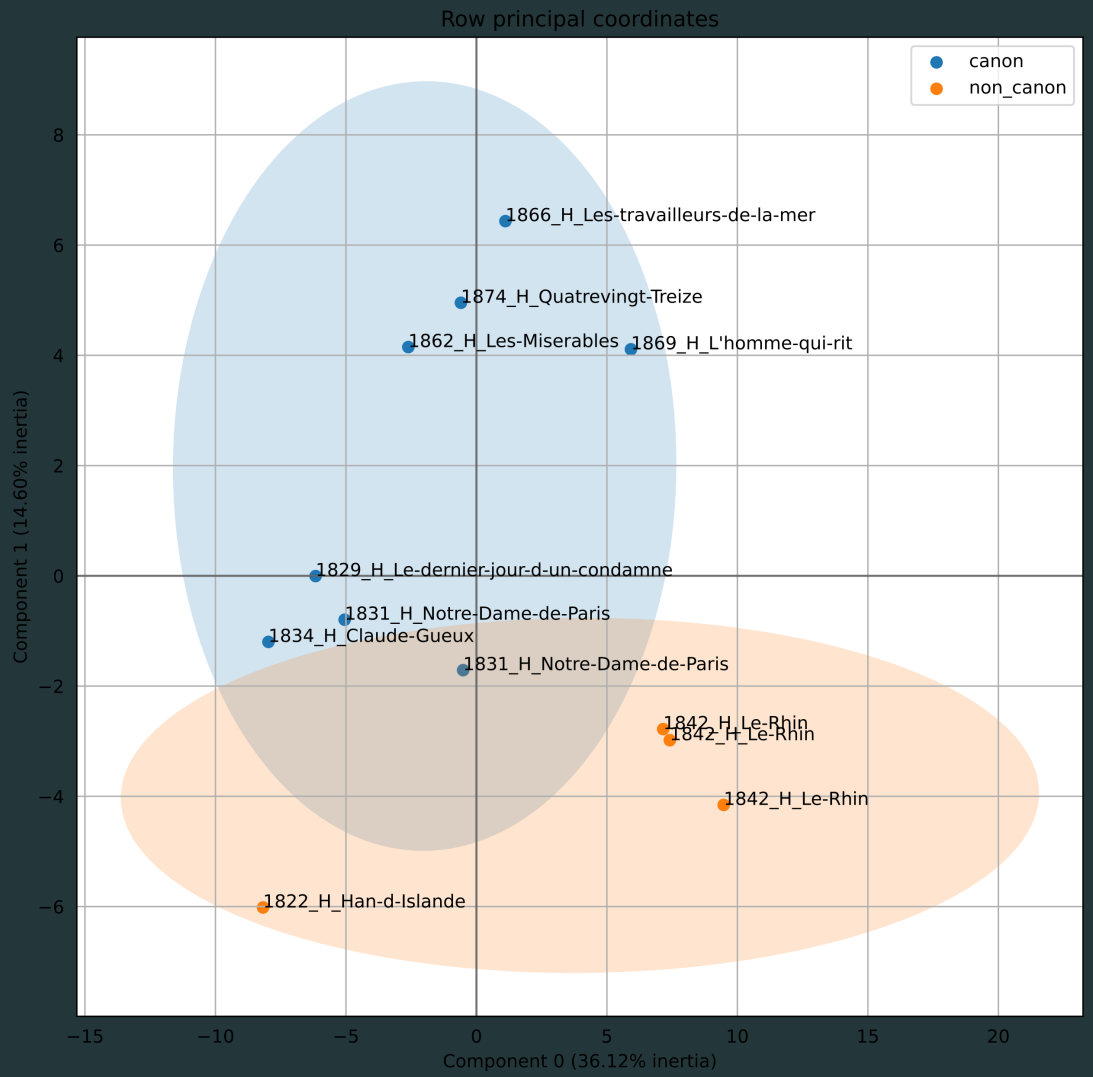
PCA Colette

- The *Claudines* series
- Not only an idiolec-tal/chronolectal drift

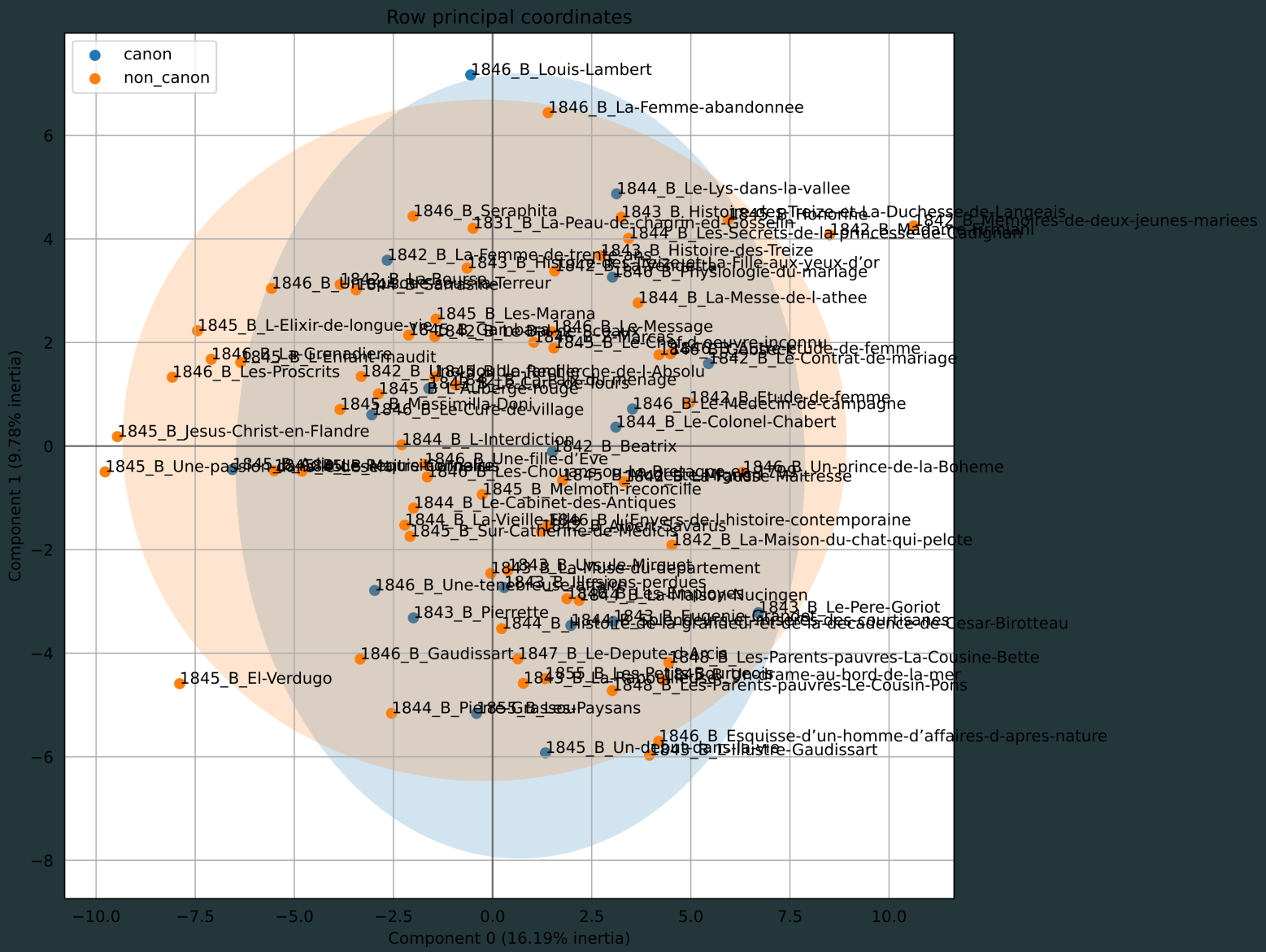


PCA Victor Hugo

- The three volumes of *Le Rhin* - A travel guide
- *Han d'Islande* The young Hugo



PCA Balzac


Row principal coordinates
canon
non_canon
1846_B_Louis-Lambert
1846_B_La-Femme-abandonnee
1844_B_Le-Lys-dans-la-vallee
1846_B_Seraphita
1831_B_La-Peau-de-chagrin-ed-Gosselin
1844_B_Les-Secrets-de-la-princesse-de-Cadignan
1842_B_La-Femme-de-trente-ans
1843_B_Histoire-des-Treize
1842_B_La-Fille-aux-yeux-d'or
1846_B_Physiologie-du-mariage
1844_B_La-Messe-de-l-athee
1845_B_Les-Marana
1845_B_L-Elixir-de-longue-vie
1846_B_Le-Message
1842_B_Le-Contrat-de-mariage
1846_B_Les-Proscrits
1842_B_Etude-de-femme
1845_B_Jesus-Christ-en-Flandre
1844_B_Le-Colonel-Chabert
1844_B_L-Interdiction
1842_B_Beatrix
1846_B_Une-fille-d'Eve
1846_B_Un-prince-de-la-Boheme
1845_B_Melmoth-reconcilie
1844_B_Le-Cabinet-des-Antiques
1845_B_Sur-Catherine-de-Medicis
1842_B_La-Maison-du-chat-qui-pelote
1843_B_Pierrette
1846_B_Gaudissart
1847_B_Le-Depute-d-Arcis
1845_B_El-Verdugo
1844_B_Pierre-Grassou
1846_B_Esquisse-d'un-homme-d'affaires-d-apres-nature
Component 0 (16.19% inertia)
Component 1 (9.78% inertia)

Conclusion

Futur Work

- Retrieve metadata diachronically: Recovering the social context of reception over time, as reception is filtered
- We could also fragment our view of the canon by agents in the literary field (editions, textbooks, academic prestige, literary journals, ...)
- Word / Paragraph embeddings ?
- Recovering and analyzing “canonical” excerpts in close reading

Conclusion

- We provide an operable definition of the notion of literary canon
- A statistical model can predict canonicity with 70% to 74% accuracy
- The model produces relatively valid criteria for specific time span, but fails for two centuries of literary production
- This detected norm support the sociocultural research on the Canonization process and add a formal aspect to it.
- We assume that this norm is the result of biased latent selection mechanisms that are producing literary value and literary “immortality”.
- Literature is also determined by its own institutions and conventions, by its own mechanisms of production and reproduction.

Thank you for your attention !

Bibliography i


Gustave Lanson.
Hommes et livres: études morales et littéraires.
Hachette livre-bnf, 1895.


Griselda Pollock.
Differencing the canon: feminist desire and the writing of art's histories.
Re visions. Routledge, 1999.


Pascale Casanova.
La république mondiale des lettres.
Number 607 in Points Série essais. Éditions du Seuil, Édition revue et corrigé e edition, 2008.

Bibliography ii


Pierre Bourdieu.
Les règles de l'art.
Éditions du Seuil, 1992.


Franco Moretti.
Conjectures on world literature.
New Left Review, 2000.


Ted Underwood.
Distant horizons: digital evidence and literary change.
The University of Chicago Press, 2019.


Aude Leblond.
Corpus chapitres, December 2022.

Bibliography iii


F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay.
Scikit-learn: Machine learning in Python.
Journal of Machine Learning Research, 12:2825–2830, 2011.

Feel free to reach us !

jean.barre@ens.psl.eu, thierry.poibeau@ens.psl.eu

https://crazyjeannot.github.io/