

Peaks, Tails, and Genre Canons

Measuring Long-Term Influence in French Popular Fiction

Jean Barré

Mapping the Canon: Quantitative Approaches to Literary History

Technical University of Darmstadt — June 19, 2026

LaTTiCe (CNRS, ENS-PSL, Sorbonne Nouvelle)

The slaughterhouse of literature

Moretti (2000)

The canon works like a **slaughterhouse**: of the thousands of novels published, only a handful survive the selection operated by critics, universities, and the market.

The slaughterhouse of literature

Moretti (2000)

The canon works like a **slaughterhouse**: of the thousands of novels published, only a handful survive the selection operated by critics, universities, and the market.

Cohen (1999): literature “hors d’usage”

Fiction that was historically central — read, serialized, shared — but now stands **outside literary memory**.

The slaughterhouse of literature

Moretti (2000)

The canon works like a **slaughterhouse**: of the thousands of novels published, only a handful survive the selection operated by critics, universities, and the market.

Cohen (1999): literature “hors d’usage”

Fiction that was historically central — read, serialized, shared — but now stands **outside literary memory**.

This talk

What happens when we look for the canon **inside the archive**?

Do popular genres run their own slaughterhouse — and can we **measure** who survives it?

A paradox: canonical authors of non-canonical genres



Alexandre Dumas

historical romance

Les Trois Mousquetaires
(1844)



Eugène Sue

roman-feuilleton

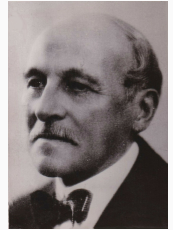
Les Mystères de Paris (1842)



Jules Verne

scientific adventure

Voyage au centre de la Terre
(1864)



Maurice Leblanc

detective fiction

Arsène Lupin (1905–)

A paradox: canonical authors of non-canonical genres



Alexandre Dumas

historical romance

Les Trois Mousquetaires
(1844)



Eugène Sue

roman-feuilleton

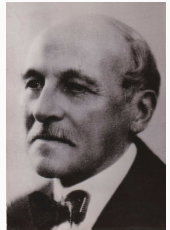
Les Mystères de Paris (1842)



Jules Verne

scientific adventure

Voyage au centre de la Terre
(1864)



Maurice Leblanc

detective fiction

Arsène Lupin (1905–)

All four emerged from **serial, genre-bound, mass-market fiction** —
yet all four are durably **canonical**: still read, adapted, translated.

A paradox: canonical authors of non-canonical genres



Alexandre Dumas

historical romance

Les Trois Mousquetaires
(1844)



Eugène Sue

roman-feuilleton

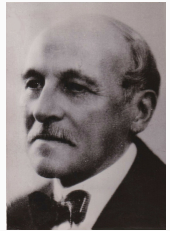
Les Mystères de Paris (1842)



Jules Verne

scientific adventure

Voyage au centre de la Terre
(1864)



Maurice Leblanc

detective fiction

Arsène Lupin (1905–)

All four emerged from **serial, genre-bound, mass-market fiction** —
yet all four are durably **canonical**: still read, adapted, translated.

Question: does that lasting canonicity leave a **measurable textual trace**
inside the genres they came from?



“Hors d’usage” corpus

10 898 French novels

(1800–1950), OCR-filtered,

deduplicated (MinHash +

metadata),

BnF/Wikidata-enriched.

JOHD 2026 • Zenodo

Infrastructure: making the archive empirically accessible



“Hors d’usage” corpus

10 898 French novels

(1800–1950), OCR-filtered,
deduplicated (MinHash +
metadata),

BnF/Wikidata-enriched.

JOHD 2026 • Zenodo



Literary embeddings

BGE-M3 fine-tuned on French
fiction; 512-token passage
embeddings, mean-pooled per
novel.

fr_literary_bge_base •

HuggingFace

Infrastructure: making the archive empirically accessible



“Hors d’usage” corpus

10 898 French novels
(1800–1950), OCR-filtered,
deduplicated (MinHash +
metadata),
BnF/Wikidata-enriched.
JOHD 2026 • Zenodo



Literary embeddings

BGE-M3 fine-tuned on French
fiction; 512-token passage
embeddings, mean-pooled per
novel.
fr_literary_bge_base •
HuggingFace



Genre labels

1 428 adventure & 830
detective novels, supervised
retrieval + graph refinement,
released with classifier
probabilities.
NLP4DH 2026

Infrastructure: making the archive empirically accessible



“Hors d’usage” corpus

10 898 French novels
(1800–1950), OCR-filtered,
deduplicated (MinHash +
metadata),
BnF/Wikidata-enriched.
JOHD 2026 • Zenodo



Literary embeddings

BGE-M3 fine-tuned on French
fiction; 512-token passage
embeddings, mean-pooled per
novel.
fr_literary_bge_base •
HuggingFace



Genre labels

**1 428 adventure & 830
detective** novels, supervised
retrieval + graph refinement,
released with classifier
probabilities.
NLP4DH 2026



Canon labels

Canonicity tags for the full
corpus, anchored in
contemporary reception proxies
(Barré et al. 2023).
1 083 canonical / 11 202 archive

Infrastructure: making the archive empirically accessible



“Hors d’usage” corpus

10 898 French novels
(1800–1950), OCR-filtered,
deduplicated (MinHash +
metadata),
BnF/Wikidata-enriched.
JOHD 2026 • Zenodo



Literary embeddings

BGE-M3 fine-tuned on French
fiction; 512-token passage
embeddings, mean-pooled per
novel.
fr_literary_bge_base •
HuggingFace



Genre labels

**1 428 adventure & 830
detective** novels, supervised
retrieval + graph refinement,
released with classifier
probabilities.
NLP4DH 2026

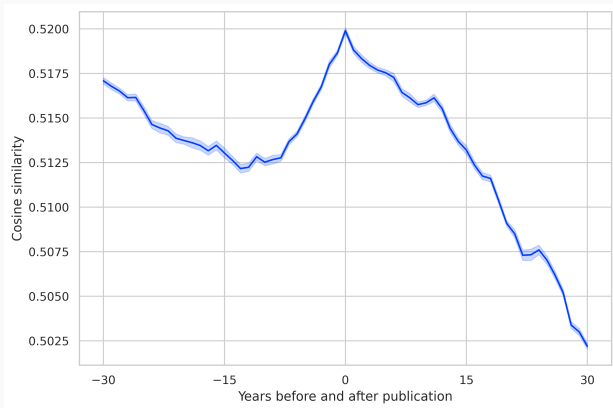


Canon labels

Canonicity tags for the full
corpus, anchored in
contemporary reception proxies
(Barré et al. 2023).
1 083 canonical / 11 202 archive

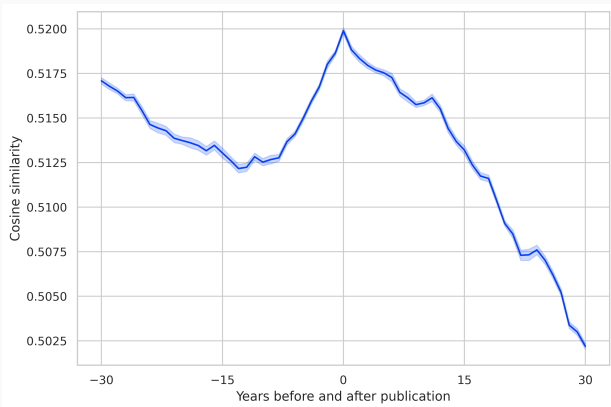
Everything used today is openly released — corpus, labels, probabilities, audit logs — and reusable for comparative canon research.

Starting point: similarity has a “style of the time”



Cosine similarity between each novel and its temporal context (Barré, CHR 2024)

Starting point: similarity has a “style of the time”

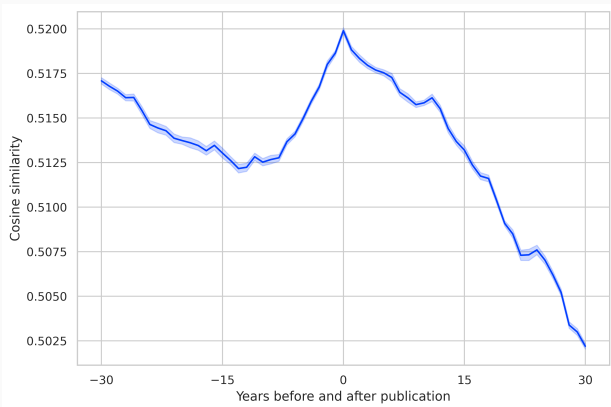


The peak

Novels resemble their **contemporaries** most: a clear similarity peak at publication year.

Cosine similarity between each novel and its temporal context (Barré, CHR 2024)

Starting point: similarity has a “style of the time”



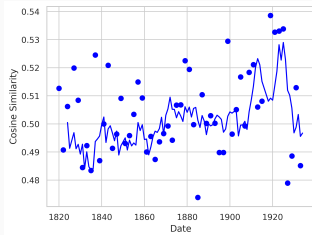
Cosine similarity between each novel and its temporal context (Barré, CHR 2024)

The peak

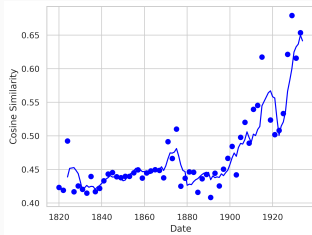
Novels resemble their **contemporaries** most: a clear similarity peak at publication year.

Protocol: fine-tuned embeddings → scaled cosine similarity → same-author pairs excluded → yearly downsampling ($\times 10$).

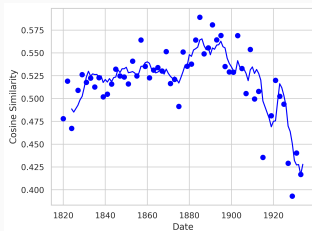
...but a few novels have long tails



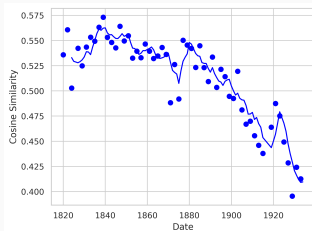
Proust, *Le temps retrouvé* (1927)



Leroux, *Chambre jaune* (1907)

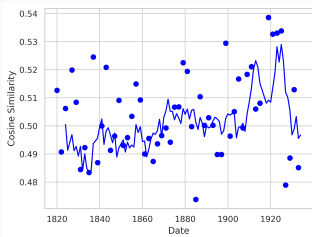


Hugo, *Les Misérables* (1862)

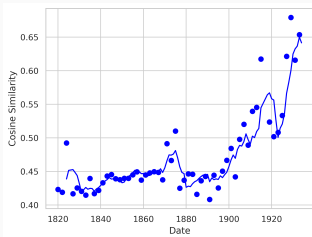


Reybaud, *Mésélie* (1839)

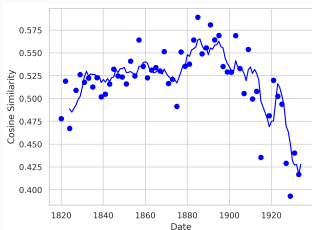
...but a few novels have long tails



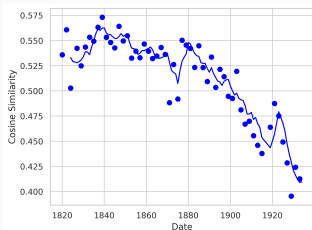
Proust, *Le temps retrouvé* (1927)



Leroux, *Chambre jaune* (1907)



Hugo, *Les Misérables* (1862)

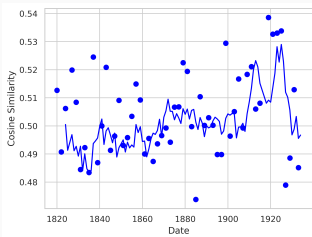


Reybaud, *Ménélie* (1839)

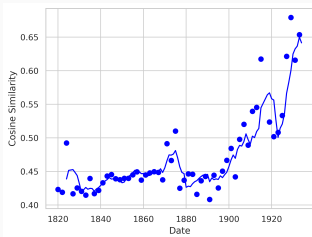
Four patterns

- **Proust**: sharp peak at publication, then fade
- **Leroux**: rising tail — the whodunit genre catches up
- **Hugo**: delayed peak, then drop (Naturalism)
- **Reybaud**: steady decline — the archive

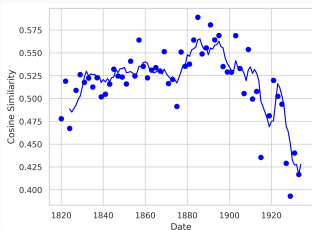
...but a few novels have long tails



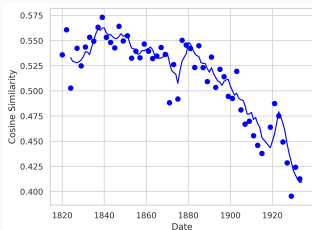
Proust, *Le temps retrouvé* (1927)



Leroux, *Chambre jaune* (1907)



Hugo, *Les Misérables* (1862)



Reybaud, *Ménélie* (1839)

Four patterns

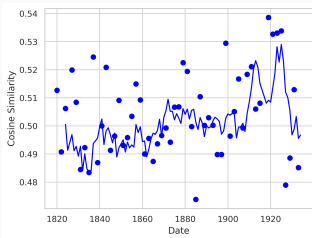
- **Proust**: sharp peak at publication, then fade
- **Leroux**: rising tail — the whodunit genre catches up
- **Hugo**: delayed peak, then drop (Naturalism)
- **Reybaud**: steady decline — the archive

The tail

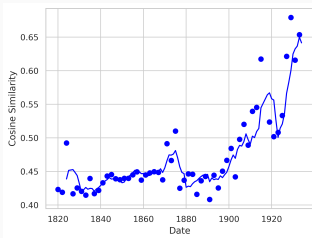
Leroux's rising curve is precisely what **genre ancestry** looks like:

influence *accumulates* as the genre crystallizes around you.

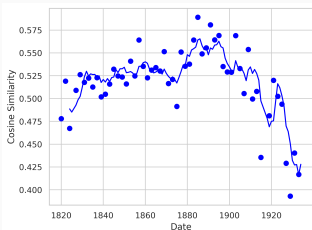
...but a few novels have long tails



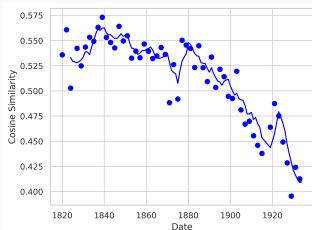
Proust, *Le temps retrouvé* (1927)



Leroux, *Chambre jaune* (1907)



Hugo, *Les Misérables* (1862)



Reybaud, *Ménélie* (1839)

Four patterns

- **Proust**: sharp peak at publication, then fade
- **Leroux**: rising tail — the whodunit genre catches up
- **Hugo**: delayed peak, then drop (Naturalism)
- **Reybaud**: steady decline — the archive

The tail

Leroux's rising curve is precisely what **genre ancestry** looks like:

influence *accumulates* as the genre crystallizes around you.

Canon-tagged novels concentrate in the **Leroux / Hugo** patterns — not Reybaud's.

From canonicity to influence *within* genres

Shift of this talk: from “does canonicity have a textual trace in the whole corpus?”
to “do genres have internal **ancestors** — and are they the authors literary history
canonized?”

From canonicity to influence *within* genres

Shift of this talk: from “does canonicity have a textual trace in the whole corpus?”
to “do genres have internal **ancestors** — and are they the authors literary history
canonized?”

The bestseller

Dominates its decade,
then **fades** with it.

High peak, short tail.

From canonicity to influence *within* genres

Shift of this talk: from “does canonicity have a textual trace in the whole corpus?” to “do genres have internal **ancestors** — and are they the authors literary history canonized?”

The bestseller

Dominates its decade,
then **fades** with it.

High peak, short tail.

The genre ancestor

Keeps shaping the genre’s narrative space
generations later.

High persistence, within-genre specificity.

From canonicity to influence *within* genres

Shift of this talk: from “does canonicity have a textual trace in the whole corpus?” to “do genres have internal **ancestors** — and are they the authors literary history canonized?”

The bestseller

Dominates its decade,
then **fades** with it.

High peak, short tail.

The genre ancestor

Keeps shaping the genre’s narrative space
generations later.

High persistence, within-genre specificity.

Two modes of textual success that external sales or reception data **cannot** distinguish — but similarity over time can.

Method: forward, genre-conditioned similarity

For each novel in a genre subcorpus:

1. measure cosine similarity to genre texts published $+t$ years later (5-year bins, up to $+50$);

Method: forward, genre-conditioned similarity

For each novel in a genre subcorpus:

1. measure cosine similarity to genre texts published $+t$ years later (5-year bins, up to $+50$);
2. downsample each bin (10–50 texts, $\times 10$ draws), exclude same-author pairs;

Method: forward, genre-conditioned similarity

For each novel in a genre subcorpus:

1. measure cosine similarity to genre texts published $+t$ years later (5-year bins, up to +50);
2. downsample each bin (10–50 texts, $\times 10$ draws), exclude same-author pairs;
3. **z-score within each calendar bin**: a novel's score is relative to all genre texts “competing” for the same future.

Method: forward, genre-conditioned similarity

For each novel in a genre subcorpus:

1. measure cosine similarity to genre texts published $+t$ years later (5-year bins, up to $+50$);
2. downsample each bin (10–50 texts, $\times 10$ draws), exclude same-author pairs;
3. **z-score within each calendar bin**: a novel's score is relative to all genre texts “competing” for the same future.

Step 3 is what prevents corpus drift and density artefacts from masquerading as persistence.

Method: forward, genre-conditioned similarity

For each novel in a genre subcorpus:

1. measure cosine similarity to genre texts published $+t$ years later (5-year bins, up to +50);
2. downsample each bin (10–50 texts, $\times 10$ draws), exclude same-author pairs;
3. **z-score within each calendar bin:** a novel's score is relative to all genre texts “competing” for the same future.

Step 3 is what prevents corpus drift and density artefacts from masquerading as persistence.

Three scores per novel

- **Peak:** similarity at ± 2 years (contemporaneity)
- **Persistence:** mean z at +20 to +40 years (the long tail)
- **Genre specificity:** persistence inside genre – outside (ancestor of the *genre*, not of the period)

Case studies: the four candidates, within their genres

img/case_studies_adventure.png

Forward similarity, adventure fiction (n=1428)

img/case_studies_detective.png

Forward similarity, detective fiction (n=830)

Case studies: the four candidates, within their genres

img/case_studies_adventure.png

Forward similarity, adventure fiction (n=1428)

img/case_studies_detective.png

Forward similarity, detective fiction (n=830)

Bestsellers vs ancestors: the quadrant

img/quadrant_adventure.png

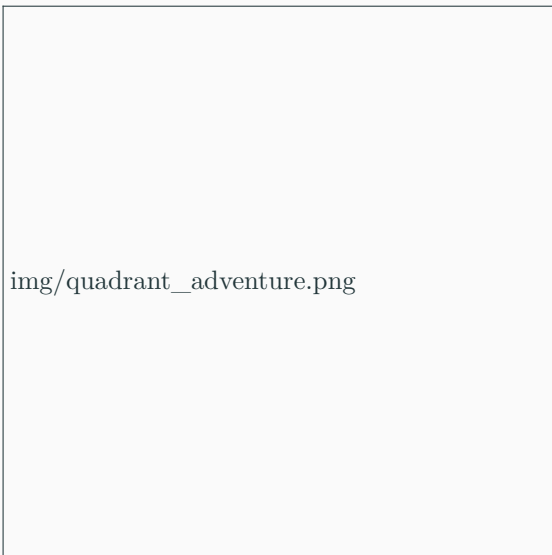
Bestsellers vs ancestors: the quadrant

img/quadrant_adventure.png

Reading the plane

- top-right: **ancestors** — of their time *and* of the genre's future
- bottom-right: bestsellers — peaked, then faded
- top-left: “sleepers” — ahead of their genre?

Bestsellers vs ancestors: the quadrant



Reading the plane

- top-right: **ancestors** — of their time *and* of the genre's future
- bottom-right: bestsellers — peaked, then faded
- top-left: “sleepers” — ahead of their genre?

Canon-tagged novels concentrate in the **high-persistence** region.

A hidden genre canon

Author	Title	Year	Canon
Verne	...	18..	✓
Dumas	...	18..	✓
???	<i>(forgotten name)</i>	18..	—
Sue	...	18..	✓
???	<i>(forgotten name)</i>	18..	—

Top persistent novels, adventure fiction (excerpt)

A hidden genre canon

Author	Title	Year	Canon
Verne	...	18..	✓
Dumas	...	18..	✓
???	<i>(forgotten name)</i>	18..	—
Sue	...	18..	✓
???	<i>(forgotten name)</i>	18..	—

Top persistent novels, adventure fiction (excerpt)

The interesting cells

The **non-canonical persistents**:
novels the genre kept reusing, but
literary history forgot.

A “hidden canon” *inside* the popular
archive.

A hidden genre canon

Author	Title	Year	Canon
Verne	...	18..	✓
Dumas	...	18..	✓
???	<i>(forgotten name)</i>	18..	—
Sue	...	18..	✓
???	<i>(forgotten name)</i>	18..	—

Top persistent novels, adventure fiction (excerpt)

The interesting cells

The **non-canonical persistents**:
novels the genre kept reusing, but
literary history forgot.

A “hidden canon” *inside* the popular
archive.

Canon vs archive persistence: median $z = [..]$ vs $[..]$, Mann-Whitney $p = [..]$

Ancestors are not origins

Genres crystallize collectively (no founding genius) —
so what is a “genre ancestor”?

Ancestors are not origins

Genres crystallize collectively (no founding genius) —
so what is a “genre ancestor”?

An ancestor is not the *inventor* of the genre:
it is the text the genre **kept reusing during its crystallization**.

Persistence is measured *in the accretion* of later texts —
Rieder’s cumulative model, not Borges’s founding one.

Ancestors are not origins

Genres crystallize collectively (no founding genius) —
so what is a “genre ancestor”?

An ancestor is not the *inventor* of the genre:
it is the text the genre **kept reusing during its crystallization**.

Persistence is measured *in the accretion* of later texts —
Rieder’s cumulative model, not Borges’s founding one.

Popular genres produce their own canons: durable textual persistence **precedes** (and predicts?)
retrospective consecration.

Limits

- **Right-censoring:** Leblanc (1905–) has ~40 forward years in a thinning corpus
- Similarity $\neq$ influence: no causal claim, a **family-resemblance** signal
- Genre labels are model outputs (released with probabilities)

Limits

- **Right-censoring:** Leblanc (1905–) has ~40 forward years in a thinning corpus
- Similarity $\neq$ influence: no causal claim, a **family-resemblance** signal
- Genre labels are model outputs (released with probabilities)

Perspectives

- What do ancestors transmit? Scene-level analysis (duel, pursuit, locked room)
- Cross-genre ancestors (Sue $\rightarrow$ detective?)
- Comparison with external canon lists (Canon Shelf!)

Thank you

Discussion

jean.barre@ens.psl.eu • <https://jeanbarre.eu>

Corpus, labels & code: <https://doi.org/10.5281/zenodo.20233179>

Backup: coverage and right-censoring

img/coverage_detective.png

- Persistence rankings restricted to novels with a **full +40y window** before 1950
- Bins with < 10 candidates are dropped, never imputed
- Detective fiction: dense only post-1890 $\Rightarrow$ 5-year bins

Backup: robustness to genre-label confidence

Genre labels are released with per-novel classifier probabilities: the ancestor ranking is stable when restricting to high-confidence members (Spearman $\rho = [..]$).

img/robustness_threshold.png

Backup: canonicity operationalization

Canon tags anchored in proxies for contemporary reception (Barré, Camps & Poibeu 2023, *Operationalizing Canonicity*): presence in school curricula, anthologies, reprints, scholarly attention.

Applied to the full corpus: ~10% canonical, 90% archive.

The point of this talk is not to defend these tags as ground truth, but to test whether **whatever they capture** correlates with within-genre textual persistence.