Modeling the Birth of a Literary Archetype

The Case of the Detective Figure in French Fiction

Jean Barré Olga Seminck Antoine Bourgois Thierry Poibeau

École Normale Supérieure - Université PSL & LaTTiCe (CNRS - UMR 8094)

Sixth Conference on Computational Humanities Research - Luxembourg 2025

The archetype we think we know

Sherlock Holmes

The archetype we think we know

Sherlock Holmes

Rouletabille

The archetype we think we know

Sherlock Holmes

Rouletabille

Hercule Poirot

The archetype we think we know

Sherlock Holmes

Rouletabille

Hercule Poirot

Jules Maigret

The archetype we think we know

Sherlock Holmes

Rouletabille

Hercule Poirot

Jules Maigret

- Strong visual code: hat, coat, pipe/cane, urban setting, solitary male figure.
- A very *recent*, audiovisual construction (late 20th - 21st c.).
- **Does this icon match what detectives actually look like in *150 years of fiction*?**

Detecting the textual archetype at scale

- **Not only a costume**
  Archetype = recurrent *textual patterns*: roles, actions, body verbs, way of speaking, network position.
- **Reasoning machine vs. other models**
  From Poe/Doyle's logician to Maigret's empathy or Lupin's trickster.
- **Unity or multiplicity of the archetype?**
  If we set aside the TV icon, do 150 years of novels still converge towards a common, scalable "detective" profile?

Nestor Burma, 1995

Research questions

General aim

Use computational tools to model the **construction** and **evolution** of the **detective archetype** in French literature.

Research questions

General aim

Use computational tools to model the **construction** and **evolution** of the **detective archetype** in French literature.

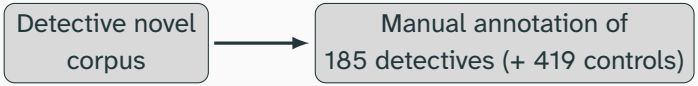
Research questions

- **RQ1 – Archetype unity**
  Can a model reliably recognize a *stable core* of the detective archetype across 150 years?
- **RQ2 – Archetype emergence**
  When does the detective move from a secondary figure to the **narrative pivot**?
- **RQ3 – Semantic trajectory**
  How does the detective evolve from a cold rational reasoner to an **emotionally engaged**, morally ambiguous figure?

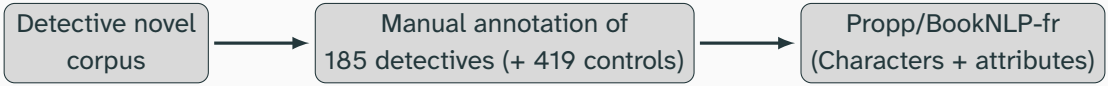
Methods pipeline

Detective novel corpus

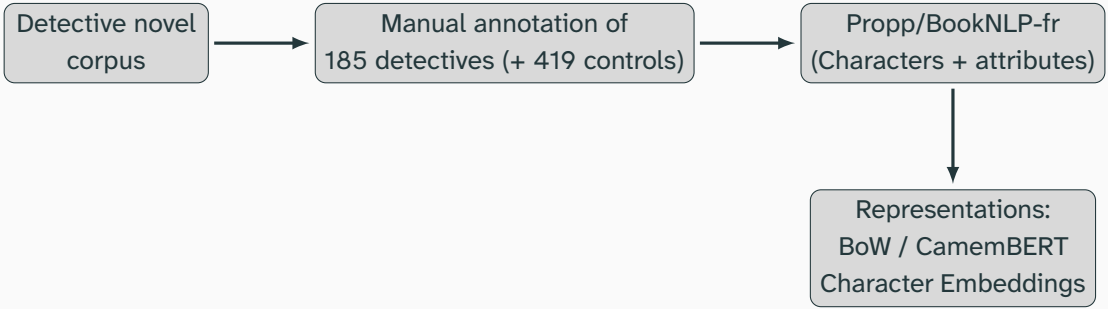
Methods pipeline



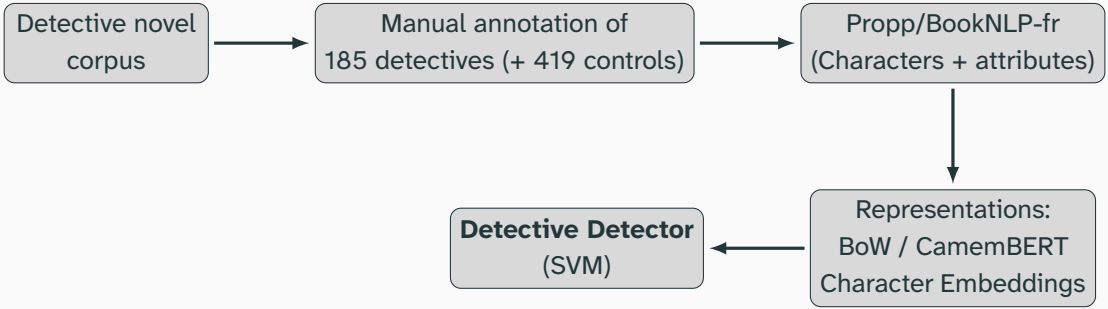
Methods pipeline



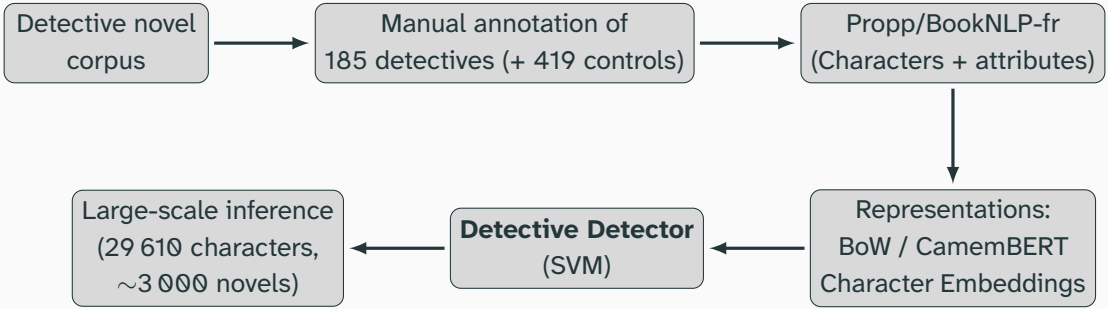
Methods pipeline



Methods pipeline



Methods pipeline



Methods focus

Character Embeddings

- **Character attributes**: agentive / passive verbs, modifiers (adjectives/predicates), possessives.
- **BoW**: 1000 most frequent attributes
- **CamemBERT$_{\text{LARGE}}$**: averaging all Bert embeddings of all attributes per character

Methods focus

Character Embeddings

- **Character attributes**: agentive / passive verbs, modifiers (adjectives/predicates), possessives.
- **BoW**: 1000 most frequent attributes
- **CamemBERT$_{\text{LARGE}}$**: averaging all Bert embeddings of all attributes per character

Learning & evaluation

- SVM and Logistic Regression (stratified 5-fold CV)
- **LOGO** (character, author, time periods) to prevent leakage
- Main metric: *balanced accuracy*

Results

Model	B.Acc.	F1 (Det.)
BoW + SVM	0.895	0.84
CamemBERT + SVM	**0.923**	**0.887**

Table 1: Detector performance (stratified 5-fold CV)

Results

Model	B.Acc.	F1 (Det.)
BoW + SVM	0.895	0.84
CamemBERT + SVM	**0.923**	**0.887**
LOGO (Author)	0.908	0.873
LOGO (10 yrs)	0.915	0.879
LOGO (50 yrs)	0.892	0.858

Table 1: Detector performance (stratified 5-fold CV)

Results

Model	B.Acc.	F1 (Det.)
BoW + SVM	0.895	0.84
CamemBERT + SVM	**0.923**	**0.887**
LOGO (Author)	0.908	0.873
LOGO (10 yrs)	0.915	0.879
LOGO (50 yrs)	0.892	0.858

Table 1: Detector performance (stratified 5-fold CV)

RQ1 – Archetype unity: high and stable accuracy → robust unity of the archetype.

Results

Model	B.Acc.	F1 (Det.)
BoW + SVM	0.895	0.84
CamemBERT + SVM	**0.923**	**0.887**
LOGO (Author)	0.908	0.873
LOGO (10 yrs)	0.915	0.879
LOGO (50 yrs)	0.892	0.858

Table 1: Detector performance (stratified 5-fold CV)

Applied to full corpus:
713 detectives predicted
(out of 29 610 characters)

RQ1 – Archetype unity: high and stable accuracy → robust unity of the archetype.

Lexical specificity of the archetype

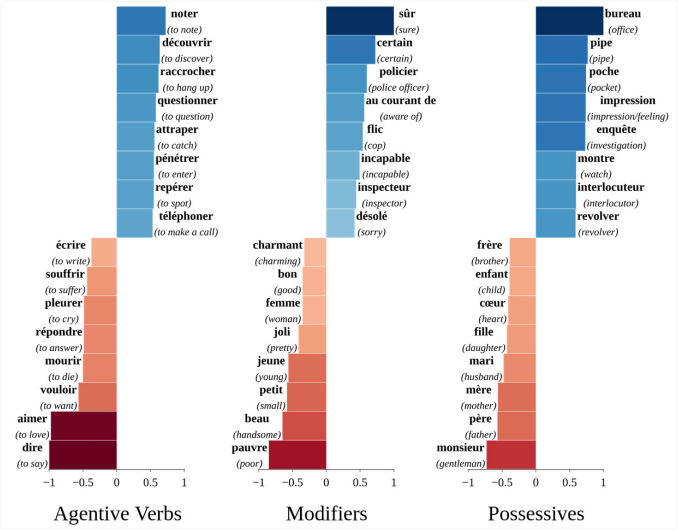


Figure 1: Most/least distinctive attributes of detectives (log-odds z-scores)

RQ2 – Archetype emergence

At scale: *When* do detectives appear?

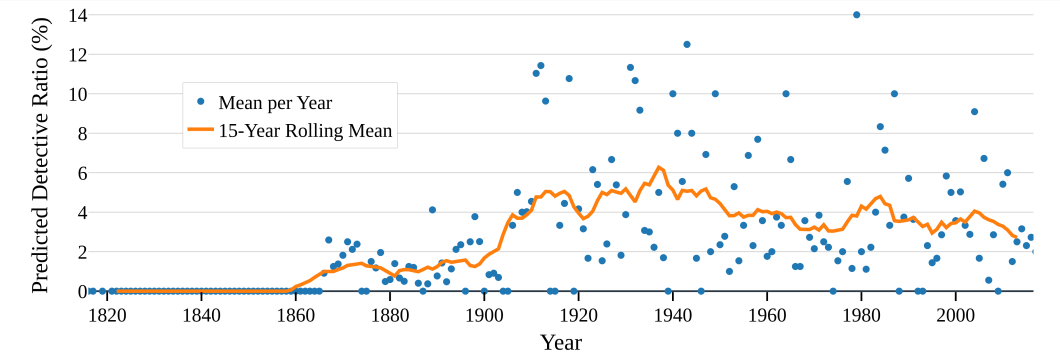


Figure 2: Ratio of detectives among all characters.

RQ2 – Archetype emergence

At scale: how *central* are they?

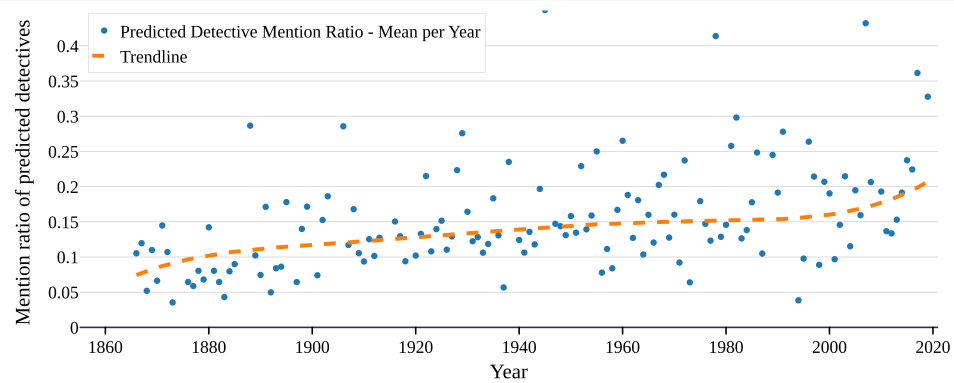


Figure 3: Mention ratio (share of text for main detective).

Three clusters: Three variations of the archetype.

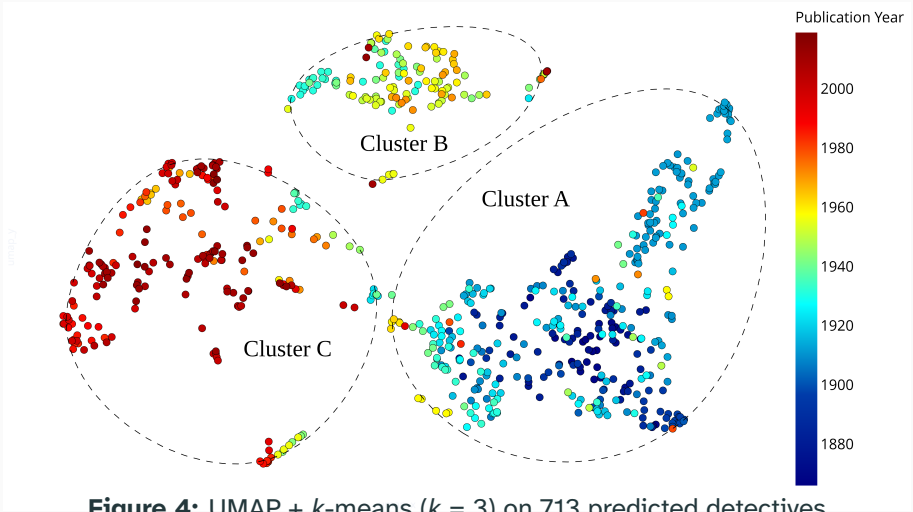


Figure 4: UMAP + *k*-means ($k = 3$) on 713 predicted detectives

Lexical specificities per period

A — Classical	B — Empathic/procedural	C — Hardboiled/neo-polar
Verbs: say, exclaim, answer, declare, murmur, reply, question	**Verbs**: question, phone, suppose, look, smoke, drink, remember	**Verbs**: feel, catch, put down, *fuck*, spot, put on, run into
Modifiers: dear, young, brave, honest, mysterious	**Modifiers**: sure, gloomy, police, drunk, tired	**Modifiers**: alone, responsible, aware, cop, guy
Possessives: friend, revolver, office, duty, master	**Possessives**: pipe, office, overcoat, hat, pocket	**Possessives**: jacket, phone, car, face, father/mother

Table 2: Distinctive attributes of the three periods

A $\rightarrow$ dialogic rationality;

Lexical specificities per period

A — Classical	B — Empathic/procedural	C — Hardboiled/neo-polar
Verbs: say, exclaim, answer, declare, murmur, reply, question	**Verbs**: question, phone, suppose, look, smoke, drink, remember	**Verbs**: feel, catch, put down, *fuck*, spot, put on, run into
Modifiers: dear, young, brave, honest, mysterious	**Modifiers**: sure, gloomy, police, drunk, tired	**Modifiers**: alone, responsible, aware, cop, guy
Possessives: friend, revolver, office, duty, master	**Possessives**: pipe, office, overcoat, hat, pocket	**Possessives**: jacket, phone, car, face, father/mother

Table 2: Distinctive attributes of the three periods

A $\rightarrow$ dialogic rationality;
B $\rightarrow$ Maigret-like empathy (pipes, offices, overcoats);

Lexical specificities per period

A — Classical	B — Empathic/procedural	C — Hardboiled/neo-polar
Verbs: say, exclaim, answer, declare, murmur, reply, question	**Verbs**: question, phone, suppose, look, smoke, drink, remember	**Verbs**: feel, catch, put down, *fuck*, spot, put on, run into
Modifiers: dear, young, brave, honest, mysterious	**Modifiers**: sure, gloomy, police, drunk, tired	**Modifiers**: alone, responsible, aware, cop, guy
Possessives: friend, revolver, office, duty, master	**Possessives**: pipe, office, overcoat, hat, pocket	**Possessives**: jacket, phone, car, face, father/mother

Table 2: Distinctive attributes of the three periods

A $\rightarrow$ dialogic rationality;
B $\rightarrow$ Maigret-like empathy (pipes, offices, overcoats);
C $\rightarrow$ immersion in bodies, violence, vulgarity.

Discussion

Revised periodization of the French detective novel

- **1866–1927** *Proto/classical rational phase*
  Import and adaptation of the English model (Poe, Doyle): investigation as a logical game, the *reasoning* detective (Gaboriau, Leroux, etc.).

Discussion

Revised periodization of the French detective novel

- **1866–1927** *Proto/classical rational phase*
  Import and adaptation of the English model (Poe, Doyle): investigation as a logical game, the *reasoning* detective (Gaboriau, Leroux, etc.).
- **1927–1945** *Classical francophone model*
  Stabilization of the classic form (Simenon's Maigret, major popular cycles): empathy, social embeddedness, institutional police.

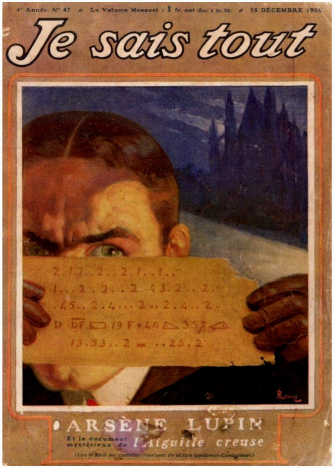
Discussion

Revised periodization of the French detective novel

- **1866–1927** *Proto/classical rational phase*
  Import and adaptation of the English model (Poe, Doyle): investigation as a logical game, the *reasoning* detective (Gaboriau, Leroux, etc.).
- **1927–1945** *Classical francophone model*
  Stabilization of the classic form (Simenon's Maigret, major popular cycles): empathy, social embeddedness, institutional police.
- **1945–2020** *American influence and generic explosion*
  Série noire, hardboiled adaptation, *néo-polar*, espionage: violent immersion, moral ambiguity, generic hybridizations.

Perspectives

- Centrality in Character Networks
- Multilingual: International exchanges in the construction of the archetype
- Towards “degrees of detectiveness” - Case Study on *Arsène Lupin*
- Other archetypes (criminal, victim, sidekick)?



Arsène Lupin, *L'Aiguille creuse*
(15 décembre 1908)

Thank you!

Code & data

github.com/lattice-8094/DETECTIVES

Contact

jean.barre@ens.psl.eu

Paper, code & viz – jeanbarre.eu

Controls & Limitations

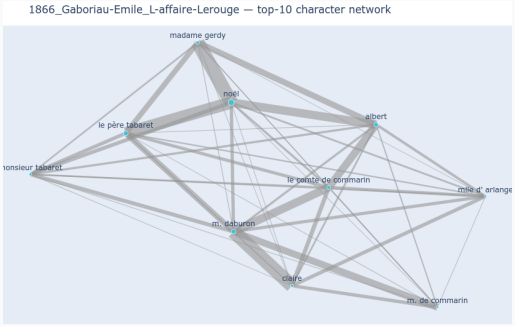
General Drift Control

- Comparing with non-detective controls shows that some drift is general (language change)
- But the middle period diverges strongly → genuine **genre-specific** evolution.

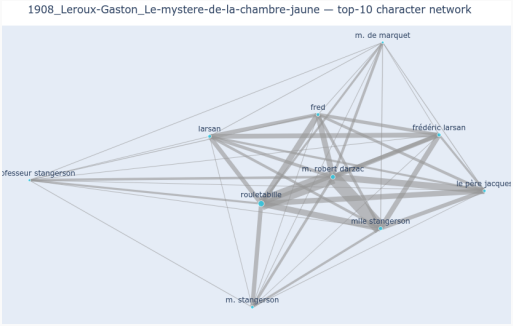
Limitations

- No negation / multi-word expressions
- Still noisy coreference chains

Examples of character networks



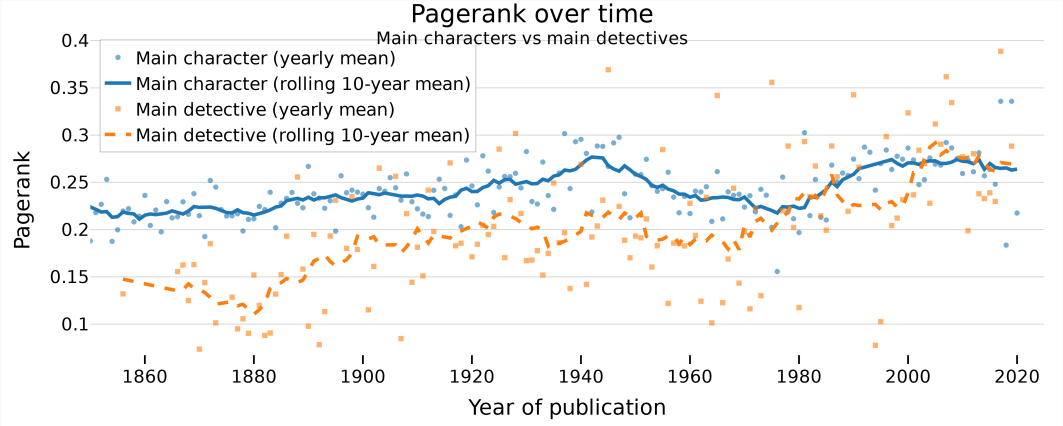
(a) Gaboriau, *The Lerouge Affair* (1866)



(b) Leroux, *The Mystery of the Yellow Room* (1908)

Two automatically extracted networks (top 10 characters)
Computing PageRank for each predicted detective

Result: detective centrality over time



Points: yearly average

Lines: 10-year moving average