Computational Analysis of the Birth of the Detective Novel

Jean Barré

June 11, 2025

École normale supérieure – Paris Science and Letters University

LaTTiCe lab

Introduction

Approaches to Literary Genres

- Formalist *vs.* contextual tension.
- Formalist : internal textual features
- Contextual : Genres emerge from interactions among authors, editors, readers, critics, etc., and are shaped by their socio-cultural enunciation contexts.
- Computational framework : Perspective Modeling (Underwood, 2019)
- Results: Genre stability, supporting transcendental and formalist approaches.

The Detective Novel

A genre with a strong identity

- Specific narrative structure: “a story devoted primarily to the methodical and gradual discovery, by rational means, of the exact circumstances of a mysterious crime” (Messac, 1929).
- Long editorial tradition : Gradual appearance from the late 19th century, stemming from the serialized novel (notably Gaboriau) to the “Collection du Masque” (Pigasse, 1927).
- Abundant, clearly defined corpus

First Research Questions - Can we detect specific textual features when the genre is clearly established ? How did the formal recipe of the genre happened ?

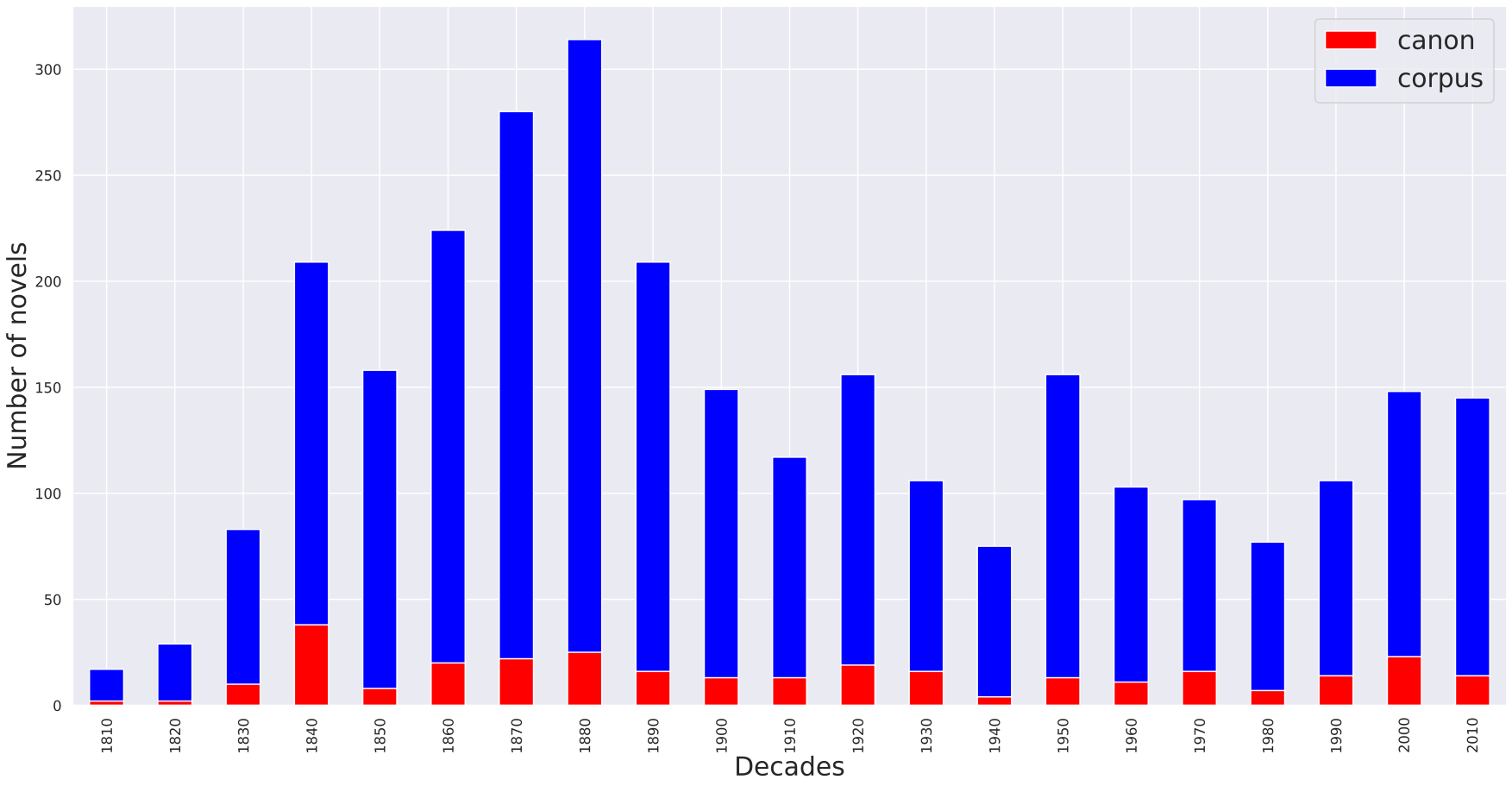


Figure 1: Temporal distribution of the corpus

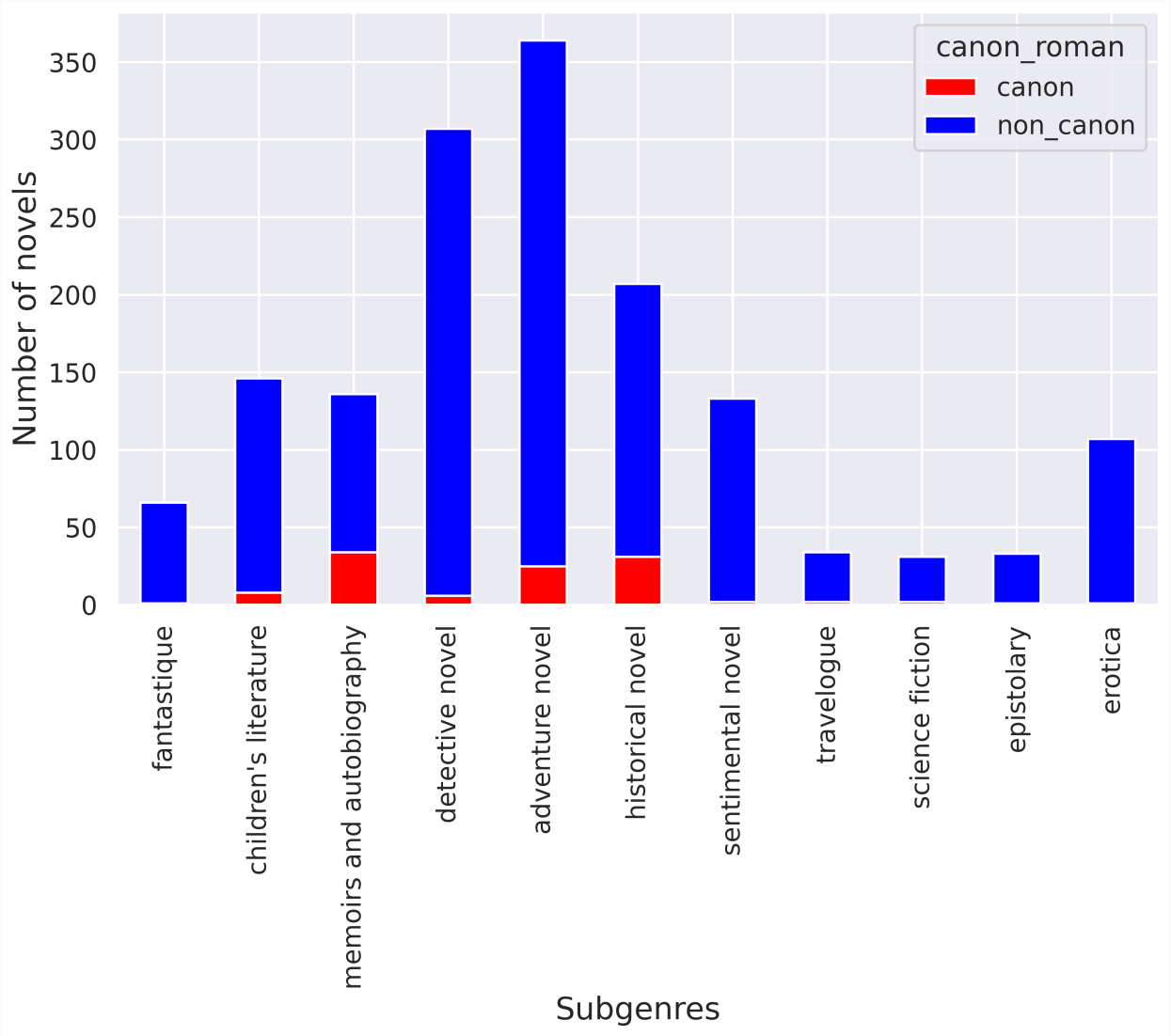


Figure 2: Literary subgenres in the corpus

Detective story repartition

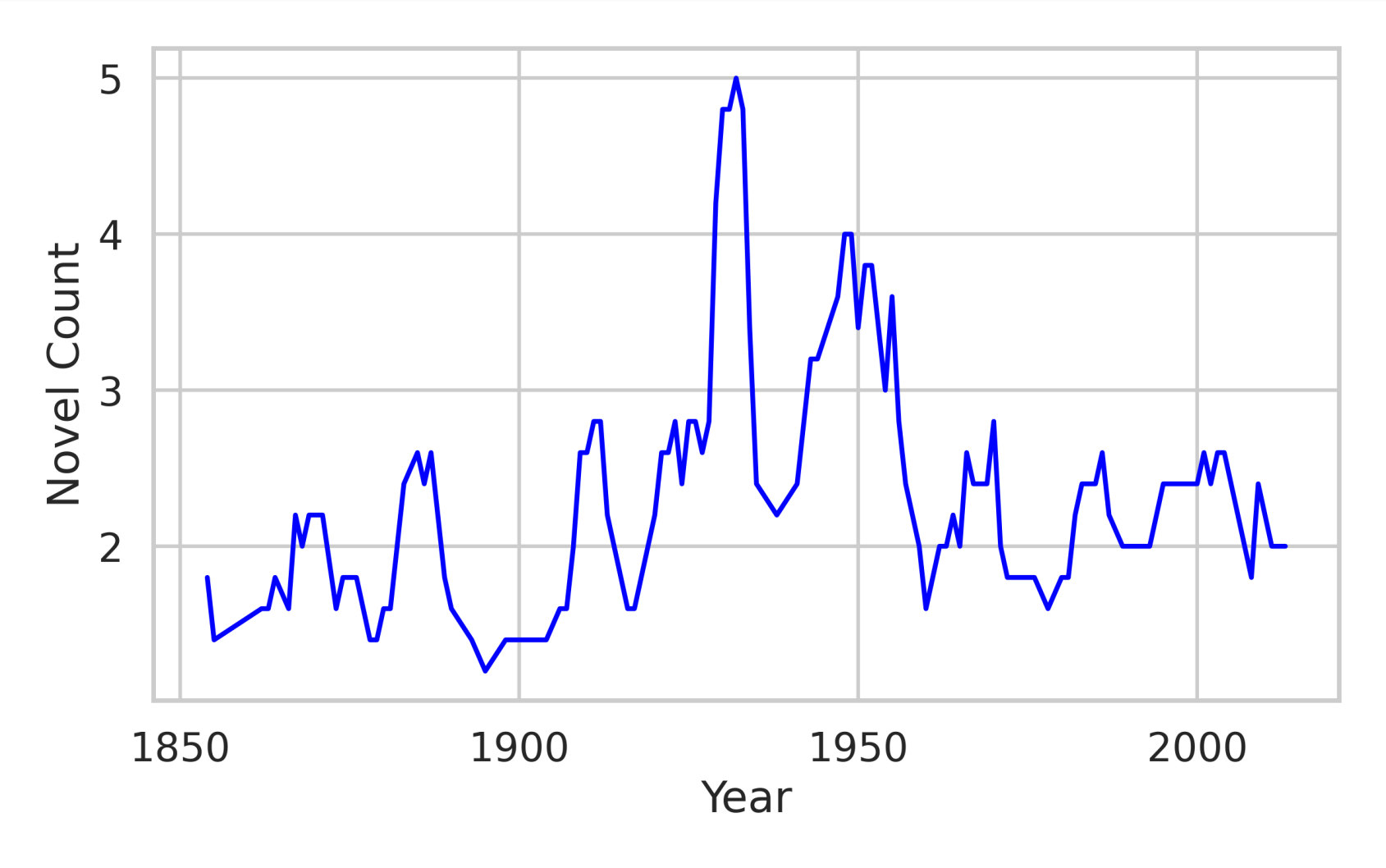


Figure 3: Rolling mean of the detective novels repartition

Classifying The Detective Novel

Three Steps Pipeline

- Extract textual features.
- Machine learning modeling – SVM.
- Error analysis and results interpretation.

Automated Classification of the Detective Novel

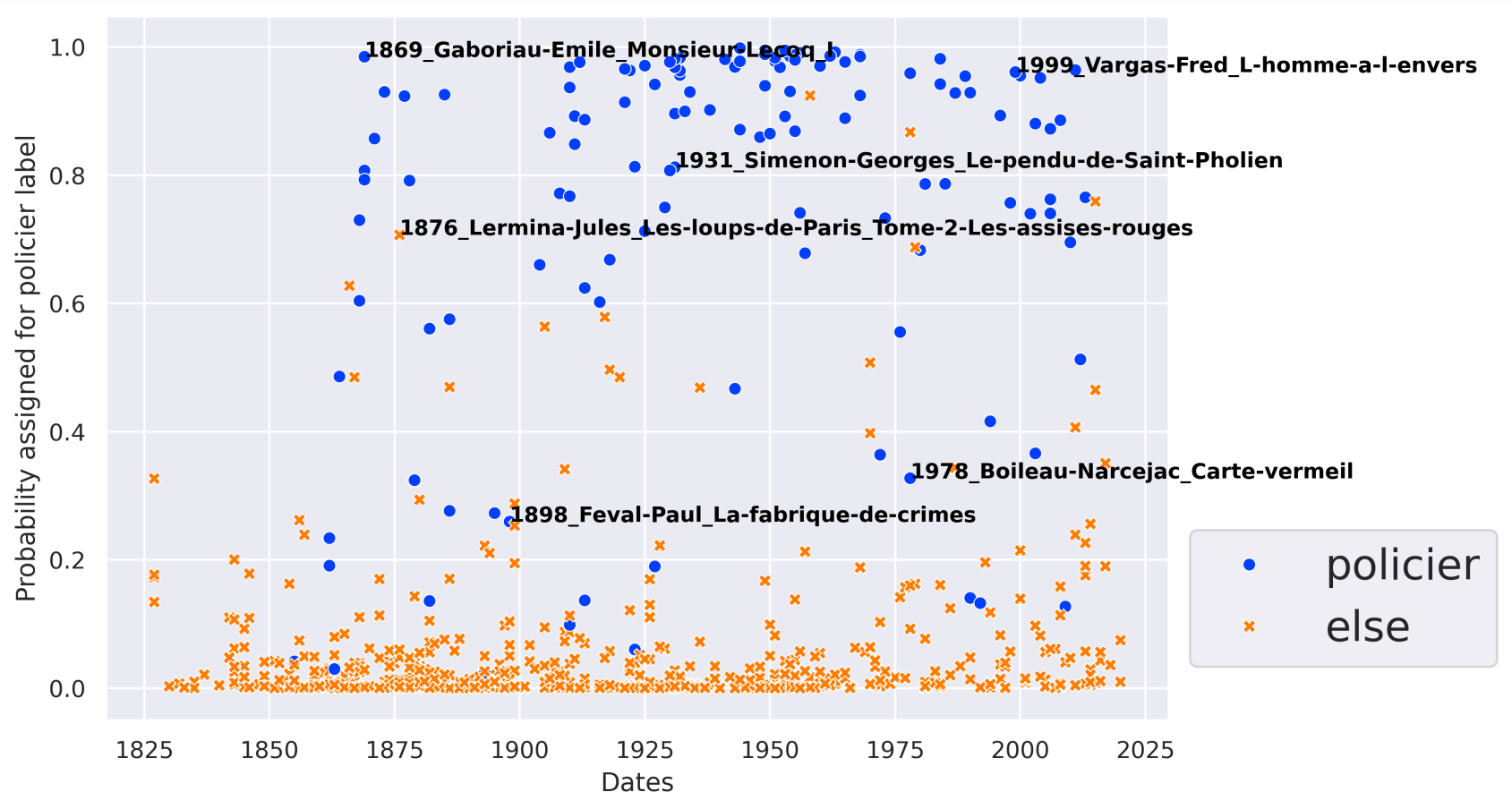


Figure 4: Automatic classification of the francophone detective novel

Discriminative Features

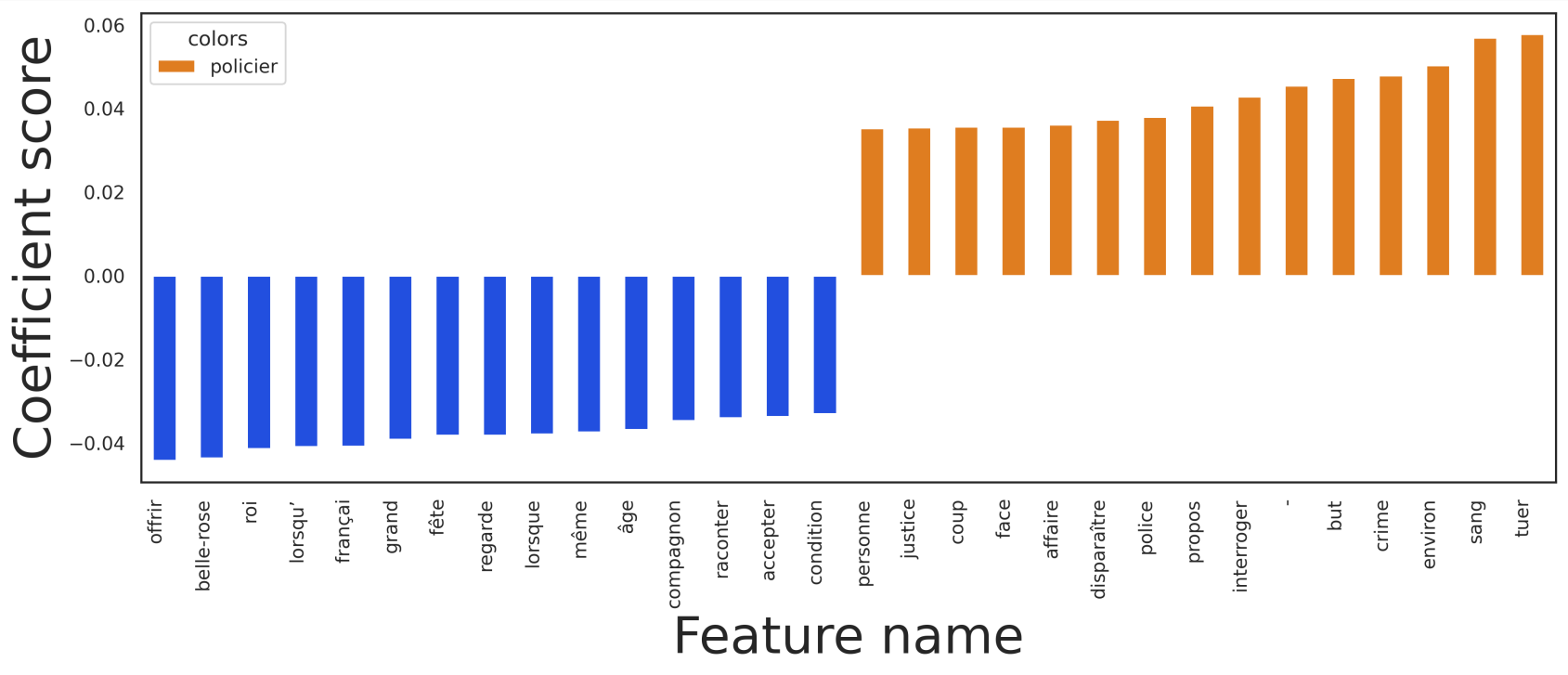


Figure 5: Discriminative features for predicting detective fiction

Limitations of the Analysis

- **Anachronistic history**: Risk of projecting our contemporary definition onto older texts.
- **Historical mutation of the genre**:
  - Two SVM models trained pre- and post-WWII yield very different results.
  - Example: A model trained on pre-1939 serialized detective novels (Gaboriau, Boisgobey) achieves 88.4% accuracy and still performs at 75% on post-war novels.
  - Conversely, a model trained on post-1939 detective novels scores 88.1% and drops to 55% on earlier works.
  - This asymmetry shows that the proto-detective novel (1860–1927) differs significantly from its modern form.
- Does the proto-detective novel really belongs to the same genre? Does it have the same textual features or just emerging ones?

Late Emergence of the Term

- The term *detective novel* appears historically very late. The genre does not yet exist (1870-1927)

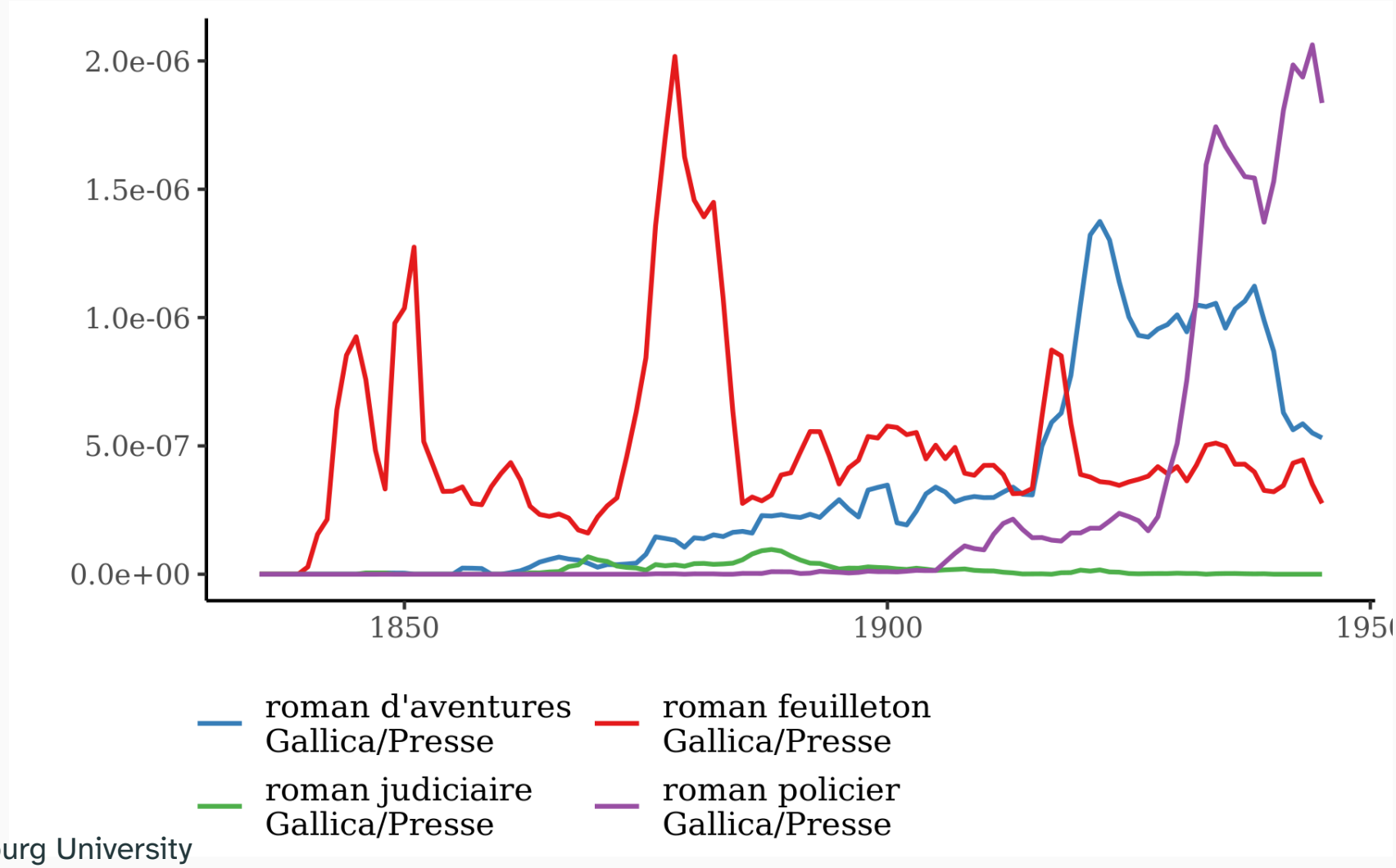


Figure 6: Gallicagram (de Courson & Azoulay, 2021)

Limitation: Evolution of the Genre Prediction

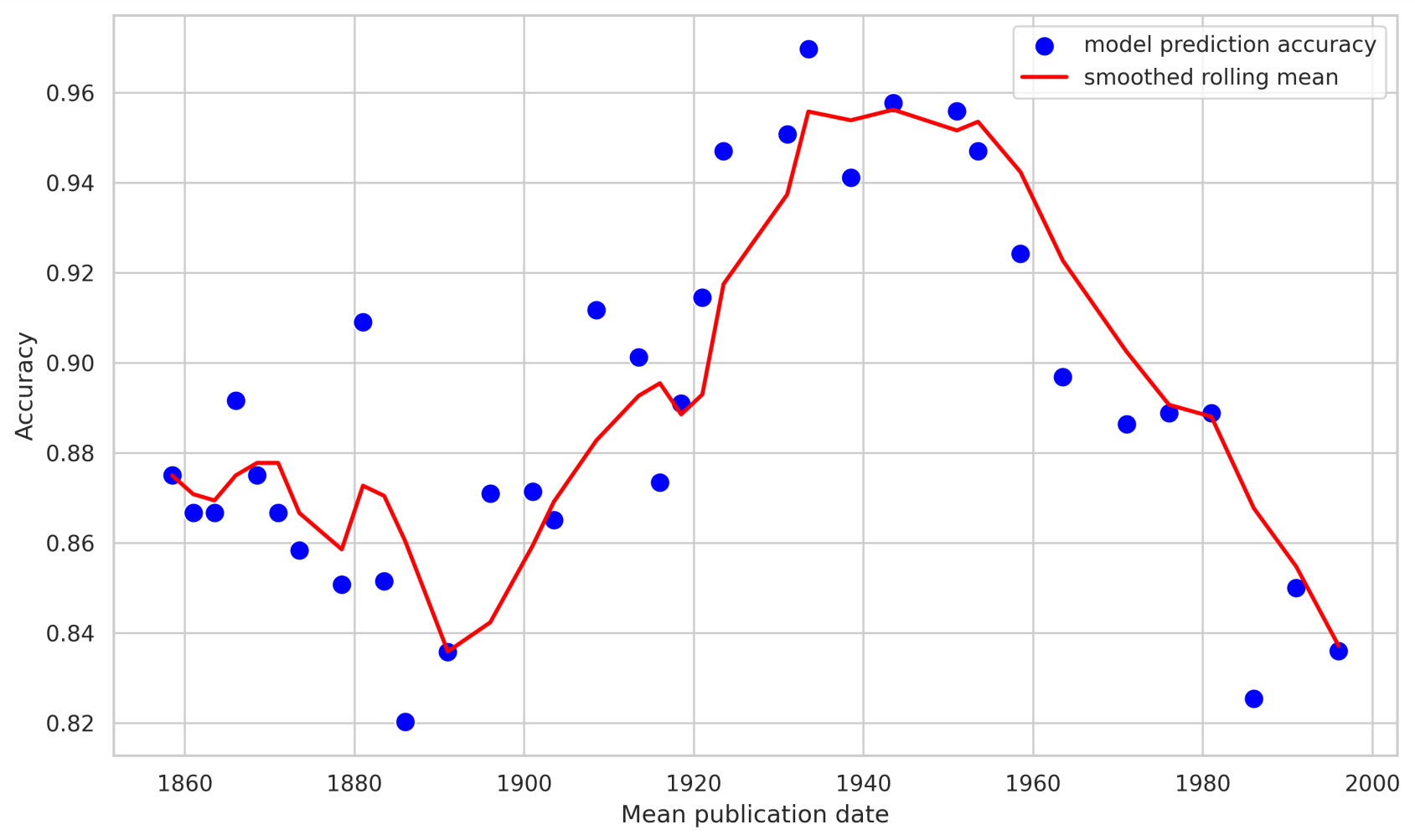


Figure 7: Genre prediction accuracy every 25 years

Research Question

Central Question

- How did the detective novel gradually distinguish itself from the serialized novel to become an autonomous genre?

“Genius” Hypothesis?

- “Émile Gaboriau is the father of the detective novel” (Messac, 1929).
- “Poe is the inventor of the detective novel” (Borges, 1936).
- “Balzac, pioneer in everything, wrote *Une ténébreuse affaire*, the first detective novel” (Fortassier, 1955).

Cumulative Approach to Genre Beginnings

- Repetition and accumulation of discursive practices enabling the genre.
- John Rieder:

> *Studying the beginnings of the genre is not about finding points of origin but observing an accretion of repetition, echoes, imitations, identifications, and distinctions that testify to an emerging awareness of a conventional network of similarities. (Rieder, 2012)*

Formal features: Before or after the contextual rise of the genre ?

The Birth of the Detective Novel (Lavergne, 2009)

The detective novel distinguishes itself from the serialized novel in three ways:

1. **Persistence of the criminal topic**: more coherent and central throughout the narrative.
2. **Centrality of the detective**: shifts from a secondary to a primary role, driving crime resolution.
3. **Narrative structure centered on investigation**: story organized around the progression of the investigation, from known clues to unknowns, punctuated by red herrings and twists.

Method: Topic Modeling (LDA)

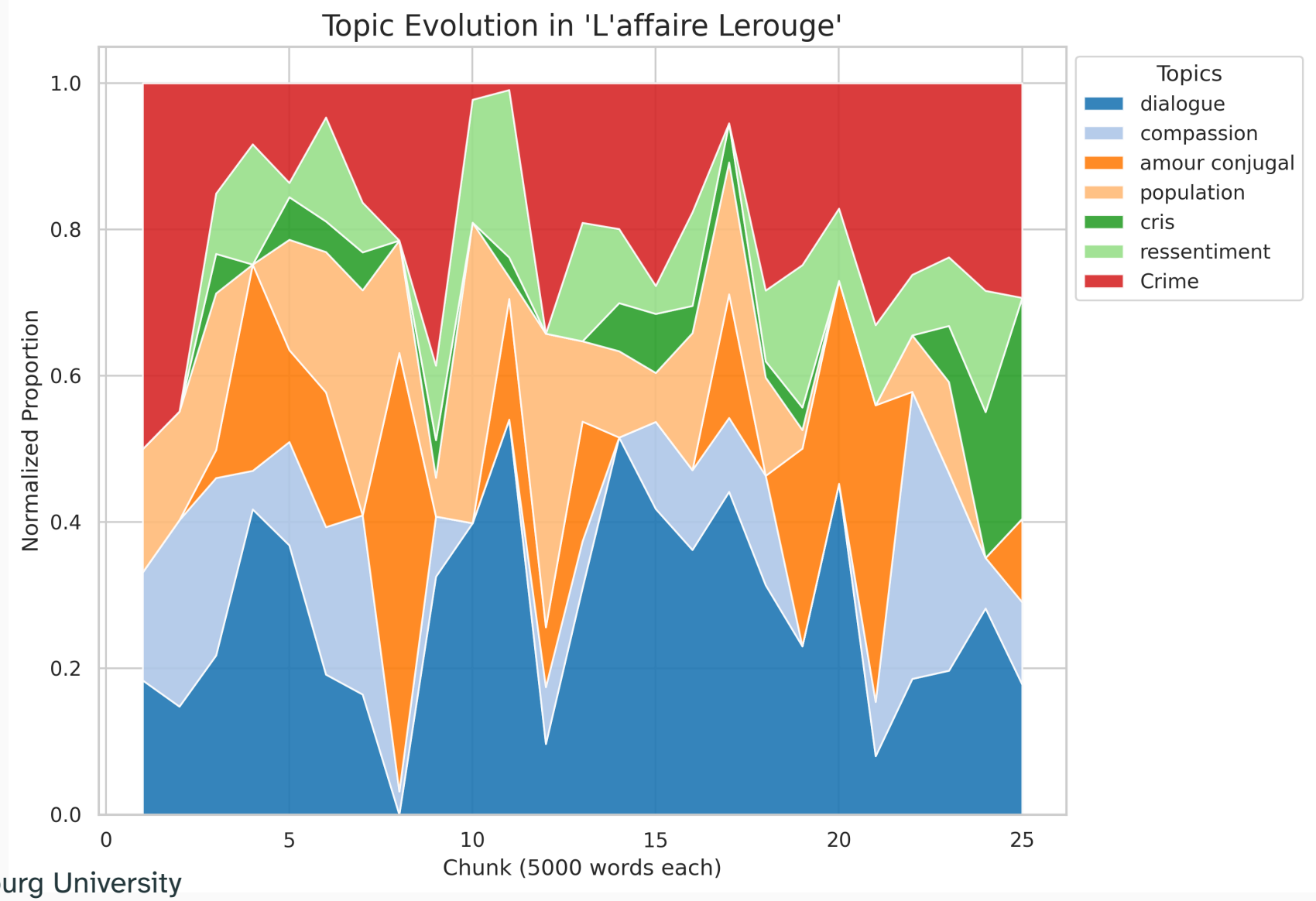
Principle of LDA Thematic Analysis

- **Latent Dirichlet Allocation (LDA)**: probabilistic method for extracting topics from a corpus.
- Documents are distributions over topics, which are distributions over words.

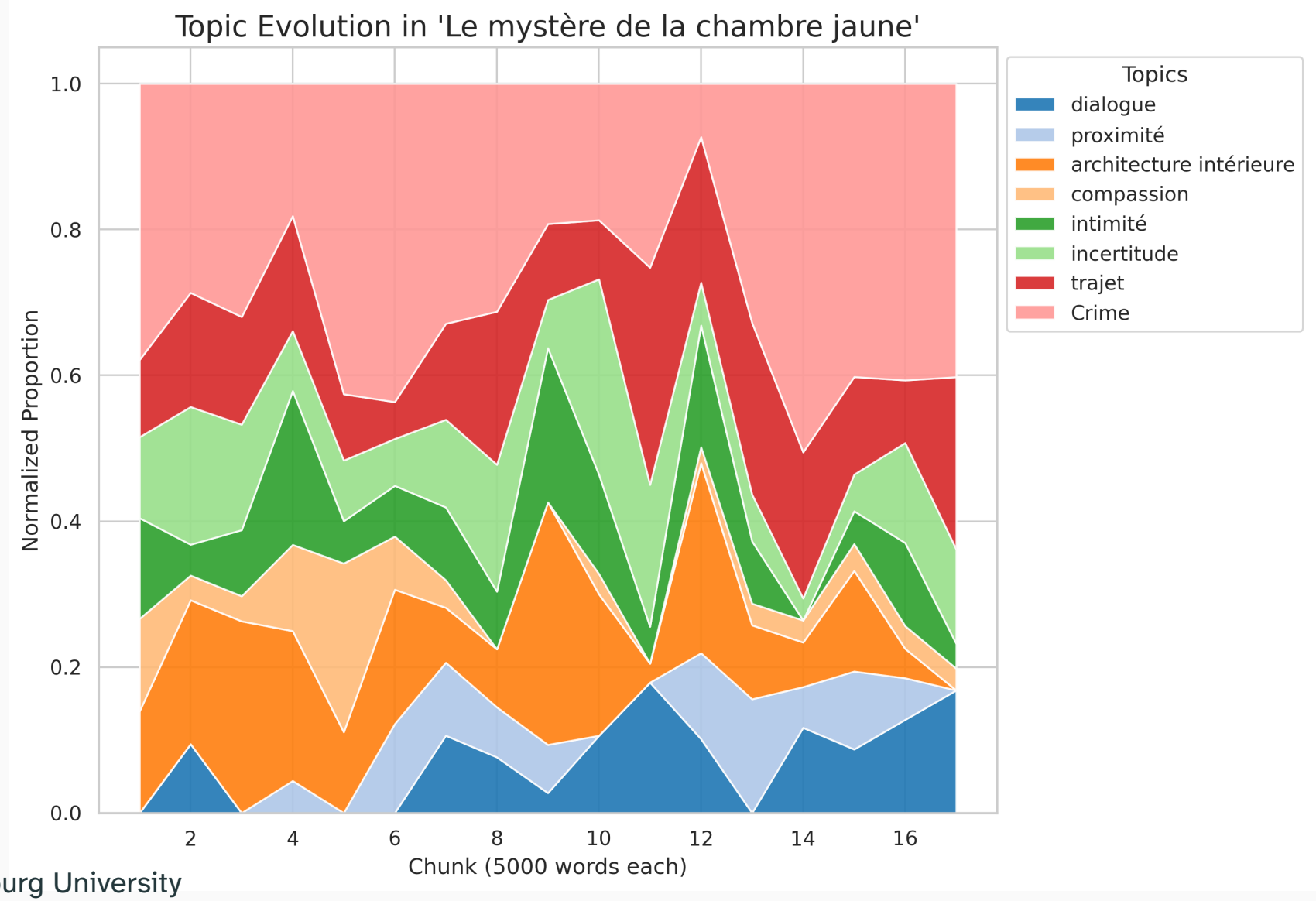
Preprocessing Applied to the Corpus

- **Lemmatization** of corpus words.
- Part-of-speech filtering (nouns, verbs, adjectives, adverbs).
- Removal of proper names, rare words, and stopwords.
- Segmentation of novels into **chunks** of 5000 words to study intra-text evolution.
- Extraction of 100 topics and their proportions in each novel/chunk.

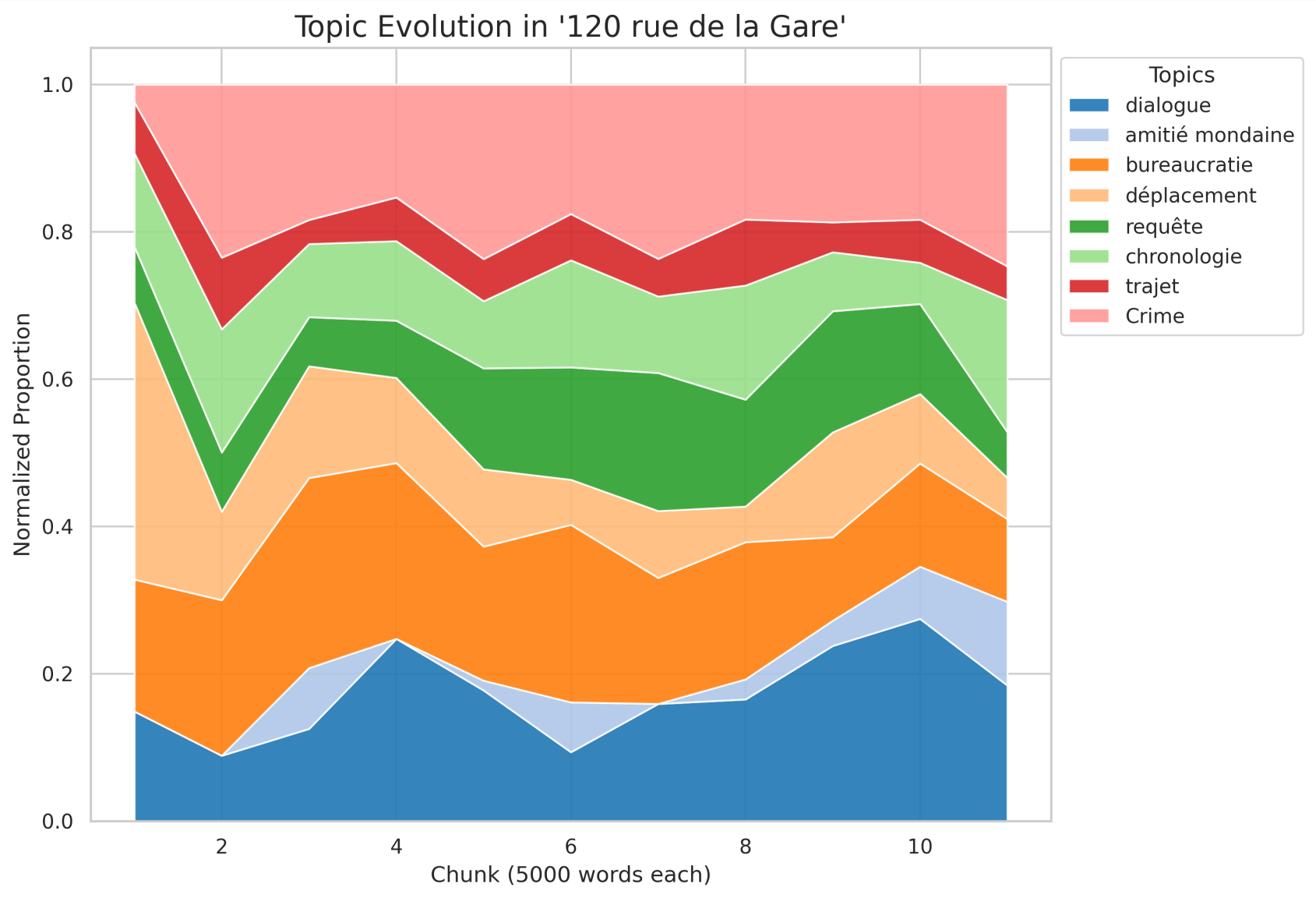
Evolution of the Criminal Topic in *L'Affaire Lerouge* (Gaboriau, 1866)



Evolution of the Topic in *The Mystery of the Yellow Room* (Leroux, 1907)



Evolution of the Topic in *120, Rue de la Gare* (Malet, 1943)



Measuring Crime Topic Persistence

Definition and Objective

- **Thematic persistence** measures a topic's stability across different parts (chunks) of the novel.
- Low variation implies stable, continuous presence of the criminal topic.

Persistence Calculation: 1 - Coefficient of Variation (CV)

$$CV = \frac{\sigma}{\mu} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}}{\mu}$$

where:

- x_i is the proportion of the criminal topic in chunk i.
- μ is the mean proportion across all chunks.
- σ is the standard deviation of proportions.

Crime Topic Persistence Over Time

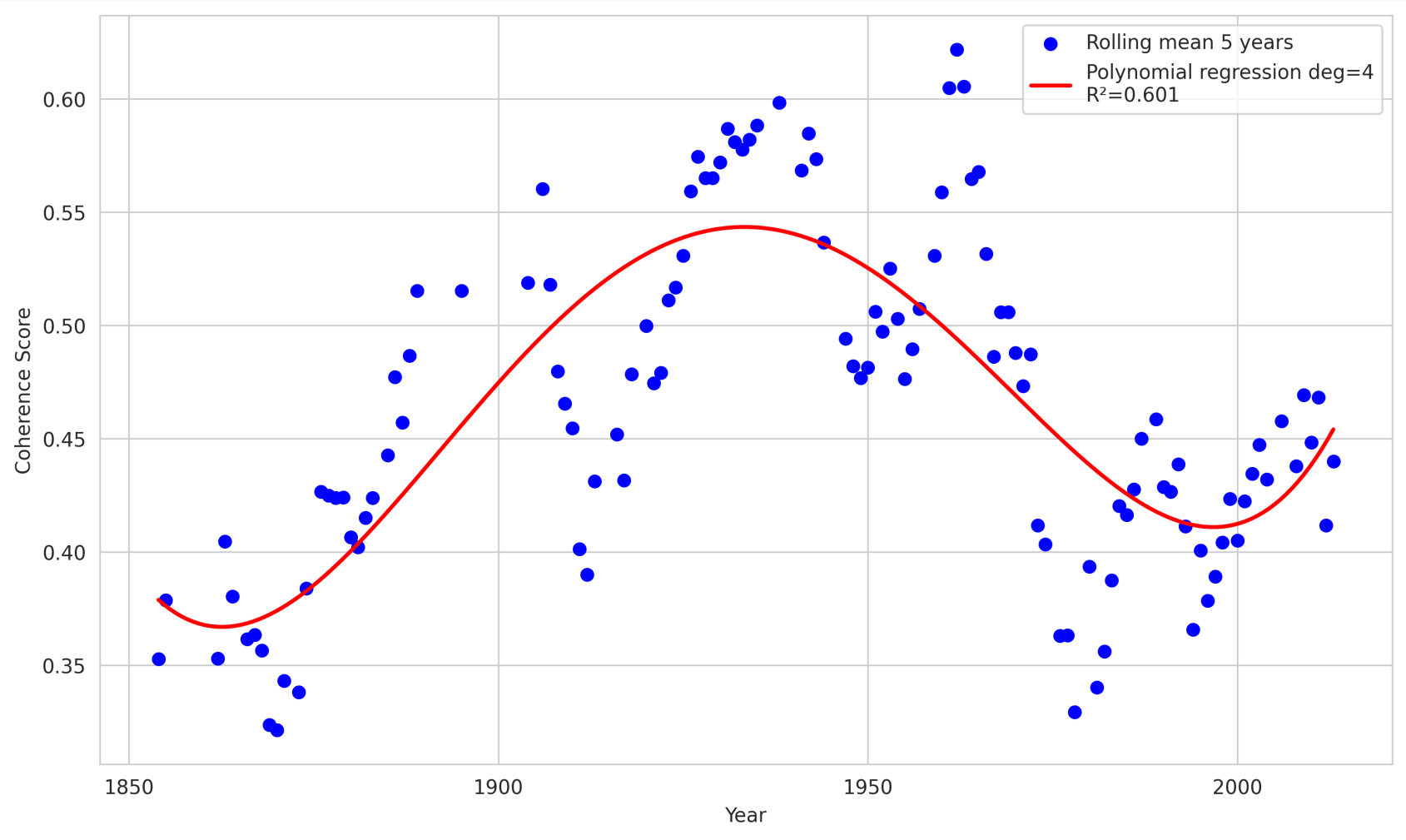


Figure 11: Persistence of the criminal topic over time

The Detective as a Character Archetype

Birth of the genre: Affirmation of the detective figure

- Origins in mid-19th c. French proto-detectives (Dantès, Rodolphe, Rocambole) and western trapper/bloodhound; real-life models (Vidocq); The archetype is solidified by Monsieur Lecoq (Gaboriau, from 1966).
- Archetype: methodical ”reasoning machine” (Symons, 1972), embodying pure rational deduction (Dupin, Holmes, Tirauclair, Rouletabille).
- 20th c. expansion: Lupin: gentleman thief and then US hard-boiled variants (Spade, Marlowe) introduce physical danger and moral complexity (corruption, murders).
- Maigret (1931-1972) marks the shift from detached ”genius” to socially embedded, relatable investigator.
- **Quantitative research aim**: Examine how the detective archetype gain importance in the narrative, how it evolves from a secondary character to the central protagonist.

Methods: Corpus, Annotation & Modeling

Corpus and Annotation

- 300 French-language crime fiction texts (e.g. *Fantômas*), from protodetective tales to hardboiled and noir.
- Automatic Character extraction of coreference chains with BookNLP-fr (Mélanie et al, 2024)
- Dataset building with the main figures of the french archetype
- Assigned binary labels: **Detective** (leads the investigation) vs. **Non-detective** (outlaws, victims, suspects, etc.).
- Annotated 175 characters as Detectives (51 unique archetypes).

Modeling

- Semantic and Lexical Representation for each Character
- Train binary classifiers to detect the detective role.
- Then looking at the detective's ratio of mentions

Classification Results

Méthode	B. Acc.	F1 Non-detective	F1 Detective
BoW + LogReg	0.827	0.86	0.74
BoW + SVM	0.881	0.92	0.83
CamemBERT + LogReg	0.914	0.93	0.86
CamemBERT + SVM	**0.925**	0.95	0.89

Table 1: F1-score per class and balanced accuracy for each model

Discriminative Features

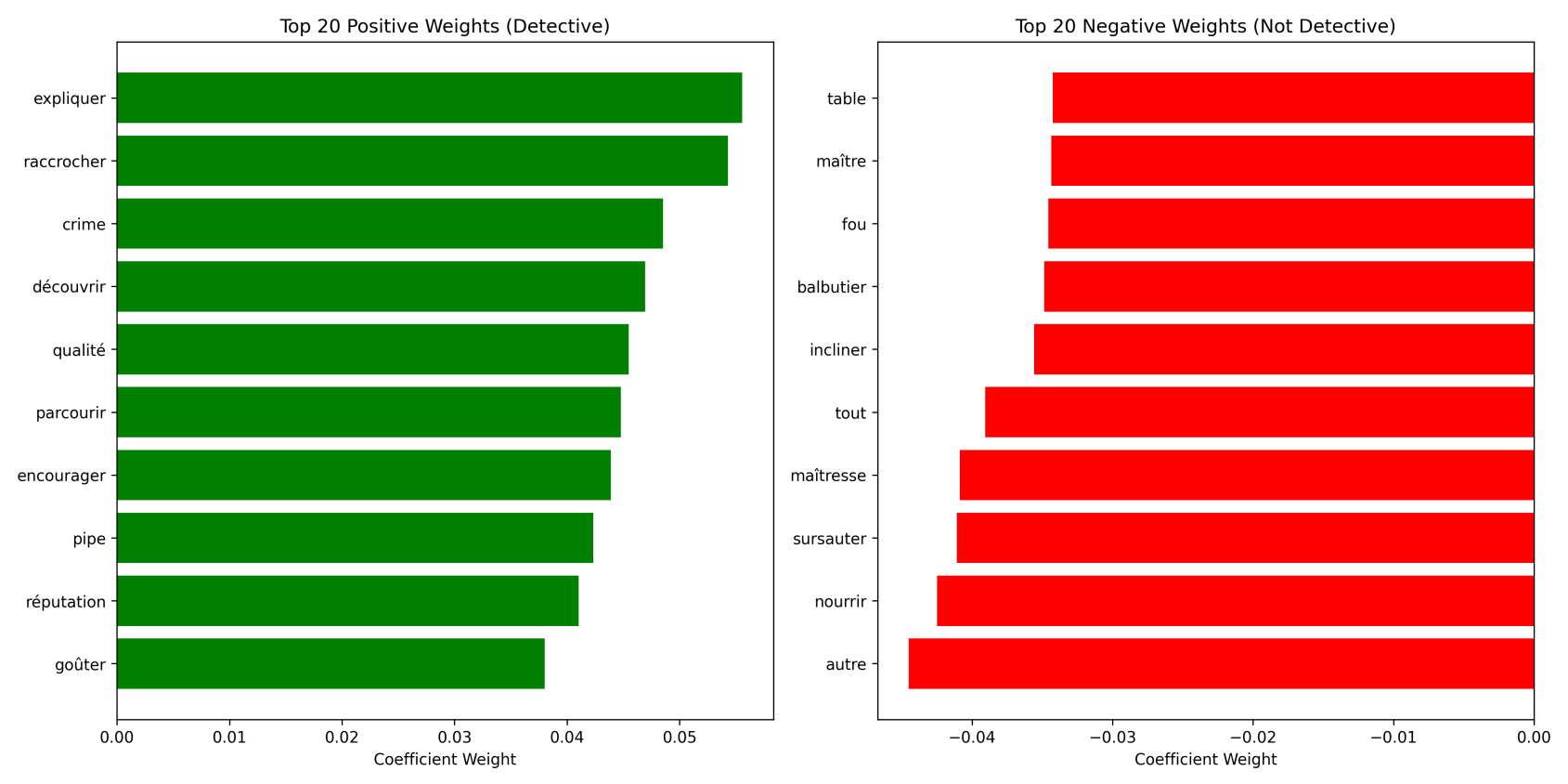


Figure 12: Discriminative features predicting a character to be a detective

Increasing Dominance of the Detective Character

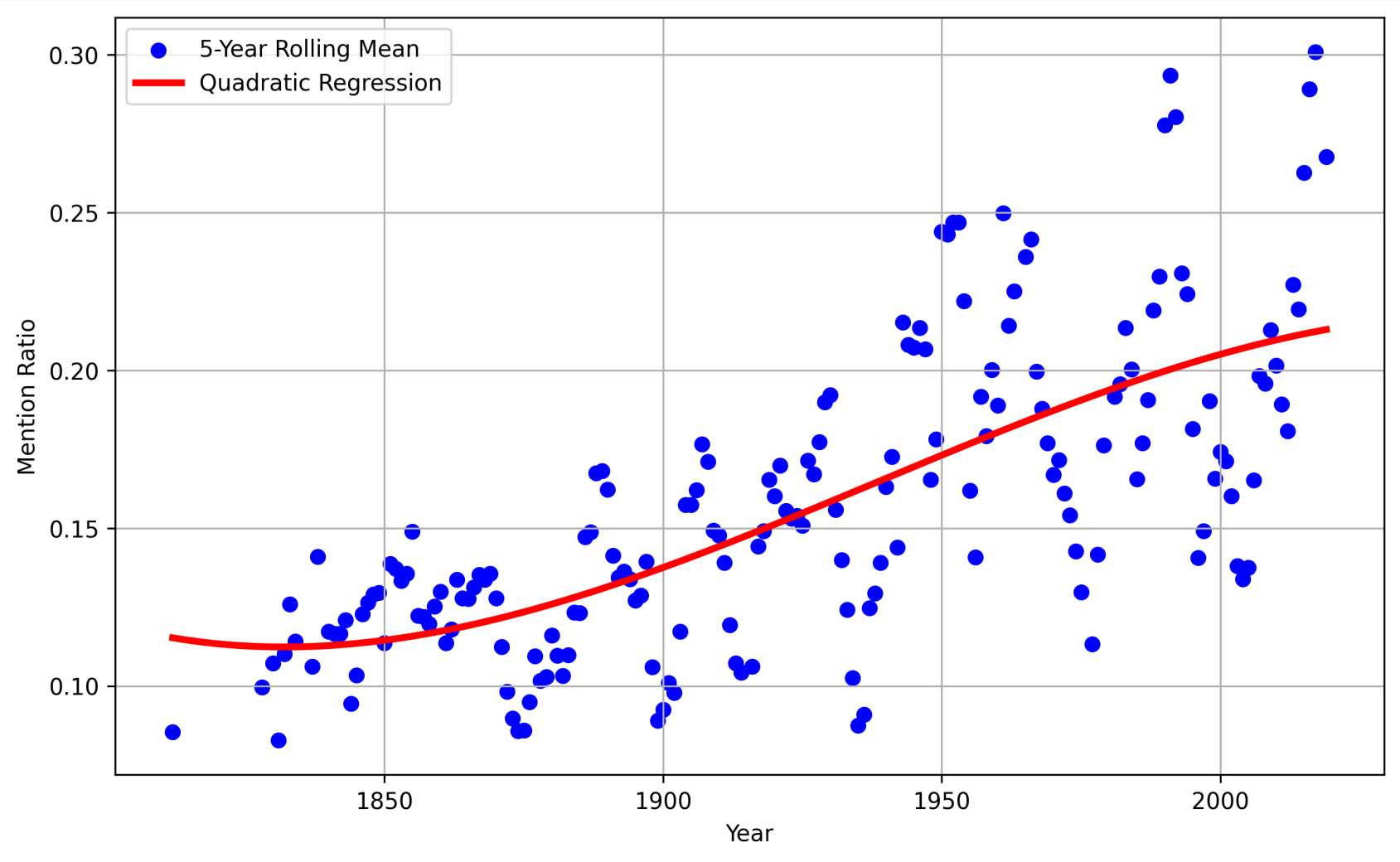


Figure 13: Evolution of detective character dominance in detective novels

Plot specific feature in the Detective Novel

- Identification of suspense mechanisms from crime to resolution.
  - Detection of key narrative sequences: crime scene, reasoning, final resolution.
  - **Automatic Annotation** DeepSeek labels each passage as *CRIME SCENE*, *REASONING ATTEMPT*, *CRIME FINAL RESOLUTION* or *ELSE*, producing one unique plot structure per novel.
  - **Hand evaluation** - Kappa: .83

Prompt

You are classifying passages from French detective novels into one of four narrative categories.

- **CRIME SCENE**: The crime or its immediate discovery is described. Use this label only when the passage truly depicts the act or finding of the crime.
- **REASONING ATTEMPT**: A character analyzes evidence, forms hypotheses, interviews suspects, or reconstructs the sequence of events.
- **CRIME FINAL RESOLUTION**: The definitive solution is revealed—who committed the crime, how, and why (not necessarily a trial).
- **ELSE**: Anything that does not fit the above (e.g., setting, unrelated dialogue, character backstory).

For each passage:

1. Give two short sentences explaining your choice.
2. On a new line, output exactly one label in UPPERCASE.

From DeepSeek Annotations to Prototypical Plot Measure

1. **Permutations Counting** For the subsequence of non-ELSE labels, count adjacent “out-of-order” pairs relative to the canonical arc Crime - Reasoning - Resolution.
2. **ELSE Penalty** Treat each *ELSE* label as an additional inversion to penalize off-topic digressions.
3. **Normalization (PPlot)**

$$\mathrm{PPlot} = \frac{\#\mathrm{permutations} + \#\mathrm{ELSE}}{\frac{n(n-1)}{2} + \#\mathrm{ELSE}} \quad \in [0, 1]$$

   where n = number of non-ELSE scenes.
4. **Aggregation & Visualization**: 5 years rolling mean of PPlot per novel

Evolution of the plot prototypicality

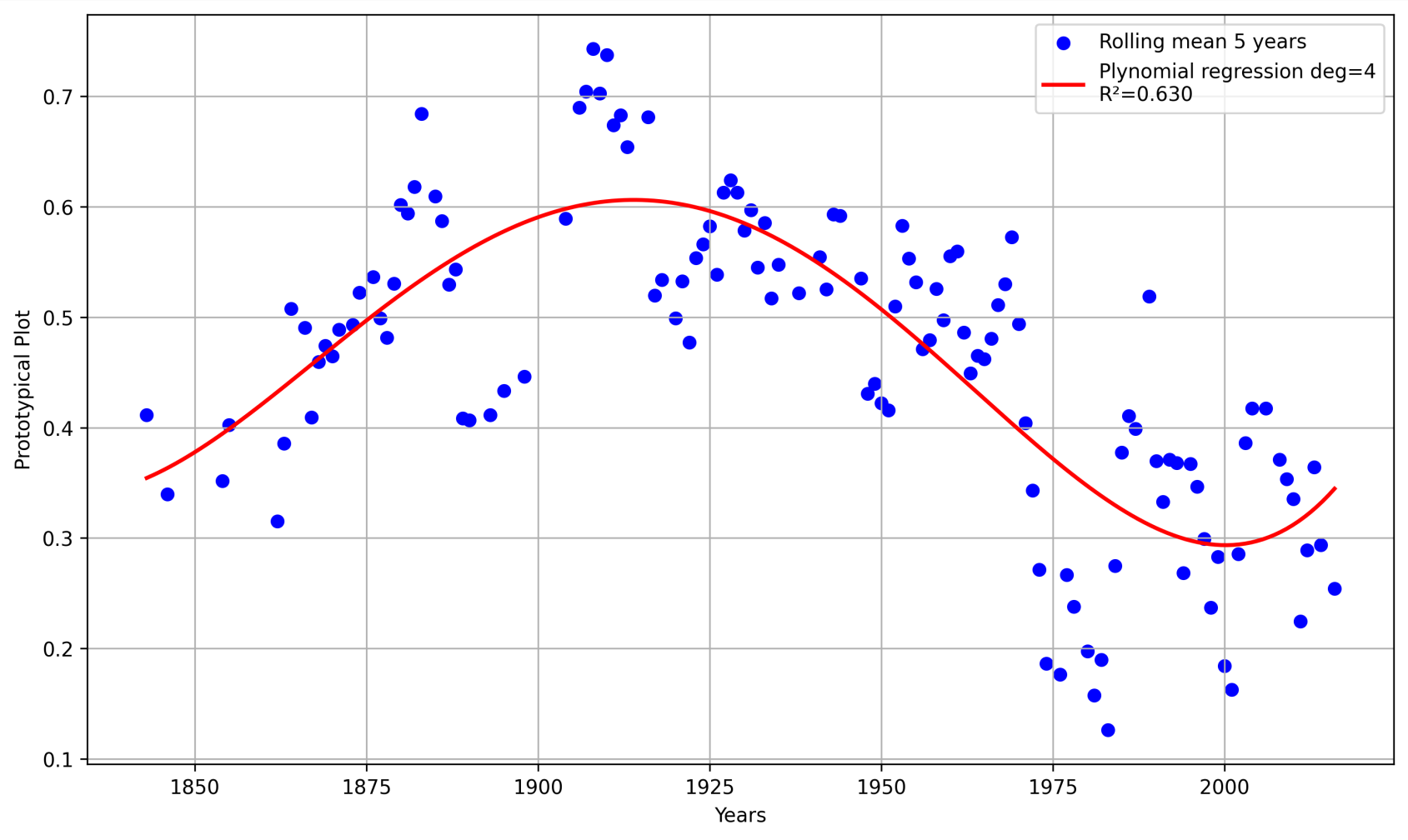


Figure 14: Evolution of the plot prototypicality

Future Work

- Map the historical emergence of our three formal features with specific authors, collections.How contextual elements helped the formal development ?
- Identify the first high level co-occurrence within the same work.
- Investigate the mechanisms of genre formation: was the “birth” of the detective novel driven by a cohort effect, an individual author’s breakthrough, or a slow crystallization process?

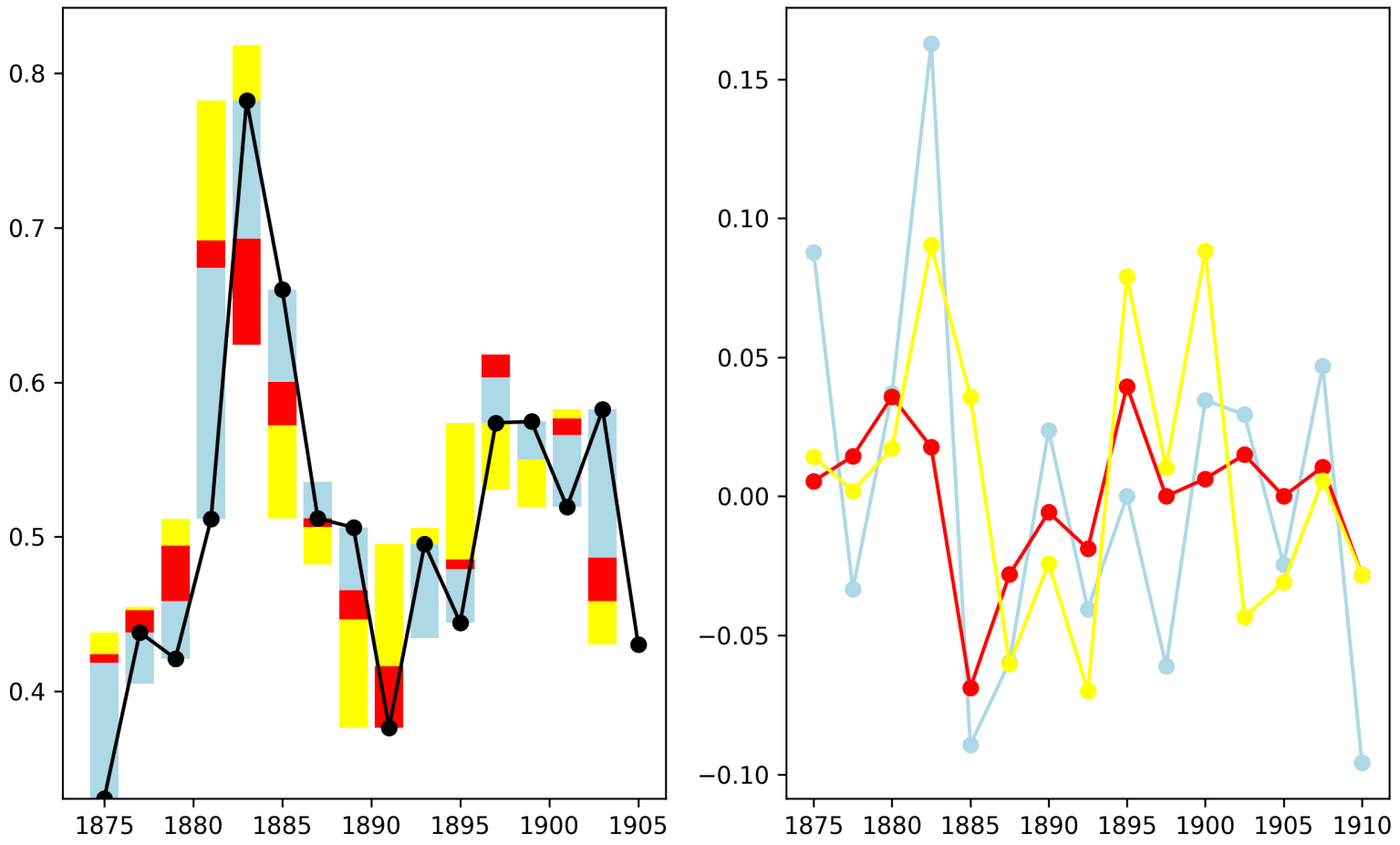


Figure 15: Decomposition equation (Sobchuk, 2025)

Thank you!

Questions?

jean.barre@ens.psl.eu

https://crazyjeannot.github.io/

Family Resemblance Approach

(Wittgenstein, 1953)

> *Take, for example, the activities we call "games" [...] you will not see anything common to all but rather a complicated network of similarities overlapping and crisscrossing [...] like the relationships between members of a family.*

Application to Literary Genres

- The genre is not a rigid category with fixed boundaries.
- It forms a dynamic network of relationships among texts.
- Similarities vary in intensity and nature from one work to another.
- No single trait defines the genre entirely, but multiple traits overlap.

Family Resemblance in the Detective Novel

- Common defining elements: presence of a crime, methodical investigation, central role of the detective.
- No detective novel possesses all genre traits simultaneously, but each shares a variable set of generic features.
- **Recurring characters**: Rouletabille (Gaston Leroux), Tirauclair and Lecoq (Émile Gaboriau), Maigret (Simenon).
- **Stylistic and narrative variations**:
  - *L'Affaire Lerouge* (Gaboriau, 1866): introduces methodical, rational investigation.
  - *The Mystery of the Yellow Room* (Leroux, 1907): revives the puzzle logic with a locked-room twist.
  - The Maigret novels (Simenon, from 1931): renew the approach while retaining the central investigator.
- Each novel contributes to a shifting definition of the genre through a network of partial resemblances.

Operationalizing Family Resemblance – An Attempt

Embeddings and Cosine Similarity as Indicators of Intertextuality

- State-of-the-art contextual encoding model for French literary text representation.
- Fine-tuning the BGE-M3-Embedding model on French literary language.
- **Query**: one paragraph; **Positive**: next 5 paragraphs; **Negative**: 5 random paragraphs.
- **Hypothesis**: embeddings capture individual style notions as well as thematic similarity.

Model Validation

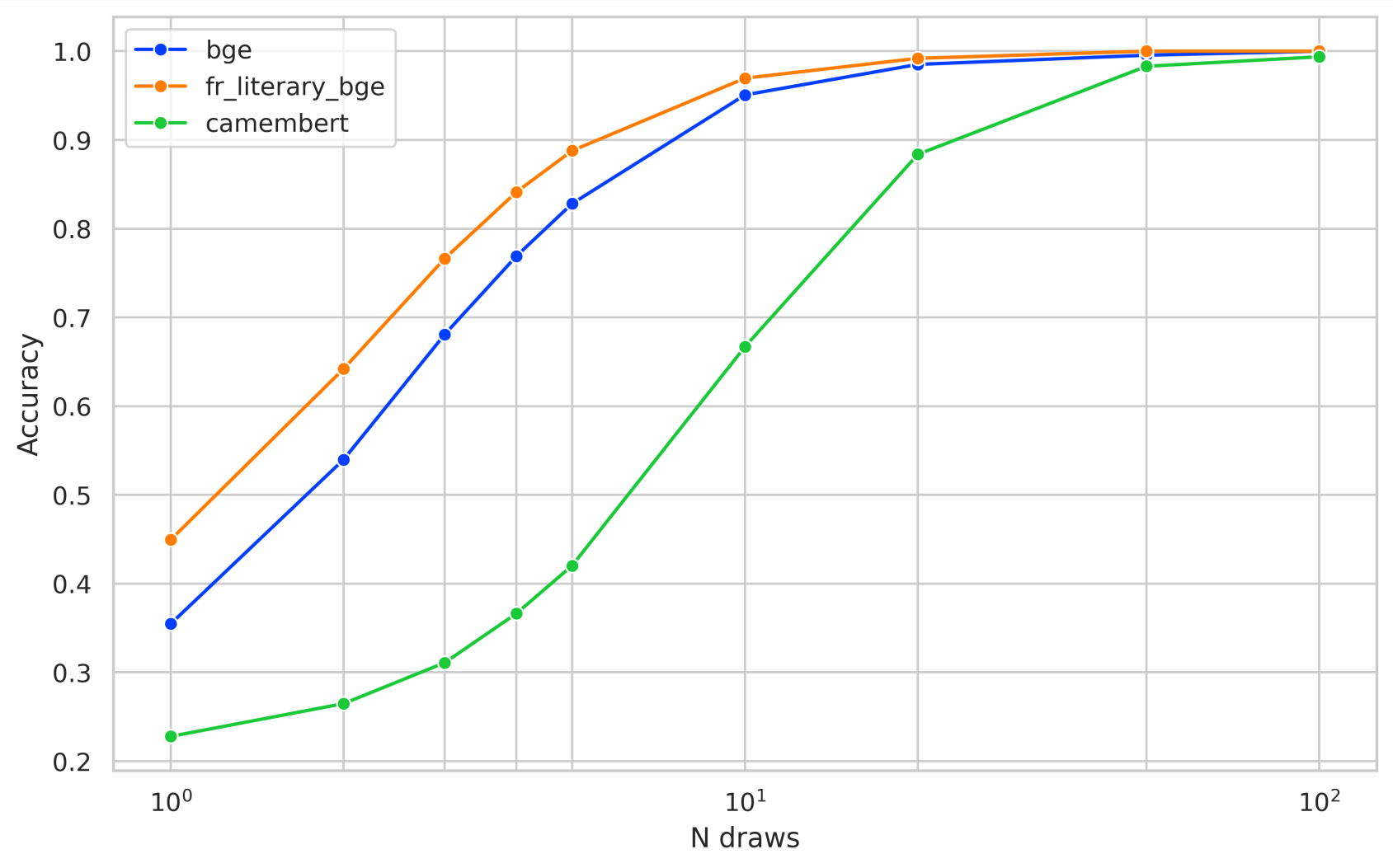


Figure 16: Encoder evaluation

Similarity Network – Novel Scale

HTML version

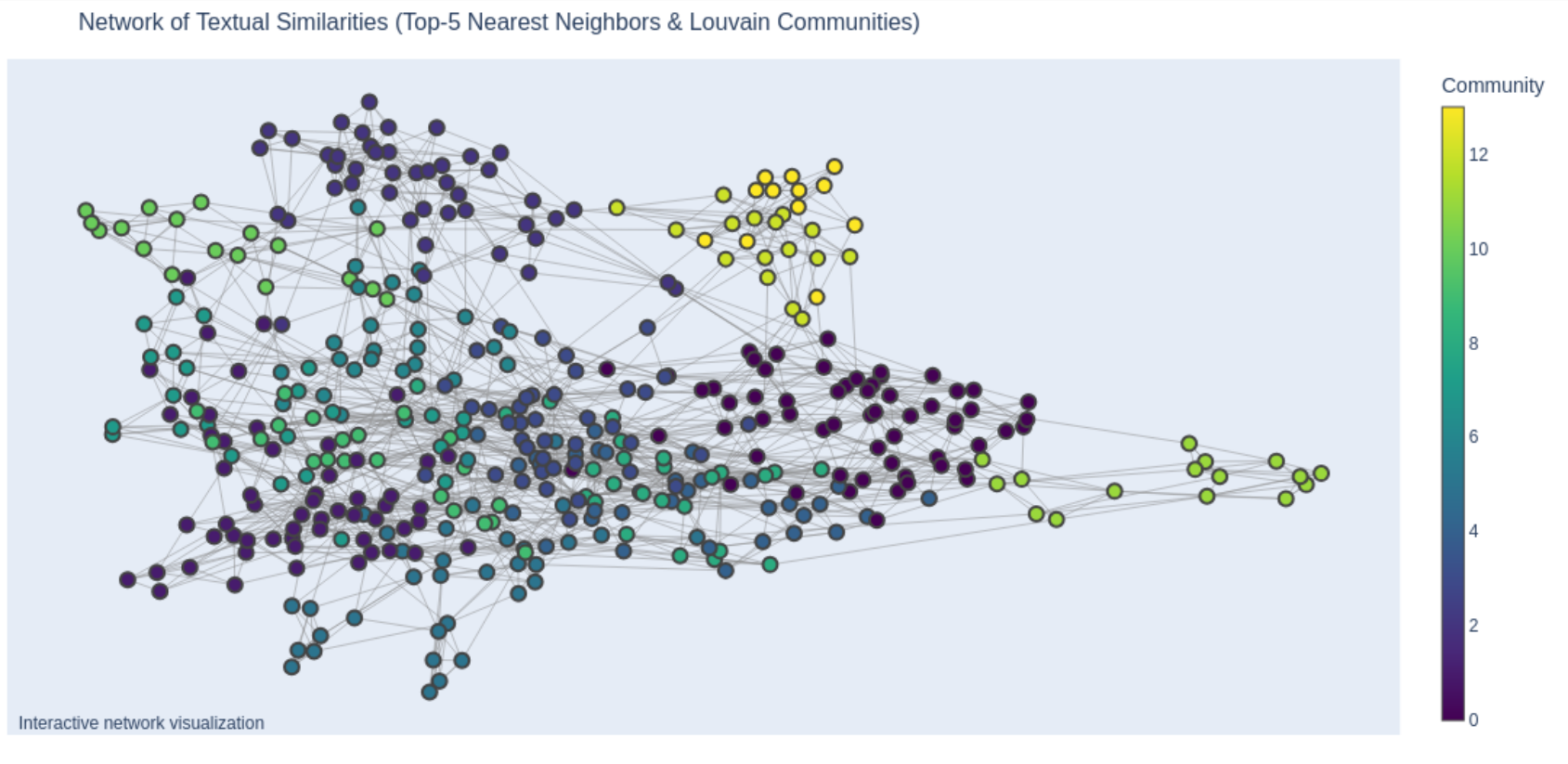


Figure 17: Similarity network for *The Mystery of the Yellow Room* (Leroux, 1908)

Example of Similarity – Excerpt 41

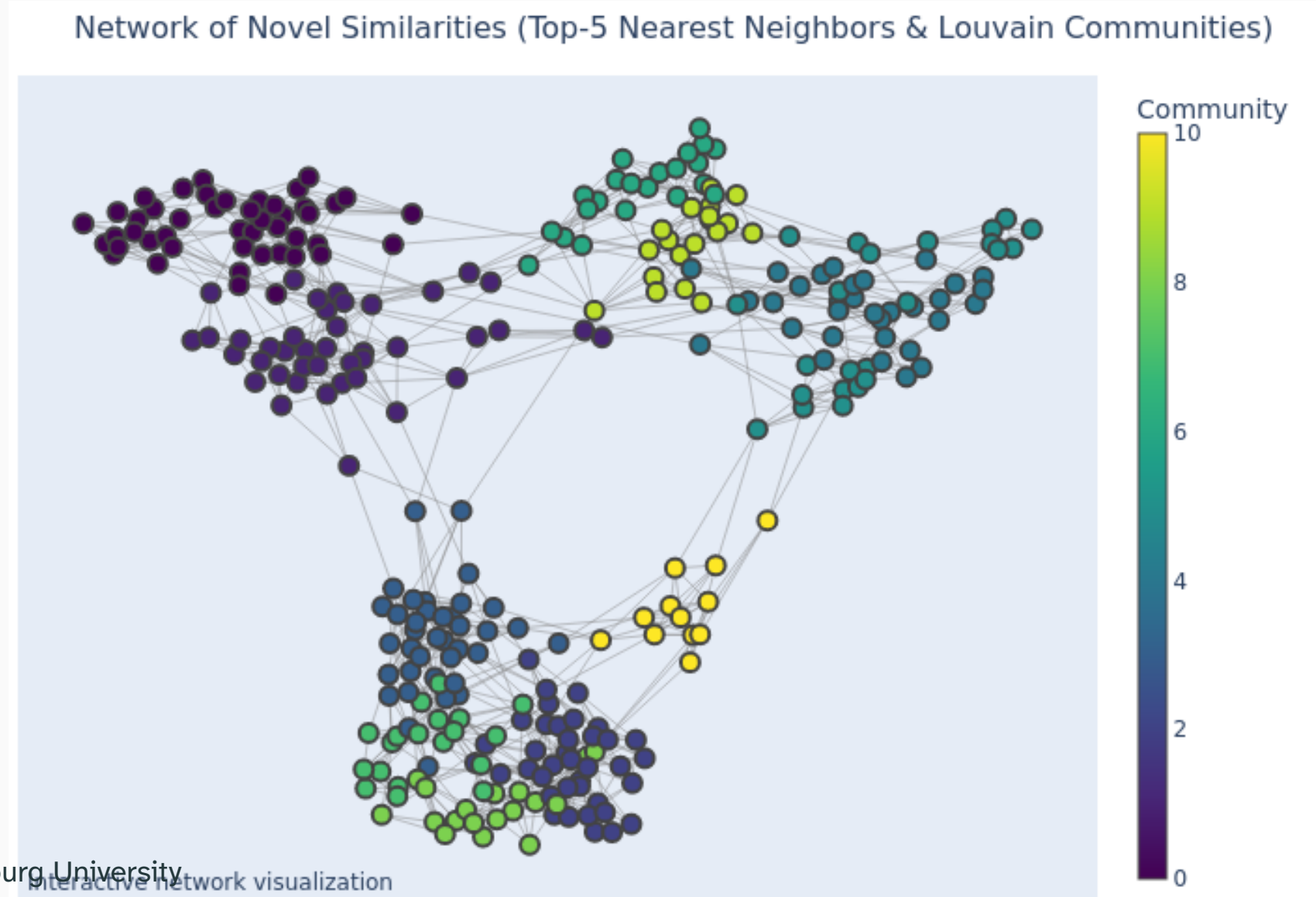
Rouletabille regarda le ciel , le trouva à sa convenance et , sans doute , à la mienne , car il me prit sous le bras et me dit : « Allons ! ... J' ai besoin de marcher . - Eh bien ! lui demandai -je . Ça se débrouille ? ... - Oh ! fit -il , oh ! Il n' y a rien de débrouillé du tout ! ... C' est encore plus embrouillé qu' avant ! Il est vrai que j' ai une idée ... - Dites - la . - Oh ! je ne peux rien dire pour le moment ... Mon idée est une question de vie ou de mort pour deux personnes au moins ... - Croyez -vous à des complices ? - Je n' y crois pas ... »

Example of Similarity – Excerpt 176

J’ en eus la preuve quand , en descendant la côte d’ mask , il me dit : « mask mask est arrivé mask mask avant moi ; il a commencé son enquête avant moi ; il a eu le temps de savoir des choses que je ne sais pas et a pu trouver des choses que je ne sais pas … Où a -t -il trouvé cette canne -là ? … » Et il ajouta : « Il est probable que son soupçon - plus que son soupçon , son raisonnement - qui va aussi directement à mask mask , doit être servi par quelque chose de palpable qu’ il palpe , lui , et que je ne palpe pas , moi … Serait -ce cette canne ? … Où diable a -t -il pu trouver cette canne -là ? … »

Similarity Network – Genre Scale

HTML version



Similarity Network – Corpus Scale

HTML version

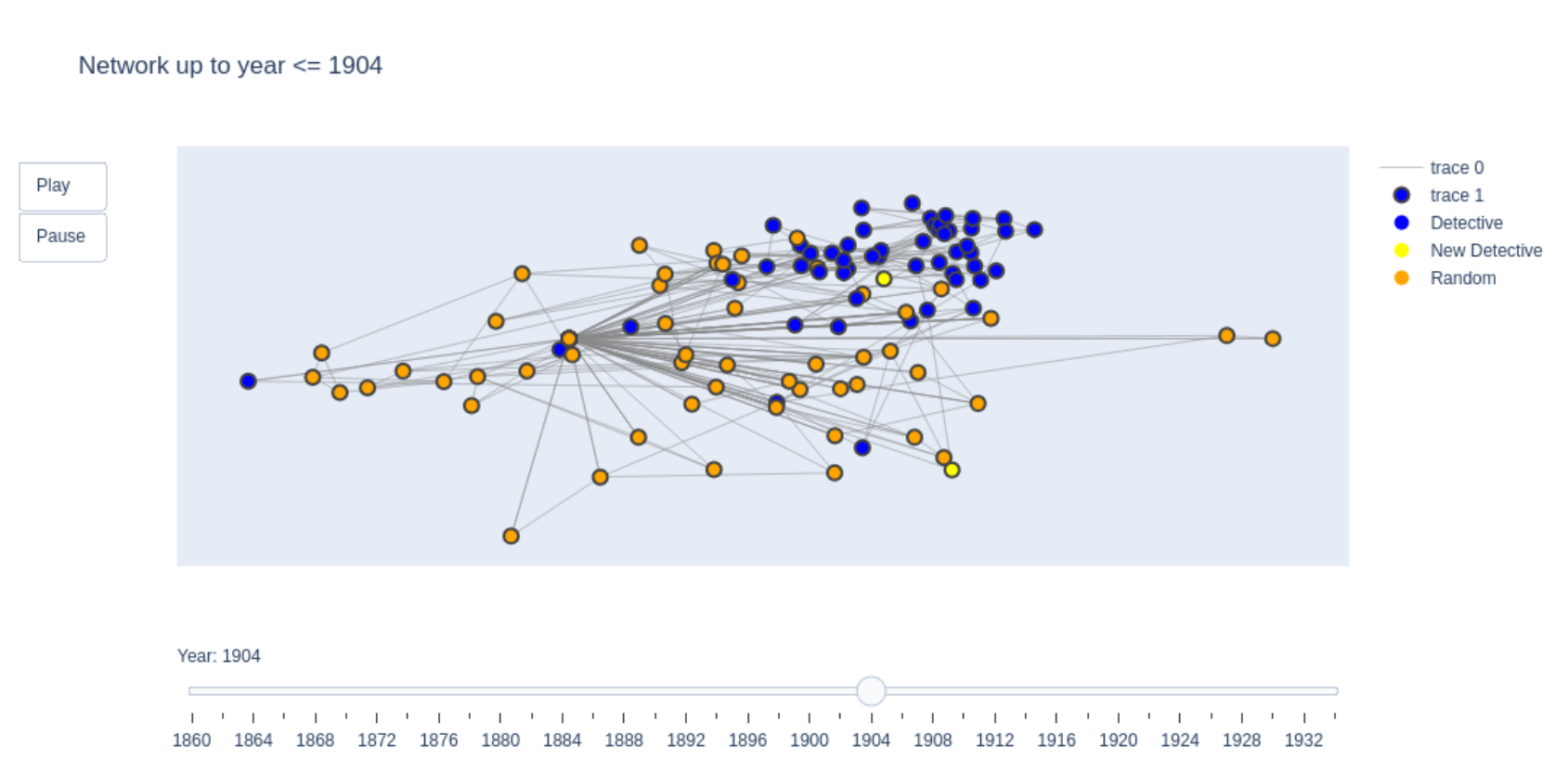


Figure 19: Cumulative similarity network approach for the detective novel corpus