Modeling the Construction of a Literary Archetype: The Case of the Detective Figure in French Literature

Jean Barré[1], Olga Seminck[1], Antoine Bourgois[1], and Thierry Poibeau[1]

[1] LaTTiCe Laboratory, CNRS-ENS-PSL, Paris, France

Abstract

This research explores the evolution of the detective archetype in French detective fiction through computational analysis. Using quantitative methods and character-level embeddings, we show that a supervised model is able to capture the unity of the detective archetype across 150 years of literature, from *M. Lecoq* (1866) to *Commissaire Adamsberg* (2017). Building on this finding, the study demonstrates how the detective figure evolves from a secondary narrative role to become the central character and the "reasoning machine" [35] of the classical detective story. In the aftermath of the Second World War, with the importation of the hardboiled tradition into France, the archetype becomes more complex, navigating the genre's turn toward social violence and moral ambiguity.

Keywords: Computational Literary Studies, Detective Fiction, Detective Figure, Character Embeddings, Genre Evolution, NLP, Machine Learning

1 Introduction

Distant reading [30] has become a key methodological approach in literary studies, enabling the large-scale analysis of digitized corpora through computational techniques. By moving beyond the traditional canon to include lesser-studied works, it offers a broader perspective on literary trends over time [4; 6; 29; 38]. While distant reading can be applied to various aspects of literary texts, the analysis of literary characters has emerged as a particularly productive area of research, allowing scholars to explore narrative structures, character portrayals, and cultural representations at scale.

A foundational work in this line of inquiry is provided by Bamman, Underwood, and Smith [3], who propose a Bayesian mixed-effects model to infer latent character types from a large corpus of English novels. Their study demonstrates that computational methods can identify underlying character personas by analyzing linguistic patterns associated with character descriptions, actions, and dialogues. Following this seminal approach, the computational literary studies community has developed specialized tools to automate the extraction of character information. Notable examples include BookNLP for English [1; 2], as well as its counterparts adapted to other languages such as French [26], German [15], and Dutch [43]. These pipelines enable researchers to systematically identify and cluster all mentions of a given character, alongside associated linguistic features including agentive and passive verbs, modifiers, and direct speech. Leveraging these computational methods within distant reading frameworks, subsequent research has explored significant literary phenomena, notably the representation of gender [31; 39; 41] and the structural

Jean Barré, Olga Seminck, Antoine Bourgois, and Thierry Poibeau. "Modeling the Construction of a Literary Archetype: The Case of the Detective Figure in French Literature." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 947–965. https://doi.org/10.63744/SMbYIWcHZj87.

analysis of plot dynamics [18; 19]. In this study, we use this character modeling approaches to study the emergence and evolution of the detective archetype in French fiction.

The figure of the detective has become a ubiquitous presence in contemporary popular culture, extending far beyond literature into cinema, television series, and various media adaptations. However, its roots lie firmly within English and French literary traditions [11, p. 12], particularly between the late 19th and early 20th centuries, a period crucial to the construction and popularization of detective fiction as a genre.

We decided to focus exclusively on the detective archetype in French literature. Although there are numerous interconnections and reciprocal influences with the Anglo-American tradition, illustrated by Poe's Dupin inspiring Gaboriau's Lecoq [34, p. 43; 32], who in turn influenced Conan Doyle's Sherlock Holmes [8, p. 394; 22, p. 110], the depth and distinctiveness of the French detective tradition alone present a substantial challenge.

To be more specific, we build a classifier that is able to automatically detect characters that are detectives (Section 3). We study the emergence of the detective figure by quantifying the percentage of detectives per year in a large corpus of French fiction *Chapitres* Corpus [21], and show that the detective characters tends to be more and more central to the novels as the literary genre is building up (Section 4). Finally, we also demonstrate that our methods support literary theory about the emergence of detective subgenres, such as the *Hard-Boiled* detectives (putting forward solitary characters fighting organized crime as well as corrupt authorities) and French *Roman Noir* or *Neo-Polar* (Section 5). Our computational analysis shows how the critical perspective from canonical whodunit figures expands to broader representations of the detective, capturing subtle shifts and variations in characterization over time.

2 The Detective Archetype

2.1 Definition

The detective archetype constitutes a foundational figure within detective fiction, evolving from a peripheral puzzle solver to a central character embodying rationality and methodical reasoning. As Dubois emphasizes, classic detectives such as Dupin, Holmes, and Rouletabille favor purely intellectual work, distancing themselves from direct action: "he solves the problems presented to him by pure analytical deduction" [14]. This approach aligns closely with the notion of a *reasoning machine* [35], which disregards human motives or psychology, focusing solely on the accuracy of deductions regarding actions. Consequently, the detective evolves into an extreme symbol of rationality, whose methods rely systematically on interpreting clues and signs to solve a mystery.

This fundamental definition of the detective is intimately linked to the *whodunit*, a subgenre of detective fiction that stands out for its puzzle-like plot, with clues and red herrings carefully scattered to invite the reader to match wits with the detective and try to uncover the truth before the final revelation. Todorov [36] analyzes the specificity of the *whodunit* by highlighting a double narrative structure: the hidden story of the crime and the visible story of the investigation. In this way, the identity and role of the detective are inseparable from the *whodunit*: the detective becomes both the reader's guide and double in logically reconstructing a crime that has already occurred.

While canonical detective figures such as Doyle's *Sherlock Holmes* or Gaboriau's *Père Tabaret* epitomize this rational genius archetype, meaning that intelligence and ingenious reasoning are key to solving the mystery by the end of the novel, the archetype is not limited to officially sanctioned investigators. *Holmes* or *Tabaret* themselves, for instance, are no professional police detectives. Rather, the archetype comprises a complex aggrega-

tion of characters, roles, and traits that evolve through time and across literary subgenres. It can be embodied by a wide variety of characters, ranging from private and amateur detectives to police officers, overly inquisitive journalists, and even criminals [9].

In France, this diversification materialized through increasingly ambiguous and complex figures, such as Arsène Lupin, simultaneously detective and criminal, pursued by genuine detectives like Commissioner Ganimard or the amateur sleuth Isidore Beautrelet. These diverse incarnations underscore the archetype's adaptability and illustrate how detective fiction explores multiple perspectives, social roles, and investigative approaches through a spectrum of distinct yet interconnected characters.

2.2 Origins

The detective archetype emerged from mid-19th-century literary developments, influenced by proto-detective figures such as Edmond Dantès (*Le Comte de Monte Cristo*), Rodolphe (*Les Mystères de Paris*), and Rocambole (Ponson du Terrail) [20]. Its origins were also shaped by figures like Eugène François Vidocq [17], a former criminal who became the first chief of the *Sûreté* in 1811 and founded the first modern detective agency, *Le Bureau de Renseignements*. Vidocq directly inspired Balzac's Vautrin, and Poe's Dupin, two proto-detective figures.

The foundation of the detective novel was laid by Edgar Allan Poe [9], through three short stories featuring Auguste Dupin, a non-professional detective who solves crimes through a combination of intuitive reasoning and scientific tools: a method Poe called *ratiocination* [16]. While Edgar Allan Poe originally formulated the detective formula in short story form [13, p. 81], Émile Gaboriau (1832-1873) was the first to adapt it to the novel format within French tradition.

The genre continued to evolve with figures such as *Rouletabille* by Gaston Leroux, Maurice Leblanc's *Arsène Lupin*, and later Georges Simenon's *Maigret*. This diversity of figures illustrates a progressive construction of the archetype through successive layers, shaped by varied sociocultural and literary contexts.

Detective fiction gradually established itself as a major literary genre by the late 19th century and, by the 1920s-30s, had become a central reference in the collective imagination. Messac [27], among the first to study the genre systematically, defined it as "a narrative dedicated above all to the methodical and gradual discovery, through rational means, of the exact circumstances of a mysterious crime". Its specificity lies in a narrative structure that moves from an initial crime or mystery to its reconstruction, generating dramatic tension through multiple suspects and building suspenseful reader expectations [9]. Beyond suspense, the detective is pivotal: their reasoning drives the story toward resolution [42]. It is through the detective's logic, insight, and analytical skill that the plot unfolds.

3 The Detective Detector

Building on this rich panorama, our goal is to extract the detective archetype as a common matrix across all these figures, from *Père Tabaret* (first appearence in 1866) to *Commissaire Adamsberg* (last appearence in 2017), and, paradoxically, as an entity that fully coincides with none of them. We aim to implement a quantitative approach[1] capable of identifying, in each text, the behavioral, narrative, and linguistic invariants that underlie the investigator's stance, despite the diversity of eras, styles, and roles (methodical, intuitive, empathetic, cynical, ...).

[1] Data and code of this paper available at `https://github.com/lattice-8094/DETECTIVES`

3.1 Corpus and Data Annotation

This study is based on a large corpus of French fiction, the *Chapitres* Corpus [21], which includes nearly 3,000 texts, including a lot of detective fiction novels.

Detecting the detective archetype during 150 years of the detective genre is by no means trivial. It requires two things: first, the ability to identify characters reliably at the scale of a novel [5; 10]. That is done using the Propp-fr pipeline[2], an extension of BookNLP-fr [26] that facilitates automatic coreference chain recognition at the scale of entire novels (over 100k tokens). Second, the ability to distinguish, among these characters, which are detectives and which are not.

Our annotation aimed to represent the detective archetype across various subgenres, from the French *whodunit* to *série noire* thrillers. We focused on paradigmatic figures with significant influence due to their recurring roles: amateur journalists (*Rouletabille*, Gaston Leroux, 1907), outlaw vigilantes (*Arsène Lupin*, Maurice Leblanc, 1908), detective-crooks (*Larsan*, from the same novel as *Rouletabille*), honest professionals (*Lecoq*, Emile Gaboriau; *Maigret*, Georges Simenon), and French adaptations of hard-boiled detectives (*Nestor Burma*, Léo Malet, 1942; *Gabriel Lecouvreur*, Didier Daeninckx, 1981; *Adamsberg*, Fred Vargas, 1991).

In total, 185 characters across 156 novels were annotated as *Detectives*, representing 47 unique figures (some appear in multiple works). We also annotated 419 characters as *Non-Detectives*. Since our goal is binary classification (*detective* vs. *not detective*), we ensured a diverse negative set so the model could learn what does not define the archetype. To this end, we randomly selected 419 non-detective characters from the Chapitres corpus, stratified by time to reflect a variety of roles and narrative functions.

We also included characters from the mid-19th century, prior to the emergence of the genre, to expose the model to periods where the *detective figure* as we know it had yet to appear.

3.2 Character Embeddings

Once the characters are annotated as Detective or Non-Detective, we created vectorized representations of them that we will refer to as *character embeddings*. These semantic representations are built upon the vocabulary extracted alongside the coreference chains by the BookNLP-fr tool. For each character, we retrieve the actions they perform or endure (grammatically agent and passive verbs), modifiers (such as adjectives and other predicatives), and nouns that come after a possessive determiner. We hypothesize that the detective archetype involves a specific textual characterization that can thus be captured quantitatively.

These elements are used as the basis for the semantic representation. Trying to represent each character in a vector space, we implemented two distinct approaches.

Bag of Words representation: The first is a baseline with Bag-of-Words (BoW). The idea is to control the dimensions of the vector space by simply selecting the 1,000 Most Frequent Words (MFW) from the character attributes retrieved by BookNLP-fr. For each character in our annotated dataset, we retrieved the relative frequency, of every MFW associated with the character.

Contextual Embeddings with CamemBERT:

The second method is based on contextualized word embeddings generated by the $\text{CamemBERT}_{\text{LARGE}}$ encoder-only language model [23]. For each attribute linked to a character, we gather contextual embeddings. The final character embedding vector is

[2] `https://github.com/lattice-8094/propp` (accessed July 11, 2025).

obtained by averaging the contextual embeddings of all attributes linked to that character. This element-wise averaging yields a dense semantic representation (1,024 dimensions) that captures the character's descriptive context beyond surface lexical frequency. Passive verbs (those for whom the character is the object) were excluded from this process, as preliminary experiments showed they provided limited discriminative power for the classification task due to their sparseness.

3.3 Supervised Classification

To predict whether a character is a detective, we used manually annotated gold-standard labels as ground truth and fed precomputed character embeddings into models. We compared two Scikit-learn classifiers [33]: Support Vector Machines (SVM) and Logistic Regression.

To handle class imbalance, we used stratified 5-fold cross-validation (`StratifiedKFold`), maintaining the detective/non-detective ratio across splits. Model performance was evaluated using balanced accuracy (`balanced_accuracy_score`).

To prevent information leakage (e.g., from characters like inspector Maigret appearing in both training and test sets) we applied a `Leave-One-Group-Out` (LOGO) strategy. When grouping by author, all books by the same author were placed entirely in either the training or test set. We define three grouping schemes:

1. *Character:* all instances of the same character are held out together.
2. *Author:* all characters by the same author are placed in one group.
3. *Time Period:* all characters originating from the same historical era form a group.

3.4 Quantitative Evaluation

Table 1 reports balanced accuracy alongside precision, recall, and F1-scores for both classes under our three main feature-model configurations, as well as under the strict LOGO validation schemes. Starting from the simple Bag-of-Words baseline, Logistic Regression achieves a balanced accuracy of 0.836, but struggles particularly on the detective class (F1 = 0.75). Replacing the classifier with an SVM yields a substantial jump (B. Acc. = 0.895), with both precision and recall improving across classes; a testament to the SVM's capacity to carve sharper decision boundaries even on sparse, high-dimensional vectors.

Moving to contextual embeddings, CamemBERT paired with Logistic Regression already outperforms both BoW setups (B. Acc.=0.906), delivering a notable boost in recall for detectives (0.919). Finally, when combining CamemBERT embeddings with an SVM, we observe the strongest results overall: a balanced accuracy of 0.923, precision and F1-scores near or above 0.94 for non-detectives, and an F1 of 0.887 for detectives. This configuration does not only maximize our core metric but also maintains high precision and recall on the under-represented detective class, minimizing both false positives and false negatives.

Given these results, we will adopt the CamemBERT + SVM model for all subsequent experiments.

Model	B. Acc.	Non-detective			Detective		
		P	R	F1-score	P	R	F1-score
BoW + LogReg	0.836	0.94	0.78	0.85	0.64	0.89	0.75
BoW + SVM	0.895	0.95	0.91	0.93	0.81	0.88	0.84
CamemBERT + LogReg	0.906	**0.961**	0.893	0.926	0.791	**0.919**	0.850
CamemBERT + SVM	**0.923**	0.959	**0.938**	**0.948**	**0.866**	0.908	**0.887**
LOGO (Character)	0.901	0.940	0.938	0.939	0.860	0.865	0.863
LOGO (Author)	0.908	0.943	0.945	0.944	0.875	0.870	0.873
LOGO (10 years)	0.915	0.950	0.943	0.946	0.872	0.886	0.879
LOGO (20 years)	0.904	0.940	0.943	0.942	0.870	0.865	0.867
LOGO (50 years)	0.892	0.928	0.952	0.940	0.885	0.832	0.858

Table 1: Detailed models performances: balanced accuracy (B. Acc.), precision (P), recall (R) and F1-score per class for the Detective Detector using different semantic representations and classification algorithms.

As can be seen in Table 1, the LOGO results are minimally different from the best non-LOGO model, leading to the conclusion that despite an unbalanced training corpus in terms of different detectives and authors, the Detective Detector is robust, and not sensitive to the over-representation of some detective figures.

We carefully examined the temporal stability of the Detective Detector to ensure high scores weren't driven by a specific period and that Detectives were consistently identified across 150 years. To this end, we performed LOGO analyses over 10-, 20-, and 50-year intervals. As shown in Table 1, results reveal no significant degradation, indicating the model is not biased toward any particular era. This is further supported by the prediction error over time (Figure 1), which remains stable despite minor fluctuations. Such diachronic robustness suggests that, beyond stylistic or editorial variation, the investigator figure retains a consistent set of linguistic and narrative features, forming a recognizable discursive imprint that transcends time and unifies 150 years of detective fiction.

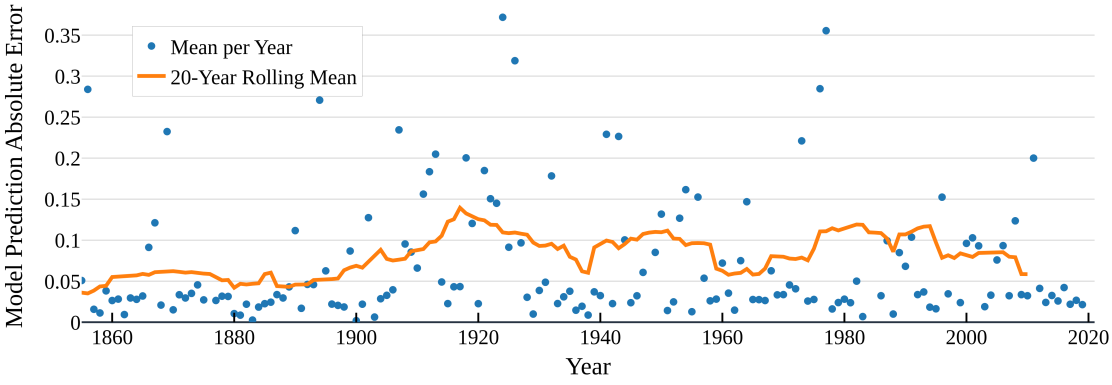

Figure 1: Model prediction error over time.

3.5 Qualitative Evaluation of the Distinctive Attributes of the Detective Archetype

While we assessed that the Detective Detector obtains a high accuracy, we also wanted to understand what distinguishes Detective characters from Non-detectives. For this, we look at the attributes (lexical items) that are associated to Detective and Non-detective

characters across the whole Chapitres corpus where we detected 713 Detectives and 29,152 Non-Detectives. To compare the attributes of the two types of characters, we used a method based on normalized z-scores in the [-1, 1] interval, where +1 indicates the most distinctive attributes for detectives.

The z-score are calculated in the following manner: the numerator is the log-odds ratio with Dirichlet prior smoothing. The denominator is the standard error estimate based on smoothed counts. This method is based on Monroe, Colaresi, and Quinn [28]'s method for informative lexical feature extraction.

$$z_{\text{attr}} = \frac{\log\left(\frac{c_1+\frac{p}{n}}{n_1+1} \Big/ \frac{c_2+\frac{p}{n}}{n_2+1}\right)}{\sqrt{\frac{1}{c_1+\frac{p}{n}} + \frac{1}{c_2+\frac{p}{n}}}}$$

$z_{\textbf{attr}}$**:** Attribute distinctiveness score (positive = more associated with group 1)
c_1**:** Attribute count in group 1 (detectives)
c_2**:** Attribute count in group 2 (non–detectives)
p**:** Attribute count in the corpus ($p = c_1 + c_2$)
n_1**:** Sum of all attributes in group 1 ($n_1 = \sum c_1$)
n_2**:** Sum of all attributes in group 2 ($n_2 = \sum c_2$)
n**:** Sum of all attributes ($n = n_1 + n_2$)

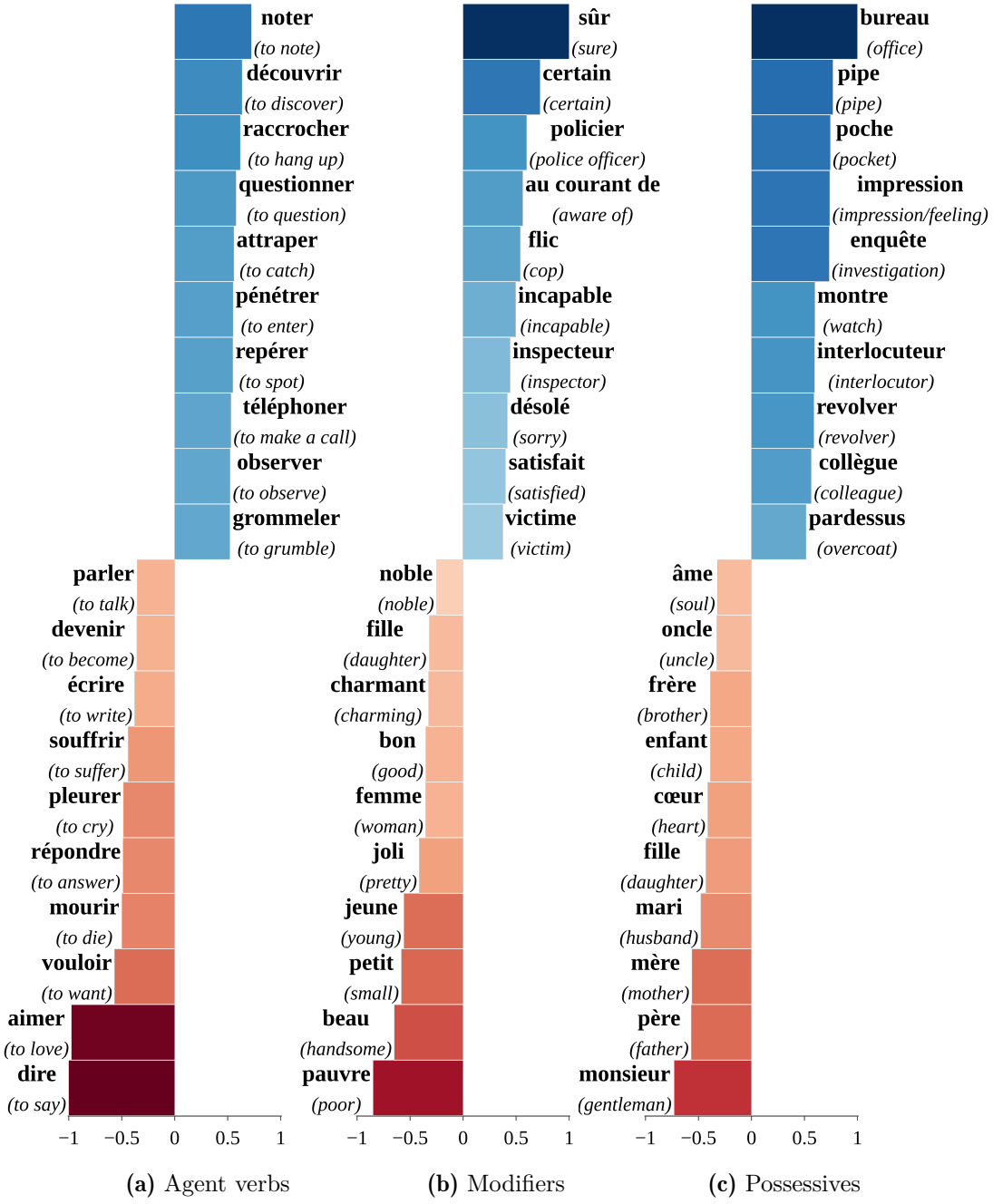

Figure 2: Attribute distinctiveness of the Detective figure, measured by normalized z-score. A value of +1 indicates the most strongly detective-associated attribute and −1 the least.

For the three retained types of attributes (Agentive verbs, Modifiers and Possessives), Figure 2a shows the vocabulary that is the most typical of the detective archetype (in blue) and the least (in red). The z-scores of agentive verbs get around terms such as *note*, *discover*, *question*, *spot*, and *observe* underscores the detective's fundamental identity as a methodical thinker. These verbs capture a relentless focus on evidence gathering and inference, painting the sleuth as a near-mechanical reasoning apparatus. Equally telling is the stark under-representation of verbs like *love*, *cry*, *suffer*, or *desire*, which reveals an emotional reserve that sets detectives apart from other characters; their narrative presence

is defined not by passion or turmoil but by an unwavering, almost clinical detachment.

Turning to possessive constructions, detectives rarely invoke familial bonds or intimate relationships. References to *mother*, *father*, or *daughter* possessions are virtually absent, while their *overcoat*, *pocket*, *revolver* and *watch* abound, which are emblems of a vocation that privileges tools of observation over the ties of kinship. This lexical pattern evokes that iconic image of the lone investigator: coat buttoned high, pockets bulging with instruments of the trade, a revolver at the hip and a timepiece ever-present, signaling both professional rigor and personal solitude.

Finally, the modifiers panel lays bare the cultural coding of the detective as neither female nor ornamental. Highly negative scores for descriptors like *woman*, *girl*, *pretty*, or *handsome* reinforce the image that the detective is rarely described as beautiful or feminine. Instead, the archetype crystallizes as a solitary, intellectually relentless figure whose practical attire and emotionally neutral posture emphasize function over form. Altogether, these lexical signatures confirm and nuance the long-standing critical portrayal of the detective as an isolated, cerebral agent whose world is built on observation, deduction, and the tools of investigation rather than on personal ties or expressive affect.

4 The Emergence of the Detective Figure

To extend our characterization of the Detective Archetype beyond the manually annotated set, we applied the CamemBERT+SVM model —our best-performing detector—to all characters in the Chapitres corpus. For each novel, we retained the ten most prominent characters (by frequency of mention), yielding a sample of 29,610 characters. 713 of them were classified as Detectives. This large-scale inference allows us to trace both the quantitative rise of detective figure and its narrative prominence over more than a century of French fiction.

Figure 3 shows that the first literary detective emerge around 1860, with a steep increase in their proportion from 1900 up to nearly 6% of all characters by the late 1930s. After 1940, the relative frequency of detectives dips to about 4%, before experiencing a modest revival in the 1980s.

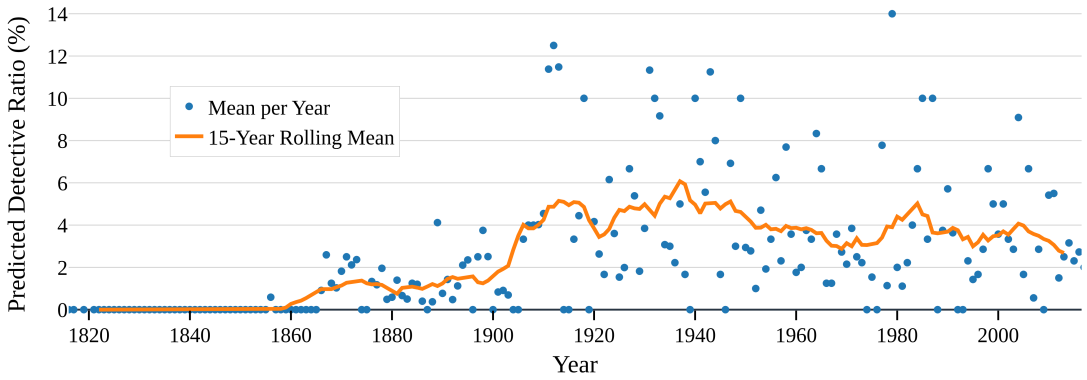

Figure 3: Model-predicted detective character ratio over time.

Lavergne [20] identifies the figure of the investigator as a key element in the emergence of the detective-novel genre. Formerly relegated to a secondary role, the investigator becomes the central character who steers the gradual reconstruction of the drama. We further assessed the detective's narrative centrality via the mention-ratio (the length of a character's coreference chain relative to the book's average). Figure 4 reveals an ongoing upward trajectory from the mid-19th century to the end of our corpus, with a first

inflection around 1900. The quadratic fit trendline underscores how, as detective fiction matured, the investigator increasingly anchored the narrative, even as subgenres diversified between 1950 and 2000 (*hard-boiled, série noir, néo-polar*), broadening the range of detective portrayals while preserving their elevated visibility.

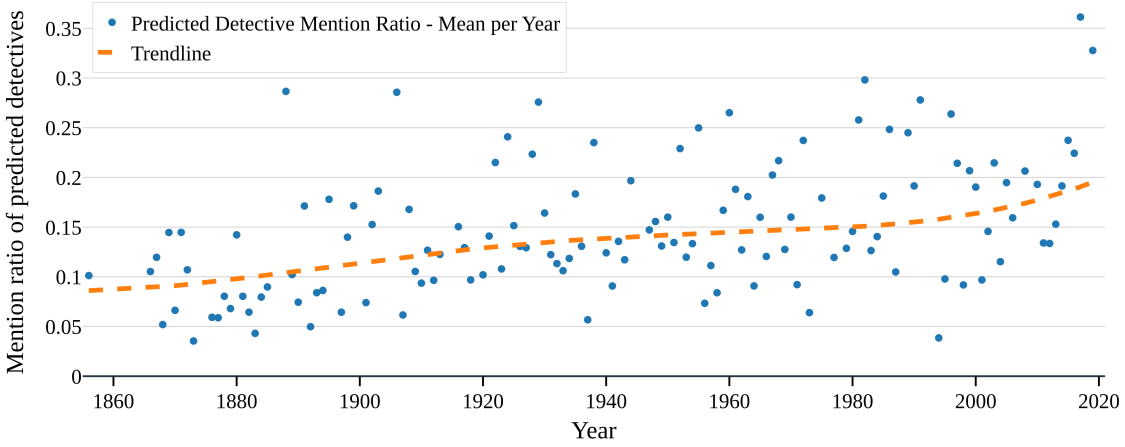

Figure 4: Progressive establishment of the detective as the narrative core.

The conjunction of these two findings, an increasing number of detective figures and their growing centrality, suggests a dual movement. On the one hand, the archetype crystallizes early enough to remain recognizable even as detective fiction experiments with new forms. On the other hand, the narrative's increasing reliance on its investigator figure indicates that the mystery's resolution depends more and more on the detective's voice and subjectivity; he is no longer a mere auxiliary to the plot, but the pivot around which narrative tension is organized.

This raises a central question for the rest of our study: to what extent does the *voice* or discursive identity of the detective archetype remain constant across 150 years of genre history? Beyond the persistence of a core set of textual markers, is the archetype's semantic imprint (its characteristic ways of speaking, perceiving, and investigating) truly stable, or does it undergo subtle shifts as it traverses successive generic and historical contexts? To address this, we turn to a fine-grained, diachronic analysis of the archetype's semantic trajectory.

5 The Semantic Trajectory of the Detective Archetype

In this section, we examine how the detective archetype evolves over time by uncovering subtle semantic shifts beyond its core textual markers. To trace this fine-grained trajectory, we take the CamemBERT contextual embeddings for the 713 characters our model flagged as Detectives in the Chapitres corpus and run a clustering experiment: first reducing these high-dimensional representations to two dimensions using the uniform manifold approximation and projection (UMAP) algorithm [25] (`n_neighbors=10`). Visually we could clearly distinguish three big clusters in the space. We therefore choose to continue with a clustering based on the K-means algorithms that identifies three clusters and our intuition about the cluster's boundaries were confirmed. The results are displayed in Figure 5. We can see clearly the influence of the publication year on the clustering: detectives from similar time periods tends to find themselves in the same clusters.

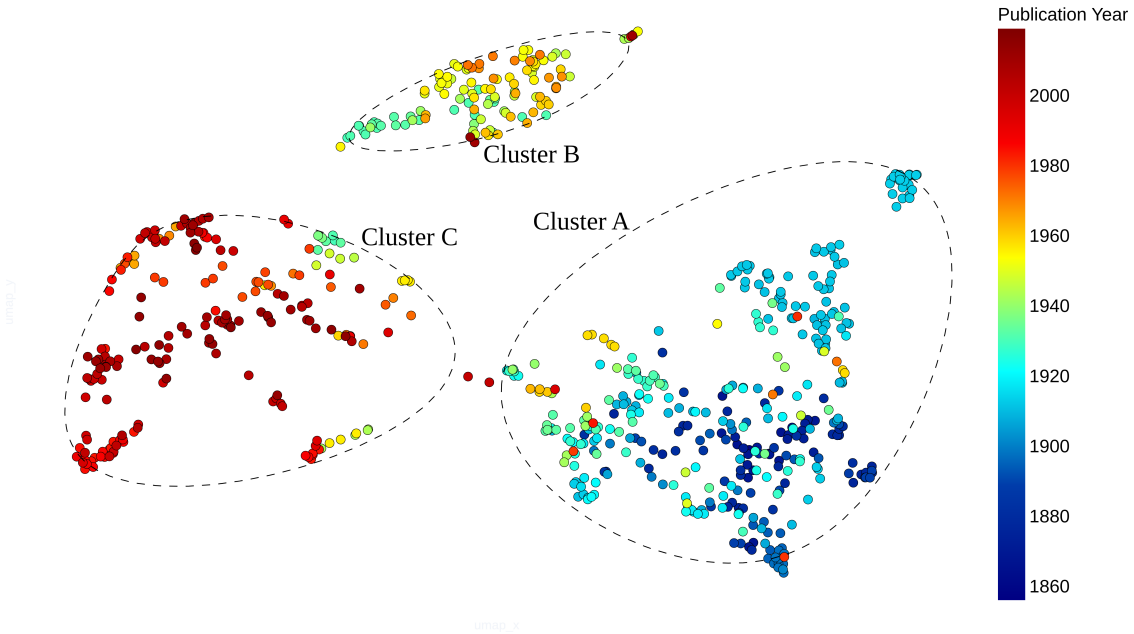

Figure 5: Predicted detective clusters.

To determine whether the three clusters reflect distinct phases in the detective subgenre's semantic trajectory, we applied the z-score method from Subsection 3.5 to extract each cluster's most distinctive lexical attributes. This time, we only took in to consideration the positively scores attributes (z-score > 0). We can find this vocabulary in Table 2. We interpret the specific traits of the three clusters by having a closer look at this vocabulary and looking it up in the context of the Chapitre Corpus.

	Cluster A	Cluster B	Cluster C
Agent Verbs	dire (to say)	questionner (to question)	sentir (to feel)
	s'écrier (to exclaim)	téléphoner (to make a call)	attraper (to catch)
	répondre (to answer)	supposer (to suppose)	poser (to put down)
	déclarer (to declare)	regarder (to look)	foutre (to not care / to make fun of)
	murmurer (to murmur)	bourrer (to stuff)	repérer (to spot)
	répliquer (to reply)	préférer (to prefer)	tenter (to attempt)
	interrompre (to interrupt)	fumer (to smoke)	lever (to get up)
	venir (to come)	connaître (to know)	croiser (to run into)
	demeurer (to remain)	boire (to drink)	acquiescer (to nod / to agree)
	reprendre (to resume)	avoir (to have)	se souvenir (to remember)
	ajouter (to add)	se souvenir (to remember)	sortir (to go out)
	interroger (to question)	savoir (to know)	enfiler (to put on (clothes))
	affirmer (to assert)	permettre (to allow)	vérifier (to check)
	crier (to shout)	soupirer (to sigh)	avaler (to swallow)
Modifiers	cher (dear)	sûr (sure / confident)	désolé (sorry)
	jeune (young)	lourd (heavy)	petit (small)
	brave (brave)	roux (red-haired)	flic (cop)
	malheureux (unhappy)	peine (sorrow / grief)	nu (naked)
	vieux (old)	commissaire (commissioner)	responsable (responsible)
	pauvre (poor)	maussade (gloomy / sullen)	certain (certain)
	excellent (excellent)	police (police)	seul (alone)
	faux (false / fake)	spécial (special)	censé (supposed)
	digne (worthy)	ivre (drunk)	conscient (aware)
	mystérieux (mysterious)	marié (married)	incapable (incapable)
	général (general)	incapable (incapable)	type (guy / dude)
	bon (good)	fatigué (tired)	proche (close / nearby)
	honnête (honest)	gros (fat)	enquêteur (investigator)
	misérable (miserable)	large (broad / wide)	au courant de (aware of)
Possessives	ami (friend)	pipe (pipe)	veste (jacket)
	parole (word / promise)	bureau (office / desk)	corps (body)
	compagnon (companion)	femme (wife)	téléphone (phone)
	dieu (god)	pardessus (overcoat)	sac (bag)
	maître (master)	chapeau (hat)	mal (pain / suffering)
	monsieur (gentleman)	inspecteur (inspector)	lunette (glasses)
	revolver (revolver)	impression (impression / feeling)	voiture (car)
	ordre (order)	envie (desire)	dos (back)
	foi (faith)	interlocuteur (interlocutor)	père (father)
	cabinet (office / cabinet)	collègue (colleague)	visage (face)
	ennemi (enemy)	veston (jacket)	ventre (belly / stomach)
	devoir (duty)	collaborateur (collaborator)	doigt (finger)
	complice (accomplice)	poche (pocket)	peau (skin)
	cheval (horse)	client (client)	mère (mother)

Table 2: Detective clusters most distinctive attributes.

Cluster A corresponds to the "classical" rational detective of the late nineteenth and early twentieth centuries: figures who resolve mysteries primarily through methodical deduction and interrogation, as in the serialized fiction of Gaboriau, Leroux, and Leblanc [40]. In addition to high z-scores for agentive verbs like *question*, this cluster is distinguished by an abundance of dialogue markers: *say*, *answer*, *declare*, *reply*, reflecting the detective's reliance on spoken questioning and verbal exchanges. Unlike the more introspective or action-oriented patterns in the other clusters, Cluster A's signature lies in its emphasis on the detective as an interlocutor whose investigative power is enacted through dialogue.

Cluster B, by contrast, moves beyond this cool rationality to embrace a more humanized investigator figure. It aligns with France's Golden Age of detective fiction, symbolically initiated in 1927 by the *Masque* collection[3] and notably symbolized by the emergence of Simenon's Inspector Maigret in the 1930s, whose introspective empathy and attention to social detail depart from the archetype's earlier detachment. Here, lexical markers

[3] See Martinetti [24] for the history of the collection.

shift toward verbs and modifiers that convey compassion, observation of daily life, and a nuanced portrayal of the detective's own emotional landscape.

We can observe this by the fact that words as *soupirer* (to sigh) and *supposer* (to suppose) are makers of Cluster B (see Examples (1) and (2)).

(1) a. Maigret, **soupira**, prit un temps pour allumer sa pipe.[4]
b. Maigret **sighed** and took a moment to light his pipe.

(2) a. Maigret **supposa** que Joseph avait amené Jaja et Sylvie à Antibes. Il ne se trompait pas.[5]
b. Maigret **guessed** that Joseph had brought Jaja and Sylvie to Antibes. He was not mistaken.

As noted by Cawelti, "Maigret's skills as a detective [...] are not ratiocinative but a result of tireless energy and irrational intuition" [13, p. 126]. Simenon's work transforms the detective narrative into *a police procedural*, moving away from the traditional *whodunit*: "there is no mystery in the inverted tale about who killed the victim and how; the mystery is whether the detective will be able to solve the crime" [13, p. 126]. No longer a distant genius, Maigret becomes a profoundly human figure: "the flow of the narrative is largely confined to the flow of Maigret's perceptions, observations, and feelings" [13, p. 127], and the reader is therefore "privy to many of the inferences and deductions he draws from the clues" [13, p. 127].

Cluster C represents the modern period, emerging post-1945 with the introduction of the American *hardboiled* style in France. However, this period also sees the continued evolution of the empathetic detective tradition. Fred Vargas's Commissaire Adamsberg (first appearing in 1991) exemplifies an alternative modern trajectory: "Unlike most data-driven detectives who center their conclusions on deliberation of facts and figures, Adamsberg, the sensitive semiotician, clearly resides in the post-World War I era where 'new methods of investigation [based] on instinct, intuition, and empathy replaced rational deduction'" [7].

Yet alongside such introspective figures, the "American moment of French literature", as described by Cadin [12], fundamentally reshaped French detective fiction with the emergence of the *Série Noire* and, later, the *néo-polar*. Authors such as Boris Vian, Léo Malet (with *Nestor Burma*), and Didier Daeninckx (with *Gabriel Lecouvreur*) adopted this style, creating investigations that were far more realistic, physically dangerous, and marked by a certain cynicism. The detective is no longer simply a rational and observant thinker, but became a character immersed in social violence, often disillusioned, and navigating a dark world where the line between good and evil was increasingly blurred.

In the vocabulary, we identified for example the word *foutre*, which is a vulgar polysemous word that originally means *to fuck*, but is used in modern language as a swear word for different matters, for example in the saying *Je m'en fous.*, meaning something such as *I don't give a shit.* and *Qu'est-ce que vous foutez ?* meaning *What the hell are you doing?*. The vulgarity of the language is typical for the violence in the subgenre (see Example (3) for an illustration).

(3) a. Le visage du commandant Verhoeven change brusquement, on ne rigole plus. Il plaque violemment le téléphone sur la table en fer. - Et maintenant, tu me **fous** un bordel noir dans la communauté. Je veux une fille, vingt-cinq-trente

[4] Commissioner Jules Maigret from *Le Port des brumes* by Georges Simenon (1932).
[5] Commissioner Jules Maigret from *Liberty Bar* by Georges Simenon (1932).

ans, pas mal mais crevée. Sale.[6]

b. The face of Commander Verhoeven changes abruptly - no more joking around. He slams the phone violently onto the metal table. - "And now you're **fucking everything up** in the community. I want a girl, twenty-five to thirty years old, decent-looking but worn out. Dirty."

From the vocabulary we identify a detective that is more part of the action, for example the verbs *enfiler* (to put on) and *croiser* (to run into) are distincitve of Cluster C.

(4) a. Camille comprend, se lève, **enfile** son manteau, prend son chapeau et sort. Au passage, il **croise** Armand.[7]

b. Camille understands, gets up, **puts on** his coat, takes his hat, and goes out. On the way, he **runs into** Armand.

6 Discussion

6.1 Literary Interpretation of Results

Our diachronic prediction across nearly 150 years of francophone detective fiction demonstrates the remarkable strength and durability of the detective archetype as the linchpin of the narrative. From its earliest affirmation, the detective has remained the central figure around whom the story pivots, resisting the vicissitudes of time even as its expressive form evolves. Although the core investigatory function (unraveling mysteries through observation and inference) remains intact, we observe a progressive enrichment of the archetype. What began as a cold, efficient "reasoning machine" gradually acquires a fuller emotional envelope and a deeper engagement with the world "on the ground," reflecting changing narrative priorities and reader expectations.

The bottom-up clustering analysis yields three distinct poles that correspond to major historical phases, each delineated by a significant temporal shift. The clarity of this tripartition might not have been obvious from traditional literary scholarship alone, yet it emerges naturally from the data: an early rational-puzzle phase, a mid-century emotionally engaged phase, and a contemporary hard-edged phase characterized by moral ambiguity. These three clusters make sense not only as statistical artifacts but also as literary realities, each shaped by broader socio-cultural and aesthetic transformations in French crime writing. These three poles are not merely artifacts of chronological periodization but resonate deeply with literary function and reader positioning. In the rational-puzzle phase, the reader is an intellectual spectator; in the empathic-procedural phase, an accomplice to the detective's intuition; and in the hardboiled-moral-ambiguous phase, a witness to the detective's struggle to impose order on a chaotic society. This development reflects broader socio-literary transformations in French detective fiction, transitioning from narratives of pure detection to novels of criminality, victimhood, and ultimately violence [37]. As the genre evolved to mirror shifting societal attitudes, values, and political realities, the detective archetype itself necessarily adapted, navigating toward representations of social violence and moral ambiguity.

6.2 Methodological Limitations and Interpretative Challenges

One limitation of our study is its reliance on the off-the-shelf BookNLP-fr pipeline, via its Propp extension, for character identification, coreference resolution, and attribute extrac-

[6] Commander Verhoeven from *Alex* by Pierre Lemaitre (2011).
[7] Camille from *Alex* by Pierre Lemaitre (2011).

tion. Though designed for long-form French literature, the pipeline introduces uncertainty, as our analysis is bound by its performance and error patterns. This affects coreference accuracy in particular: the system may mistakenly merge distinct characters or split one into several chains. Such errors introduce noise, potentially distorting character representations. While we thoroughly assessed our classification models and attribute relevance, the dataset's scale prevents full manual validation of coreference chains. Consequently, some detective profiles may include attributes from unrelated characters, impacting both archetype identification and its diachronic analysis.

Another limitation lies in our attribute extraction method, which relies on lemmatized unigrams. While this reduces vocabulary size and facilitates analysis, it also introduces key drawbacks. The approach does not account for negation (for instance, treating "intelligent" and "not intelligent" as equivalent) which can invert or alter meaning. It also overlooks gradation (e.g., "very clever" vs. "somewhat clever") and the broader context in which attributes appear. Furthermore, polysemous terms are treated in isolation, without disambiguation, which risks conflating distinct meanings and misrepresenting characters. Addressing these issues would require more advanced linguistic processing, such as incorporating multi-word expressions, explicitly handling negation, and using context-aware embeddings or disambiguation models to improve the precision and nuance of character profiling.

7 Conclusion

This paper has proposed a novel method for tracking character archetypes within literary genres over time. We have applied this approach to the figure of the detective, a particularly well-suited case for such an investigation. Our results show a remarkable degree of stability in the textual signature associated with the detective archetype. An analysis of the linguistic features most strongly associated with the detective role confirms the persistence of a core investigative lexicon, comprising characteristic modifiers and verbs across the corpus. Yet, a more detailed examination of model outputs reveal both the figure's permeability and historical contingency. Newer subgenres introduce additional elements, including subjectivity, cynicism, and existential doubt. This evolution is consistent with theoretical perspectives that frame the detective not as a fixed type, but as a cultural site for negotiating the tensions between reason, intuition, and moral ambiguity in changing historical contexts.

Future research may benefit from moving beyond binary classifications (Detective vs. Non-detective) to explore a spectrum of detectiveness, as well as its interaction with other character roles. Such an approach could further elucidate the detective's dual status as both a narrative anchor and a locus of transformation within the genre. In addition, this methodology opens new avenues for examining archetypal characters across other genres. It enables a dual perspective: not only can we track the defining traits of central characters, but we can also investigate points of similarity and overlap among secondary figures. By integrating *distant reading* techniques (such as large-scale feature extraction and clustering) with *close reading* of selected examples, we aim to identify salient features, assess their interpretability, and situate them within broader literary traditions. We anticipate that further applications of this method across diverse genres will offer valuable insights into character typology and contribute to the refinement (or reassessment) of existing literary theories.

Acknowledgements

This research was funded in part by PRAIRIE-PSAI (Paris Artificial intelligence Research institute-Paris School of Artificial Intelligence, reference ANR-22-CMAS-0007). This work has received support under the Major Research Program of PSL Research University "CultureLab" launched by PSL Research University and implemented by ANR with the references ANR-10-IDEX-0001.

References

[1] Bamman, David. "BookNLP". 2021. URL: https://github.com/booknlp/booknlp.

[2] Bamman, David, Lewke, Olivia, and Mansoor, Anya. "An Annotated Dataset of Coreference in English Literature". English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, May 2020, pp. 44–54. ISBN: 979-10-95546-34-4.

[3] Bamman, David, Underwood, Ted, and Smith, Noah A. "A Bayesian Mixed Effects Model of Literary Character". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 370–379. DOI: 10.3115/v1/P14-1035. (Visited on 04/10/2023).

[4] Barré, Jean. "Latent Structures of Intertextuality in French Fiction". In: *Proceedings of the Computational Humanities Research Conference 2024*. CHR 2024. Aarhus, 2024, pp. 21–26. URL: https://ceur-ws.org/Vol-3834/paper97.pdf.

[5] Barré, Jean, Cabrera Ramírez, Pedro, Mélanie, Frédérique, and Galleron, Ioanna. "Pour une détection automatique de l'espace textuel des personnages romanesques". In: *Humanistica 2023*. Corpus. Association francophone des humanités numériques. Genève, Switzerland, June 2023. URL: https://hal.science/hal-04105537.

[6] Barré, Jean, Camps, Jean-Baptiste, and Poibeau, Thierry. "Operationalizing Canonicity: A Quantitative Study of French 19th and 20th Century Literature". In: *Journal of Cultural Analytics 8*, no. 3 (2023). DOI: 10.22148/001c.88113.

[7] Becker, Lucille F. *Georges Simenon*. eng. Twayne's world authors series France 456. Boston, Mass: Twayne, 1977. ISBN: 978-0-8057-6293-8.

[8] Bonniot, Roger. *Émile Gaboriau ou La naissance du roman policier*. fre. Paris: Vrin, 1985. ISBN: 978-2-7116-9277-4.

[9] Bork, G.J. van et al. "Algemeen Letterkundig Lexicon". In: Den Haag: Digitale Bibliotheek voor de Nederlandse Letteren (DBNL), 2012. Chap. Detectiveroman, n.p. URL: https://www.dbnl.org/tekst/dela012alge01_01/dela012alge01_01_00775.php.

[10] Bourgois, Antoine and Poibeau, Thierry. "The Elephant in the Coreference Room: Resolving Coreference in Full-Length French Fiction Works". 2025. arXiv: 2510.15594 [cs.CL].

[11] Brownson, Charles. *The figure of the detective: a literary history and analysis*. eng. Jefferson: McFarland, 2014. ISBN: 978-1-4766-1272-0.

[12] Cadin, Anne. *Le moment américain du roman français: 1945-1950*. fre. PhD thesis. Paris: Classiques Garnier, 2018. ISBN: 9782406077510.

[13] Cawelti, John G. *Adventure, mystery, and romance: formula stories as art and popular culture*. eng. pbk. ed., [Nachdr.] Chicago, Ill.: The Univ. of Chicago Press, 1997. ISBN: 978-0-226-09867-8.

[14] Dubois, Jacques. "Naissance du récit policier". fre. In: *Actes de la Recherche en Sciences Sociales 60*, no. 1 (1985), pp. 47–55. DOI: 10.3406/arss.1985.2287. (Visited on 05/16/2025).

[15] Ehrmanntraut, Anton, Konle, Leonard, and Jannidis, Fotis. "LLpro: A Literary Language Processing Pipeline for German Narrative Texts". In: *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, ed. by Munir Georges, Aaricia Herygers, Annemarie Friedrich, and Benjamin Roth. Ingolstadt, Germany: Association for Computational Lingustics, Sept. 2023, pp. 28–39. URL: https://aclanthology.org/2023.konvens-main.3/.

[16] Ellison, Murray S. *Edgar Allan Poe and Science: Unraveling the Plot of the Universe*. Available at https://scholarscompass.vcu.edu/cgi/viewcontent.cgi?article=5047&context=etd. Master's thesis. Virginia Commonwealth University, 2015.

[17] Gerson, Noel B. *The Vidocq dossier: the story of the world's first detective*. eng. Boston: Houghton Mifflin, 1977. ISBN: 978-0-395-25176-8.

[18] Konle, Leonard, Hilger, Agnes, and Jannidis, Fotis. "On Character Perception and Plot Structure of German Romance Novels". English. In: *Proceedings of the Computational Humanities Research Conference (CHR 2023)*. Vol. 3518. CEUR Workshop Proceedings, Dec. 2023, pp. 592–605.

[19] Konle, Leonard and Jannidis, Fotis. "Modeling Plots of Narrative Texts as Temporal Graphs". English. In: *Proceedings of the Computational Humanities Research Conference (CHR 2022)*. Vol. 3304. CEUR Workshop Proceedings, Dec. 2022, pp. 318–330.

[20] Lavergne, Elsa de. *La naissance du roman policier français: du Second Empire à la Première Guerre mondiale*. Études de littérature des XXe et XXIe siècles 7. OCLC: ocn436637927. Paris: Classiques Garnier, 2009. ISBN: 978-2-8124-0028-5.

[21] Leblond, Aude. "Corpus Chapitres". Version v1.0.0. 2022. DOI: 10.5281/zenodo.7446728.

[22] Lycett, Andrew. *The man who created Sherlock Holmes: the life and times of Sir Arthur Conan Doyle*. 1st Free Press hardcover ed. New York: Free Press, 2007. ISBN: 978-0-7432-7523-1.

[23] Martin, Louis, Muller, Benjamin, Suárez, Pedro Javier Ortiz, Dupont, Yoann, Romary, Laurent, La Clergerie, Éric Villemonte de, Seddah, Djamé, and Sagot, Benoît. "CamemBERT: a tasty French language model". In: *arXiv preprint arXiv:1911.03894* (2019).

[24] Martinetti, Anne. *Le Masque: histoire d'une collection*. Références 3. Amiens: Encrage, 1997. ISBN: 978-2-906389-82-3.

[25] McInnes, Leland, Healy, John, and Melville, James. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". 2020. arXiv: 1802.03426 [stat.ML].

[26] Mélanie-Becquet, Frédérique, Barré, Jean, Seminck, Olga, Plancq, Clément, Naguib, Marco, Pastor, Martial, and Poibeau, Thierry. "BookNLP-fr, the French Versant of BookNLP. A Tailored Pipeline for 19th and 20th Century French Literature". In: *Journal of Computational Literary Studies 3* (1 Nov. 2024), pp. 1–34. ISSN: 2940-1348. DOI: 10.48694/jcls.3924.

[27] Messac, Régis. *Le detective novel et l'influence de la pensée scientifique*. fre. Travaux 55. Amiens: Encrage, 2011. ISBN: 978-2-251-74246-5.

[28] Monroe, Burt L., Colaresi, Michael P., and Quinn, Kevin M. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict". In: *Political Analysis 16*, no. 4 (2017), pp. 372–403. DOI: 10.1093/pan/mpn018.

[29] Moretti, Franco. "Conjectures on world literature". In: *New left review 2*, no. 1 (2000), pp. 54–68.

[30] Moretti, Franco. *Distant Reading*. London: Verso, 2013.

[31] Naguib, Marco, Delaborde, Marine, Andrault, Blandine, Bekolo, Anaïs, and Seminck, Olga. "Romanciers et romancières du XIXème siècle: une étude automatique du genre sur le corpus GIRLS (Male and female novelists: an automatic study of gender of authors and their characters)". In: *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier TAL et Humanités Numériques (TAL-HN)*. 2022, pp. 66–77.

[32] Pandzic, Maja. "E. A. Poe and A. A. Shkljarevskij: Foregrounding Deduction and/or Social Commentary – A Comparative Study of Early Detective Fiction". In: *[sic] - a journal of literature, culture and literary translation* , no. 1.11 (Dec. 2020). DOI: 10.15291/sic/1.11.lc.5.

[33] Pedregosa, F. et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research 12* (2011), pp. 2825–2830.

[34] Schutt, Sita. "Rivalry and Influence: Nineteenth-Century French Detective Narratives". In: *The Art of Murder*, ed. by Gustav Klaus and Stephen Knight. 1998, pp. 38–49.

[35] Symons, Julian. *Bloody murder: from the detective story to the crime novel*. eng. 3., rev. ed. New York: Mysterious Pr, 1993. ISBN: 978-0-89296-496-3.

[36] Todorov, Tzvetan. "Typologie du roman policier". fre. In: *Poétique de la prose*. Points Littérature 120. Paris: Éd. du Seuil, 1980. ISBN: 978-2-02-005693-9.

[37] Tourteau, Jean-Jacques. *D'Arsène Lupin à San-Antonio: Le roman policier français de 1900 à 1970*. Français. FeniXX réédition numérique, Jan. 1970.

[38] Underwood, Ted. *Distant horizons: digital evidence and literary change*. University of Chicago Press, 2019.

[39] Underwood, Ted, Bamman, David, and Lee, Sabrina. "The Transformation of Gender in English-Language Fiction". en. In: *Journal of Cultural Analytics 3*, no. 2 (Feb. 2018). DOI: 10.22148/16.019. (Visited on 10/06/2022).

[40] Vareille, Jean-Claude. "Roman policier archaïque et aventure archaïque". fr. In: *L'Aventure dans la littérature populaire au xixe siècle*, ed. by Roger Bellet. Littérature & idéologies. Code: L'Aventure dans la littérature populaire au xixe siècle. Lyon: Presses universitaires de Lyon, 1985, pp. 185–196. ISBN: 978-2-7297-0999-0. DOI: 10.4000/books.pul.1232. (Visited on 05/14/2025).

[41] Vianne, Laurine, Dupont, Yoann, and Barré, Jean. “Gender Bias in French Literature”. In: *Proceedings of the Computational Humanities Research Conference*, ed. by Artjoms Sela, Fotis Jannidis, and Iza Romanowska. Vol. 3558. CEUR Workshop Proceedings. 2023, pp. 247–262.

[42] Williard Huntington Wright. “The Great Detective Stories”. In: *The Art of the Mystery Story*. New York: Grosset & Dunlap, 1947, pp. 33–70.

[43] Zundert, Joris van, Cranenburgh, Andreas van, and Smeets, Roel. “Putting Dutchcoref to the Test: Character Detection and Gender Dynamics in Contemporary Dutch Novels”. In: *Computational Humanities Research Conference*. CEUR Workshop Proceedings (CEUR-WS. org). 2023, pp. 757–771.