Modèles de Langues et Intertextualité dans le Roman Policier

Promesses et Limites des Approches Computationnelles

Jean Barré

02 Juillet 2025

École Normale Supérieure - Université PSL

LaTTiCe lab

Fondements théoriques :

Intertextualité

- Kristeva : « Chaque texte est une mosaïque de citations ; chaque texte est l'absorption et la transformation d'un autre texte. »
- Barthes : « Tout texte est un intertexte ; d'autres textes y sont présents à des niveaux variables, sous des formes plus ou moins reconnaissables [...] tout texte est un tissu de citations. »

Architextualité

- L'*architextualité* : la perception générique, qui oriente et détermine la réception de l'œuvre.
- Genette :

  > *L'objet de la poétique n'est pas le texte, considéré dans sa singularité (qui est plutôt l'affaire de la critique), mais l'architexte, [...] c'est-à-dire l'ensemble des catégories générales, ou transcendantes (types de discours, modes d'énonciation, genres littéraires, etc.) dont relève chaque texte singulier.*

Les Études Littéraires Computationnelles

Apports

- Opérationnalisation de concepts de la théorie littéraire
- Corpus massifs
- Lecture Distante - Évolution de dynamiques textuelles sur le temps long.

Approches quantitatives de l'intertextualité

- **Approche stricte** : repérage explicite de citations et de références identifiables.
- **Approche large** : allusions, résonances thématiques ou stylistiques plus diffuses - Similarité textuelle.

Cas d'étude : Le Roman Policier

Approches des Genres Littéraires

- Tension entre approche formaliste *et* approche contextuelle.
- Formaliste : caractéristiques textuelles internes
- Contextuelle : les genres émergent des interactions entre auteurs, éditeurs, lecteurs, critiques, etc., et sont façonnés par leurs contextes d'énonciation socio-culturels.
- Cadre computationnel : *Perspective Modeling* (Underwood, 2019)
- Résultats : stabilité des genres, soutenant les approches transcendantales et formalistes.

Le Roman Policier

Un genre à forte identité

- Structure narrative spécifique : « un récit consacré principalement à la découverte méthodique et progressive, par des moyens rationnels, des circonstances exactes d'un crime mystérieux » (Messac, 1929).
- Longue tradition éditoriale : apparition progressive dès la fin du XIX^{e} siècle, issue du roman-feuilleton (notamment Gaboriau) jusqu'à la « Collection du Masque » (Pigasse, 1927).
- Corpus abondant et clairement défini

Classification des romans policiers

Pipeline en trois étapes

- Extraction des caractéristiques textuelles.
- Modélisation par apprentissage automatique – SVM.
- Analyse des erreurs et interprétation des résultats.

Classification automatisée du roman policier

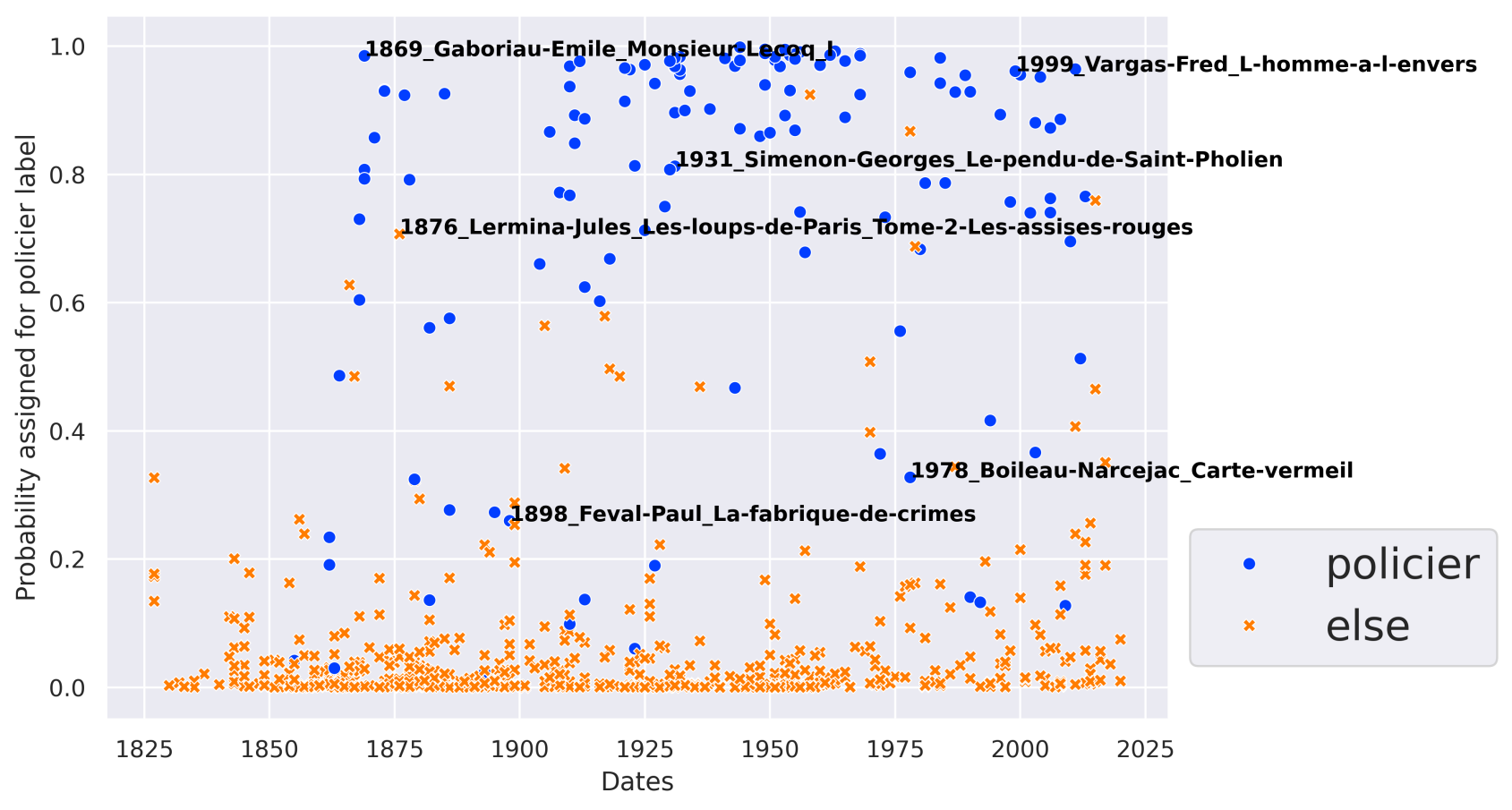


Figure 1 – Classification automatique du roman policier francophone

Caractéristiques discriminantes

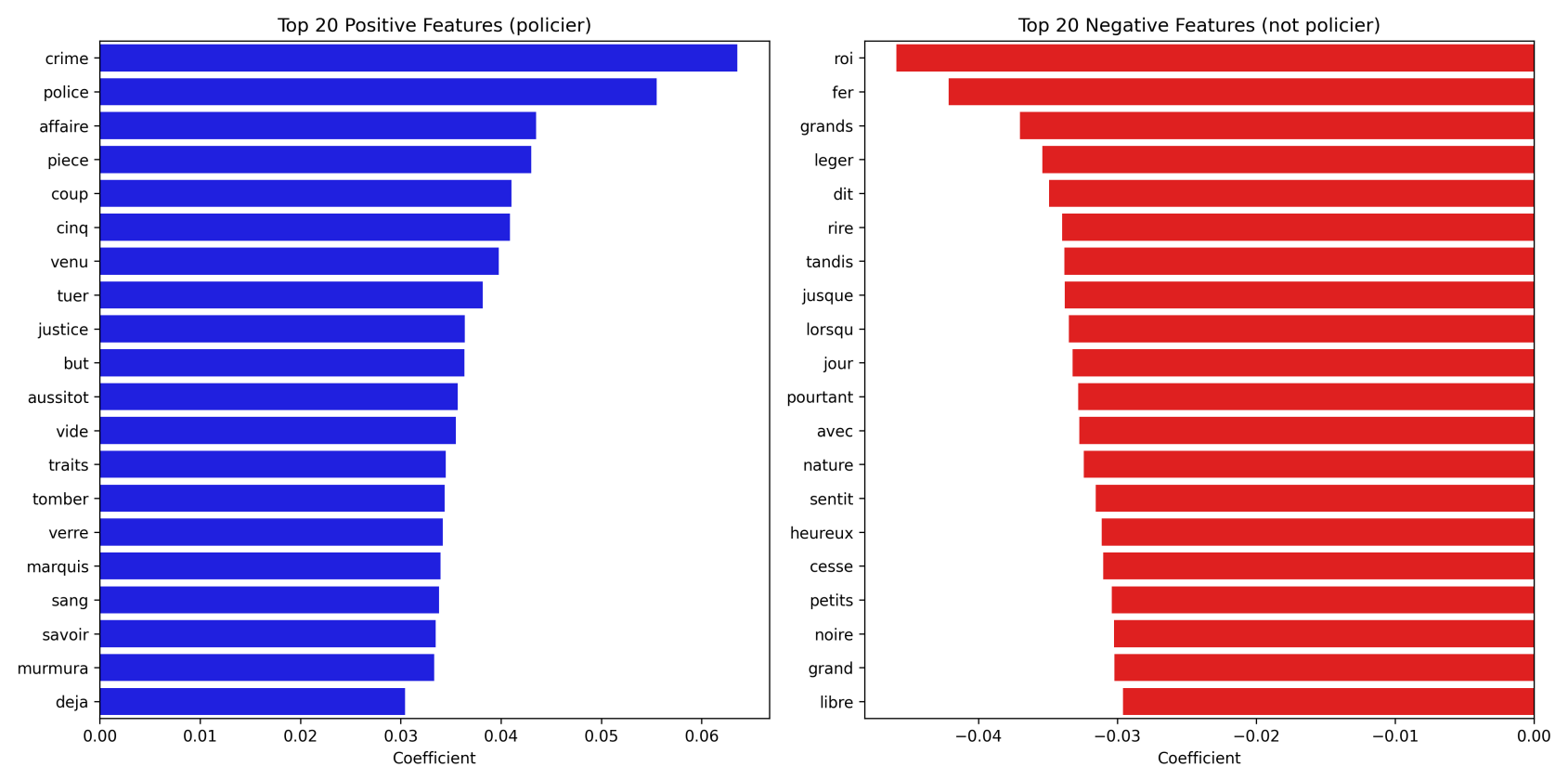


Figure 2 – Caractéristiques discriminantes pour la prédiction du roman policier

Problèmes : Émergence Tardive du Terme

- Le terme *roman policier* apparaît historiquement très tard. Le genre n'existe pas encore (1870–1927).

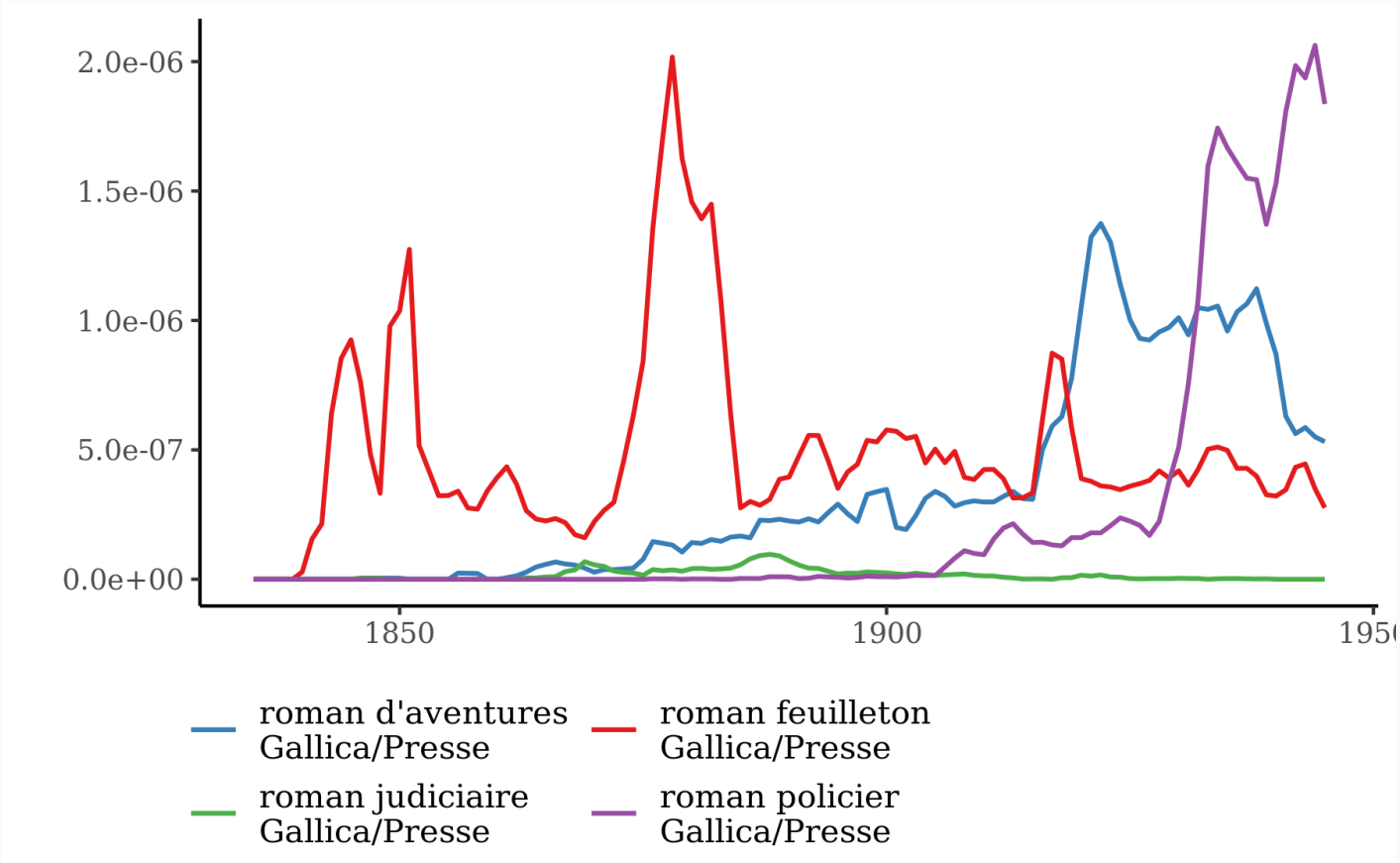


Figure 3 – Gallicagram (de Courson & Azoulay, 2021)

Limitation : Évolution de la Prédiction de Genre

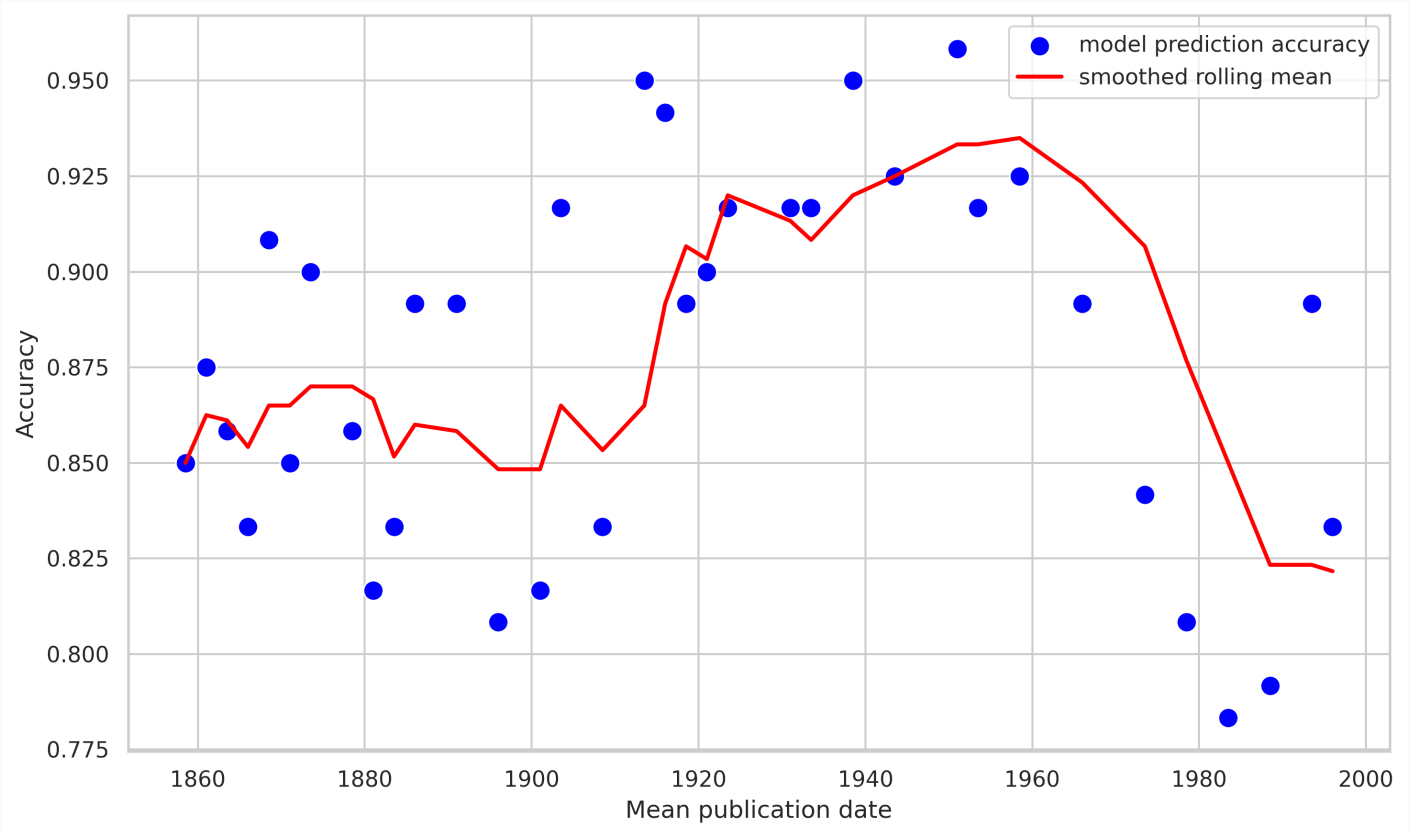


Figure 4 – Précision de la prédiction de genre tous les 25 ans

Approche Cumulative des Débuts du Genre

- Répétition et accumulation des pratiques discursives qui rendent possible le genre.
- John Rieder :

> *Étudier les débuts du genre ne consiste pas à trouver des points d'origine, mais à observer une accumulation de répétitions, d'échos, d'imitations, d'identifications et de distinctions qui témoignent d'une prise de conscience émergente d'un réseau conventionnel de ressemblances. (Rieder, 2012)*

Approche Cumulative des Débuts du Genre

- Répétition et accumulation des pratiques discursives qui rendent possible le genre.
- John Rieder :

> *Étudier les débuts du genre ne consiste pas à trouver des points d'origine, mais à observer une accumulation de répétitions, d'échos, d'imitations, d'identifications et de distinctions qui témoignent d'une prise de conscience émergente d'un réseau conventionnel de ressemblances. (Rieder, 2012)*

Est-il possible de réaliser le constat quantitatif de l'accumulation de ressemblance lors de la naissance du roman policier ?

Opérationnaliser l'intertextualité – Une Tentative

Objectif : Construire un réseau de ressemblances et observer son évolution

Embeddings et Similarité Cosine comme Indicateurs d'Intertextualité

- Modèle d'encodage contextuel de pointe pour la représentation des textes littéraires français.
- Affinage du modèle BGE-M3-Embedding sur la langue littéraire française.
- **Query** : un paragraphe ; **Positif** : les 5 paragraphes suivants ; **Négatif** : 5 paragraphes aléatoires.
- **Hypothèse** : les embeddings capturent des notions de style individuel ainsi que la similarité thématique.

Validation du Modèle

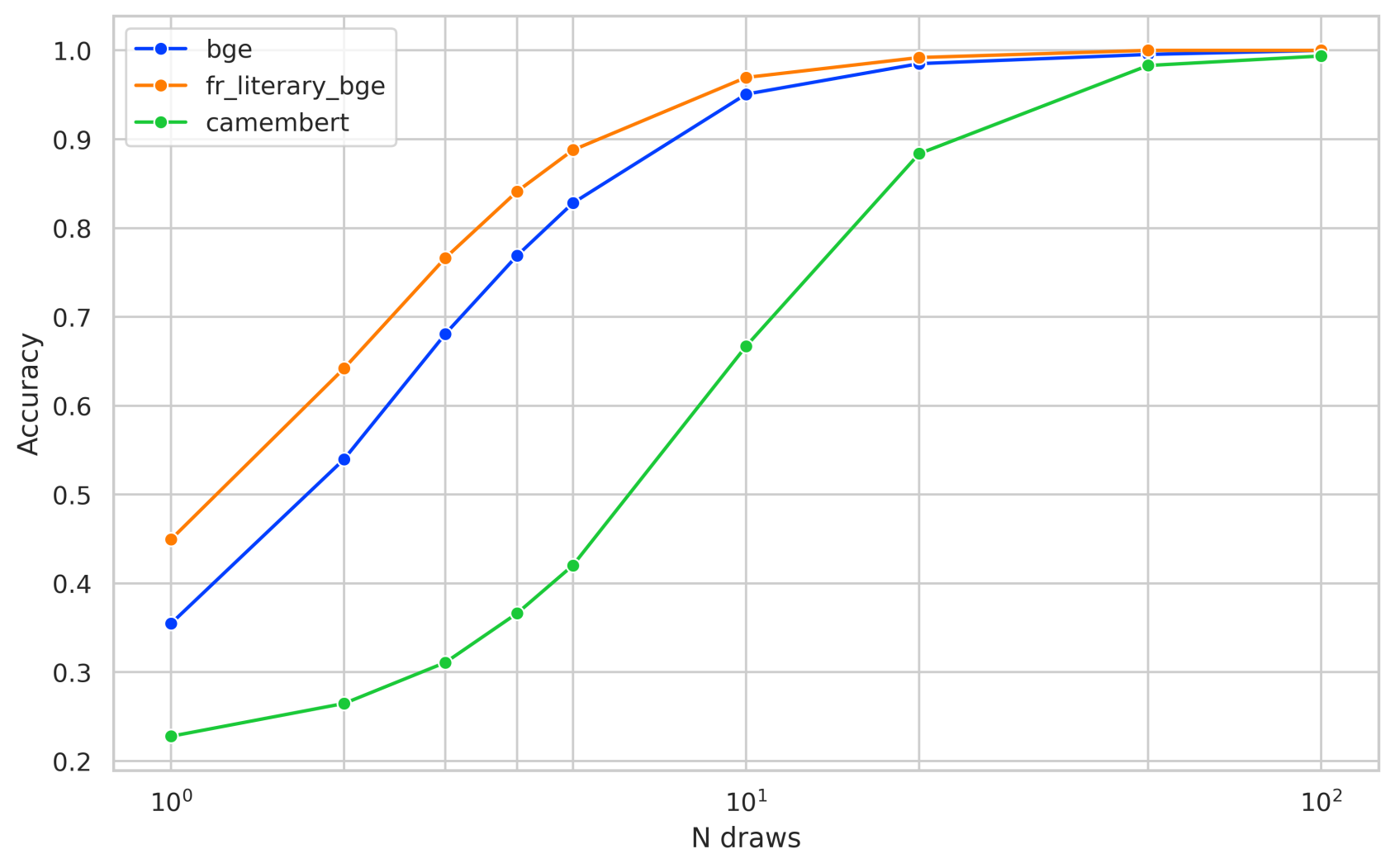


Figure 5 – Évaluation de l'encodeur

Réseau de Similarité – Échelle du Roman

Version HTML

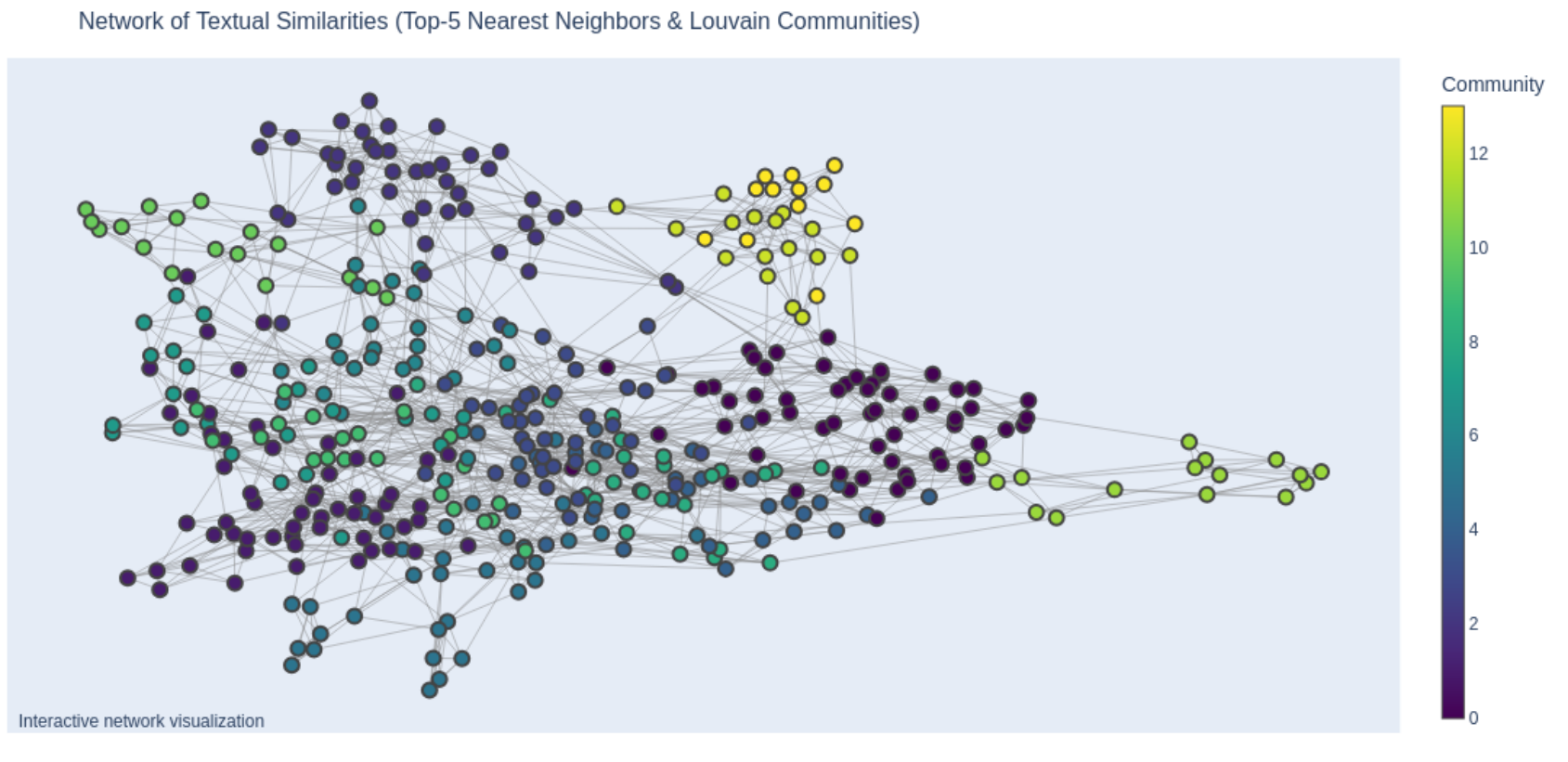


Figure 6 – Réseau de similarité pour *Le Mystère de la chambre jaune* (Leroux, 1908)

Réseau de Similarité – Échelle du Roman

Version HTML

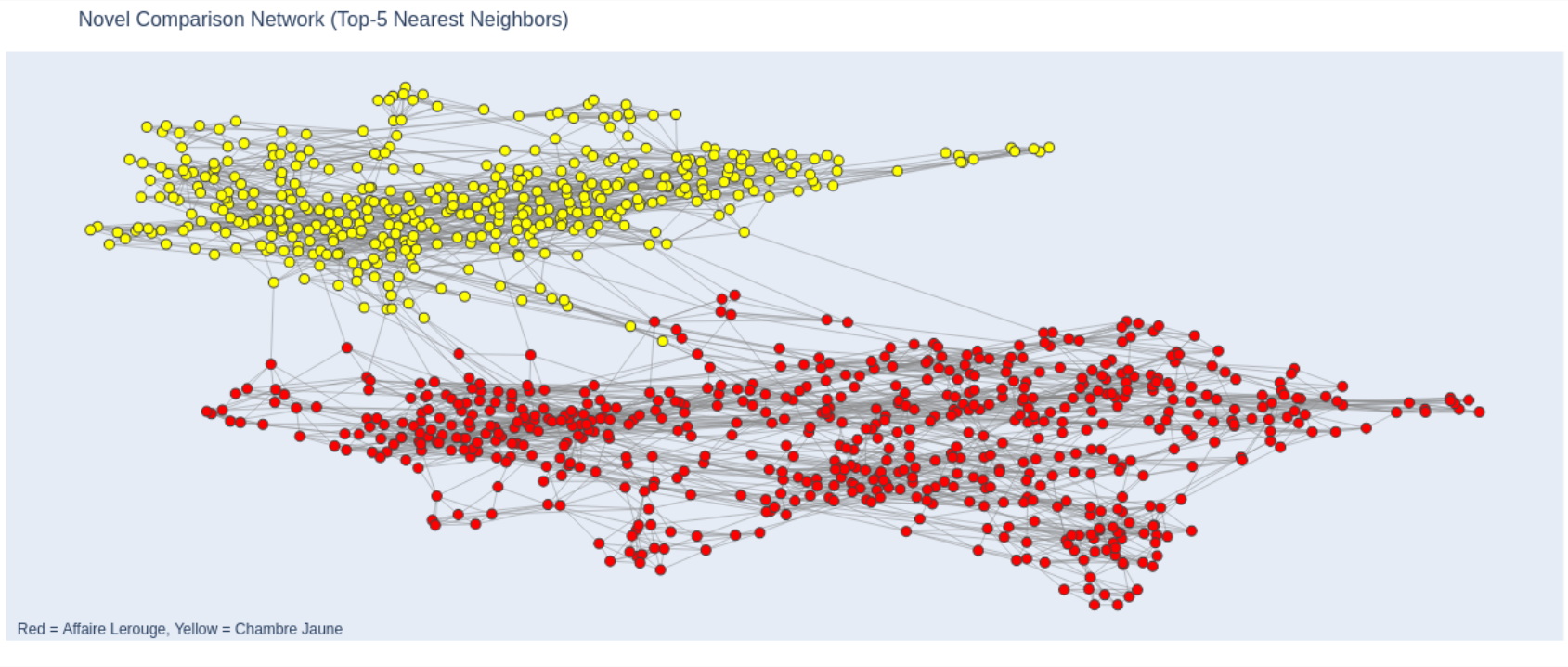


Figure 7 – Réseau de similarité pour *Le Mystère de la chambre jaune* (Leroux, 1908) et L'Affaire Lerouge (Gaboriau, 1866)

Exemple de Similarité – Extrait 374 de l'Affaire Lerouge (Gaboriau, 1866)

Ah ! c' est trop fort ! dites -vous . Eh bien ! daignez jeter un regard sur ces morceaux de plâtre humide . Ils vous représentent les talons des bottes de l' assassin dont j' ai trouvé le moule d' une netteté magnifique près du fossé où on a aperçu la clé . Sur ces feuilles de papier j' ai calqué l' empreinte entière du pied que je ne pouvais relever ; car elle se trouve sur du sable . " Regardez : talon haut , cambrure prononcée , semelle petite et étroite , chaussure d' élégant à pied soigné , bien évidemment . Cherchez - la , cette empreinte , tout le long du chemin , vous la trouverez..."

Exemple de Similarité – Extrait 330 du Mystère de la Chambre Jaune (Leroux, 1907)

”Voilà, dit-il, les souliers que chaussait l’assassin ! Les reconnaissez-vous, père Jacques ?” Le père Jacques se pencha sur ce cuir infect et, tout stupéfait, reconnut de vieilles chaussures à lui qu’il avait jetées il y avait déjà un certain temps au rebut, dans un coin du grenier ; il était tellement troublé qu’il dut se moucher pour dissimuler son émotion.

Exemple de Similarité – Extrait 37 de l'Affaire Lerouge (Gaboriau, 1866)

— Que voulez-vous dire ? interrompit le juge.
— Rien d'autre que ce que je dis, monsieur.
— Ainsi vous persistez à nier ?
— Je suis innocent.
— Mais c'est de la folie…
— Je suis innocent.
— C'est bien, fit M. Daburon, pour aujourd'hui en voilà assez. Vous allez entendre la lecture du procès-verbal et on vous reconduira au secret. Je vous exhorte à réfléchir. La nuit vous inspirera peut-être un bon mouvement ; si le désir de me parler vous venait, quelle que soit l'heure, envoyez-moi chercher, je viendrai. Des ordres seront donnés. Lisez, Constant.

Exemple de Similarité – Extrait 165 du Mystère de la Chambre Jaune (Leroux, 1907)

Seul, Frédéric Larsan avait une figure rayonnante et montrait la joie d'un chien de chasse qui s'est enfin emparé de sa proie. M. de Marquet dit, montrant à M. Darzac le jeune employé à la barbiche blonde :
" Vous reconnaissez monsieur ?
— Je le reconnais, fit Robert Darzac d'une voix qu'il essayait en vain de rendre ferme. C'est un employé de l'Orléans à la station d'Épinay-sur-Orge.
— Ce jeune homme, continua M. de Marquet, affirme qu'il vous a vu descendre de chemin de fer, à Épinay...
— Cette nuit, termina M. Darzac, à dix heures et demie... c'est vrai ! Mais je suis innocent !"

Ablations sur le texte : pour comprendre le modèle

Quels éléments du texte sont encodés dans le modèle ?

- On enlève différentes catégories de mots dans les textes :
  - **Noms propres** : pour tester l'importance des personnages ou lieux.
  - **Mots thématiques** : pour voir si le modèle s'appuie sur le sujet du texte.
  - **Stopwords** (mots outils) : pour évaluer le rôle du signal auctorial.
  - **Ponctuation** : pour vérifier leur impact sur la forme.
  - **Ordre des mots** (par mélange) : pour tester la sensibilité à la syntaxe.
- On mesure ensuite comment la similarité entre paragraphes change selon l'ablation.

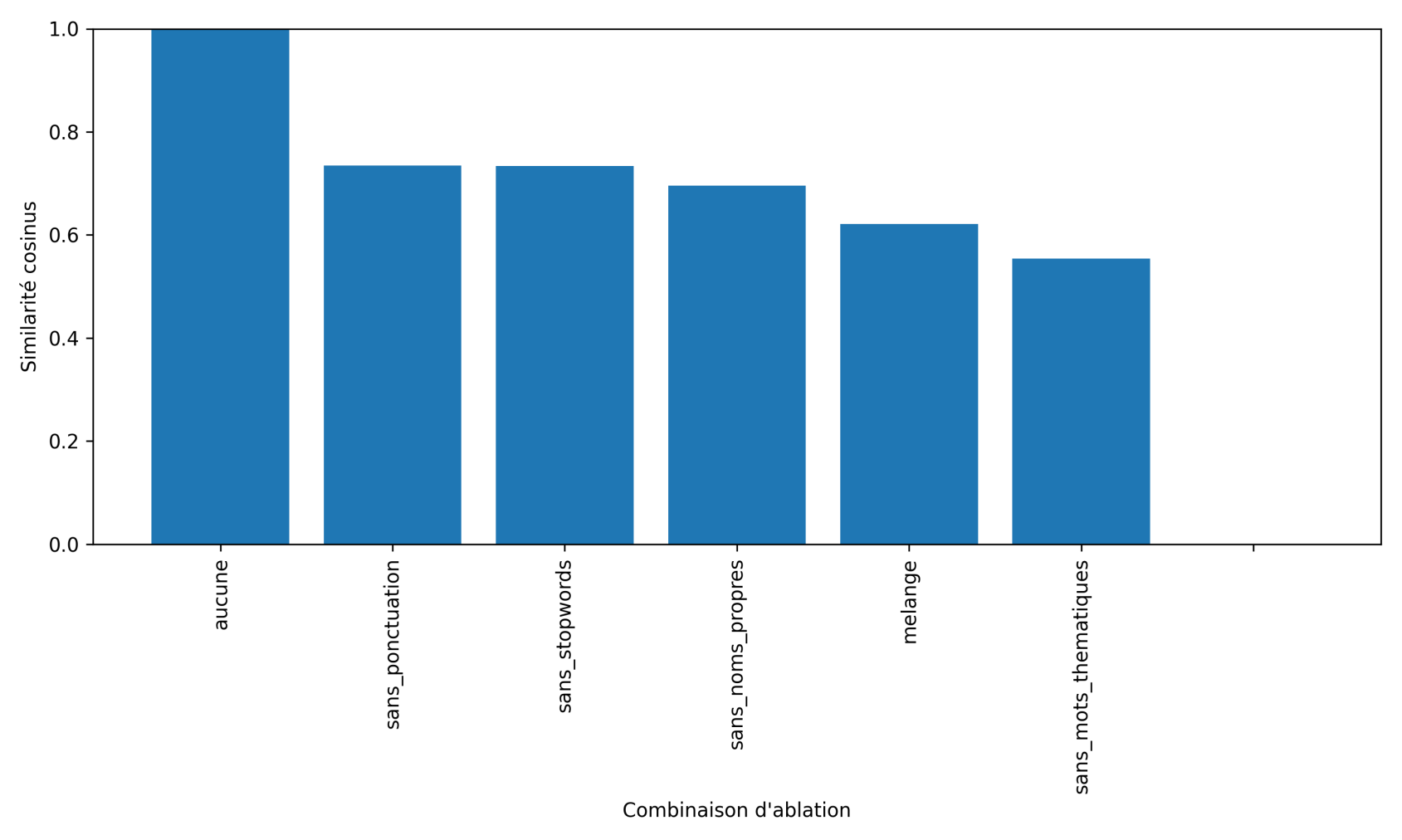
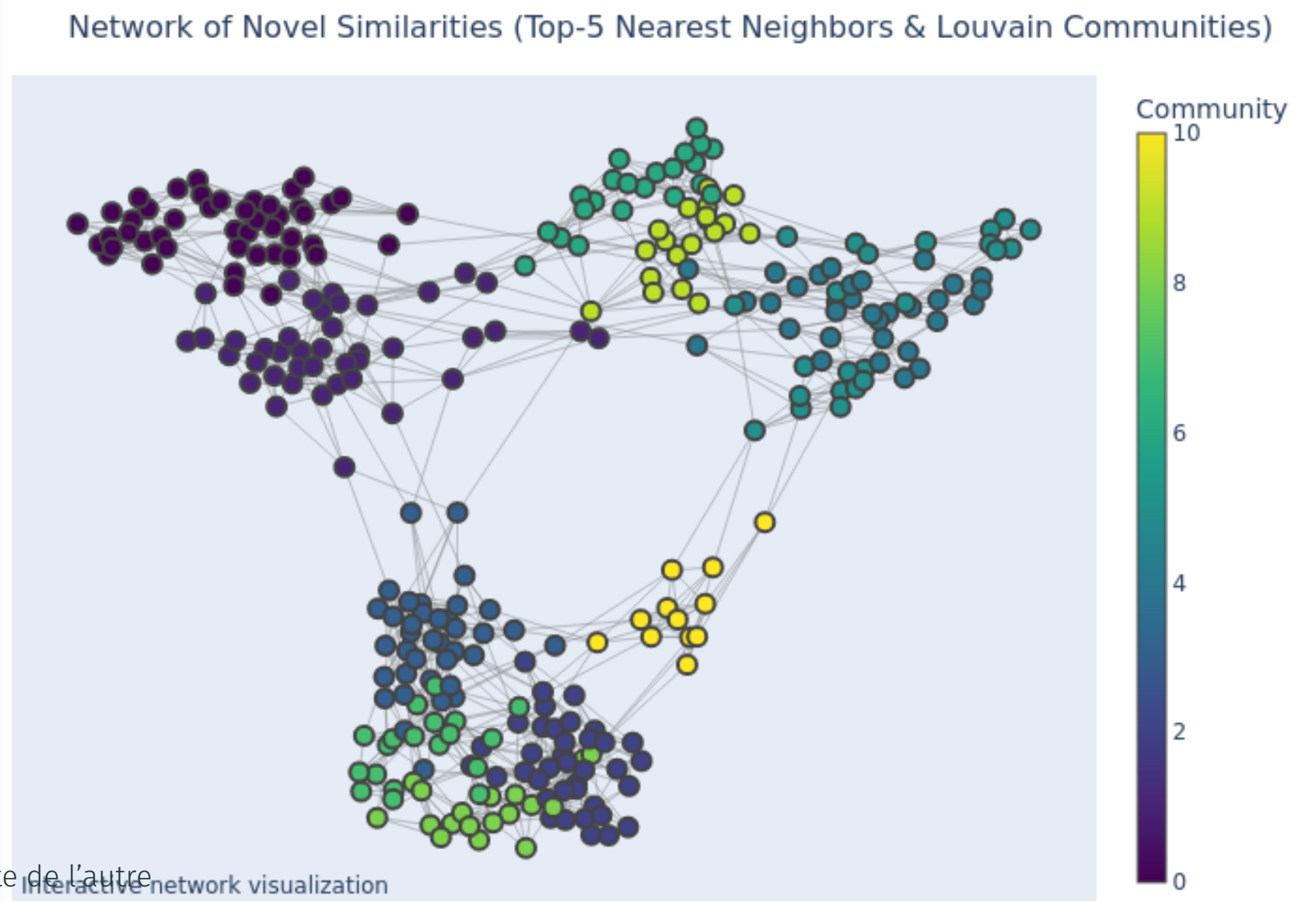


Figure 8 – Résultats des différentes ablations

Réseau de Similarité – Échelle du Genre

Version HTML



Réseau de Similarité – Échelle du Corpus

Version HTML

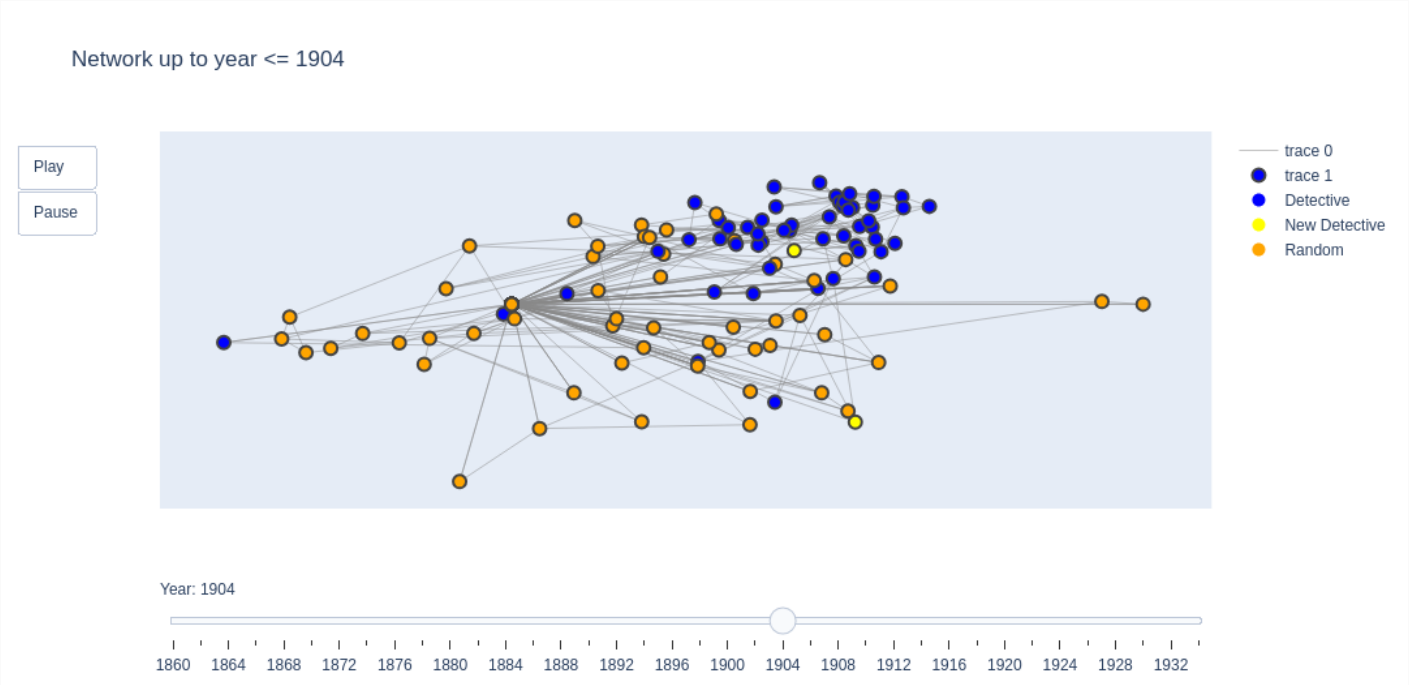


Figure 10 – Approche cumulative du réseau de similarité pour le corpus de romans policiers

Tendances individuelles

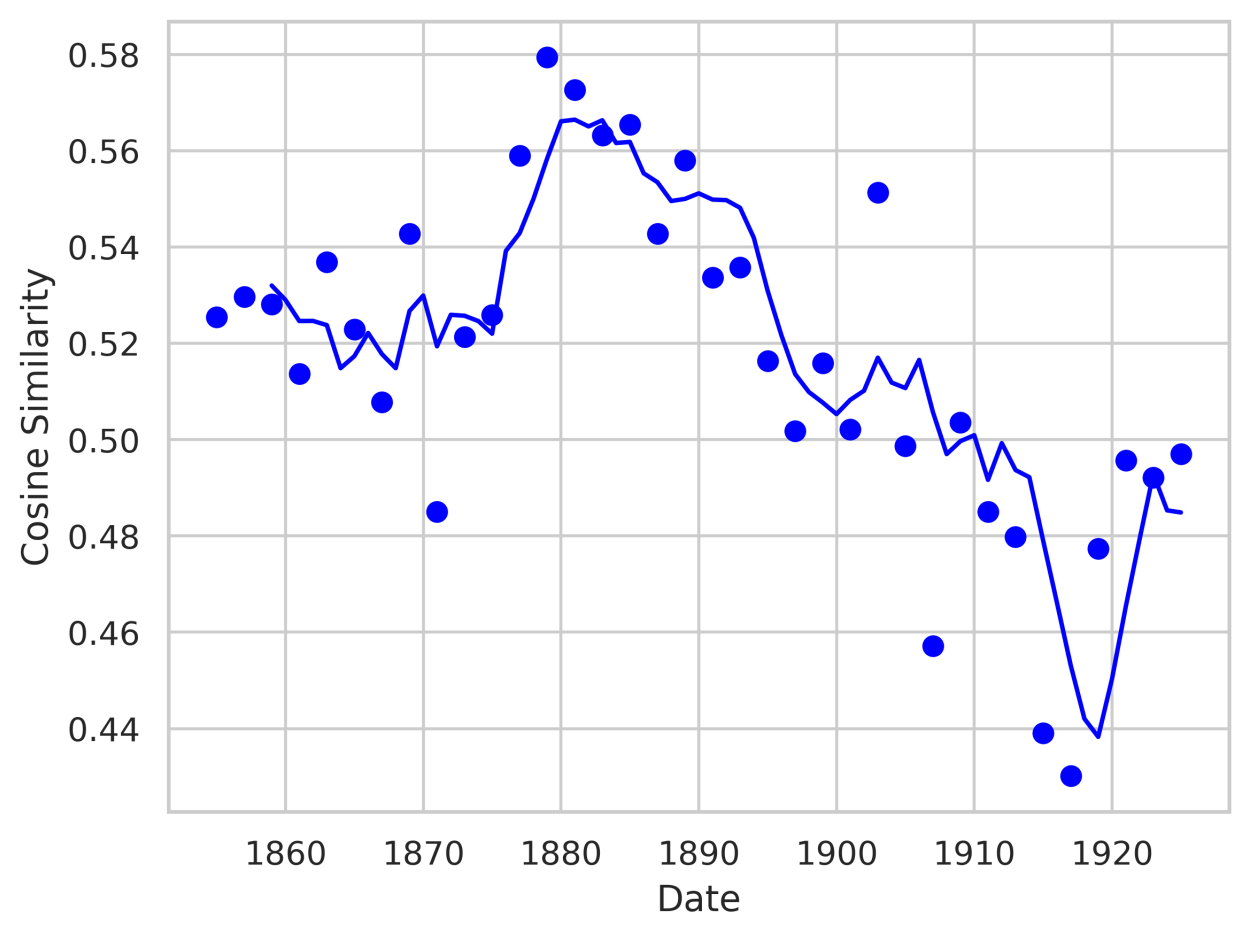


Figure 11 – Similarité cosinus dans le temps pour *L'affaire Lerouge* (Gaboriau, 1866)

Tendances individuelles

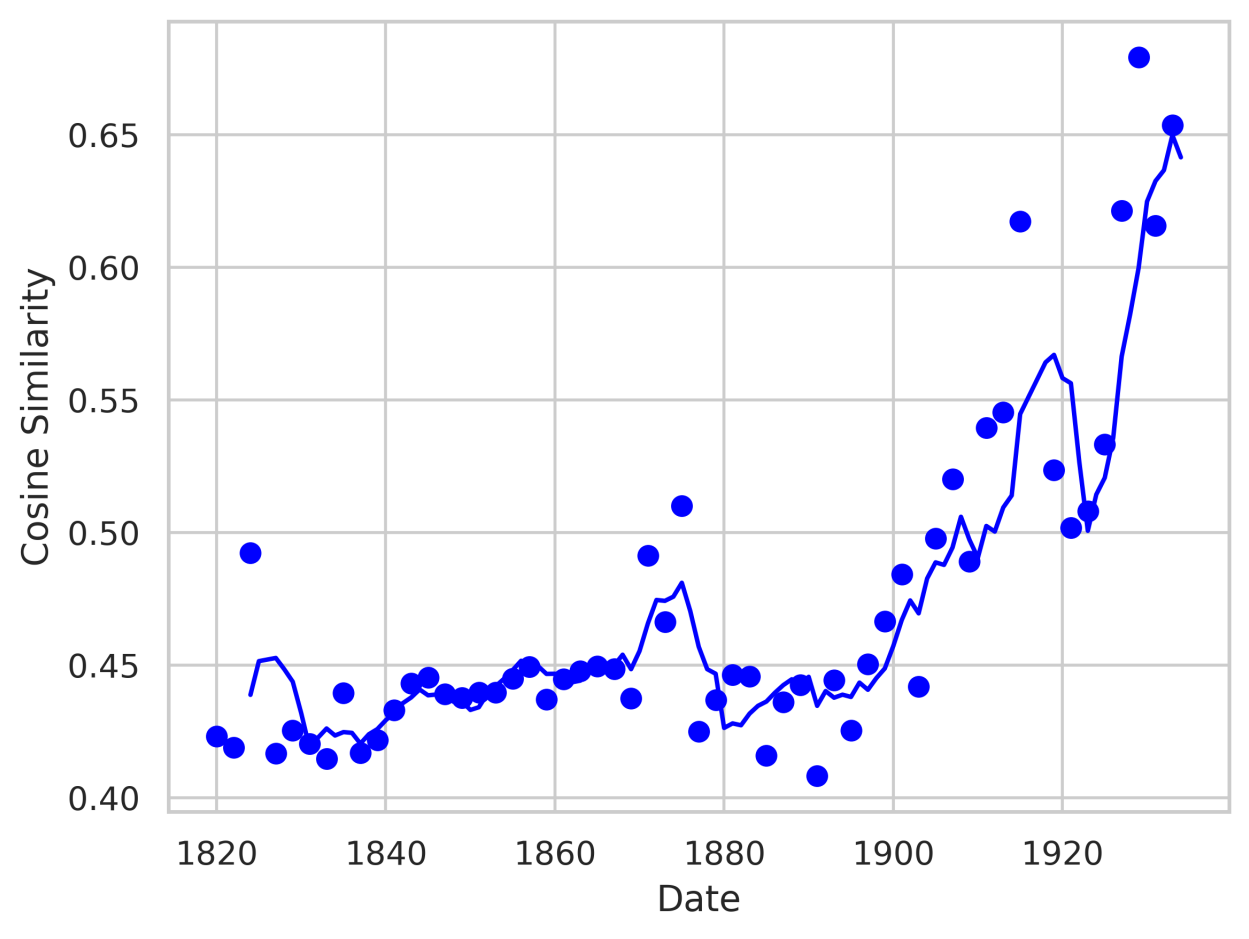


Figure 12 – Similarité cosinus dans le temps pour *Le mystère de la chambre jaune* (Leroux, 1907)

Perspectives

- Naissance du roman policier par accrétion du réseau de ressemblance
- Similarité textuelle quantitative == intertextualité ?
- L'intertextualité comme cadre théorique pour les études littéraires computationnelles
- Travail en lecture proche à poursuivre

Merci !

jean.barre@ens.psl.eu - jbarre.bsky.social

Données & Code :
https://github.com/crazyjeannot/CHR_latent_structures
Modèle :
https://huggingface.co/crazyjeannot/fr_literary_bge_base

La Naissance du Roman Policier (Lavergne, 2009)

Le roman policier se distingue du roman-feuilleton en trois points :

1. **Persistance de la thématique criminelle** : plus cohérente et centrale tout au long du récit.
2. **Centralité du détective** : passage d'un rôle secondaire à un rôle principal, moteur de la résolution de l'énigme.
3. **Structure narrative centrée sur l'enquête** : récit organisé autour de la progression de l'investigation, des indices connus vers les inconnues, ponctué de fausses pistes et de rebondissements.

Limitation : Modélisation de l'Évolution du Genre

Setup Expérimental

- Objectif : Mesurer la capacité des modèles à reconnaître le genre policier à travers trois périodes historiques.
- **Périodes distinguées** :
  1. **Période émergente** (1860–1927) :
     - Roman policier feuilleton.
     - Gaboriau, Leroux, Leblanc.
  2. **Période canonique** (1927–1945) :
     - Stabilisation des codes narratifs.
     - Collection du Masque, Simenon.
  3. **Période moderne** (après 1945) :
     - Importation du *hardboiled* américain.
     - Roman noir, polar.
- Méthode : entraînement d'un modèle SVM sur chaque période, puis test croisé sur les deux autres.

Résultats et Analyse

	émergente	canonique	moderne
émergente	id	0.675	0.550
canonique	0.929	id	0.832
moderne	0.689	0.710	id

Table 1 – Précision des modèles entraînés sur chaque période

Analyse des Résultats :

- Le modèle **canonique** est le plus stable :
  - 0,929 sur la période émergente.
  - 0,832 sur la période moderne.
- Le modèle **émergent** est peu généralisable :
  - 0,675 sur la période canonique.
  - 0,550 sur la période moderne.
- Le modèle **moderne** obtient des performances intermédiaires.

Tendances Collectives : Similarité dans le temps

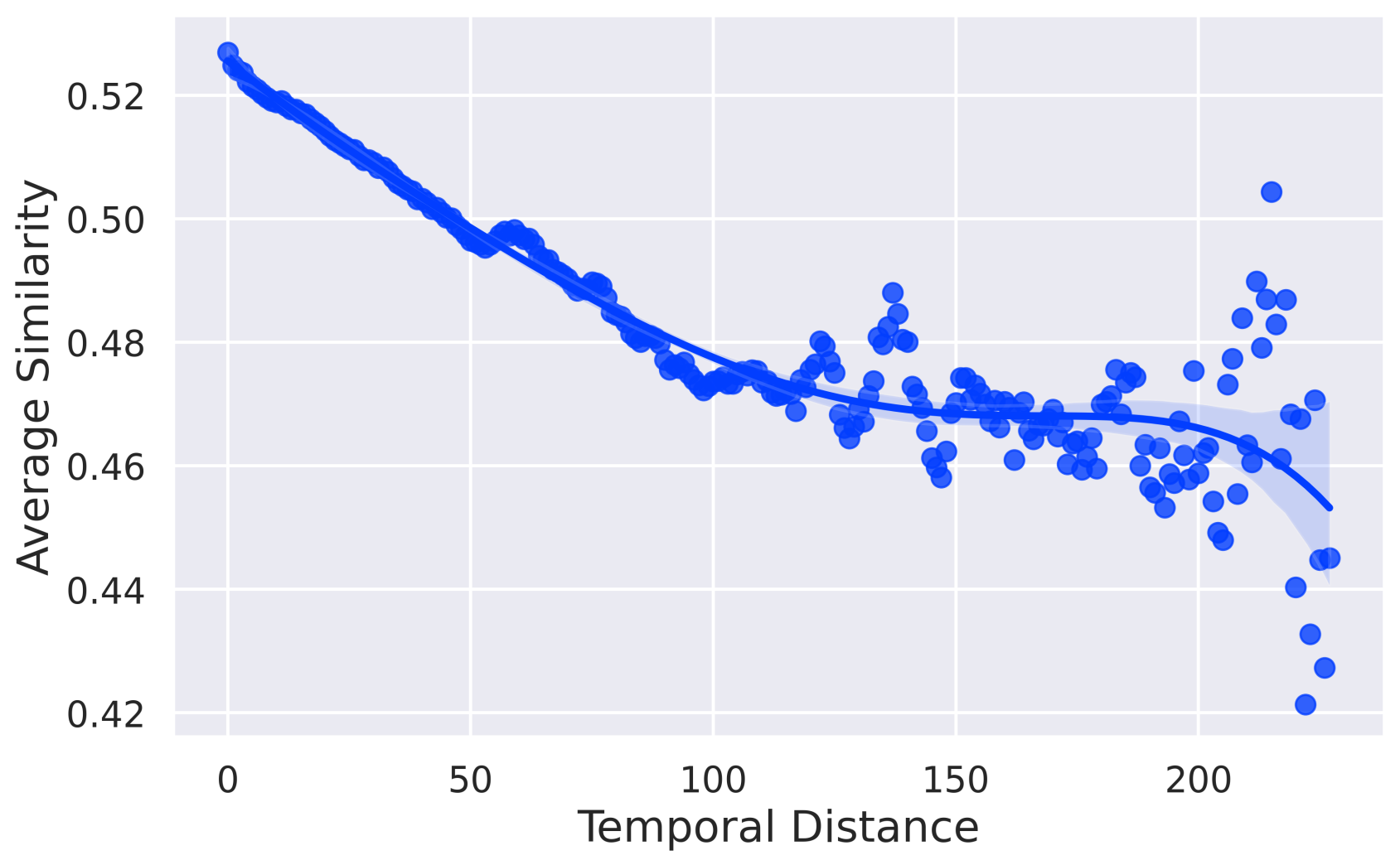


Figure 13 – Tendances de similarité temporelle dans le corpus

Tendances Collectives : Similarité dans le temps

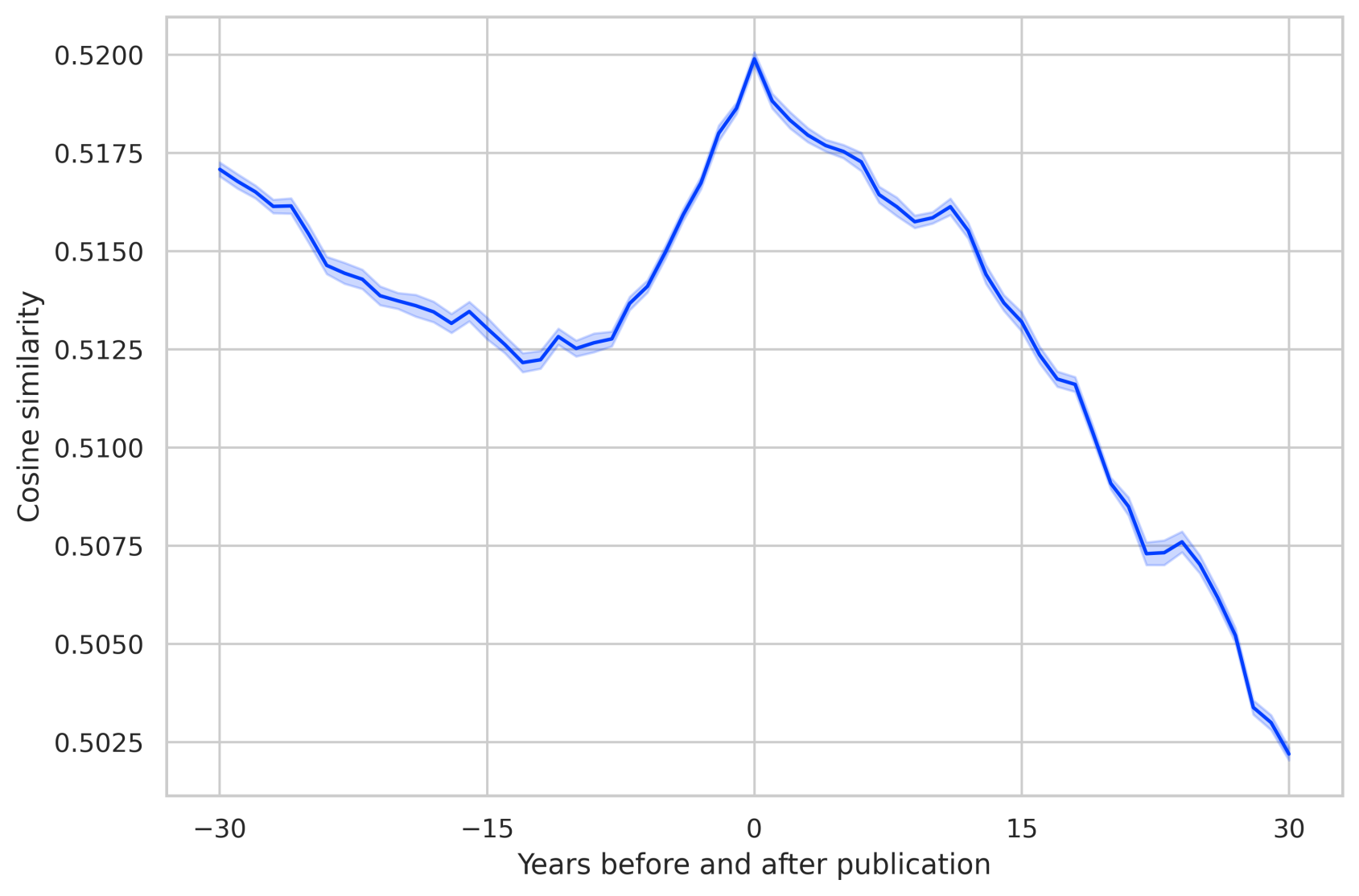


Figure 14 – Tendances de similarité temporelle dans le corpus

Résultats : Canon vs. Archive

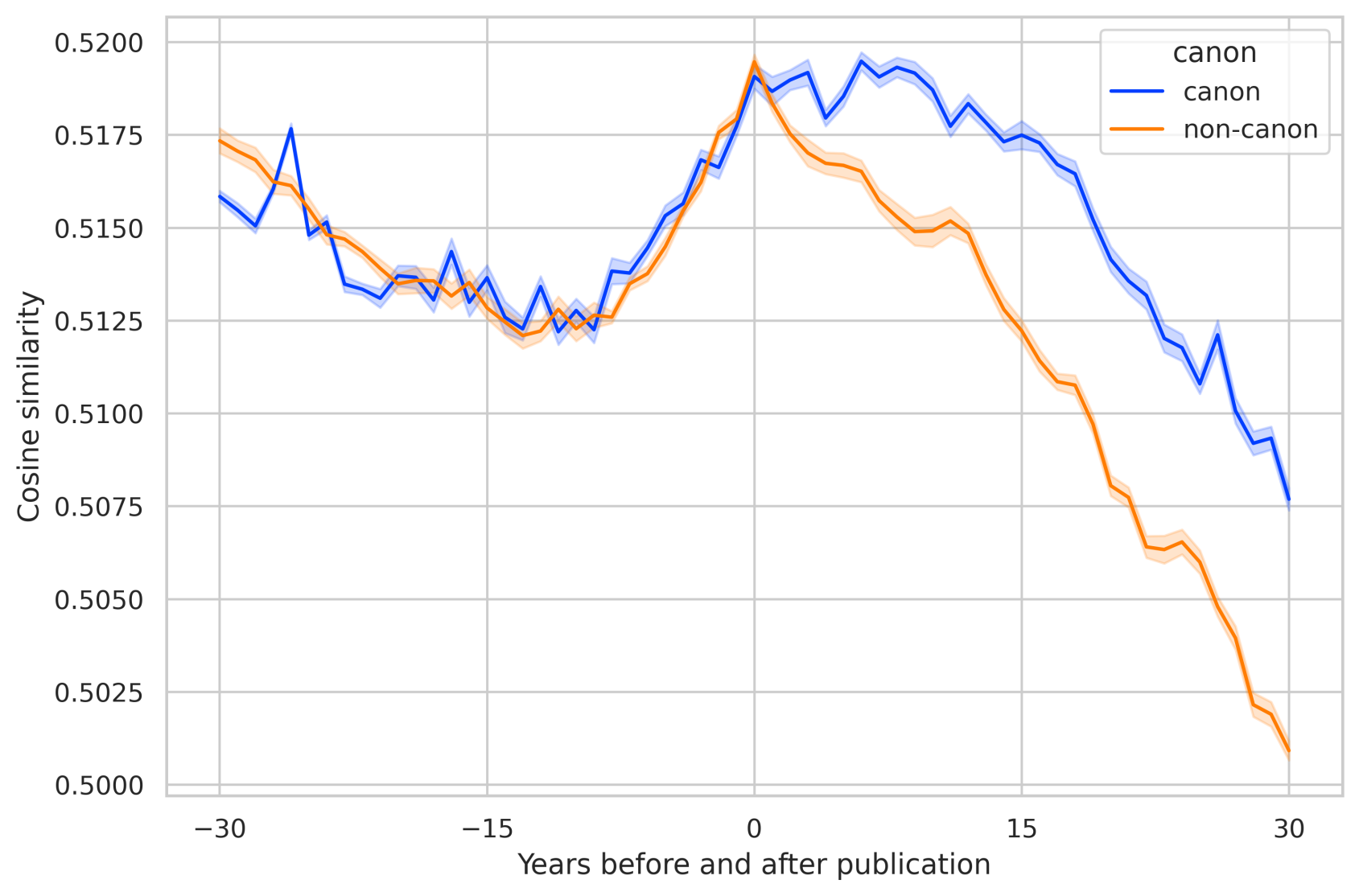


Figure 15 – Impact du canon sur la similarité textuelle au fil du temps

Résultats : Dynamique des genres

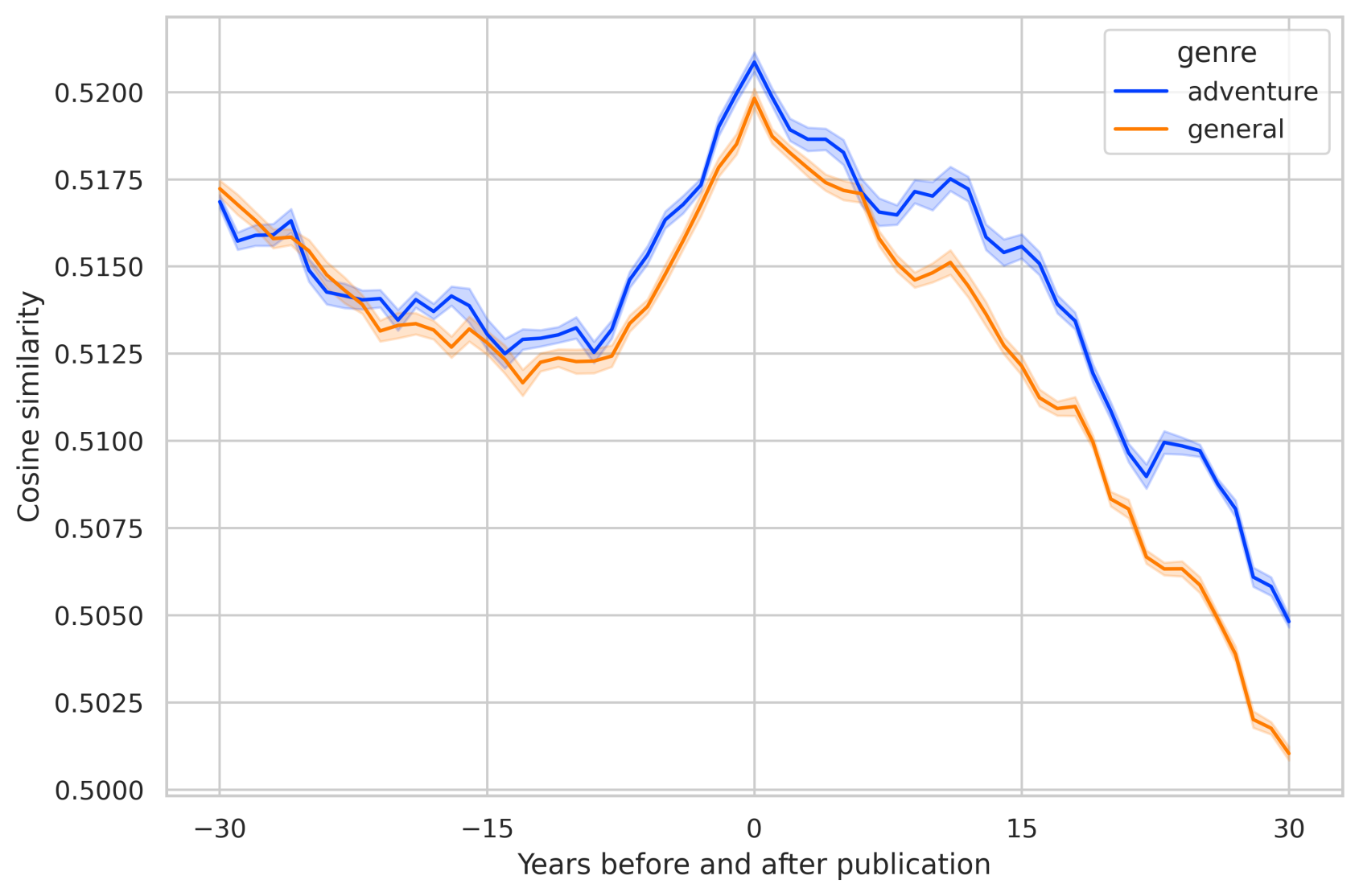


Figure 16 – Impact des genres sur la similarité textuelle au fil du temps