Computational Cultural Science Workshop

Paris, 18th-19th May 2026

This booklet contains the timetable and the abstracts of all the talks and presentations (posters, lightning talks, keynote talks and long papers) given at the Computational Cultural Science (C^2S) workshop. For more details on the workshop, please visit `https://c2s.sciencesconf.org`

This template originates from `LaTeXTemplates.com` and is based on the original version at: `https://github.com/maximelucas/AMCOS_booklet`

Contents

Introduction **4**

Timetable **5**
Monday 18th May 2026 5
Tuesday 19th May 2026 6

Abstracts - Monday 18th May 2026 **7**
Keynote Lecture I 7
Long Papers I 8
Poster Session I 10
Lightning talks I: Modelling Cultural Artefacts 44
Lightning Talks II: Manuscripts and Stylometry 49

Abstracts - Tuesday 19th May 2026 **57**
Long Papers II 57
Poster Session II [Students of the Master of Digital Humanities (ENC-PSL)] 59
Long Papers III 68
Keynote Lecture II 71

Useful Information **72**
How to get to the workshop? 72

Acknowledgments **73**

Introduction

We are excited to welcome you to the first edition of the Computational Cultural Science Workshop. Like the PSL Major Program CultureLab inaugural event which took place last September, this workshop aims to bring together scholars applying computational methods to the study of cultural artefacts. We are particularly committed to encouraging the next generation of researchers and are proud to host a poster session showcasing the work of students of the Master of Digital Humanities (École nationale des Chartes).

We are thrilled by the enthusiastic response the workshop has received and have assembled a wide-ranging program which, we hope, reflects the diversity of existing approaches and the various challenges related to the automatic processing of cultural datasets. This desire to represent different strands of computational cultural studies also guided us while choosing our two keynote speakers, Giles Bergel and Simon Carrignon, whom we warmly thank for accepting our invitation.

We are, of course, extremely grateful to all the authors who submitted lightning talks, posters and long papers to this workshop. We would like also to extend our sincere thanks to all our reviewers for their valuable feedback. Finally, we would like to thank Humanistica, the francophone association for digital humanities, for reaching out to us and agreeing to include an English-speaking workshop among the satellite events of its main conference.

The workshop organizers

Timetable

LP: Long Paper, KL: Keynote Lecture, LT: Lightning Talk.

Monday 18th May 2026

10:00–10:30	**Welcome with coffee and croissants**		
10:30–11:00	**Opening Speeches**		
11:00–12:00	KL	**Giles Bergel** (University of Oxford)	The Wandering Jew's Chronicle and the Phylogenetics of English Print Culture and Oral Tradition
12:00–14:00	**Lunch Break and Poster Session I**		
14:00–14:30	LP	**Hazim Baroudi, Wassim Ammar, Farid Bouchiba, Shadha Karoumi, Christian Müller, Fabrice Rossi** (Université Paris Dauphine-PSL; IRHT)	Automated Annotation via Structural Alignment for Hierarchical Classification of Classical Fiqh Texts
14:30–15:00	LP	**Vy Cao** (Université du Luxembourg)	Health, Vitality, and Trust: Embedding-Based Cross-Lingual Retrieval of Medicinal Advertisements in European and Colonial Press
15:00–15:30	LP	**Kirill Maslinsky** (INALCO)	Measuring the Effect of Censorship on Translation Flow
15:30–15:45	**Coffee break**		
15:45–16:45	LT	**Lightning talks** Session 1	Modelling Cultural Artefacts
16:45–17:45	LT	**Lightning talks** Session 2	Manuscripts and Stylometry
19:30	**Dinner (by invitation only) - Le Comptoir des Petits Champs (75001 Paris)**		

Tuesday 19th May 2026

10:00–10:30	**Welcome with coffee and croissants**		
10:00–10:30	LP	**Mark Hill** (King's College London)	Discursive Signatures: A Computational Method for Mapping How Meaning Varies Across Communities
10:30–11:00	LP	**Mathilde Ducos, Frédéric Landragin** (Université Sorbonne Nouvelle)	A NER Model for Science Fiction's Lexical Innovation
11:00–11:30	LP	**Simon Gabay, Florian Cafiero, Jean-Luc Falcone** (University of Geneva; EPITA)	A Gueuloir of One's Own: Computing the Acoustic Signature of Flaubert
12:00–14:00	**Lunch Break and Poster Session II**		
14:00–14:30	LP	**Matti La Mela, Yunting Xie** (Uppsala University)	Is There a Patent Genre? A Textual Analysis of French Patents, 1903-1940
14:30–15:00	LP	**Alexandre Lionnet-Rollin, Florian Cafiero** (EPHE; ENC-PSL)	Neutral Drift and Cumulative Memory: A Wright–Fisher Model of Thematic Evolution in Online Horror Fiction
15:00–15:30	LP	**Jean Barré** (ENS-PSL)	In Search of Lost Adventure Novels. A Two-Stage Pipeline to Retrieve Genre Literature from the Large Scale National French Library Archive
15:30–15:45	**Coffee break**		
15:45–16:45	KL	**Simon Carrignon** (University College London)	Using Modelling and Simulation to Understand Change in Past Societies across Scales
16:45–17:30	**Closing Remarks**		
19:30	**Dinner (by invitation only) - Le Mesturet (75002 Paris)**		

Abstracts - Monday 18th May 2026

Keynote Lecture I

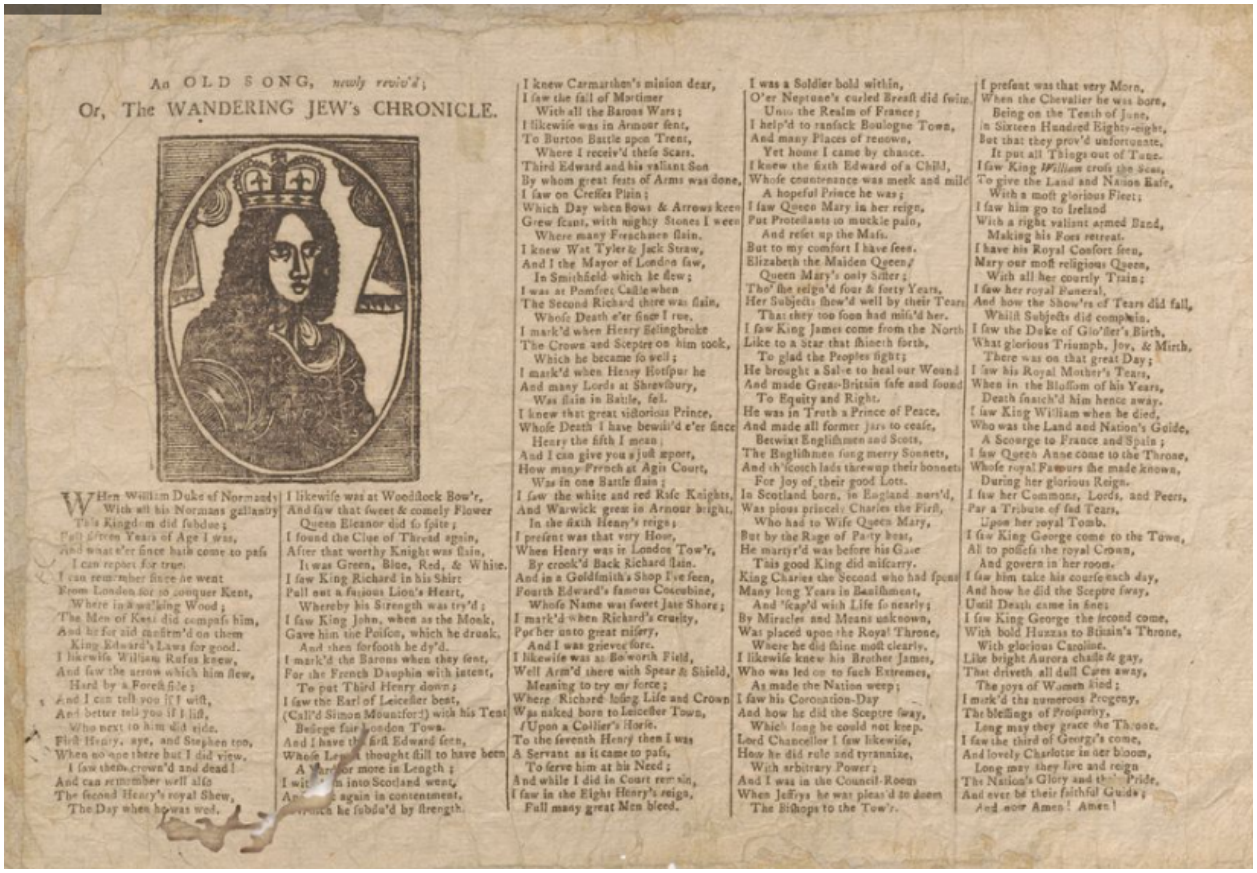
The Wandering Jew's Chronicle and the Phylogenetics of English Print Culture and Oral Tradition

Giles Bergel

KL

University of Oxford

The Wandering Jew's Chronicle, first published in England around 1634, is an appropriation of the well-known European myth of the Wandering Jew, to tell a distinctly English story - that of the succession to the throne of the kings and queens of England from William of Normandy to Charles I. Republished over two centuries, it has been more recently made available as a TEI-XML parallel-text edition, with visualisations, by the speaker. This talk will offer some reflections on how digital editing can usefully demonstrate the dynamics of cultural transmission in printed and oral traditions.



An OLD SONG, *newly revis'd*;
Or, The WANDERING JEW's CHRONICLE.

Long Papers I

Automated Annotation via Structural Alignment for Hierarchical Classification of Classical Fiqh Texts

Hazim Baroudi[1], ***Wassim Ammar***[2], ***Farid Bouchiba***[2],***Shadha Karoumi***[2],***Christian Müller***[2], ***Fabrice Rossi***[1]

LP

[1] Centre de Recherches en Mathématiques de la décision, Université Paris Dauphine-PSL
[2] Institut de recherche et d'histoire des textes, Centre National de la Recherche Scientifique

This work presents an automated pipeline for the structural organization and classification of classical Arabic legal corpora. We develop a hierarchical classification framework that segments unpunctuated legal prose into fine-grained semantic units and categorizes them according to the inherent taxonomy of Islamic jurisprudence.To address the scarcity of available segment level annotations for Islamic texts, we use the highly structured Hanafi legal compendium *Mukhtasar al-Qudrī* as a reference. Other texts are aligned to this manual using optimal transport. This process is used together with the SAT segmenter to annotate segments, using the structure of the reference text (chapters and sections) as hierarchical labels. This enables us to train a deep learning model which incorporates hierarchical constraints to ensure consistency across different levels of legal categorization. This approach provides a scalable solution for the large-scale analysis of historical jurisprudence.

Health, Vitality, and Trust: Embedding-Based Cross-Lingual Retrieval of Medicinal Advertisements in European and Colonial Press.

Vy Cao

LP

Centre for Contemporary and Digital History (C²DH),Université du Luxembourg

This paper presents a case study of medicinal advertisements published in the early 20th century across European and colonial press. It offers a critical reflection on the retrieval and evaluation of multilingual corpora within large-scale media archives, with particular attention to their validity and relevance for historical research. The study develops a reproducible workflow for embedding-based retrieval, focusing on keyword-level queries as a first stage of cross-lingual corpus construction. Retrieved materials are evaluated through a two-step protocol combining cosine similarity ranking and expert close reading, enabling a comparative assessment of semantic precision and discursive coherence across selected keywords. The paper argues that embeddings provide a powerful alternative to string-based search in multilingual and orthographically unstable corpora, while demonstrating that quantitative cluster stability is a necessary but insufficient condition for meaningful retrieval. This finding has particular implications for languages such as Vietnamese, where compound terms are prone to lexical and semantic drift in cross-lingual vector spaces.

Measuring the Effect of Censorship on Translation Flow

Kirill Maslinsky

Institut National des Langues et Civilisations Orientales (INALCO)

The intensity of translation flows between national literatures is defined not only by symbolic value, but by politics as well. State censorship is the most obvious limiting factor for the incoming translations. In this paper, I propose a method to estimate the strength of the negative effect of censorship on translation flow using bibliographic data on published translations. The foundational idea of the suggested measurement is that engagement with a foreign literary field is reflected in the publication of works by contemporary writers. Higher political pressure would mean less publications of contemporary works. I argue that *vitality coefficient* — a proportion of works by authors who were alive at the time of publication of the translation — is indicative of the strength of the political censorship. Translations of children's books into Russian in the USSR during the period of the Cold War will serve as a case study. Quantitative analysis is enabled by the availability of comprehensive bibliographic data. The effect of the censorial barriers associated with the country of provenance on the share of works by living authors in the translation flow was estimated with the help of a Bayesian generalized additive model. Results show that vitality coefficient responded to the political alliances and hostility dynamics during the period in the predicted way: higher levels of confrontation led to lower share of the contemporary works in the translation flow. This fact helps realize a side effect of the political censorship barriers that is not much talked about: by hindering the publication of current authors, censorship reinforces the position in the book market of the already well-established and familiar authors. If we assume that reprints contribute to the canonical status of a writer, censorship is then an indirect but real factor for canon consolidation.


emotion
literature
media
metaphors
troubadours
song
autofiction
multilingualism
mothers
childbirth
archeology
ina
llms
phylogentics
archives
tv
stemmatology
data
narratology
women
skyblogs

Autofiction in English: Generic and Narrative Characteristics.

Jeffrey Clapp[1], ***Cheng Tsz Chung***[2], ***Ralph Fung***[1], ***Lau Chaak Ming***[1]

[1] The Education University of Hong Kong
[2] Kyushu University

1 Introduction

1.1 Autofiction

While the term "autofiction" has been in use in French since the 1970s, it has only recently become common in English [4]. Its introduction has been contested, and more than a few have suggested that the concept is modish, vacuous, or both.

So what is autofiction in English? Why did English come to require this term, and what has it been used to describe?

Our project seeks to model an emergent "literary system" [3]—insofar as this is possible with an inchoate and disputed genre category like this one.

1.2 From Ascriptions…

We begin by compiling an "ascription corpus."

According to our codebook, "A text contains an ascription if a work is explicitly described as 'autofiction' or 'autofictional' in that text."

Sources of ascriptions (English language):

- academic books / journal abstracts
- Goodreads
- Common Crawl

The result:

- ∼12,000 ascriptions
- ∼2,800 unique English-language books
- ∼300 unique English-language books ascribed in two or more source categories

1.3 …to our Corpus

Corpus as analyzed below:

- English books published since 2000.
- 195 well-attested works.
- 153 unique authors:
  - 80 F / 67 M / 6 others;
  - 80 US / 31 UK / 42 others.

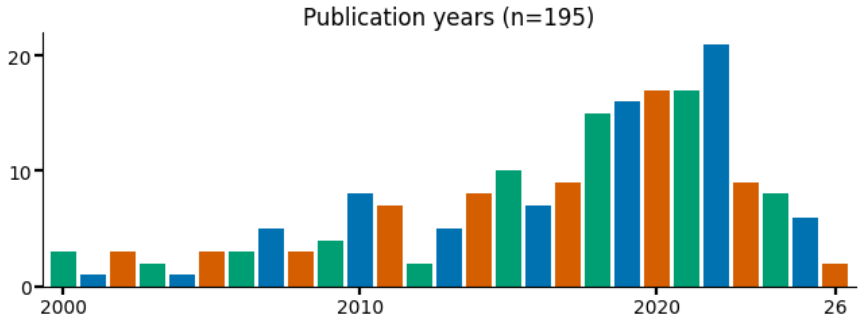


Figure 1: Distribution of corpus publication years.

2 Results

2.1 Autofiction has few events

We process our corpus via BookNLP [2], which tags text features including tokens, token supersenses, and entities, for comparability with the CONLIT corpus [7]. An immediate discovery is that autofiction is short, so it is important to normalize for length.

As Figure 2 shows, autofiction also has dramatically fewer "events" than fictional forms [8,9]. On this metric, autofiction is more like nonfiction.

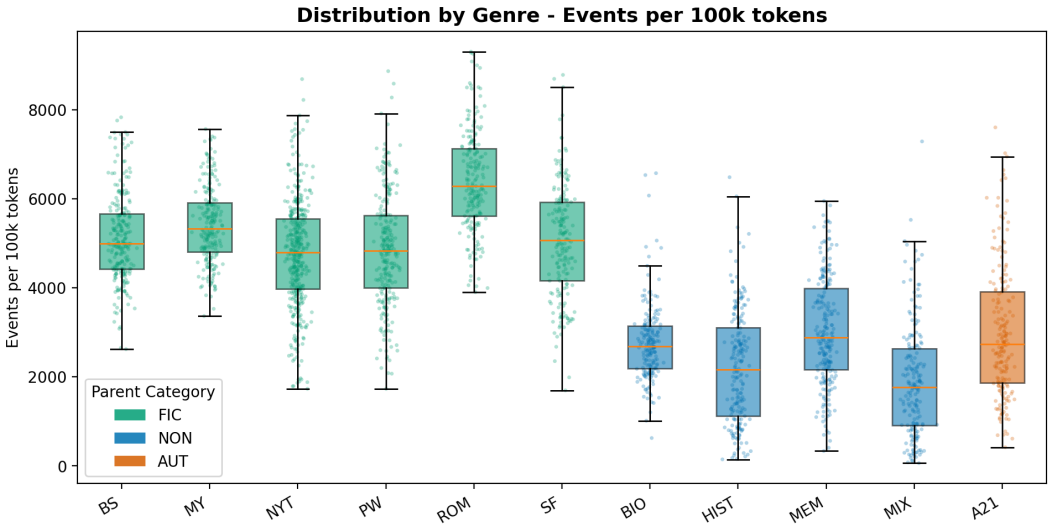


Figure 2: Events per 100k tokens by genre.

2.2 Autofiction has few characters

Fiction has fewer characters than nonfiction (Figure 3). Although autofiction has even fewer characters, on this metric autofiction is more like fiction.

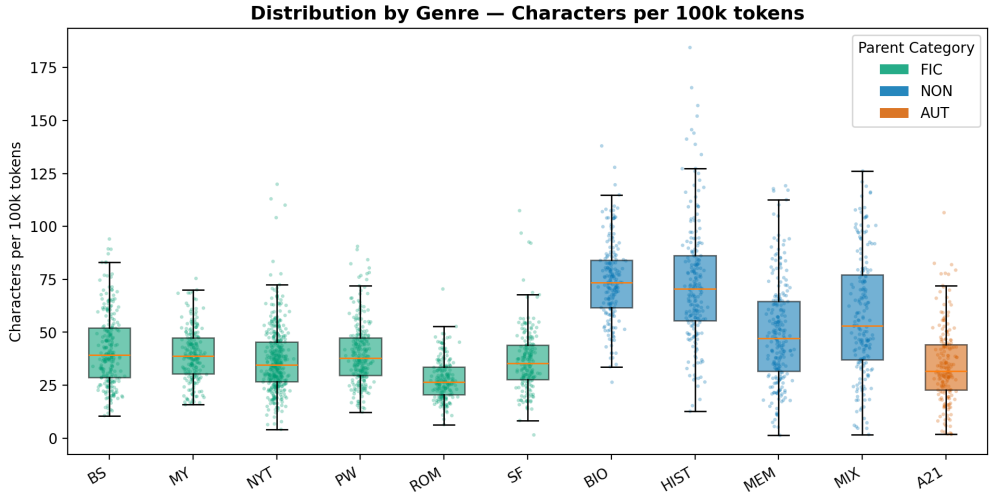


Figure 3: Characters per 100k tokens by genre.

2.3 Fiction, Nonfiction, Autofiction

Autofiction has often been understood as a "hybrid" genre. So it presents an interesting test case for commonly-employed genre classification methods [1, 6, 10, 5]. We show that by training SVMs on Wordnet supersenses, autofiction can be distinguished from nonfiction (accuracy = 89.5% on size-balanced datasets) and from fiction (82.6%).

Then, we project autofiction into the feature space defined by CONLIT fiction and nonfiction (Figure 4). The results show that autofiction clusters around the fiction-nonfiction decision boundary, while a majority of our texts (68%) lean toward fiction. The distribution of autofiction in this space, although wide, is unimodal.

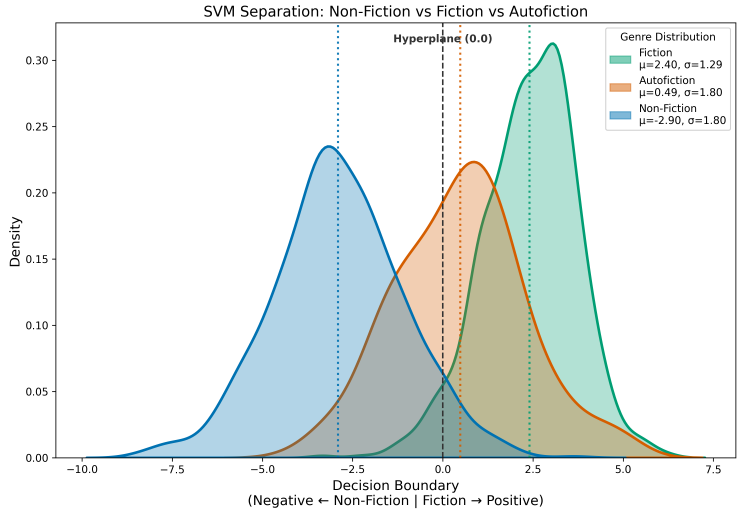


Figure 4: Autofiction projected into the fiction-nonfiction feature space.

The supersense features that most strongly distinguish fiction and nonfiction (e.g. the verbs of perception discussed by Piper [6], are not the features that distinguish autofiction from fiction. On this definition, English language autofiction is not a hybrid of fiction and nonfiction.

2.4 Autofiction and Other Genres

Against other kinds in the CONLIT corpus, autofiction is most readily distinguished from romance, and least distinguishable from prizewinning literary fiction. See Table 1.

Autofiction versus	Accuracy	Top Single Feature (41 features)	Top Leave One Out
Prizewinners	0.779	verb.stative(0.690); noun.person(0.677); verb.creation(0.669)	verb.change(+0.0203); noun.communication(+0.0078); noun.person(+0.0077)
NYT-reviewed	0.811	noun.person(0.705); verb.change(0.682); noun.Tops(0.681)	verb.change(+0.0391); noun.location(+0.0349); noun.person(+0.0229)
Memoirs	0.839	verb.social(0.818); noun.act(0.762); noun.group(0.749)	verb.communication(+0.0206); noun.feeling(+0.0182); noun.plant(+0.0155)
Bestsellers	0.886	verb.creation(0.795); noun.cognition(0.760); noun.state(0.749)	verb.social(+0.0195); verb.motion(+0.0157); noun.motive(+0.0078)
Nonfiction Mix	0.897	noun.group(0.855); noun.act(0.853); verb.perception(0.839)	verb.perception(+0.0053); verb.weather(+0.0053); noun.animal(+0.0026)
Science Fiction	0.937	noun.time(0.800); verb.competition(0.765); noun.food(0.731)	noun.time(+0.0274); noun.quantity(+0.0157); verb.social(+0.0157)
Mystery	0.938	verb.creation(0.855); noun.time(0.770); noun.state(0.739)	noun.time(+0.0078); noun.Tops(+0.0000); noun.animal(+0.0000)
Biography	0.961	noun.person(0.911); noun.group(0.904); verb.perception(0.896)	noun.person(+0.0079); verb.creation(+0.0052); noun.act(+0.0052)
History	0.984	noun.group(0.948); noun.act(0.935); verb.body(0.919)	noun.animal(+0.0078); noun.group(+0.0078); noun.object(+0.0078)
Romance	0.984	verb.creation(0.879); noun.body(0.878); noun.location(0.843)	noun.motive(+0.0075); noun.Tops(+0.0038); noun.communication(+0.0038)

Table 1: Classification accuracy and feature importance for autofiction versus other genres. Balanced datasets.

2.5 Discussion

The event of autofiction is reflected in some commonly-employed genre classification measures. The use of the term is not simply catachresis, nor does it simply denote a way of reading paratext, nor is it just a new means of selling books. Time will tell if this new species of expression will flourish in the evolving ecosystem of contemporary literature in English.

Across a series of categories verb.creation is among the strongest single discriminator. Much more can be learned. For example, when compared with the memoir, verb.social is the strongest discriminator; memoirs are much more "social" than autofictions.

This analysis also reveals complexity in the fiction-nonfiction comparison. Autofiction is most like highbrow fiction; it is unlike genre fiction. These "prestige" signals are as strong as those related to fictionality.

References

[1] Allison, S. et al. (2011). Quantitative Formalism: an experiment. Stanford Literary Lab #1.
[2] Bamman, D. (2020). BookNLP: A Natural Language Processing Pipeline for Books. (Version 1.0) [Computer software]. https://github.com/booknlp/booknlp
[3] Bode, K. (2018). *A World of Fiction: Digital Collections and the Future of Literary History*. University of Michigan Press.
[4] Dix, H. (Ed.). (2018). *Autofiction in English*. Springer.
[5] Hatzel, H. O. et al. (2023). Machine Learning in Computational Literary Studies. *it - Information Technology*, 65(4-5), 200-217. https://doi.org/10.1515/itit-2023-0041.
[6] Piper, A. (2018). *Enumerations: Data and literary study*. University of Chicago Press.
[7] Piper, A. (2022). The CONLIT dataset of contemporary literature. *Journal of Open Humanities Data*, 8, 1-24.
[8] Sims, M., Park, J. H. and Bamman, D. (2019). Literary Event Detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3623-3634.
[9] Solissa, N. V., van Cranenburgh, A. and Pianzola, F. (2025). Event Detection between Literary Studies and NLP. A Survey, a Narratological Reflection, and a Case Study. *Journal of Computational Literary Studies*, 4(1). doi: https://doi.org/10.48694/jcls.421.
[10] Underwood, T. (2019). *Distant horizons: Digital Evidence and Literary Change*. University of Chicago Press.

Computational Research on Media using INA's Archives

Cassandra Gorin*, *Arthur Lezer P

Institut National de l'Audiovisuel

1 Introduction

Following the dismantling of the Office de Radiodiffusion Télévision Française (ORTF) in the 1970s, the Institut National de l'Audiovisuel (INA) was created, with the responsibility to preserve and manage the French national broadcasting archives.

A major change occurred in the early 1990s, as the legal deposit law was extended to include to radio and television broadcasting. Rooted in a principle dating back to the 17th century, the legal deposit aims to preserve a copy of all cultural productions. In addition to archiving the content of its institutional clients, INA was therefore tasked with recording all radio and television broadcasts aired in France.

This went along with a complementary mission: facilitating access to these archives to support academic research. This led to the creation of INA's consultation center, the Inathèque, where researchers can view INA's archives. In 2007 , this mission was expanded to include web archiving, in response to the digitization of media and the growing importance of online platforms in the audiovisual ecosystem.

Regarding tooling, academics have traditionally relied on the Inathèque search engines and visualization interfaces to view and manually annotate archival material. However, as the size of the legal deposit increased (representing 24 million hours, 3,400 years of TV back-to-back today), a growing number of research projects shifted from this document focused approach to data centered approaches. In fields such as digital humanities and computational social sciences, researchers were now seeking to apply quantitative methods and AI-supported mining techniques to vast audiovisual corpora.

The creation of the lab responded to this need to provide the academic community with dedicated appropriate services and tools, beginning with access to large-scale audiovisual corpora. By combining media archives, web archives and automated enrichment technologies, the lab supports computational and quantitative research based on INA's collections.

2 Datasets and Corpora from INA's Collections

The INA archives are structured around two sources: broadcast archives and web archives

2.1 Traditional Media Archive

INA's media archives spans the entire history of French radio, with a scope that has expanded over time. As of today, INA continuously captures and archives broadcasts from more than 135 television channels and over 85 radio stations.

Each broadcast is accompanied by documentary metadata, such as broadcast dates, times and channel. For high-profile or widely viewed programs, detailed summaries, thematic classifications,

participants and audience metrics can also be provided.

2.2 Datasets and Corpora from INA's Collections
Since 2007, the web archive follows the online activity of key actors of the French audiovisual landscape, including television and radio channels, journalists, content creators, and podcast producers. It includes:

- 30,000 websites, each captured at multiple points in time;
- 30,000 channels across platforms such as YouTube, Twitch and podcast platforms;
- 16,000 social media accounts, mainly on Twitter, representing approximately 1 billion posts between 2007 and 2023.

Together, broadcast and web archives allow researchers to build complex corpora, enabling longitudinal analysis of programming trends, thematic evolutions, and media event coverage across decades and media ecosystems.

3 Automated Processing and Data Enrichment

In addition to descriptive record metadata, automatically generated features extracted from audiovisual content can be used to enhance audiovisual content and further research possibilities.

- **Automatic Speech Recognition (ASR**): Speech-to-text technologies enable the automatic transcription of spoken content in radio and television programs. These transcriptions allow for natural language processing driven analysis such as topic modeling and discourse analysis.
- **Named Entity Recognition (NER)**: From these transcriptions, named entity recognition tools extract references to people, organizations, places, and events. This facilitates network analysis or tracking of media visibility.
- **Speaker Analysis and Gender Detection**: Audio processing tools allow for speaker classification. For example, INA has developed InaSpeechSegmenter,a tool capable of estimating speaker gender, allowing research on gender representation and speaking time distribution. These automatically generated data layers complement the documentary metadata created by archivists. The result is a richly structured, multimodal dataset that supports computational exploration.
- **Optical Character Recognition (OCR) of embedded Text**: OCR is applied to texts embed- ded in audiovisual programs, such as the display the main headlines of the day or participants name on a infobar at the bottom of the screen.

4 Modalities of access and data provision

By combining exhaustive national broadcast capture, curated web archives, and automated enrichment, the INA archive constitutes a unique, large-scale, longitudinal data source for computational analysis of media and culture.

Custom datasets, tailored to specific research projects, can therefore be provided to researchers. These corpora may focus on:

- A particular historical period
- A specific theme (climate change, presidential elections, sports coverage)
- A media genre (news, talk shows, entertainment)

Researchers wishing to work with on such corporas may reach the INA lab at lelab@ina.fr.

References

David, D. et al. (2018). An Open-Source Speaker Gender Detection Framework for Monitoring Gender Equality. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Do LLMs get Earworms? A Study of Song Memorization in LLMs

Oumaya Chelbi, Olga Seminck, Yoann Dupont P

Langues, Textes, Traitements informatiques, Cognition (Lattice), CNRS – École Normale Supérieure – Sorbonne Nouvelle

Large Language Models (LLMs) trained on web-scale corpora have been shown to reproduce copyrighted and culturally salient texts such as widely circulated song lyrics or iconic verses that are repeatedly encountered online, raising major legal and ethical concerns. Recent studies have demonstrated that verbatim memorization correlates with model size and with the frequency of documents in the training data [1, 9]. However, much less is known about the poetic and linguistic properties of the memorized material itself such as repetition patterns, rhyme schemes, or formulaic expressions typical of popular music. While we now have robust evidence that LLMs memorize, we still lack an understanding of which kinds of cultural productions are more likely to be retained and why. Preliminary observations further support this question: in exploratory tests conducted on public domain poetry, only highly ubiquitous texts such as “Twinkle, Twinkle, Little Star” (~7M hits on the Google search engine) were consistently memorized, suggesting a strong link between memorization and large-scale web exposure.

As generative models are increasingly used to produce cultural content, including popular music and lyrics, understanding how training data biases shape what models remember and reproduce has become a major societal and cultural challenge.

This project proposes a computational analysis of verbatim memorization using song lyrics from the Wasabi corpus [4] as a case study, with the broader goal of understanding how structural and cultural factors shape memorization in large language models. Building on prior work on literary fiction memorization [9], we extend the investigation to popular music, a domain characterized by high repetition, strong genre conventions, and large-scale circulation, including recurring chorus hooks and repetitions, and standardized verse structures that may facilitate memorization. We combine model scaling experiments, corpus metadata, and musicological structure analysis to test whether a song’s formal organization (structure), genre, historical period, popularity, and frequency in the training corpus influence its likelihood of being memorized and reproduced by a model.

Our experiments focus on the OLMo family [6, 7] of open-source models (OLMo-2 and OLMo-3, ranging from 7B to 32B parameters), trained on the Dolma corpus [8]. Because both the models and the training data are fully documented and openly available, this setting enables a direct connection between memorization behavior and properties of the pretraining corpus. This transparency makes it possible to empirically verify hypotheses about how data exposure and cultural structure shape model memory. Additionally, these models have shown state of the art performance on AI-benchmarks [6, 7].

The study is guided by five research questions. We first examine whether larger models exhibit higher levels of lyric memorization than smaller ones. We then test whether formal song structure, such as verse-chorus repetition patterns (AABA, ABABCB, where A denotes verse, B chorus, and C bridge), affects memorization probability. Third, we analyze whether musical genre (pop, rock, hip-hop..) influences memorization rates. Fourth, we investigate temporal effects by comparing songs released before and after 2010. Finally, we assess the impact of corpus exposure and

cultural popularity by combining verse-level frequency counts from Dolma with chart performance indicators as proxies for social circulation.

Following the protocols of Chang et al. [3] and Zhang et al. [9], memorization is evaluated through prompt based reconstruction tasks. The models are queried using the first line of a verse, partial choruses, or randomly selected fragments, and their outputs are compared to the original lyrics for instance prompting a model with a well-known opening line and observing whether it continues with the exact original verse. Verbatim memorization is quantified using exact match rates, longest common subsequence scores, and token overlap metrics. These measures are then statistically analyzed.

While memorization has primarily been framed as a technical failure or a legal risk [1, 2], we adopt a cultural perspective. If some cultural artefacts are systematically more likely to be memorized than others, this may reflect the hierarchies, repetitions, and conventions that structure digital culture itself. Drawing on cultural analytics [5] and computational humanities, we treat memorization not only as a limitation of language models, but also as a signal of cultural salience and formal regularity encoded in training data.

If our results demonstrate that memorization is not random but culturally and structurally patterned, it would contribute to a better understanding of how digital culture is represented inside large language models. It would also show how poetic and musical forms support algorithmic memory, extending prior work on literary texts to popular music, offering a generalizable framework for studying memorization across cultural domains.

References

[1] Carlini, N. et al. (2021). Extracting Training Data from Large Language Models. *30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650.
[2] Carlini, N. et al. (2023). Extracting Training Data from Diffusion Models. *32nd USENIX Security Symposium (USENIX Security 23)*, 5253–5270.
[3] Chang, K. et al. (2023). Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7312–7327. https://aclanthology.org/2023.emnlp-main.453/.
[4] Fell, M. et al. (2020). Love Me, Love Me, Say (and Write!) that You Love Me: Enriching the WASABI Song Corpus with Lyrics Annotations. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2138–2147.
[5] Manovich, L. (2020). *Cultural analytics*. MIT Press.
[6] OLMo, Team et al. (2025). 2 OLMo 2 Furious. https://arxiv.org/abs/2501.00656.
[7] Olmo, Team et al. Olmo 3. (2025). https://arxiv.org/abs/2512.13961.
[8] Soldaini, Luca et al. (2024). Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. https://arxiv.org/abs/2402.00159.
[9] Zhang, X., Seminck, O., and Amsili, P. (2024). Remember to Forget: A Study on Verbatim Memorization of Literature in Large Language Models. *Proceedings of the Computational Humanities Research Conference*, 961–981.

Emotional Heritage in Digital Ruins: Fine-Grained Analysis of Skyblogs as Patrimonial Object.

Damien Conceicao **, *Emmanuelle Bermes*** P

École nationale des chartes - PSL

The advent of born-digital heritage has profoundly transformed conservation and analytical practices, requiring innovative methodologies to process vast volumes of cultural data. Skyblogs, the iconic French blogging platform of the 2000s, represents a particularly compelling case: before its closure in 2023, it hosted over 12 million blogs, now archived by the Bibliothèque nationale de France (BnF), amounting to approximately 40 TB of data. While these materials may appear trivial or ephemeral, Daniel Fabre's concept of "patrimonial emotions" [1] highlights how affective attachments can elevate everyday cultural productions to heritage status. Skyblogs thus constitute a key site for investigating how emotional expression participates in the construction of collective memory and digital heritage.

As part of the SkyTaste project (PSL Young Researcher Starting Grant 2024), this experiment addresses the following research question: what can the study of emotions in Skyblogs tell us about these digital objects and the socio-cultural practices they embody? Answering this requires moving beyond coarse-grained sentiment analysis, which typically reduces affect to binary or ternary categories (positive/negative/neutral), and instead developing tools capable of capturing more nuanced emotional expressions adapted to the specificity of early web vernacular practices.

A sampling strategy was implemented by selecting from a curated corpus extracted from the BnF web archives, a total of 1,000 blogs at random, from which up to 10 articles per blog were drawn (depending on availability). Only articles containing textual content and at least one comment were retained. After HTML cleaning and extraction of both posts and associated comments, the final corpus consists of 7,117 articles and their comments. This dataset aims to balance representativeness and feasibility, given both computational constraints and legal restrictions associated with copyrighted web archives.

A central methodological contribution of the project lies in the development of a tailored emotion annotation grid, designed to reflect the specific affective landscape of Skyblogs. Following an initial annotation workshop involving researchers and students familiar with the corpus, an inductive annotation scheme was established. The resulting grid includes eleven categories: *Appréciation* (Appreciation), *Amour* (Love), *Irritation*, *Tristesse* (Sadness), *Mépris* (Contempt), *Empathie* (Empathy), *Excitation* (Excitement), *Joie* (Joy), *Neutre* (Neutral), *Peur* (Fear), *Étonnement* (Surprise). This grid departs from standard emotion taxonomies by foregrounding categories such as *Appréciation*, which emerged as particularly salient in the corpus.

A second annotation phase was conducted to produce training data. Annotators worked on 1,600 textual excerpts, each limited to 256 characters and carefully truncated at sentence boundaries to preserve semantic coherence. Nine annotators participated in this task, working in independent groups of two or three without intra-group communication. To ensure high-quality labels, only annotations with optimal agreement were retained: for pairs, full agreement was required; for groups of three, at least two annotators had to agree. This strict filtering resulted in a final training set of 792 annotated excerpts.

The computational pipeline was designed under significant constraints: data could not be exported outside the BnF environment, internet access was restricted, and available computational resources were limited. These conditions led to the adoption of relatively lightweight ("low-tech") yet robust approaches. Given the highly non-standardized nature of Skyblogs language—characterized by vernacular French, creative orthography, abbreviations, and emoticons, a normalization step was introduced. A custom dictionary was used to map frequent idiomatic forms to standard French, while emoticons were converted into descriptive syntagms enclosed in delimiters (e.g., ":smile:"). The normalized texts were then embedded using CamemBERT-base, and classification was performed using a Support Vector Machine (SVM). The resulting model reflects the inherent imbalance of the dataset, with strong prevalence of the categories *Neutre*, *Amour*, and *Appréciation*.

Evaluation yields an accuracy of 0.663, a macro-averaged F1-score of 0.402, and a weighted F1-score of 0.647. While these results remain modest (particularly for underrepresented classes) they are consistent with the difficulty of fine-grained emotion classification in noisy, informal textual data. To scale the analysis to the full corpus, each article was segmented into smaller textual chunks, to which the model was applied individually. All predicted labels were then aggregated at the article level. This approach enables distant reading of the corpus while preserving sensitivity to local emotional variation. The obtained results reveal several notable patterns. First, the overall distribution of predicted emotions in the full corpus closely mirrors that of the annotated training data, suggesting a degree of internal consistency in the modeling approach. Second, when comparing emotional distributions between blog posts and their associated comments, a striking regularity emerges: regardless of the initial emotion expressed in a post, approximately one-third of responses fall under the category of *Appréciation*. This suggests that appreciation functions as a dominant interactional mode within the Skyblogs ecosystem.

Further insight is gained through analysis of the model's confusion matrix. A large number of misclassifications involve confusion with *Appréciation*. However, qualitative examination of these cases reveals that such "errors" are often meaningful: many excerpts simultaneously express appreciation alongside another emotion (e.g., sadness, irritation, or even contempt). This indicates that appreciation operates less as a discrete category than as a pervasive affective layer, co-occurring with other emotional states. In this sense, model errors become analytically productive, highlighting the limits of single-label classification and pointing toward the fundamentally multi-label nature of emotional expression in the corpus. These findings contribute to a reassessment of Skyblogs as a socio-emotional space.

Contrary to certain retrospective narratives emphasizing conflict or superficiality, the platform appears to have been structured around practices of mutual appreciation and affective reinforcement, particularly among teenagers. At the same time, the study underscores important methodological limitations. The current annotation scheme may lack sufficient granularity to fully capture the complexity of emotional expression, especially given the pervasive overlap between categories. Future work should explore multi-label annotation strategies and larger textual units to better account for this phenomenon. Finally, the scale of the analyzed sample remains limited relative to the full archive. Although the observed distributions are stable within the current dataset, further validation on larger or differently sampled subsets is necessary to assess the robustness of these findings. More broadly, this project demonstrates how constrained computational environments and copyrighted data can still support meaningful analysis through carefully designed, resource-efficient methods. By combining close and distant reading, and by treating model limitations as epistemological opportunities, it contributes to the development of new frameworks

for studying emotion in digital heritage corpora and, more generally, for accessing the socio-cultural dynamics of early web archives.

References

[1] Fabre, D.(2013). *Émotions patrimoniales*. Éditions de la Maison des sciences de l'homme.

Gröber Revisited: Bayesian Phylogenetics of the Troubadours' Chansonniers

Jean-Baptiste Camps, Ulysse Godreau **P**

École nationale des chartes - PSL

1 Introduction

Since the 19th century, and the seminal work of Gröber [5], it has proven difficult to offer a single unified evolutionary tree (stemma) for the tradition of the compilations of troubadour poems known as *chansonniers*. The story of the transmission of these compilations, between France, Italy and the Iberian peninsula, has proven very complex, likely involving numerous operations of 'collection' of texts, as well as multiple instances of lateral transmission ("contamination") between branches of the tradition.

Gröber's work offered the first global theory on the transmission of the troubadour poems. According to him, the transmission has known several phases, and used different types of written vectors (even setting apart potential oral transmission). In addition, the earliest forms of troubadour manuscripts appeared as *Liederblätter* – isolated song sheets, sometimes called *breu de parchemina* in Occitan, which may have been used directly used by performers. These sheets would have enable an initial, non-exclusively oral transmission of troubadour works. Later, texts began to be compiled into *Liederbücher* (songbooks), individual scrolls, or books containing the works of a single troubadour, potentially organized by the troubadour themselves or by others. The first collections to emerge were likely *Gelegenheitssammlungen* (occasional compilations) or *Zusammengesetztensammlungen* (factitious collections) — ad hoc assemblages of juxtaposed texts without editorial reworking. These early collections lacked the defining features of the most significant chansonniers: the *einheitliche geordneten Sammlungen* (unified, ordered collections). The latter were systematically organized according to unifying principles (genre, author sections, chronology, aesthetic or sociologic judgment, or even, as in the case of chansonnier *E*, simple alphabetical order). The great retrospective anthologies of the Venetian chansonniers, such as *A*, *I*, and *K*, belong to this category, blending authorial and generic classifications, with further internal ordering that remains less transparent (possibly tied to chronology or the prominence of individual troubadours and their works). Also attested are specialized collections: *Coblassammlungen* and *Sentenzensammlungen* (florilegia of strophes or extracts) [4, 5].

Then, based on the particular assemblages of texts in each manuscript, as well as their particular order, he established groupings, list of shared sources, and partial stemmata for the chansonniers. His method was soon used by Schwan [10] to establish the stemma of the manuscripts containing the work of the *trouvères*, the French counterparts of the Occitan troubadours. Yet, in contrast to Schwan's results, that established a form of base consensus on the transmission of the *trouvères*, it has proven hard to reach a consensus regarding the transmission of the troubadours. Gröber's results have been criticised and debated since the 19th century, while reaching a unified solution has proven elusive.

We offer here to address this question, in an approach similar to Gröber's, but using the framework offered by Bayesian phylogenetics, that has been recently transposed in stemmatology [7, 8].

Like Gröber, we will use the particular assemblages of texts in the books in order to draft a transmission history of the compilations themselves, possibly partially independent from the transmission of individual texts. For now, we will use the presence or absence, in each individual chansonnier, of a given poem as binary traits and proxies to infer genealogical relationships.

2 Material and Methods

In the framework of bayesian phylogenetics, the stemma of the tradition is inferred from the simulation of a large number of candidate trees generated according to a given stochastic model. The output of the inference process is a posterior probability distribution on tree topologies compatible with given data on the tradition. In our case, these data consist in binary arrays indicating, for each individual poem and *vida* existing in the whole set of chansonniers, if the the work is present in the corresponding witness. The stochastic model itself is a combination of the three following components

Tree model it is used to generate random tree topologies. In our case, we use the Birth-Death Skyline (BDSKY) model, in which new branches of the tradition are created with rate λ and become inactive at rate μ. Living branches are moreover continuously sampled over time (resulting in extent witnesses) at rate ψ.

Substitution model it rules how the binary features switch values from one manuscript to another. We take here the same rate for substitutions in both directions. Although this is unrealistic, it allows for a diverse sampling of evolutionary histories of the tradition.

Clock model We used a strict clock model, i.e. we assumed that the average rate of change – whose distribution of value is determined during the inference process – is uniform on all branch of the traditions. This is also an oversimplification that harms the precision of the dating of internal nodes of the stemma, but not its overall topology.
The likelihood of the randomly generated trees is then systematically checked against the input data, and a Monte-Carlo Markov Chain algorithm, implemented with the software BEAST2 [2], is then used to sample the space of topologies and construct the maximum credibility stemma.

Variants (sites) We use the registration of the presence or absence of individuals poems in the chansonniers provided by the *Bibliografia Elettronica dei Trovatori* [1], in the form of a binary incidence matrix.

Witness dates Witness dates have been recorded, based on those provided in the bibliography [3, 6].

3 Results and Discussion

Current results (Figure 1) show local groupings that have been long identified (for instance, between the two "twin manuscripts" *I* and *K*), as well has a general outlook that matches some of the main regions of manuscript production, Languedoc (*CR*) or Venetia (*DAIK*). Future work may explore the inclusion of more manuscripts, as well as enriching the priors taken into account, regarding for

instance the order of the poems, the sections, illumination or other codicological features, as well as the dates of the poems themselves. Finally, multispecies-coalescent models [9] may be introduced to study the interplay between text-level variants and compilation-level text transmission.

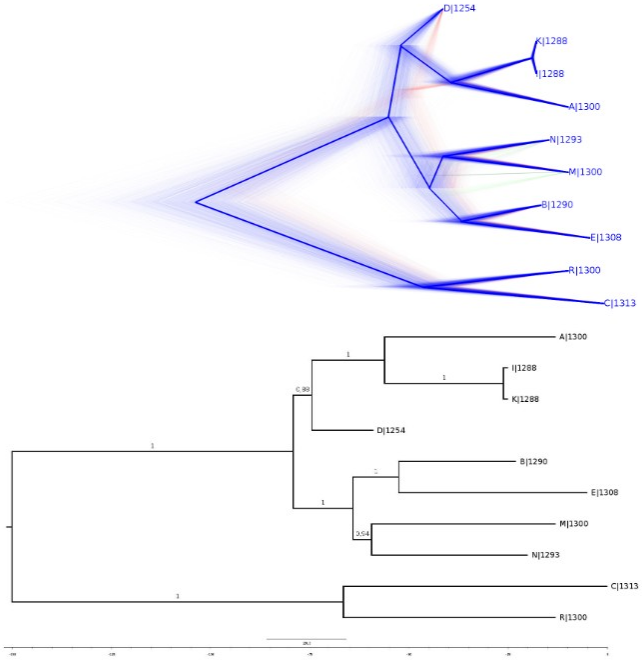


Figure 1: Densitree and maximum credibility tree resulting for the largest *chansonniers*

References

[1] Asperti, S., and De Nigro ,L. (2012). *Bibliografia Elettronica dei Trovatori* (BeDT). v.2.5. Università di Roma La Sapienza. http://www.bedt.it.

[2] Remco, B. et al. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology* 15(4), 1–28. https://doi.org/10.1371/journal.pcbi.1006650.

[3] Camps, J.-B. (2010). Les Manuscrits occitans à la Bibliothèque nationale de France. Mém. de conservateur de bibl. Lyon: ENSSIB. http://coreac.uk/download/pdf/12438620.pdf.

[4] Camps, J.-B. (2009). *Vocabulaire du texte, vocabulaire de l'image: la représentation des troubadours dans les chansonniers occitans A (BAV Vat. Lat. 5232), I (BnF Fr.854) et K (BnF Fr. 12473)*. fre. mém. de master, dir. Françoise Vielliard et Fabio Zinelli. Paris: École nationale des chartes.

[5] Gröber, G. (1877). Die Liedersammlungen der Trobadours. *Romanische Studien* 2, 337–670.

[6] IRHT. *Jonas: Répertoire des textes et des manuscrits médiévaux d'oc et d'oïl*. http://jonas.irht.cnrs.fr/.

[7] McCollum, J., and Turnbull, R. (2024). Using Bayesian phylogenetics to infer manuscript transmission history. *Digital Scholarship in the Humanities*, 39(1), 258–279. https://academic.oup.com/dsh/article-abstract/39/1/258/7477852.

[8] Moors, S., and McCollum, J. (2025). From a Computer-Assisted Stemma to a Phylogenetic Tree: The Medieval Dutch Martijn Trilogy by Jacob van Maerlant. *Anthology of Computers and the Humanities* 3, 189–210. https://doi.org/10.63744/vBRBHo3Hn4fX.

[9] Pekka, P., and Nei, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution* 5(5), 568–83. https://api.semanticscholar.org/CorpusID:7936117.

[10] Schwan, E. (1886). *Die Altfranzösischen Liederhandschriften, ihr Verhältniss, ihre Entstehung und ihre Bestimmung. Eine litterarhistorische Untersuchung*. Berlin.

Identifying Metaphorical Sense Extensions Using Large Language Models

Gabriel Zerguit[1]**, *Rowan Hall Maudslay***[2]

[1] École nationale supérieure - PSL
[2] École nationale des chartes - PSL

1 Conceptual metaphor in language

Human language is heavily influenced by metaphor. A metaphor is a piece of language in which one entity is framed in terms of another which it resembles, as with "the *seed* of an idea" or "the *march* of progress". In *Metaphors We Live By*, Lakoff and Johnson [6] argue that everyday reasoning is structured by systematic patterns of metaphor known as *conceptual metaphors*. A conceptual metaphor maps an abstract domain onto a more concrete one. For example, in English and other European languages, argument is often conceptualised metaphorically in terms of warfare: speakers *attack* opponents, *defend* claims, *shoot down* ideas, and *campaign* for support. These instances of metaphorical language are all evocations of the same conceptual metaphor, ARGUMENT IS WAR.

Lakoff and Johnson [6] claim that variation of conceptual metaphor systems between languages could influence how members of different speaking communities think or feel. This theory, known as *conceptual metaphor theory* (CMT), is the dominant theoretical framework in metaphor research, and has motivated extensive work across linguistics, cognitive science, and the humanities [3, 4]. However, large-scale cross-linguistic evidence for CMT is limited due to the cost of manual analysis. We do not know how conceptual metaphors vary between languages, let alone whether variation in conceptual metaphor relates to cultural variation.

2 Dictionaries as window onto metaphor

Our goal is to design methods that can be used to automatically identify the conceptual metaphors in a particular language, in order to enable a large-scale analysis of conceptual metaphor variation. The methods we are designing exploit data that is readily available in dictionaries. A dictionary provides a compact representation of all the lexicalised meanings in a language. Among the senses contained within a dictionary are many widely conventionalised metaphors. For example, the popular computational dictionary WordNet [2] defines *attack* both (i) as a military offensive against an enemy, and (ii) as intense adverse criticism in a debate. The latter is a metaphorical extension of the former. Similar literal and figurative sense pairings recur across many lemmas, capturing the presence of ARGUMENT IS WAR in the English lexicon. It follows that dictionaries could be used to recover the conceptual metaphors of a language.

However, while many dictionaries contain literal and metaphorical senses, they do not identify which senses are literal and which are metaphorical extensions. Recent work has shown that information of this nature can be manually annotated with high inter-annotator agreement [8], but thus far only partial data for English has been collected, covering 6,500 nouns. The resource containing this annotation is called ChainNet. Other parts of speech and other languages have not been treated. If dictionaries are to be used to infer the conceptual metaphors of different languages, we need a method to automatically identify metaphorical sense extensions in dictionaries at scale.

3 Polysemy parsing with LLMs

In this work, we investigate whether large language models (LLMs) can be used to automatically identify metaphorical sense extensions. The input to the model is the list of sense definitions for a single lemma (e.g. all the nominal senses of *attack*), which are sourced from the Open English Wordnet (OEW). The output is a *polysemy parse* [8], a forest of directed out-trees in which each node corresponds to a sense. In a polysemy parse, a root corresponds to a *prototypical sense*, which is a cognitively central and representative meaning of a word [5, 9]. Directed edges from a prototype to other senses indicate sequential sense extensions, labelled either *metaphor* or *metonymy*, where metonymy is a meaning extension based on contiguity rather than cross-domain mapping [4]. An example of metonymy is when the word *glass* is used to refer to a drinking container ("a glass"), as opposed to the material from which the container is made.

We have implemented an end-to-end pipeline that loads sense definitions from the OEW 2025 release, generates an LLM prompt, and submits requests asynchronously through the OpenAI Batch API. We compare three system-prompt designs of increasing specificity:

- **(A)** a minimal description of the labels;
- **(B)** the full annotation guidelines, including diagnostic patterns for metonymy (product–producer, container–content, part–whole, etc.) and edge constraints on the parse;
- **(C)** the same guidelines extended with one in-context example.

We benchmark several models, including gpt-4.1-mini, gpt-4o, and o3-mini, and, for reasoning models, we additionally experiment will different amounts of reasoning tokens. Predictions are evaluated against the ChainNet gold standard annotation using three standard graph-parsing metrics: Label-Only Score (LOS), Unlabelled Attachment Score (UAS), and Labelled Attachment Score (LAS). Results will be summarised on a performance–cost tradeoff plot, allowing us to identify configurations that strike the best balance between accuracy and large-scale feasibility.

4 Outlook

Our aim is to deploy our chosen method to achieve full coverage on the OEW, and then to investigate whether it can also be successfully applied to other languages. The Japanese Wordnet [1] will be used as a test case. Japanese provides an interesting test case because it is typologically distant from English and has been argued to exhibit distinctive conceptual metaphors [7]. If successful,our approach will make it possible to identify conceptual metaphors in a scalable manner, and will thus play a part in facilitating a large-scale cross-linguistic analysis of conceptual metaphor variation

References

[1] Bond, F., et al.(2009). *Proceedings of the 7th Workshop on Asian Language Resources*, 1–8.
[2] Fellbaum, C. (1988). *WordNet: An Electronic Lexical Database*. MIT Press.
[3] Gibbs, R. W. (2008). *The Cambridge Handbook of Metaphor and Thought*. Cambridge University Press.
[4] Kövecses, Z. (2010). *Metaphor: A Practical Introduction*. Oxford University Press.
[5] Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press.

[6] Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.
[7] Matsuki, K. (1995). Metaphors of Anger in Japanese. *Language and the Cognitive Construal of the World*, ed. by John R. Taylor and Robert E. MacLaury. Berlin: Mouton de Gruyter, 137–151.
[8] Maudslay, R. H et al. (2024). ChainNet: Structured Metaphor and Metonymy in WordNet. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2984–2996.
[9] Rosch, E. (1975). Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology: General* 104(4), 192–233.

LatticIA: Childbirth and Mother Characters in Big Literary Corpora

Olga Seminck, Nathan Ducros P

Langues, Textes, Traitements informatiques, Cognition (Lattice), CNRS – École Normale Supérieure – Sorbonne Nouvelle

1 Context

Inspired by previous studies that have highlighted stereotypes associated with women in literature [16, 22, 23], using distant reading techniques [15, 21], this project focuses on childbirth scenes and the figure of the mother.

Beyond the initial objective of quantifying the recurrence of childbirth scenes across genres and periods using large literary corpora, the study also aims to provide qualitative analyses. It examines the metaphors employed [20] and recurring topoi, such as death in childbirth1, in order to explore how literature represents birth and motherhood. It further seeks to determine whether an archetype of the mother exists, which stereotypes are associated with it, and how these have evolved over time. This research complements existing literary studies that address these questions without a computational component [4, 10, 11, 18].

Recent studies have demonstrated the possibility of detecting prototypical scenes—such as those involving danger [1, 19]. Frameworks have also been developed to support largescale qualitative analyses of fictional characters, for example to describe their social status or physical traits [3, 17]. The LatticIA project builds on these approaches.

As the project is still in its early stages, the remainder of this article focuses on the pipeline currently being developed for the automatic detection of childbirth scenes.

2 Method

We employ a processing pipeline whose stages are detailed in Figure 1.

1. First, we compile a training corpus based on the Birth(ing) Stories database [7], developed within the Birth(ing) Stories project: a project that aims of uncovering the issues of aesthetic and cultural representation surrounding childbirth in literature. It contains the titles [6] of 143 books from the 12th to the 21st century, each featuring at least one childbirth scene. This database serves as a starting point for building the training corpus. We search for full-text versions of these works in large literary corpora (Chapitres [13], Gallica, and Majinbook [14]).

2. The training corpus for the childbirth scene detection model is built using Retrieval-Augmented Generation (RAG) [9] in a zero-shot framework, following a strict set of annotation guidelines [8] developed within our project LatticIA. These guidelines address the following questions:

    - Is this a childbirth scene?

    - What is the phase of labor?

- Which characters are present?
- Is pregnancy mentioned?
- Is there medical intervention?
- Are there complications?

3. We then conduct an iterative process combining manual corrections and new automated annotations using the INCEpTION semantic annotation platform [12]. At this stage, human annotators play a crucial role: the interaction between the LLM, the annotation guidelines, and inter-annotator discussions enables the refinement of a gold-standard annotated dataset.

4. Based on this gold-standard dataset, we train a classification model to automatically detect childbirth scenes across a larger corpus, using contextualized embeddings using models with a BERT-based architecture [2].

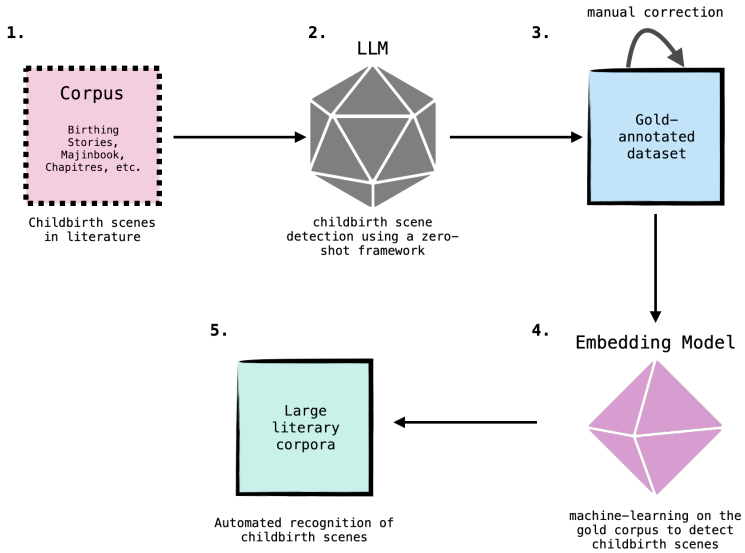


Figure 1: Overview of the Processing Pipeline.

3 Conclusion

Computational methods make it possible to quantify and analyze the role of childbirth and more specifically the mother character—in literature. Within close reading, it is often evident that female characters are frequently marginalized or stereotyped,particularly in relation to childbirth [4, 11, 18]. Approaching this topic through computational methods and distant reading [15] not only offers a new perspective but also reveals long-term evolutionary patterns in literary representation and provides new analytical tools. A seemingly minor element in a single work can acquire systematic significance when examined at scale.

The stereotypes and representations found in literature reveal much about how childbirth and the role of the mother have been perceived in society. Beliefs about childbirth profoundly influence women's personal experiences [5]. By identifying large-scale patterns in literary representations, this study aims to contribute to a clearer understanding of the origins and persistence of ideas surrounding childbirth.

References

[1] Barré, J. (2024). Détection automatique de l'architextualité dans le roman d'aventures. *Humanistica 2024*.

[2] Barré, J. et al. (2025). Modeling the Construction of a Literary Archetype: The Case of the Detective Figure in French Literature. *Proceedings of the Computational Humanities Research Conference*.

[3] Bourgois, A. et al. (2026) Toward an Ontological Representation of Fictional Characters". *Computational Humanities Research* 2 .

[4] Braun, A. (2022). Representing Childbirth in Literature: From Third to First Person. *Mothers and Writers*. https://doi.org/10.58079/rolv.

[5] Danino, C. (2020). Au commencement était le verbe… Explorations linguistiques de la périnatalité. *Sages-Femmes* 19(1), 48–52.

[6] Danino, C. et al. (2020). Birth(ing) Stories Project. https://sites.google.com/view/birthing-stories/.

[7] Danino, C.et al. (2024). Birth(ing) Stories - Zotero dataset. https://www.zotero.org/groups/4896649/childbirth_scenes/items/Q8V5XTEW/library

[8] Ducros, N. (2026). scenes_accouchements_inception: Annotation guidelines for childbirth scenes on INCEpTION. https://github.com/nthdcr/scenes_accouchements_inception.

[9] Gao, Y. et. (2023). Retrieval-augmented generation for large language models: A survey.https://arxiv.org/abs/2312.10997.

[10] Huet, M.-N. (2018). *Maternité, identité, écriture: discours de mères dans la littérature des femmes de l'extrême contemporain en France*. Thèse de doctorat, Université du Québec à Montréal.

[11] Jolivet, V. (2017) Accouchements libertins : Sade et la femme enceinte. *Enfanter dans la France d'Ancien Régime*, ed. by Laetitia Dion and al. Artois Presses Université.https://doi.org/10.4000/books.apu.11011.

[12] Klie, J.-C. et al. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5–9. http://tubiblio.ulb.tu-darmstadt.de/106270/.

[13] Leblond, A. (2022). Corpus Chapitres. (Version v1.0). doi: 10.5281/zenodo. 7446728.

[14] Mazières, A. and Poibeau, T. (2025). MajinBook. https://ens.hal.science/hal-05365551.

[15] Moretti, F. (2013). *Distant Reading*. London: Verso.

[16] Naguib, M. et al. (2022). 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". *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles, Atelier TAL et Humanités Numériques (TAL-HN)*, 66–77.

[17] Pagel, J. and Reiter, N. (2025). Automatic Detection and Classification of Literary Character Properties in German Narratives. *Anthology of Computers and the Humanities* 3, 1493–1508. doi: 10.63744/UkAh6gmT12av.

[18] Poston, C. H. (1978). Childbirth in Literature. *Feminist Studies* 4(2), 18–31.

[19] Seminck, O., and Barré, J. (2024). DELiRAnt : Danger et Exploration dans le Roman d'Aventures. Version 1. https://doi.org/10.5281/zenodo.14520608.

[20] Uludag, E. and Cerit, E. (2022). The Women's Emotions about Experience of Vaginal Birth Based on the Metaphors: A Phenomenological Study. *Journal of Nursing and Midwifery Sciences* 9(4), 273–281.

[21] Underwood, T. (2019).*Distant Horizons: Digital Evidence and Literary Change*. The University of Chicago Press.

[22] Underwood, T., Bamman, D., and Lee, S. (2018). The Transformation of Gender in English-Language Fiction. *Journal of Cultural Analytics* 3(2). doi: 10.22148/16.019.

[23] Vianne, L., Dupont, Y., and Barré, J. (2023). Gender Bias in French Literature. *Proceedings of the Computational Humanities Research Conference* Vol. 3558. CEUR Workshop Proceedings, 247–262.

Propp: Multilingual Structured Narrative Extraction for Computational Humanities

Denise Atzori[1], ***Jean Barré***[1], ***Antoine Bourgois***[1], ***Maria Kirbasova***[1], ***Kirill Maslinsky***[2], ***Frédérique Mélanie-Becquet***[1], ***Olga Seminck***[1], ***Thierry Poibeau***[1] **P**

[1] Langues, Textes, Traitements informatiques, Cognition (Lattice), CNRS – École Normale Supérieure – Sorbonne Nouvelle
[2] Institut National des Langues et Civilisations Orientales (INALCO)

1 Introduction

Computational humanities increasingly rely on large-scale textual analysis to investigate narrative structure, character typology, and literary form. Following Moretti's formulation of distant reading as a methodological approach [8], research in the field has sought to move beyond isolated close readings toward corpus-level analysis of literary phenomena. This shift has raised a central methodological question: how to develop linguistically grounded tools capable of extracting interpretable narrative structure at scale [9].

Systems such as BookNLP [1, 2] have shown that character references and coreference chains can be extracted from English fiction in a systematic way, enabling downstream analyses such as automatic construction of character networks. More recent work in computational literary history [10] has further demonstrated how large-scale literary analysis can be anchored in explicit modeling frameworks that support interpretability and theoretical reflection. The Propp project, extending BookNLP-fr [7], contributes to this line of research by providing a multilingual framework for structured narrative extraction centered on fictional entities and their attributes.

2 Structured Representation of Characters

Propp is developed to support research in literature and the social sciences. Its core objective is to transform narrative texts into structured representations in which fictional entities are associated with systematically extracted attributes.

The system includes named entity recognition for person names, coreference resolution (including pronouns and definite noun phrases), and extraction of character attributes. Rather than treating verbs, modifiers, and related constructions separately, Propp models them under a unified notion of character attributes, divided into three categories:

1. Modifiers, including adjectival epithets and descriptive phrases attached to nouns.

2. Possessive constructions, capturing relations such as "his house" or "her brother."

3. Verbal predicates attached to entities, with a distinction between cases where the entity functions as subject and cases where it functions as object.

This representation produces structured character profiles. For each character, the system records what properties are attributed to it, what it possesses, what actions it performs, and what actions it undergoes. These features can then be aggregated within a novel or across corpora.

Such structured outputs enable research that moves beyond character co-occurrence toward semantic modeling of roles and archetypes. Barré et al. [3] use Propp-derived features to model the emergence of the detective figure in French fiction. Bourgois et al. [4] propose an ontological representation of fictional characters grounded in systematically extracted attributes and relations. In this respect, Propp extends the character-centered paradigm established by BookNLP while emphasizing structured narrative semantics and cross-linguistic applicability.

3 Architecture and Evaluation

Propp combines neural and rule-based components. The architecture integrates dependency parsing and transformer-based encoders, including BERT-style models. Coreference resolution relies on dedicated algorithms described in recent work [5].

Evaluation follows established coreference metrics, including F-measure and standard scoring protocols. Quantitative evaluation is currently most extensive for French literary texts, where Propp has been compared with existing systems. Annotated datasets used for development and evaluation are publicly documented in the project repository (`https://lattice-8094.github.io/propp/`).

While multilingual benchmarking remains ongoing, the evaluation framework aligns with established practices in narrative information extraction and literary NLP [1].

4 Multilingual Narrative Extraction

A distinctive contribution of Propp lies in its multilingual design. Beyond French, the current implementation supports English, Italian, and Russian. The annotation guidelines are generally consistent across the target languages in order to obtain comparable and homogeneous annotated corpora. However, adaptations remain necessary to account for language-specific phenomena.

In Italian, subject pronouns are frequently omitted, since person and number are encoded in verbal morphology:

“Entrò nella stanza.” (“(He) entered the room.”)

Predicate assignment therefore requires recovering implicit subjects. Italian also exhibits clitic constructions where pronouns are attached to verbs within a single token:

“Vederlo” (“To see him.”)

Such constructions affect tokenization and coreference resolution.

Russian presents related challenges. Subject omission can occur in subordinate (and, sometimes, even main) clauses:

“Когда вошёл в комнату, сел.” ("When (he) entered the room, (he) sat down.”)

Here, no overt subject appears, yet both predicates must be linked to a coherent entity. Russian also makes extensive use of predicative adjectives:

“Отцов дом” (“The father’s house”)

The adjective must therefore be analyzed as expressing a possessive relation between entities, even though this relation is realized morphologically as an adjective rather than as a nominal genitive phrase.

These examples illustrate that multilingual narrative extraction cannot rely solely on surface alignment of modules. Propp minimizes language-specific engineering but incorporates targeted adaptations where required. Current work explores multilingual transformer models in order to move toward a unified cross-linguistic architecture.

Evaluation across languages relies on F-measure and established coreference scoring schemes. Developing a fully systematic cross-linguistic evaluation framework remains a central objective of the project.

5 Applications in Computational Humanities

Propp is designed to support research in digital humanities and computational cultural analysis. Its structured outputs enable character network construction [6], archetype detection, diachronic modeling of narrative roles, and comparative cross-linguistic literary analysis. Recent work demonstrates how structured character attributes can model the formation of literary archetypes [3] and support ontological formalization of fictional entities [4]. By extracting structured narrative features rather than surface co-occurrence patterns, Propp provides interpretable representations suitable for corpus-scale narratological inquiry.

6 Conclusion

Propp provides a multilingual framework for structured narrative extraction centered on fictional entities and their attributes. By integrating named entity recognition, coreference resolution, and systematic modeling of modifiers, possessives, and predicate relations, it enables fine-grained and scalable characterization of characters across large corpora. Its multilingual design extends structured narrative extraction beyond French and English and supports comparative analysis across linguistic traditions.

Ongoing developments include the extraction of reported speech and quotation structures, enabling the identification of who said what and to whom, thereby enriching the modeling of dialogic interactions within narrative texts. Beyond literary analysis, the framework is also being adapted to address research needs in the social sciences, where structured representations of actors, attributed properties, and reported statements are central to the study of political discourse, media corpora, and other large-scale textual datasets. In this way, Propp contributes to the development of linguistically grounded and transferable methods for computational humanities and social science research.

References

[1] Bamman, D., Lewke, O., and Mansoor, A. (2020). An Annotated Dataset of Coreference in English Literature". *Proceedings of the Twelfth Language Resources and Evaluation Conference*, ed. by Nicoletta Calzolari et al., 44–54. https://aclanthology.org/2020.lrec-1.6/.

[2] Bamman, D., Underwood, T., and Smith, N. A. "A Bayesian Mixed Effects Model of Literary Character. (2014). *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ed. by Kristina Toutanova and Hua Wu, 370–379. ttps://aclanthology.org/P14-1035/.

[3] Barré, J. et al. (2025). Modeling the Birth of a Literary Archetype: The Case of the Detective Figure in French Fiction. *Proceedings of the Conference on Computational Humanities Research (CHR 2025)*.

[4] Bourgois, A. et al. (2026). Toward an Ontological Representation of Fictional Characters. *Journal of Computational Humanities Research*.

[5] Bourgois, A. and Poibeau, T. (2025). The Elephant in the Coreference Room: Resolving Coreference in Full-Length French Fiction Works. *Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference*, ed. by Maciej Ogrodniczuk, Michal Novak, Massimo Poesio, Sameer Pradhan, and Vincent Ng, 55–69. https://aclanthology.org/2025.crac-1.5/.

[6] Chen, Newman, et al. (2024). Network Analysis, Plot Theory: Revisiting French Literature through Character Networks". *Digital Humanities (DH2024)*. https://hal.science/hal-04855204.

[7] Mélanie-Becquet, F. et al. (2024). BookNLP-fr, the French Versant of BookNLP. A Tailored Pipeline for 19th and 20th Century French Literature. *Journal of Computational Literary Studies* 3, 1–34. https://jcls.io/article/id/3924/.

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

[9] Piper, A., So, R. J., and Bamman, D. (2021). Narrative Theory for Computational Narrative Understanding. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, 298–311. https://aclanthology.org/2021.emnlp-main.26/.

[10] Underwood, T. (2019). *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.

Quantitative Stemmatology and OpenRecensio: A Test Case on the Bible Historiale's Book of Exodus

Valentina Modolo[1,2]

[1] École nationale des chartes - PSL
[2] Scuola Normale Superiore di Pisa

1 Introduction

The translations of the Bible occupy a central position within the medieval European literary heritage. In recent decades, philological and codicological research has produced major contributions to the study of their transmission and documentary history. Despite these advances, comprehensive critical editions are still lacking. The scale of the corpus and the number of witnesses (often several dozen for a single book) make traditional *recensio* demanding and the fundamental reference for the field remains Berger's pioneering survey [1].

The case of the *Bible historiale* (BH) is emblematic. Conceived as a composite work and progressively intertwined with the *Bible du XIIIe siècle*, its textual stratigraphy and redactional phases complicate attempts at a stable reconstruction. Although partial editions exist for individual manuscripts, Michel's edition of the *Genèse* remains the only large-scale intervention so far [10], leaving open several stemmatological questions. Moreover, the BH exemplifies what Pasquali defined as an "*open recensio*" [13], where contamination, interpolation, and rewriting render purely mechanical methods insufficient.

This complexity renders computational approaches particularly relevant. Computer-assisted stemmatology has a well-established scholarly history [14, ch. 5], as demonstrated by recent applications to stratified and contaminated traditions [3]. This contribution applies these approaches to the transmission of *Exodus* in the BH, beginning with a subset of the primitive witnesses. The long-term aim is to extend the analysis to the seventy-four manuscripts transmitting Exodus and assess to what extent computational modelling can support stemmatic reconstruction in biblical traditions.

2 Corpus, Methodology, and Encoding

The corpus comprises eight witnesses: five *manuscrits primitifs* of the first redaction (Maz312, BnF 152, BR 987, JenEl 9596, Bein 129), two witnesses of the 1297 *Seconde Édition* (BnF 155,BL19DIII), and Arsenal 5059 (1317) as an early representative of the *Bible historiale complétée* (see 1). The *Historia Scholastica* [11] and the *Vulgata* [2] were included as external comparanda. Manuscript images were processed in eScriptorium [8] using CATMuS Medieval [Large] (v1.0.0 [5]; CER 98.5%), followed by manual post-correction. The output is natively exportable to ALTO-XML, ensuring interoperability with TEI-based workflows. Relevant variant locations were subsequently encoded in TEI P5 following Camps and Cafiero [4], preserving philological control before the inferential stage — a step that general-purpose software such as PHYLIP does not address intrinsically [14, ch. 5.5]. Each locus is classified by transformation type (lexical, syntactic, semantically driven replacement, macro-structural); purely graphic alternations were excluded. Variant locations were then analysed with Maximum Parsimony (PS [12]), Neighbor-Joining (NJ; [15]), and Unweighted Neighbor-Joining (UNJ [7]).

Table 1: Composition of the corpus. Dates follow Komada [9].

Sigla	Manuscript	Redaction	Date
Maz312	Paris, Bibl. Mazarine, ms. 312	*Première Éd.*	1330–1360
BnF152	Paris, BnF, fr. 152	*Première Éd.*	1347
BR987	Bruxelles, KBR, ms. II 987	*Première Éd.*	1350
JenEl9596	Jena, ThÜLB, El. f. 95–96	*Première Éd.*	1465-1473
Bein129		*Première Éd.*	1460 c.
BnF155	Paris, BnF, fr. 155	*Seconde Éd.*	1310-1315
BL19DIII	London, British Lib., Royal 19 D III	*Bibles à prologues*	1411
Ars5059	Paris, Bibl. de l'Arsenal, ms. 5059	*BH complétée*	1317

For instance, at Col. 1155A (*Tulit Moyses scriptum in lamina aurea nomen Domini tetragrammaton*), the witnesses diverge in the rendering of *lamina*: the *Historia Scholastica reads lamina*; BnF 155 and Ars5059 read *tombe*; Maz312, BnF152, BR987, JenEl9596, Bein129, and BL19DIII read *lame*. This semantic lexical variant is encoded in TEI P5 as a nested parallel-segmentation structure:

```
<p xml:id="Col.1155A" part="NV">
  non mie en
  <app type="semlex:nonsens" xml:id="a_511">
    <rdg wit="#Maz312 #BnF152 #BR987 #JenEl9596 #Bein129 #BL19DIII #hs">
      <app type="translation" xml:id="a_512">
        <rdg wit="#hs">lamina</rdg>
        <rdg wit="#Maz312 #BnF152 #BR987 #JenEl9596 #Bein129 #BL19DIII">lame</rdg>
      </app>
    </rdg>
    <rdg wit="#BnF155 #Ars5059">tombe</rdg>
  </app>
  la ou il fust
</p>
```

This encoding imposes no base text, is compatible with an open *recensio*, and enables decomposition into minimal operations essential for quantitative stemmatology.

3 Results and Discussion

All three methods converge on four stable clusters:
$\{Maz312, BnF152, Bein129\}$, $\{BR987, JenEl9596\}$, $\{BnF155, Ars5059\}$, $\{BL19DIII\}$.

The contrast between the primitive cluster of the *Première* and the revised cluster of the *Seconde Édition* is clear. BnF155 is consistently followed by Ars5059, corresponding to the β–γ families;

BL19DIII matches the α branch of the *Bibles à prologues*.

The methods diverge in the internal configuration of the *Première Édition*. UNJ most closely reflects the qualitative analysis of the tradition: Maz312 and Bein129 form a subgroup opposed to BR987–JenEl9596, with BnF152 slightly more external. NJ is close, but separates Bein129 from Maz312–BnF152. MP, in contrast, clusters JenEl9596 with Bein129 against the nucleus made up of Maz312, BnF152, and BR987. Maz312 and BR987 are regarded as the most conservative witnesses of the primitive recension, and their position will require further investigation on the broader corpus. Bootstrap proportions [6] confirm the stability of the main clusters.

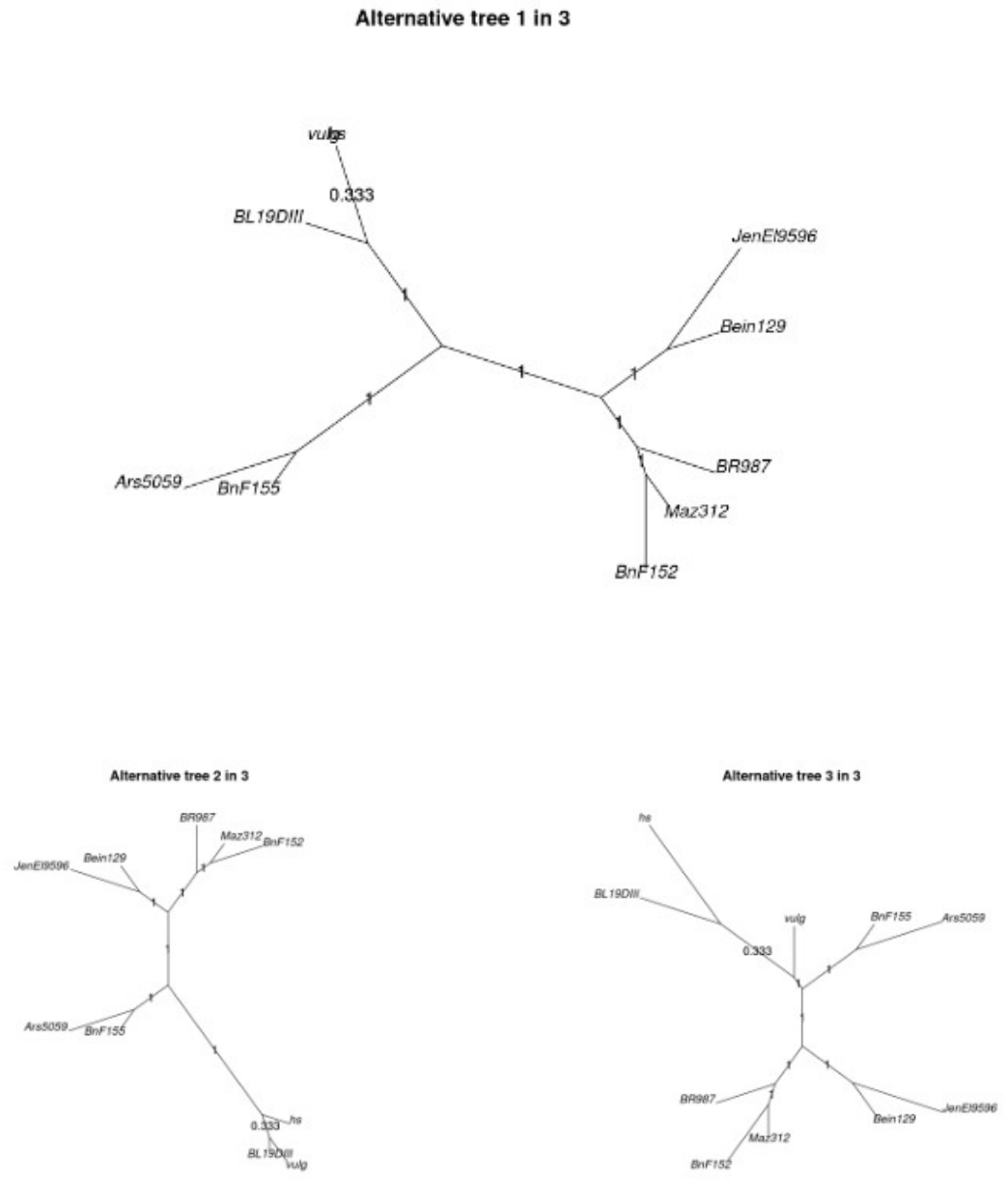


Figure 1: Three equally parsimonious MP trees. Edge values = bootstrap proportions [6]. Only the placement of **hs**, **BL19DIII**, and **vulg** differs.

UNJ proves the most promising method, recovering all major groupings supported by the qualitative analysis of the tradition. The affinity between JenEl9596 and Bein129 — the two latest copies of

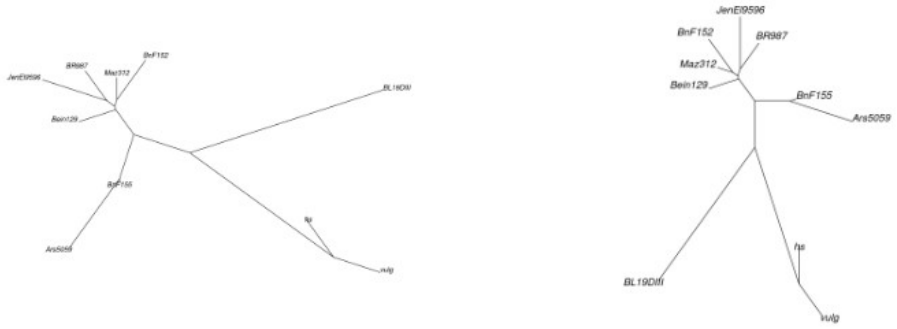


Figure 2: Unrooted NJ (left) and UNJ (right) trees based on ML distances.

the *Première Édition*, captured together by MP and partially by NJ — warrants further investigation, and may find compatibility with Michel's identification of both witnesses as contaminated from the same source in the Genèse [10]. Future work will extend the sample to 45 witnesses.

3 Results and Discussion

The source data and code are available at
`https://github.com/bh-anon-repo/stemmatology-exodus`.

References

[1] Berger, S. (1884). *La Bible française au Moyen Âge*. Vol. 1. Imprimerie nationale.
[2] *Biblia Sacra Vulgata*. (1592). Editio Romana. Editio Clementina promulgated by Clement VIII.Roma: Typographia Apostolica Vaticana.
[3] Buzzoni, M. et al. (2016). Open versus closed recensions (Pasquali): Pros and Cons of Some Methods for Computer-Assisted Stemmatology. *Digital Scholarship in the Humanities* 31(3), 652–669.
[4] Camps, J.-C. and Cafiero, F. (2018). Stemmatology: an R package for the computer- assisted analysis of textual traditions. *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities (CRH-2)*, 65–74.
[5] Clérice, T. et al. (2024). CATMuS Medieval: A Multilingual Large-Scale Cross- Century dataset in Latin Script for Handwritten Text Recognition and Beyond. *2024 International Conference on Document Analysis and Recognition (ICDAR)*, 174–194.
[6] Felsenstein, J. (1985) Confidence limits on Phylogenies: An Approach using the Bootstrap. *Evolution* 39(4), 783–791.
[7] Gascuel, O. (1996). Concerning the NJ algorithm and its unweighted version, UNJ. *Mathematical Hierarchies and Biology* 37, 149-170.
[8] Kiessling, B. et al. (2019). eScriptorium: An Open Source Platform for Historical Document Analysis. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 19-19. doi: `10.1109/ICDARW.2019.10032`
[9] Komada, Akiko. (2004). *Les Illustrations de la Bible historiale: les manuscrits réalisés dans le Nord*. PhD thesis. Paris 4.
[10] Michel, B. (2004). *La "Bible historiale" de Guiart des Moulins : édition critique de la Genèse d'après le manuscrit Bruxelles II 1987*. Supervisor: Jean-Marie Fritz. PhD thesis.Université de Dijon.
[11] Migne, J. P. et al. (1855). *Adami Scoti Canonici regularis ordinis praemonstratensis Opera omnia… accedunt Magistri Petri Comestoris Historia Scholastica*, *Sermones… Godefridi Viterbiensis*

Chronicon... Vol. 198.
[12] Nixon, K. C. (1999). The Parsimony Ratchet, A New Method for Rapid Parsimony Analysis. *Cladistics* 15(4), 407–414.
[13] Pasquali, G. (1952). *Storia della tradizione e critica del testo*. Felice Le Monnier, 1952.
[14] Roelli, P. (2020). *Handbook of Stemmatology: History, Methodology, Digital Approaches*. De Gruyter.
[15] Saitou, N. and Nei, M. (1987). The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution* 4(4), 406–425.

The Other Dinner Party: Representing Women in Archaeology through an Immersive Digital Installation

Valentin Cotta P

Trajectoires - UMR 8215, Université Paris 1 Panthéon-Sorbonne

Since the creation of Wikipedia in 2001, a significant portion of contemporary biographical production has been developed collaboratively. While this open encyclopedia has become an essential tool for accessing knowledge, it also reflects, through its structure and content, the persistent systemic biases regarding the visibility of women in scientific disciplines. Archaeology is no exception: many women (pioneers, researchers, volunteers, or technicians) remain under-represented, poorly documented, or even completely absent.

In this context, I have developed an algorithm for extracting, classifying, and relating biographical data sourced from Wikidata, Inrap or VIAF, aimed at cataloging all the women who have contributed to archaeology since the 19th century [1]. The tool analyzes several hundred pages, identifies relevant categories, extracts available metadata (related to countries, periods, specialties, institutions, publications), and thus generates a structured corpus allowing for the objectification of the presence and diversity of female trajectories within the discipline. Fictional women archaeologists are also taken into account.

This collection work serves as the foundation for a digital installation in 3D virtual reality, inspired by Judy Chicago's The Dinner Party (1974-79)[2]. The original work, an icon of feminist art, sought to restore a symbolic place to women erased from history. The installation I propose transposes this gesture into the archaeological field: in an immersive space, each identified woman is represented by a digital "place," an artifact, a text, or a visual trace that materializes her contribution. The visitor can navigate within a circular environment enhanced by dynamic networks of connections, making visible communities, moments of rupture, specialties, and strategies for adaptation and sociability. This experience is also available through a standard desktop internet browser.

Beyond an artistic approach, this project serves as a sociological analysis tool: it illustrates the distribution of careers, the diversity of roles (for example, excavation, documentation, post-excavation, conservation), the evolution of institutional recognition, as well as the persistent gaps in the digital memory of the discipline, which are emphasized once statistics are made. As a true learning-through-play tool, it proposes a crossover between data feminism, the history of science, digital epistemology, and immersive mediation. VR cardboard headsets are provided alongside the poster for symposium attendees to test. Finally, this work presents an opportunity to introduce the #Wiki4Women event and invite contributions to Wikipedia through the 'Women in Red' project.

References

[1] Cotta, V. (2026). Les archéologues invisibles : recenser, représenter et rendre visibles les femmes de l'archéologie par une installation numérique immersive - Archéologie au féminin III. `https://www.youtube.com/live/qQB4WCL2kZg?si=CHxF4AFt\T1\textbackslash_Sr-eAV4`

[2] Chicago, J. and Borzello, F. (2014). The Dinner Party - Judy Chicago: Restoring Women to History. Monacelli Press.

Lightning talks I: Modelling Cultural Artefacts

Digitized, Transcribed, Analyzed, and Published Excavation Archives for an Emotional History of Archaeology

Christophe Tufféry[1,2]

LT

[1] Ministère de la Culture
[2] UMR 8068 Technologie et Ethnologie des Mondes Préhistoriques (CNRS)

The Rivaux site, located in Espaly Saint-Marcel (Haute-Loire), was excavated almost continuously between 1970 and 1990 under the direction of Jean-Pierre Daugas (1946–2011). The site primarily dates to the Middle Neolithic period, and more specifically to the Chasséen culture. As part of our PhD research, we were able to consult and digitise part of the site's archives. We digitally transcribed the textual content of the excavation notebooks, testing several artificial intelligence (AI) solutions (Transkribus, Kraken, Gemini). We then carried out various types of analyses: lexicometric, textometric, 3D visualisation on ArcheoViz, etc. The observations recorded in the transcribed excavation notebooks and certain archaeological finds were spatialised by grid square. This work was initially carried out using GIS software, then in the form of a web application based on an AI solution. The study of the excavation diaries revealed that field notation practices evolved over the course of the excavation campaigns. Other primary and secondary sources (interviews, personal memoirs, other archives) were used to supplement the documentary corpus. Together, these archives have enabled us to present an emotional history of the excavation site and to reconstruct part of social life over two decades, whilst acknowledging the subjectivity of the author of this research. In addition to our PhD dissertation [1], some of the archives used for this research have been published under a free license on the Nakala platform. A website [2] and several articles provide an overview of our work and new research in progress.

References

[1] Tufféry, C. (2023). Ce que le numérique fait à l'archéologie et aux archéologues.Contribution historiographique et épistémologique l'étude des évolutions d'une discipline et de ses pratiques en France depuis les années 1970. Thèse de doctorat Patrimoine. Études patrimoniales sous la direction de Julien Longhi et Boris Valentin, soutenue le 12 décembre 2022, CY Cergy Paris Université. `http://www.theses.fr/2022CYUN1129/document`
[2] Tufféry, C. (2025). En fouillant dans les archives de fouille des Rivaux. Contribution à la reconstitution de l'histoire et de la vie du chantier de fouilles des Rivaux à partir de ses archives (1970-1990). `https://arcg.is/10XDPm1`

Phyto-Vision: A Reproducible Workflow for the Computational Excavation of Global Botanical Iconography

Guang Yang LT

University College Cork

This lightning poster presents Phyto-Vision, a reproducible workflow for the computational analysis of botanical iconography across heterogeneous global archives. Situated within Digital Plant Humanities, the system integrates metadata normalisation, archival deduplication, and automated illustration detection using YOLOv11. Trained on a manually annotated and culturally diverse corpus of botanical imagery, the model achieves robust performance across varied historical print styles. By translating historical image–text structures into machine-readable form, Phyto-Vision enables large-scale, cross-cultural analysis of visual standardisation in botanical works. The project contributes to the history of botany by providing a computational framework for examining how illustrated pages structured and circulated botanical knowledge across imperial and non-European traditions.

References

[1] Bleichmar, D. (2012). *Visible Empire: Botanical Expeditions and Visual Culture in the Hispanic Enlightenment*. University of Chicago Press.
[2] Jocher, G., and Qiu, J. Ultralytics YOLOv11 (Version 11.0). [Computer Software]. `https://github.com/ultralytics/ultralytics`
[3] Kaoua, R. et al.(2021). Image Collation: Matching Illustrations in Manuscripts. *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, 351–366.
[4] Mei, X., et al. (2025). *ZuantuSet: A Collection of Historical Chinese Visualizations and Illustrations. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–15.
[5] Nickelsen, K. (2006). *Draughtsmen, Botanists and Nature: The Construction of Eighteenth-Century Botanical Illustrations*. Springer. doi:`10.1007/978-1-4020-4820-3`.
[6] Smits, T., and Wevers, M. (2023). A Multimodal Turn in Digital Humanities: Using Contrastive Machine Learning Models to Explore, Enrich, and Analyze Digital Visual Historical Collections. *Digital Scholarship in the Humanities* 38(3), 1267–1280. doi: `10.1093/llc/fqad008`.
[7] Zhao, Z., and Liang, Z. (2017). The Original Source of Modern Research on Chinese Medicinal Materials: Bencao Texts. *Journal of Alternative, Complementary and Integrative Medicine* 3(4). doi: `10.24966/ACIM-7562/100045`.

Looking at the Eyes: A Computational Study of Eye Representation in Paintings.

Chloé Jollivet-Courtois[1], ***Marion Charpier***[1], ***Daniel Stockholm***[2]

[1] École nationale des chartes - PSL
[2] École Pratique des Hautes Études

LT

The increasing availability of digitized artworks and detection models opens new possibilities for building structured visual datasets and exploring cultural artefacts beyond traditional stylistic classification. In this context, faces appearing across a wide range of paintings—regardless of genre—offer a rich domain for investigating how localized visual elements evolve across artistic practices. The depiction of eyes, long considered central to expression and presence, provides a promising entry point for studying stylistic variation through computational means

We developed a large-scale processing workflow based on a refined combination of two YOLO models: one dedicated to face detection and a second trained specifically to localize eyes within detected faces. This hierarchical detection strategy improves robustness in the context of painted imagery, where stylization often challenges standard facial detection approaches.

In a first stage, we rely on convolutional deep learning models and their latent representations to explore visual structures without predefined stylistic or genre-based labels. Using dimensionality reduction and clustering techniques, we investigate whether patterns in eye depiction emerge directly from the data and whether historical shifts in representation may be reflected in localized anatomical variation across paintings.

In a second stage, we introduce an interpretable feature-based analysis to support the interpretation of these latent structures. We extract morphometric descriptors such as relative eye size, interocular distance, symmetry, orientation, and gaze alignment, allowing us to relate patterns observed in the latent space to explicit geometric properties of eye representation.

By articulating latent modeling with interpretable descriptors, this study contributes to cultural analytics by combining automated dataset construction with multi-level analysis, and illustrates how computational methods can support the study of visual representation in large cultural image collections.

Statistical Modeling of the Wave Model for the Spread of Innovations

Grégoire Clarté LT

School of Mathematics, University of Edinburgh

We propose a mathematical formalisation of the "wave model" to reconstitute the spread of innovations between populations. The model we present, WaST, represents interactions between the populations as a metric graph, whose vertices are the populations and edge length their capacity to interact. At the vertices of the graph, innovations appear and spread to neighbouring populations. Innovations can propagate or not according to a random process parametrized by the edge length and can then disappear in any given population with time.

The goal is then to infer the graph in which the spread of wave occurs and the parameters of the evolution (rate of disparition of the innovations mostly). We propose numerical methods to infer those parameters in a fully Bayesian setting, and test the method on several datasets among which Greek dialects [1], North Vanuatu [2]. The method reconstructs the generally recognised dependency structures, the novelty of the approach being in its formalism and theoretical bases, which ensures quantitative results and estimation of uncertainty. This model is only a first iteration, and many simplifying assumptions have been made to ensure computational feasability.

The paper can be found at `https://arxiv.org/html/2604.08220v2`.

References

[1] Skelton, C. M. (2015). Borrowing, character weighting, and preliminary cluster analysis in a phylogenetic analysis of the ancient greek dialects. *Indo-European Linguistics* 3(1), 84–117.
[2] Kalyan, S. and François, A. (2018). Freeing the comparative method from the tree model: a framework for historical glottometry. *Senri Ethnological Studies* 98, 59–89.

Towards an Atlas of Narrative Motifs across the Pacific

Sandrine Bessis[1], ***Alexandre François***[1], ***Luc Massip***[1], ***Sébastien Christian***[2]

[1] Langues, Textes, Traitements informatiques, Cognition (Lattice), CNRS – École Normale Supérieure – Sorbonne Nouvelle
[2] Université de la Polynésie française

LT

The Pacific is home to nearly 1,500 languages, including 1,420 in Melanesia, representing nearly a quarter of the world's linguistic heritage. However, with cultural globalization, these languages are being absorbed ever more rapidly by larger ones, and with them the cultural traditions that they have preserved for generations. Yet, when we are fortunate enough to collect them in time, the oral traditions of Oceanic cultures provide us with treasures – valued by their communities, but also as part of humanity's intangible heritage.

One of the aims of the HéLiCéO (`https://heliceo.huma-num.fr/axes.htm`) project is to create a database for Pacific oral literature. Its objectives will be to inventorise what already exists in terms of documentation and publications; and also, to identify what areas remain to be documented. A related project is to analyse existing narratives into "motifs" (e.g. *rising into the sky*) and "tale types" (e.g. *Fighting the ogre*). Rather than use reading grids that were defined around European traditions (e.g. the 1961 Aarne-Thompson index of folktales), we hope to identify those narrative elements bottom-up, so as to recognise what patterns are specific to the Pacific.

The first steps in analysing narratives were taken manually, by reading and summarising stories, before identifying individual motifs one by one. One possible alternative would be to use machine learning to read and summarise large numbers of texts, and/or suggest a list of potential narrative motifs. With this talk, our hope is to open a discussion with specialists in computational linguistics, to assess what can be realistically done with current tools.

References

[1] Aarne, A., and Thompson, S. (1961). *The Types of the Folktales. A Classification and Bibliography*. Helsinki: Suomalainen Tiedeakatemia.
[2] Bessis, S. (2022). Corpus of audio recordings and annotated texts in the Namakura language. Pangloss Collection. CNRS Huma-num, Paris. `https://pangloss.cnrs.fr/corpus/Namakura`.
[3] Bessis, S. (2023a). *Souvenir des ancêtres et histoire orale au Vanuatu : Les récits de chefferies anciennes aux îles Shepherd*. PhD Dissertation in Linguistics, Université Sorbonne nouvelle, Paris.
[4] Bessis, S. (2023b). Kuwae: A living Tradition of the Shepherd Islands. Paper read at the Vanuatu Language Conference (VLC).
[5] François, A. (2018). In Search of Island Treasures: Language Documentation in the Pacific. *Language Documentation and Conservation Special Publications* 15, 210–223.
[6] François, A. (2019). Comparative narratology in Oceanic folk traditions: A pilot study from north Vanuatu. Paper read at 11th Conference On Oceanic Linguistics (COOL11).
[7] François, A. (2021). Fonds Alexandre François: Archive of fieldwork recordings made in Vanuatu and the Solomon Islands.`https://tiny.cc/Francois-archives`.

Lightning Talks II: Manuscripts and Stylometry

Building an HTR Model for Medieval Swedish Manuscripts Transmitting Old Norse Romances

Paola Peratello LT

École nationale des chartes - PSL

The transmission and preservation of medieval literature is a complex process shaped by linguistic, cultural, material, and social factors. Old Norse romances, including *riddarasögur* (chivalric sagas), *fornaldarsögur* (legendary sagas) and *rimur* (related narrative poems), represent a key literary tradition that spread from Iceland across Scandinavia during the Middle Ages. Many of these works have been lost over time, while others survived in medieval Danish and Swedish manuscripts [3]. During the Middle Ages, some Old Norse texts were translated into Old Swedish and Old Danish: this points to a shared courtly literary culture across the medieval Nordic regions, in which chivalric narratives circulated and were adapted to meet the tastes of Scandinavian aristocratic audiences.

This two-year postdoctoral research is part of an ANR-funded project "Digital Approaches to the Survival and Loss of Old Norse Romances" (PI: CPJ Katarzyna Anna Kapitan). The first phase of the project involves applying HTR methods in eScriptorium [7] to a corpus of Old Swedish handwritten sources spanning from the beginning of the 14th century to the first half of the 16th century. The digital surrogates used are available on manuscripta.se [6] and Handrit.is [5] under a Public Domain licence. Notably, no HTR models are currently available for Old Swedish: for this reason, this phase builds on the methodological framework of *Bifrost* [1,2], an HTR model for Old Norse, published under a CC BY license. The dataset consists of 4 medieval manuscripts (2 in *cursiva gothica recentior*, 2 in *gothic hybrida/semicursiva*), 10 pages (5 folios) for each manuscript for a total of 40 pages, and between 35 and 40 lines per side of the folio. The finetuning of the model, along with the obtained performance will be briefly addressed during the workshop, as the project is still in progress. The dataset and the model will be made freely available online upon completion of the project.

References

[1] Kapitan, K. A., and Vidal-Gorène, C. (2025a). Bifrost: A Handwritten Text Recognition Model for Old Norse. Released. Kraken. `https://doi.org/10.5281/zenodo.15366732`.
[2] Kapitan, K. A., and Vidal-Gorène, C. (2025b). Crossing the Bifrost Towards an Open Access FAIR HTR Model for Old Norse Manuscripts'. HTR ground truth. `https://doi.org/10.5281/zenodo.15366896`.
[3] Kapitan, K.A. (2024). *Lost but Not Forgotten: the Saga of Hrómundur and Its Manuscript Transmission*. Taylor Institution Library.
[4] Kapitan, K.A. (2019). A Danish Collection of Old Norse Sagas: Material-Philological and Textual Studies of Acc. 61. *From text to artefact: Studies in honour of Anne Mette Hansen*, ed. Katarzyna Anna Kapitan, Beeke Stegmann and Seán Vrieland, p39–46.
[5] Handrit.is. `https://handrit.is/`.
[6] Manuscripta.se. `https://www.manuscripta.se/`.
[7] Stokes, P. et al. (2021). The eScriptorium VRE for Manuscript Cultures. Classics@ Journal, 18 (Ancient Manuscripts and Virtual Research Environments). `https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/`

Document-Based Modelling of Medieval Texts: Between Philology and Interoperable Cultural Data

Katarzyna Anna Kapitan LT

École nationale des chartes - PSL

Many studies have shown that Automatic Text Recognition (ATR) for manuscripts offers significant potential for large-scale, data-driven research into textual traditions. For medievalists working with Old Norse (ON) material, however, the promise of machine-readable corpora must be balanced against the specific philological requirements. A key point of reference in this landscape is the body of editions produced within the Medieval Nordic Text Archive (MENOTA) framework [1]. MENOTA editions provide a crucial starting point for digital ON scholarship and model a structured approach to encoding medieval Nordic texts. At the same time, they reveal the heterogeneity inherent in scholarly transcription practices: editorial principles, levels of normalisation, and representation of manuscript features differ across projects. While this diversity reflects legitimate research priorities, it also complicates the development of unified datasets suitable for AI-based analysis.

Old Norse manuscripts illustrate particularly well how philological detail intersects with computational requirements. Orthographic variation and the use of special characters, such as thorn (þ), eth (ð), or specialised abbreviations, carry important information for linguistic analysis and manuscript dating. Yet predictive text models and ATR pipelines often favour simplified character sets or normalised forms in order to improve recognition accuracy. This creates a methodological tension: preserving philologically meaningful detail while ensuring that computational tools remain functional and adaptable. One challenge is therefore to design transcription and prediction workflows that remain useful and accessible for individual researchers while still producing data that can be reused in broader infrastructures.

This paper presents one such a pipeline, following the recent developments in pre-trained Handwritten Text Recognition (HTR) models which have begun to mitigate these problems. Generic medieval models developed on platforms like Transkribus and eScriptorium provide robust baselines that can be fine-tuned with relatively small training datasets tailored to a specific manuscript tradition. Importantly, such workflows no longer require specialised computing infrastructure: fine-tuning and inference can be performed locally on consumer hardware. This paper explores how personalised HTR pipelines can be combined with existing MENOTA-based resources to support both the detailed philological work of individual scholars and the long-term goal of interoperable digital corpora for Old Norse studies.

References

[1] Medieval Nordic Text Archive. `https://menota.org`.

New Models of Automatic Transcription for Medieval Latin Manuscripts and 19th-century Italian Archival Documents

Denise Ugliano, Stefano Giustino, Lorenza Laccetti, Sohail Maqsood, Yahya Momtaz LT

Università degli studi di Napoli Federico II

Digital Humanities are increasingly moving beyond passive preservation toward active digital restoration recovering not just the physical form of a document but its legibility and cultural meaning. Ancient manuscripts present some of the most challenging material in this regard. Over centuries,chemical agents, humidity, and biological degradation compromise ink and parchment to the point where entire lines of text or marginal annotations become practically unreadable. Recovering even partial legibility opens significant new avenues for scholarship.

The first phase of this project involved digitising documents, manuscripts, and early printed books across the participating institutions. Raw scans alone, however, are rarely sufficient particularly for manuscripts affected by bleed-through, where ink from one side of a leaf interferes with the text on the other. Image enhancement followed a multi-stage adaptive pipeline. Paper background was estimated using morphological closing (r = 60 px) with Gaussian smoothing (= 25). A normalised darkness map was derived via local contrast normalisation relative to this background. Pixel sharpness was quantified by a composite descriptor (gradient magnitude, local standard deviation, Laplacian) applied to the darkness map, isolating ink from paper texture. A Euclidean distance transform from detected text edges defined a spatial influence zone. Sharpness and proximity drive an adaptive threshold: $\dot{T}(x,y) = T_high - \alpha * z(x,y) - \beta * S(x,y)$. This assigns each pixel an individualised boundary, suppressing blurry bleed-through while preserving sharp text. Residual artefacts are removed via three iterations of fixed-point correction.

The approach extends Sauvola's [2] and Niblack's [1] thresholding techniques by adding a sharpness dimension. The cleaned images feed directly into Transkribus, where manually produced transcriptions are aligned at line level to create training data for two dedicated HTR models one for 19th century scripts (like the "Copialettere" and other documents from the State Archives of Caserta), and another model for Beneventan script, which we will use to transcribe manuscripts from the monastery on Montecassino.

References

[1] Niblack, W. (1986). *An Introduction to Digital Image Processing*. Prentice-Hall.
[2] Sauvola, J., and Pietikainen, M. (2000). Adaptive Document Image Binarization. *Pattern Recognition* 33(2), 225-236. doi:10.1016/S0031-3203(99)00055-2.

Attributing Authorship in a Nineteenth-Century Atelier: Stylometry and Collaborative Practices around Alexandre Dumas

Florian Cafiero, Jean-Baptiste Camps, Denise Atzori, Juliette Grenier, Chloé Jollivet-Courtois **LT**

École nationale des chartes - PSL

In 19th literature, authorship attribution has often treated collaboration as noise to be filtered out of computational models. However, in the case of Alexandre Dumas, collaboration was structural rather than incidental: many of his novels and plays were conceived, drafted and revised within a workshop system involving partners such as Auguste Maquet, Paul Meurice, Gustave Cherville and Paul Bocage. Rather than seeking definitive verdicts about 'true authorship', this study adopts a layered model of literary production, attempting to distinguish between dossier preparation, drafting, orchestration and serial remediation.

Based on this framework, we compile a French-language corpus of novels associated with Dumas and his documented collaborators, using edition-homogeneous texts and detailed metadata (date, OCR quality and source) where possible. The analysis operates at the level of rolling windows rather than whole texts, enabling section-level attribution. We extract stylistic features (primarily functional words and affixes) and topic-based features to examine the interplay between narrative content and stylistic signal.

As the list of potential co-authors identified by previous literature might not be complete, we start by treating the problem as open set, thus performing unsupervised analyses (PCA/t-SNE and cosine-distance heatmaps) to reveal potential clustering inside the corpus.

Using the Python SuperStyl package [1], we then implement a rolling supervised classification task (linear SVM) to predict whether a text section is by Dumas or one of his potential collaborators. To disentangle the effects of style and topic, we compare models trained on style-only, topic-only and combined feature sets. We diagnose topic leakage by removing named entities and balancing windows per work.

References

[1] Camps, J.-B., Moins T., and Cafiero, F. (2024). SUPERvised STYLometry (SuperStyl) (Version v1.0) [Computer software]. `https://doi.org/10.5281/zenodo.14069799`.

Robust Rhymes or Stable Stresses? Finding the Authorial Features Most Resilient to Variation

Valentina Modolo[1,2]**,** ***Théo Moins***[1]**,** ***Marie Bizais-Lillig***[1,3]**,** ***Cecile Vermaas***[1]**,** ***Jean-Baptiste Camps***[1]**,** ***Mathias Garnier***[1] LT

[1] École nationale des chartes - PSL
[2] Scuola Normale Superiore di Pisa
[3] Université de Strasbourg

Authorship attribution in medieval French literature poses specific methodological challenges due to manuscript transmission, orthographic variation, and the instability of textual corpora. While computational stylometry has shown that such texts can be analysed using supervised methods such as Support Vector Machines, the reliability of stylometric analyses depends critically on the robustness of the features employed. In this perspective, Kestemont, Dalemans and Sandra [1] suggested that metrical constraints, rather than surface lexical forms, may preserve authorial signals more effectively. Building on this insight, recent work by Camps et al.[2] has shifted the focus toward prosodic features, using stress patterns as quantifiable stylistic markers. Approaches based on prosodic annotation, such as the Metronome framework [3], allow rhythmic structures of medieval verse to be modelled independently of lexical variation. The present study situates itself within this line of research and examines the behaviour of stress-based features under conditions of textual variation in medieval octosyllabic verse.

The corpus consists of texts written by Chrétien de Troyes (ca. 1130-1190), Robert Wace (ca. 1100-1175) and Gerbert de Montreuil (born ca. 1200). For each author two texts are selected and for each text two witnesses (see appendix). Therefore it is possible to compare relative distances between different witnesses, texts and authors. The samples taken from each witness overlap in halfth their length with the other witness from the same text, to enable us to see the influence of scribal intervention.

To process the corpus we transcribe three hundred verses of each witness and apply lemmatization and POS-tagging to the selection. Finally stress patterns are created using Anochre [4] and the distances between the stress patterns of each witness are computed (see appendix for examples).

Corpus processing on overlapping segments reveals that lexical features and rhymes are highly sensitive to textual variation and therefore poorly suited for capturing authorial signals in medieval verse. By contrast, stressed-based features display a remarkable stability across witnesses and texts, confirming their resilience to scribal intervention and supporting their relevance as a foundation for authorship-oriented analyses. On the other hand, the experiments based on non-overlapping segments suggest a less clear-cut authorial fingerprint.

References

[1] Kestemont, M., Daelemans, W., and Sandra, D. (2012). Robust Rhymes? The Stability of Authorial Style in Medieval Narratives*. *Journal of Quantitative Linguistics*, 19(1), 54–76. `https://doi.org/10.1080/09296174.2012.638796`
[2] Camps, J.-B.et al. (2025a). Style in Eight Syllables: Metric Annotation and Stylometry of Chrétien de Troyes and Contemporaries. Digital Humanities Benelux 2025 Conference, 1-12.
[3] Nagy, B. et al. (2025). Metronome: Tracing Variation in Poetic Meters via Local Sequence

Alignment. *Computational Humanities Research*, 1(1).doi:10.1017/chr.2025.1
[4] Camps, J.-B.et al. (2025b).Anochre. https://github.com/PoidsPlume/AnoChre.

Appendix

List of corpus

Table 1: Composition of the corpus (all texts are romances in octosyllabic verse).

Author	Title	Date (Work)	Manuscript(s)	Date (MS)
Chrétien de Troyes	*Erec et Enide*	ca. 1170	fr. 01376, BnF; fr. 24403, BnF	1288; mil. XIII c.
Chrétien de Troyes	*Cligès*	ca. 1176	fr. 1374, BnF; fr. 01450, BnF	mil. XIII c.; 1st q. XII c.
Robert Wace	*Rou*	ca. 1170	BnF, fr. 794 (C)	c. 1235
Robert Wace	*Brut*	ca. 1155	Ottob. Lat. 1869 (Vatican Library); BL Add. 45103 (British Library)	XII c.; XIII c.
Gerbert de Montreuil	*Violette*	ca. 1227–1229	fr. 1553, BnF; fr. 1374, BnF	1284; mil. XIII c.
Gerbert de Montreuil	*Continuation*	start XIII c.	fr. 12576, BnF; nouv. acq. fr. 6614, BnF	beg. XIII c.; XIII c. (frag.)

Numerical analysis

The two plots below are preliminary results to support the arguments. See https://github.com/Bizais-Lillig/robust_rythm for the code and more detailed examples.

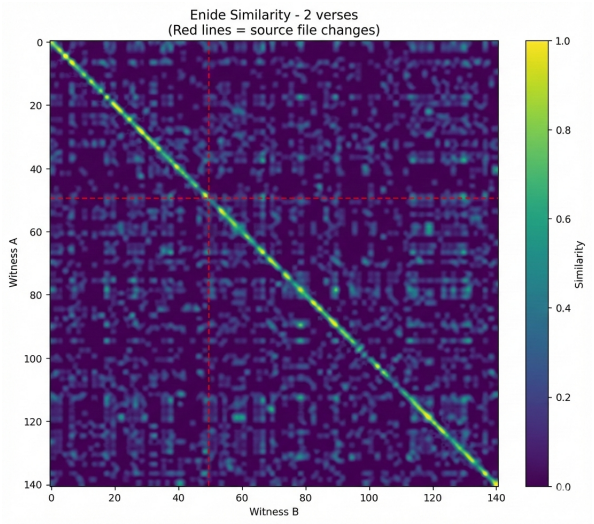


Figure 1: Similarity between witnesses for *Erec et Enide*. A high intensity on the diagonal means proximity between the two witnesses.

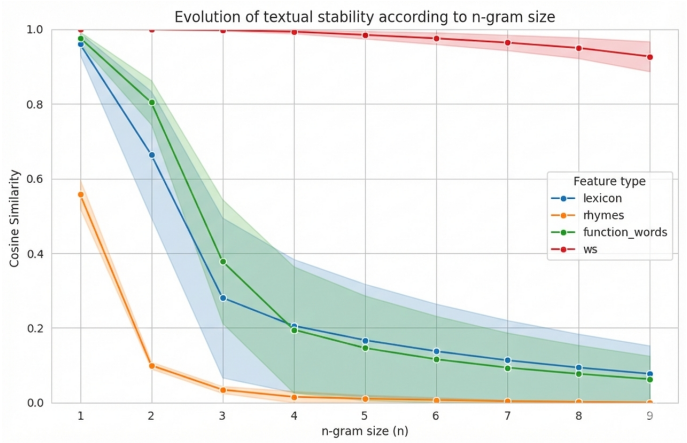


Figure 2: Looking for the most stable feature in Wace's opera. It appears that the weak and strong patterns are the most stable feature for Wace, as for the other authors.

Automatic Metrical Scansion for Comparative Research on an Emergent Tradition

Pablo Ruiz Fabo[1,2]**, *Anxo Alonso***[1]**, *Pablo Gamallo***[1]

[1] Centro Singular de Investigación en Tecnoloxías Intelixentes
[2] Université de Strasbourg – LiLPa UR 1339

LT

1 Context

Our project seeks to perform a large-scale analysis of 19th-century poetry in Galician, a Romance language close to Portuguese and Spanish. The extant record consists of ca. 165,000 lines (based on our analysis of the *Tesouro informatizado da lingua galega*[7]).

Galician had prestigious literary production in the Middle Ages, shared with Portuguese. After centuries of decreased activity, its use in cultivated and literary settings increased towards the second half of the 19th century, with the *Rexurdimento* (reemergence) cultural movement [8].

We are interested in the features that Galician poetry as an emergent 19th-century literature adopted, comparing a large sample to corpora in the two traditions surrounding it, Portuguese and Spanish, inspired by work like Nagy et al. [4]. Comparative studies exist linking individual Galician authors to other European traditions, e.g. Carballo Calero [1, 2], but no large-sample studies.

A challenge for Galician is the lack of resources supporting a computational approach. We started by developing automatic metrical scansion tools and a TEI-encoded corpus (Section 2).

We would like to outline our experience developing first automatic scansion systems from scratch for a tradition for which no related resources preexisted, and interact with the community about good practices for this research scenario.

2 Automatic Metrical Scansion

Our first scansion system was rule-based, with ca. 89% accuracy at exact per-line stress-pattern match, tested on 800 lines with metrical variety [6]. This helped annotate a 4,000 example corpus to fine-tune neural models: two encoder-decoders (mT5 and ByT5, cf. Glaser [3]) and one LLM

(decoder only). The encoder-decoder systems (but not the LLM) attained higher accuracy than the symbolic one (>90-93% in most setups)[5]. However, the symbolic system outperformed all neural models with rare metrical licenses infrequent in training data (dieresis and syneresis). Thus, both approaches are complementary and can be combined. Besides, although LLMs outperform earlier architectures at challenging semantic tasks, for our scansion task, largely form-oriented, smaller encoder-decoder models performed significantly better than the fine-tuned LLM.

We TEI-encoded 35,000 poetry lines and will present metrical trends based on automatic scansion.

Acknowledgements
This work was supported by the European Union, under the Marie Skłodowska-Curie Actions, HORIZON MSCA-2023-PF, Grant ID 101149659, COMPEL – Computational Analysis of Peripheral Literatures.
The work was also supported by Xunta de Galicia – Consellería de Cultura, Educación, Formación Profesional e Universidades (Centro de investigación de Galicia accreditation 2024–2027 ED431G-2023/04 and Reference Competitive Group accreditation 2022–2026, ED431C 2022/19) and by the European Union's European Regional Development Fund – ERDF.
We are grateful to Elisa Fernández Rei (Instituto da Lingua Galega, ILG), for full-text access to 19th-century electronic sources in Tesouro informatizado da lingua galega (Version 4.1), directed by Antón Santamarina, Ernesto González Seoane, and María Álvarez de la Granja (http://ilg.usc.gal/TILG/).

References
[1] Carballo Calero, R. (2020). *Contribución ao estudo das fontes literarias de Rosalía*, ed. by Real Academia Galega. 2nd ed. Real Academia Galega. (Original work published 1959]. doi:10.32766/rag.366.
[2] Carballo Calero, R. (1957).Rosalía y Otros. *Cuadernos de Estudios Gallegos XII* 36.
[3] Glaser, B. (2025). TrochAIc: Metrical Tools for AI Interpretability.*Anthology of Computers and the Humanities*, 3, 1438–1453. doi:10.63744/K9Mwiiszu7QL.
[4] Nagy, B. et al. (2025). Metronome: tracing variation in poetic meters via local sequence alignment. *Computational Humanities Research*, 1 (1).doi:10.1017/chr.2025.1
[5] Ruiz Fabo, P. et al. (2026) Automatic Metrical Scansion of Poetry in a Low-Resource Setting". *LLMs4SSH @LREC 2026: Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities*. doi:10.5281/zenodo.19701826.
[6] Ruiz Fabo, P., Moreau, P., and Pérez, A. A. (2026). Automatic Metrical Scansion of Galician Poetry: First Results". *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)*, 994–1004. https://aclanthology.org/2026.propor-1.101/.
[7] Santamarina, A., González Seoane, E., and Álvarez de la Granja, M. (2018). Tesouro informatizado da lingua galega. Version 4.1. URL: http://ilg.usc.gal/TILG/
[8] Vilavedra, D. (1999). *Historia da literatura galega*. Editorial Galaxia..

Codes
1. Rule-based system : https://github.com/compellit/gama-sym
2. Transformers-based systems: https://github.com/compellit/gama-trf

Long Papers II

Discursive Signatures: A Computational Method for Mapping How Meaning Varies Across Communities

Mark Hill LP

King's College London

This paper presents a computational method for studying how the meaning of cultural artefacts varies across online communities. Rather than treating topic model outputs as thematic labels, we propose using full topic-proportion vectors as discursive signatures that capture meaning at both the level of individual expression and the community contexts that shape it. By aggregating these signatures at the community level, validating groupings with PERMANOVA, applying hierarchical clustering, and projecting the resulting space via PCA, we produce a topology of discourse: a spatial representation of how meaning-making varies systematically across contexts. We demonstrate the method through a case study of mental health technology discourse on Reddit (289,039 posts across 445 subreddits), showing how the same technologies are conceptualised differently depending on community context. The method is designed to be transferable to any domain where researchers seek to understand how technologies, practices, or concepts are variably constituted across online spaces.

A NER Model for Science Fiction's Lexical Innovation

Mathilde Ducos, Frédéric Landragin LP

Langues, Textes, Traitements informatiques, Cognition (Lattice), CNRS – École Normale Supérieure – Sorbonne Nouvelle

Speculative fiction, including science fiction, is distinguished by its ability to generate an invented lexicon composed of terms called novums. However, these lexical units are absent from Named Entity Recognition (NER) models applied to computational literary studies, mainly because of their complexity in detection (highly varied syntactic distributions, strong semantic ambiguity). We present a CamemBERT-based NER model, adapted to the analysis of science fiction narratives and enriched with a dedicated class (NOV for novum) to identify these lexical inventions. Evaluated on a specific annotated corpus, our model achieves an F1-score of 63% for the detection of the very rare NOV class, with a macro F1-score of 79% across all entities. This model thus provides a tool for studying how these lexical creations reflect or even anticipate the technological and social imaginaries of their time. The resources, including the annotated corpus and code, are available online.

A Gueuloir of One's Own: Computing the Acoustic Signature of Flaubert

Simon Gabay[1], ***Florian Cafiero***[2,3], ***Jean-Luc Falcone***[1]

[1] University of Geneva
[2] EPITA
[3] Centre Jean Mabillon, École nationale des chartes - PSL

Can the traces of an ear-driven literary practice be detected in prose? This paper addresses the question through the case of Gustave Flaubert, whose gueuloir, his habit of declaiming sentences aloud and revising them by ear, has long been assumed to shape his writing. We compare Flaubert's novels against those of three contemporaries (Aimard, Chevalier, Ponson du Terrail). We first show that phonetic transcriptions consistently outperform orthographic text for authorship classification, suggesting that French spelling obscures part of the authorial signal and that phonemes are more reliable stylometric features than graphemes. Building on this finding, we examine suprasegmental representations to characterise each author's phonological profile. At the phonemic level, Flaubert displays a consonant-heavy profile: higher cluster density, a lower vowel-to-consonant ratio, and reduced bigram entropy point to dense, acoustically controlled prose. Rhythmically, his writing is also distinctive, with sparser stresses, longer unstressed runs, and more predictable local sequences, though rhythm alone provides a weaker basis for attribution. These results suggest that the gueuloir shapes both fine-grained acoustic texture and large-scale metrical structure, and make a broader case for phonetic-based stylometry as a tool for studying phonological phenomena in prose composition, especially for French.

Poster Session II [Students of the Master of Digital Humanities (ENC-PSL)]

A Two-Channel Pipeline for Similarity-Based Comparison of Bronze Inscription Rubbings Using Object Detection and Visual Similarity Analysis

Quanwen Long P

École nationale des chartes - PSL

Bronze inscription rubbings constitute a crucial category of visual documents for the study of Shang and Zhou writing systems, historical institutions, and material culture. Their scholarly value has depended on stable physical preservation conditions and on the interpretative skills of individual researchers. Prior to the widespread adoption of digital technologies, the preservation and use of rubbings were largely constrained by the fragility of their material support and by unstable environmental conditions. As paper-based artifacts, rubbings are highly sensitive to temperature, humidity, and light exposure. Excessive heat and uncontrolled humidity can lead to fiber degradation and mold growth, while prolonged exposure to light accelerates paper aging, progressively blurring or obscuring inscriptional traces. Moreover, many collecting institutions, particularly local museums and archival repositories, lack the financial resources and technical infrastructure required for precise long-term environmental control. Combined with frequent exhibition and manual handling, these constraints significantly increase the risk of irreversible physical deterioration and information loss.

At the level of scholarly research, these material limitations translate directly into systematic difficulties in the identification and comparative analysis of rubbings. Long-term weathering of the original bronze surfaces, together with uneven or poor-quality early rubbings, often results in incomplete or ambiguous character forms. Cracks, stone textures, soil residues, and other forms of visual noise frequently overlap with inscriptional strokes, substantially complicating manual reading, comparison, and paleographic interpretation. Furthermore, under traditional research paradigms, rubbings have tended to remain dispersed across institutions in their physical form. Issues related to copyright, access restrictions, and storage costs have limited researchers' ability to consult high-quality images and to conduct large-scale, cross- institutional comparisons. This fragmentation has constrained the systematic reconstruction of content relationships among rubbings and has more broadly restricted the dissemination of research results.

In response to these challenges, this study proposes a quality-aware, two-channel automated processing pipeline designed to support content-level comparison and alignment across heterogeneous bronze inscription rubbings, without relying on character transcription or manually annotated textual data. Rather than forcing all rubbings into a single processing pathway, the proposed approach explicitly accounts for differences in acquisition conditions and noise structures by first distinguishing input images at the visual level and routing them into adapted processing channels

Discursive Representations of Prostitution in the Francophone Press: Domain-Specific NLP, Spatial Analysis, and Web Network Cartography

Damien Conceicao P

École nationale des chartes - PSL

This poster presents a large-scale study of how prostitution is represented in Francophone news media, combining corpus linguistics, natural language processing, spatial analysis, and web network cartography. Focusing on 2000–2025, it examines three competing frameworks – regulationism, abolitionism/prohibitionism, and decriminalisation – through their preferred terminologies (e.g. “travailleur·euse du sexe”, “personne en situation de prostitution”), with particular attention to France, where the legal paradigm shifted twice (2002, 2006). Using a corpus of about 250,000 articles, a domain-specific NER schema and a hybrid spaCy–Stanza pipeline capture both lexical choices and syntactic roles of key actors. Spatial mapping and a hyperlink network (via Hyphe) then relate these discursive patterns to geographic focus, legal regimes, political alignment, and migration narratives.

French Television Channels' Framing of Immigration

Chloé Jollivet-Courtois P

École nationale des chartes - PSL

The 2015 refugee crisis and the increase in migration flows have made immigration a highly controversial issue in Europe. Immigration's place at the centre of the media arena has coincided with the rise of nationalist and populist movements [1,3]. Researchers have shown that increased media coverage of immigration on French television pushed moderate individuals to adopt extreme opinions[2,3]. Thus, when it comes to immigration, the quantity of content polarises opinions. But what influence does the media framing of immigration have on public opinion? This study examines the persuasive power of the media on public opinion regarding immigration.

The study is based on a cross-reference between the ELIPSS 'Dynamics of Mobilisation' (DYNAMOB) surveys (2013-2017) [4] and transcripts of programmes and television news broadcasts from TF1, France 2, France 3, M6, CNews and BFMTV, made available by the INA. Only news programmes and television news broadcasts identified as discussing immigration were included in the corpus.

Three questions in the survey provide an indicator of respondents' ideological position on immigration [3]. The Dynamob survey also includes a question on respondent's preferred television channel for viewing informative content. Together, these two indicators make it possible to cross-reference individuals' opinions with the content to which they have previously been exposed. The primary objective of the study is to observe changes in opinion over a short period of time in order to determine whether these changes coincide with those of the framing of the preferred channel. Does media exposure to a certain framing predict changes in opinion?

References

[1] Chouliaraki, L. and Zaborowski, R. (2017). Voice and Community in the 2015 Refugee Crisis: A Content Analysis of News Coverage in Eight European Countries. *International Communication Gazette*, 79(6-7), 613-635.
[2] Dennison, J. and Geddes, A. (2019). A Rising Tide? The Salience of Immigration and the Rise of Anti-Immigration Political Parties in Western Europe. *The Political Quarterly*, 90(1), 107-116.
[3] Schneider-Strawczynski, S., Valette, J. (2025). Media Coverage of Immigration and the Polarization of Attitudes. *American Economic Journal: Applied Economics*, 17(2),337-368.
[4] Tiberj, V., and Gougou, F. (2020).Dynamiques de Mobilisation - Vague 13 (ELIPSS 2016). doi:10.21410/7E4/RNEMLQ.

From China to Korea: Tracing a Divine Persona through Motif-Based Visual Analysis of the Yushu Baojing and Its Recompilations.

Tianjie Yin P

École nationale des chartes - PSL

Daoism is a native religious tradition of China organized around self-cultivation practices, ritual performances, and scriptural-register systems. Puhua Tianzun is a major figure in Daoist Thunder Ritual traditions and serves in the capacities of presiding over thunder, exorcism, and moral transformation. The Yushu Baojing is a central text associated with the teachings of Puhua Tianzun that conveys the cosmology and ritual logic of thunder rites. During the 16th to 18th centuries, the Yushu Baojing circulated across in China and Korea through a process of textual re-composition and pictorial re-production that routinely redesigned the divine persona of Puhua Tianzun. Illustrations across different editions create a representational thunder realm through visual components that are modular, separable, and replicable, and these same motifs occur in other media, including printed scriptures, ritual manuals and talismans. This study reconsiders the cross-edition transmission of Yushu Baojing imagery as a system of recombinable cultural units. Focusing on an eighteenth-century Korean re-edited version, it examines how Korean compilers inherited, trans- formed, and reconfigured visual vocabularies from earlier Chinese editions. The dataset is derived from a 16th century Chinese polychrome edition and an eighteenth-century Korean edition, comprising 27 illustrations. Using supervised learning and YOLO-based image detection, 198 visual motifs (225 image instances) were extracted to form a small, structured dataset.

Georgian Press in Exile (1921–1939): Automatic Text Extraction and Discourse Analysis

Nana Maglakelidze P

École nationale des chartes - PSL

This Master's thesis focuses on the automated analysis of Georgian diaspora newspapers in the 20th century, with particular attention to the critical period following the annexation of Georgia by the Red Army on February 25, 1921, despite its prior recognition as a sovereign state by Soviet Russia on May 7, 1920. In this respect, the present study focuses on the newspaper Independent Georgia (1926–1939), which functioned as the official voice of the Georgian government in exile and reported extensively on the activities of the emigrants' administration. The main objective of this thesis is threefold. First, it aims to develop a model capable of automatically identifying texts, authors, and titles in Georgian newspapers published in exile during the 20th century. Second, it seeks to determine the most suitable Optical Character Recognition (OCR) method for reliably extracting textual content from historical documents written in Georgian. Third, the extracted texts will be analyzed in order to better understand the dominant themes addressed during the period of Georgian emigration. This analysis will make it possible to observe the evolution of these themes over time and to identify the individuals most frequently mentioned in the articles, using Topic Modeling and Named Entity Recognition (NER) techniques.

Modeling Bourgeois Childhood: A Computational Study of Class and Gender in French Children's Literature (1833–1941

Denise Atzori P

École nationale des chartes - PSL

This poster outlines an ongoing research project exploring the portrayal of children in French children's literature published between 1833 and 1941. The main goal of the research is to examine the concept of childhood as depicted in books originally intended for the bourgeoisie. The focus is on how different social classes, particularly popular groups, and gender are portrayed in these stories. The study will analyse literary representations of childhood through the lens of the 'sentiment de l'enfance', a concept theorised by Philippe Ariès in L'Enfant et la vie familiale sous l'Ancien Régime (1960) and later synthesised by Nathalie Prince as 'the social, cultural and historical evolution of adult perceptions of childhood'. By applying natural language processing techniques, this research seeks to provide a quantitative and computational perspective on a topic that has traditionally been examined through qualitative literary analysis.

Patriarchal Structures in Postwar Film Criticism: A Comparative Computational Study of French and Italian Cultural Journalism (1945 1959)

Susanna De Luca P

École nationale des chartes - PSL

This research seeks to examine how film criticism published in French and Italian daily newspapers between 1945 and 1959 contributed to the construction of gender hierarchies within the cultural field. While scholarship has analyzed the emergence of the "politique des auteurs" and the consolidation of the director as a central figure of artistic authority, its gender implications require further investigation. The project envisages the selection of three French and three Italian newspapers. For France: Le Figaro, Combat, and the historical Libération. For Italy: L'Unità, Corriere della Sera, and La Stampa, considering that many postwar newspapers, although formally national, retained strong regional anchoring. The final corpus is estimated at between 300 and 800 articles and will be preceded by a pilot corpus (40 articles) to refine selection criteria. The computational methodology is defined in relation to the sub-questions and to the project's comparative, diachronic, and discursive nature. Comparative lexical analysis will constitute the foundation of the study, relying on normalized frequencies, keyword analysis, concordances, and collocation analysis. This approach will examine whether cinema was defined as "art" or "industry," whether artistic quality was associated with male directors, and how moral or national vocabulary varied over time.

Sailing in the Wind Preliminary Analyses of Slave Ship Routes in the Early 18th Century

Juliette Grenier P

École nationale des chartes - PSL

This study focuses on the logbooks of slave ships from the first half of the 18th century in order to examine the influence of weather and seasonal conditions encountered at sea on the route and duration of transatlantic voyages. This approach combines quantitative methods and historical analysis to illustrate the uncertainties associated with navigation at a time back when instruments were still imperfect. The study corpus consists of twenty logbooks from ships belonging to the *Compagnie des Indes*, written between 1733 and 1743. The *Compagnie* was among the leading shipowners involved in maritime commerce, and especially in the slave trade, during the 18th century. Each logbook contains an account of the exact course of the journey, with a daily report of events, varying in detail depending on the author. The writing style is very formal and has been regulated since 1681 – the date of the first ordinance regulating the keeping of records for all sea voyages – and leaves no place for anecdotes. The main purpose is to provide a series of standardized information such as wind strength and direction, sea, sky, and weather conditions such as storms and precipitation. The initial results highlight the fact that since navigation instruments were not yet perfected at the beginning of the 18th century, position readings were quite inaccurate.

Self-Presentation Strategies of Airbnb Hosts during the 2024 Paris Olympic Games: A Psychometric Zero-Shot Analysis

Ye Liu P

École nationale des chartes - PSL

This study investigates how Airbnb hosts present themselves through profile self-descriptions across market contexts surrounding the 2024 Paris Olympic Games, and how these strategies relate to short-term performance outcomes. Grounded in Goffman's impression management and signaling theory, the study quantifies self-presentation tactics using zero-shot classification of host profile texts. It then compares the distribution of these tactics across cities (Paris vs. London), time periods, and host types, and employs econometric modeling to examine how signaling strategies operate and interact under varying contextual conditions in short-term rental markets.

Spatio-temporal analysis of the suffix -(V)$\lambda(\lambda)$-

Gaëtan Drouet P

École nationale des chartes - PSL

This project investigates the spatial and temporal variation of the Greek anthroponymic suffix -(V)$\lambda(\lambda)$- across the ancient Greek world. By combining linguistic, geographical, and chronological data from two different databases and analysing them through computational and cartographic methods, it aims to identify patterns in the formation and diffusion of personal names and to highlight the potential of digital approaches for large-scale onomastic research.

Staged Hospitality and Everyday Tensions. Digital Traces of the Reception of Ukrainian War Refugees in Russia.

Maria Kirbasova P

École nationale des chartes - PSL

The war in Ukraine, which began in 2014 with an outbreak of armed conflict in the eastern regions of the country, has gradually led to a dynamic of forced migration, notably towards Russia. The large-scale invasion of February 2022 has greatly exacerbated these dynamics, intensifying conflict-related forced displacement and further sharpening the issue of forced migration from Ukraine to Russia. Although much of the academic research has focused on the movement of the Ukrainian population to European countries since 2022, forced migration to Russia remains relatively understudied. This imbalance is partly explained by limited access to this area, a lack of reliable empirical data, and the high political sensitivity associated with this subject of research. In this context, the reception of Ukrainian refugees in Russia is publicly framed through narratives of humanitarian aid, solidarity, and hospitality, emphasizing an organized and supportive reception infrastructure. However, digital traces on online platforms and testimony from independent journalists show a more complex and diverse reality. They point to practical difficulties and raise more sensitive issues concerning evacuation conditions and the fate of certain categories of displaced persons.

This apparent discrepancy raises the central question of the study: *To what extent does the comparison between Russia's state migration discourse and Telegram discussions among refugees and volunteers reveal tensions and difficulties in the reception process?*

This study answers this question by analyzing the gap between institutional narratives and the traces of this forced migration reproduced by migrants and volunteers, which reflect the real trajectories of the people involved.

The Representation of Women in Travel Narratives by Arab Authors: A Study Using Natural Language Processing Tools

Radjaa Benabdallah P

École nationale des chartes - PSL

This research focuses on a corpus of works by Arab authors in order to examine the portrayal of French women in travel literature. Natural Language Processing (NLP) tools are employed to support the manual identification and annotation of female figures in the texts. Once these figures are identified and annotated, sentiment analysis is conducted to assess how Arab writers perceive women in Europe.

Video Games as Literary Art: The Role of the Protagonist in Narrative Games, A narrative and Stylistic Study

Côme Parrinello P

École nationale des chartes - PSL

This Master's thesis contributes to ongoing discussions on the legitimacy of video games as an art form and, more specifically, on the place of narrative video games within literary studies. It first revisits the emergence of ludology, which focuses on the mechanics and structure of games, and points out how this approach often overlooks the narrative,visual, and emotional dimensions that make games immersive artistic experiences. Next, the study takes a more technical turn, aiming to demonstrate, through data analysis, that video game texts (dialogues, choices, descriptions) have distinctive features: a high frequency of imperatives, emotional variation, and direct interaction with the player. Computational tools are developed to detect these patterns, including models for identifying commands, sarcasm, and sentiment. It also aims at positioning narrative video games through a wide corpus next to other forms of literature in two different ways: stylistic hierarchy, ranking using word embeddings, machine and deep learning. It also provides a visual by thanks to combining content and stylistic similarities through unsupervised learning.

Speech Development in 1-year-olds: Vocal Maturity Social Interaction

Julie Duhesme P

École nationale des chartes - PSL

-

Workers and Their Machines. Machinism and Working-Class Literature (19th - 21st Century

Martin Houlier P

École nationale des chartes - PSL

This Master's thesis project examines the role the machine occupies in the working-class world from the 1830s to the turn of the 21st century. The objective is to propose a quantitative modeling of written sources by mobilizing natural language processing as well as machine learning. This study seeks to identify the existence of working-class "sociolects" and to make visible linguistic trends that are difficult to identify through linear reading, in order to observe how the daily manipulation of technical tools crystallizes a distinct textual identity.

Long Papers III

Is There a Patent Genre? A Textual Analysis of French Patents, 1903-1940

Matti La Mela[1], ***Yunting Xie***[2]

LP

[1]Institut d'études avancées de Paris, Uppsala University
[2]Department of Business Studies, Uppsala University

Patents are legal documents that grant temporary exclusive rights over inventions and describe their technical features. As such, they have become an important source for studying innovation and technological change, including through computational analyses of patent texts. Yet patent applications are not only repositories of technical knowledge, but are also shaped by institutional and linguistic conventions that structure their writing. This paper approaches patents as a rhetorical genre by analysing a corpus of French patent descriptions from 1903–1940. Using computational text analysis, it examines both the stability and variation of patent language—a characteristic feature of genre. First, we analyse the persistence of frequent bigrams over time to identify repeating lexical patterns in patents. Second, we measure lexical differences between patents from different countries filed in France using Jensen–Shannon distance. The results show a stable core of word constructions in patent descriptions. Interestingly, we see differences between patents from different countries, but also persistent closeness by country groups, e.g. between English-language patenting countries. The findings highlight the importance of approaching patents as historically situated textual artefacts.

Neutral Drift and Cumulative Memory: A Wright–Fisher Model of Thematic Evolution in Online Horror Fiction

Alexandre Lionnet-Rollin[1], ***Florian Cafiero***[2,3]

[1] École Pratique des Hautes Études
[2] EPITA
[3] Centre Jean Mabillon, École nationale des chartes - PSL

Cultural evolutionary models offer a principled framework for understanding how digital genres stabilise over time. In this paper, we apply a modified Wright–Fisher model to the thematic evolution of creepypasta narratives published on the Creepypasta Fandom wiki between 2010 and 2015. Drawing on a corpus of approximately 13,000 stories tagged across 169 moderated thematic categories, we test whether the observed power-law distribution of themes can be reproduced by a neutral drift process augmented with cumulative archiving. Standard Wright–Fisher simulations fail to replicate the empirical long tail, as minor categories extinguish too rapidly. When a mechanism of cumulative memory is introduced—whereby each generation contributes a sampled archive that persists alongside the living population—the model successfully reproduces both the dominance of major themes and the stable persistence of rare ones. Log-log analysis confirms that the simulated final distribution follows a power law, consistent with the empirical data. These results suggest that the thematic structure of the Fandom corpus is best understood not as a snapshot of active production, but as a sedimented historical record whose composition reflects processes of neutral drift operating over a cumulative substrate. We discuss implications for platform studies and the modeling of digital literary evolution.

In Search of Lost Adventure Novels. A Two-Stage Pipeline to Retrieve Genre Literature from the Large Scale National French Library Archive

Jean Barré LP

Langues, Textes, Traitements informatiques, Cognition (Lattice), CNRS – École Normale Supérieure – Sorbonne Nouvelle

This paper proposes a practical method to retrieve adventure novels from a large digitized collection of French fiction where genre metadata are sparse and unreliable. We combine (1) a supervised classifier (SVM) trained on a historically situated seed list of 101 adventure novels cited by a literary scholar [1], with (2) an unsupervised, graph-based correction step that exploits the structure of a k-nearest-neighbor (k-NN) similarity graph over the whole corpus. The correction removes SVM-labeled adventure novels that are isolated among non-adventure neighbors (false positives) and adds unlabeled novels embedded in adventure neighborhoods (false negatives). On the 12,176 novels from the filtered Fictions littéraires de Gallica collection [2], the supervised modeling identifies 1375 as adventure novels, and the unsupervised one 113 outliers and 129 infiltrators, yielding a final corpus of 1,391 adventure novels. Quantitatively, the corrected corpus is more cohesive (3.6% higher mean cosine similarity) and more homogeneous (18.4% lower standard deviation). Qualitatively, the corrections expose interpretable failure modes (parody, non-fiction, adjacent genres; missed canonical authors and subgenres), which we read as a computational trace of a central issue in genre theory: the tension between categorical definitions and graded, family-resemblance membership.

References

[1] Letourneux, M. (2010). Le roman d'aventures: 1870-1930. fre. PhD thesis. Limoges: PULIM.
[2] Langlais, P.-C. (2021). Fictions littéraires de Gallica / Literary fictions of Gallica. Version 1. https://doi.org/10.5281/zenodo.4751204.

Keynote Lecture II

Using Modelling and Simulation to Understand Change in Past Societies across Scales

Simon Carrignon KL

University College London

Computational approaches, combined with easy access to large amounts of computational power and data, have enabled us to understand how humans interact and how culture spreads through time in ways that were unthinkable before. Nonetheless, these datasets are often biased toward relatively modern societies, and specific aspects of culture. To find general factors behind the spread of cultural transformations that can reshape how humans live around the world, one needs to understand the mechanisms that enable culturally coherent groups to emerge, persist for centuries, or disappear. This requires data that span large geographical and temporal scales, sometimes over multiple countries and millennia, that only be provided by archaeology. These data, unless modern large datasets, are scarce, not documented evenly across regions and time period and need to be carefully handled when used as evidence of change in past behaviour.
To bridge the gap between the knowledge modern datasets provide and the general mechanisms behind the emergence and evolution of past societies, we propose to use computational modelling and simulation. Modelling allows us to simulate behaviour at local levels and inform these interactions with observations from very diverse fields of study. These models, combined with the right statistical tools, can then be used to test hypotheses about the processes that gave birth to modern societies and try to answer some of the questions raised before: what makes culturally consistent groups emerge and persist, why and when some cultures blend while others replace one another and so on? Drawing on case studies from archaeology, this talk will present how agent-based modelling, combined with machine learning and Bayesian inference, can help us address these questions and the challenges such approaches raise

Useful Information

Talks (keynote lectures, long papers and lightning talks) will be held in the **Léopold Delisle Room** (65 rue de Richelieu, 75002 Paris). The workshop room is located on the ground floor of the building on the left behind the elevators.

The **poster sessions** will take place in **Jules Quicherat room** which is located on the first floor of the building.

Coffee breaks and lunches will be served in the hall outside the workshop room. During the lunch break, workshop participants are invited to visit the poster session or enjoy their meal in the terrace (second floor of the building).

How to get to the workshop?

- **Metro:** Pyramides (lines 7 and 14), Bourse (line 3), Palais Royal - Musée du Louvre (lines 1 and 7), Richelieu-Drouot (lines 8 and 9);
- **RER:** Opéra (A);
- **Bus**: 20, 21, 27, 29, 39, 74, 85, 95.

Acknowledgments

The Computational Cultural Science (C^2S) workshop is organized in collaboration with the Master of Digital Humanities (ENC-PSL) and funded by the PSL Major Research Programme CultureLab (`https://www.culturelab.psl.eu/en`).