From Roland to Conan: First results on the corpus of French literary fictions (1050-1920)

Pierre-Carl Langlais, Jean-Baptiste Camps, Nicolas Baumard, Olivier Morin

From *Roland* to *Conan*: First results on the corpus of French literary fictions (1050-1920)

Pierre-Carl Langlais[1], Jean-Baptiste Camps[2], Nicolas Baumard[3], and Olivier Morin[4]

[1]Université de Montpellier Paul-Valéry
[2]École nationale des chartes, Université PSL
[3]Institut Jean-Nicod, CNRS; Département d'Études Cognitives, École normale supérieure, Université PSL
[4]MPI für Menschheitsgechichte, Jena

1 A corpus of French literary fictions: 1050-1920

A decade ago, when he set out to explore the evolution of English novels in the period 1740–1850, F. Moretti had to focus on the titles as the main available data, yet noting that:

> in a few years, we will have a digital archive with the full texts of (almost) all novels ever published (Moretti, 2009).

Today, we propose to embark on a journey to model the evolution of French literary fictions, from their epic and chivalric origins up to the more modern productions of genres such as heroïc fantasy or historical novels. We offer to do that not from the titles, but by analysing a corpus, currently in construction, whose aim is to cover all French literary fiction digitized from the earliest sample of French literature to the 20th century (or more precisely, to the point were a significant share of published literature is not yet in the public domain).

This project is at the intersection of two major trends in computational literary analysis: the creation and documentation of large literary corpora and the analysis of literary genre and discourse through machine learning classification. In comparison with previous examples of French literary corpus (like *théâtre classique* (Fièvre, 2007)) or the French corpus of the European literary text collection (Odebrecht et al., 2021), French novels make up a massive amount of texts (80,000 registered work in the French National Library before the 20th century), a large share of which is non-canonical and little documented. Text mining techniques make it possible to explore and document large digitized corpora with little editorial work. Classification is not simply used as cataloguing tool: its limitations can in fact inform in a more complex way the development of genres and the intertextual interplay between one genre and another.

The French fiction corpus initially results from the collocation of three different collections:

1. A collection of medieval fictions and *chansons de gestes* (1050-1450) - (cf. Camps, 2019)

2. A collection of printed fictions of Gallica from the modern period and the 19th century (Langlais, 2021b).

3. A new collection comprising most of the digitized fictions from the early modern period (1450-1700).

The second corpus relied on one of the oldest bibliographic database: the catalog of the French National Library. The classification scheme used by the library until 1996 was developed between 1684 and 1688 by the librarian of Louis XIV, Nicolas Clément. A specific category for novels (Y2) has been existing since 1730 as a duplication of the category for poetry (Y). The novels are now classified in Y2 or Y Bis. For cultural history, the BNF catalog presents a major interest: the categories are often (nearly) contemporary of the documents they aim to classify.

This corpus was made possible thanks to the policy of open data and open content which the BNF has been engaged in for several years. While the Clément classification ceased to be used for the classification of physical collections in 1996, it has become available in 2017 for data analysis when the BNF opened a new catalog access service, the SRU[1]. It was originally limited to novels digitized by the French National Library with a strong focus on the 19th century.

The third corpus was created specifically for the project: it aims to cover all the digitized versions of early modern novels published in France between 1473 (publication date of our earliest entry, *Le Roman de Jason et Médée*) and 1700 (our arbitrary cutoff, for now). This collection originally aimed to bridge the two previous corpora which were focused on two extreme temporal points of the history of French literature (the medieval period and the post-1800 period). It became a more ambitious experiment: a systematic harvesting of all available digitizations. The collections of digital documents already available in *Fictions littéraires de Gallica* have been significantly expanded thanks to the semi-automatical retrieval of documents digitized by Google Books and other online digital library. The creation of this composite corpus underline that digitized collections may be already more representative than expected, although nobody knew it or could measure the scale of it.

Using the combined collections of several digital libraries made it possible to cover a large amount of the novels registered in the catalog of the French National Library. For the 1470-1600 period, we have retrieved 275 novels out 349, that is 78.8% of the corpus. For the 1600-1700 period, we have 724 novels out of 1058, that is 68.4% of the corpus. Representativeness ratio is not only high but consistent on the entire time period as shown in figure 2 as the size of the available digitized corpus remains proportional to the total amount of novels identified by the French National Library.

Such high coverage seems to alleviate most concerns related to the representativeness of digital corpora: it becomes less likely that important genres or themas are neglected. Yet the novels registered on the French National Library catalog do not encompass the total sum of literary fiction published or circulated on the period. The comparison with Google Books showed that numerous editions were not recorded by Gallica. In the context of the project, we only checked novels recorded on the catalogue, but this discrepancy shows that some published novels could definitely have been overlooked, although by definition these documents are not necessarily expected to belong to those with the largest impact, as they failed to be noticed by bibliographic records. After the 19th century, nearly all published monographs are expected to be indexed in the catalogue of the French National Library due to progress in the implementation of the policy of legal deposit (*dépôt légal*, established in 1537). Yet, at this time, a large amount of literature was increasingly being published in periodicals and newspapers.

In short as representativeness of existing library catalogues becomes less of a concern, digital collections are bound to raise more complex issues and address a more critical perspective on existing bibliographic resources: what is a "novel"? what is a "publication"? what about piracy edition or periodical fictions?

The creation of the third corpus also aims to benefit from unprecedented progresses in the detection of historical OCR. The OCR17+ model trained by Claire Jahan and Simon Gabay on a collection of 17th century prints already yields a usable text for most of the 16th to 18th century

[1] *Search/Retrieve via URL*, `https://api.bnf.fr/fr/api-sru-catalogue-general`.

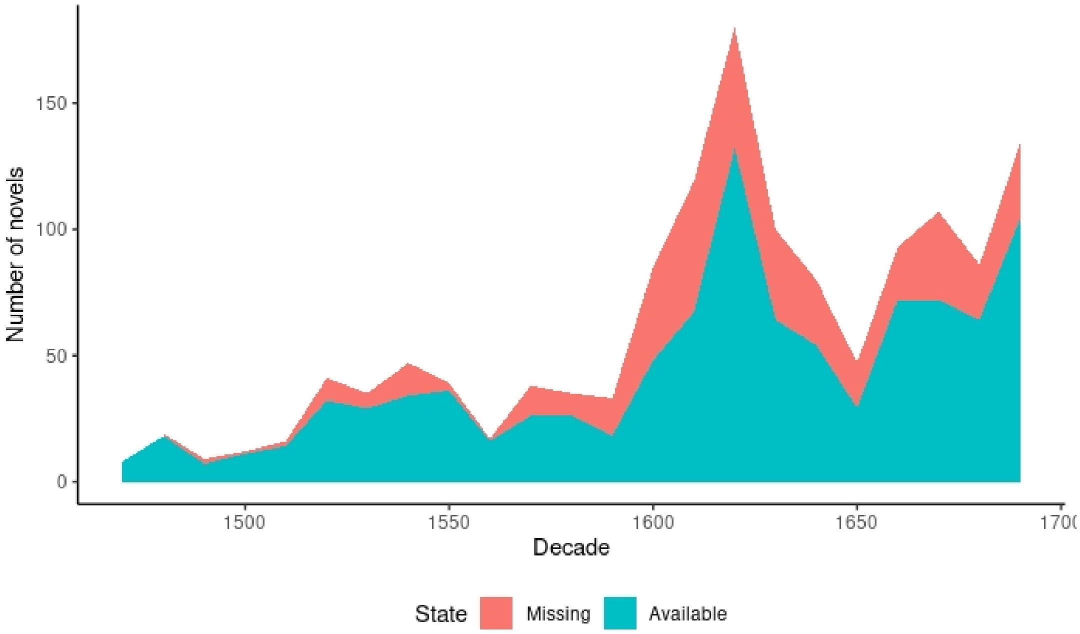

Figure 1: Composition of the corpus by decade

documents[2].

2 Modelling literary genre: a back and forth approach

Chivalric romance is one of oldest and most enduring genre of European literature. It finds its origin in Old French epics, known as *chansons de geste*, whose first attestations go back to the 11th century, and in the genre of *roman* that emerges in the second half of the 12the century. Continuously modified, adapted and rewritten during the Middle Ages and the Early Modern times, this repertoire of fictions provided numerous tropes and patterns that arguably live on in more modern and even contemporary literature, especially in popular literature and even fantasy novels.

Preliminary exploration of our large corpus of French literary fictions suggests that the chivalric novel, rather than ceasing to exist, has been continually evolving and gradually morphed into forms close to the contemporary fantasy novels. This structural transformation has been largely overlooked as forms of chilvalric romances have largely disappeared from the high literary canon

[2] Jahan and Gabay (2021). This model will be applied on the entire corpus. Preliminary results on a random selection of 60 pages from the corpus show that OCR quality is already high for the 17th century (nearly 99% character accuracy). The 16th century corpus is more challenging due to specific issues with Gothic fonts and segmentation, although we intend to solve these limitations by fine tuning a model for Gothic fonts in incunabula and 16th century prints. In comparison, the mean OCR quality in the collections held by Gallica is at roughly 80% for the 16th and 17th centuries (according to available metadata) and a large part is currently unavailable for full-text search, possibly because OCR quality was too low. The project will also benefit from the development made by the ongoing Gallic(orpor)a project, that aims to provide a fully reusable pipeline for Gallica documents in French from the Middle Ages to the Revolution. By the end of this project, we plan to create a small search engines of the entire collected corpus of 15th to 17th century documents. This would be the first resource to give access to a large share of the published literature in France in the early modern period. An initial version of this search engine may be available by the time of the conference and serve as a demonstration of our methodology of systematic collection and enhancement of available digitized collections.

after the Renaissance and continued to evolve in the production of lesser known and more obscure authors.

Our analysis relies on a anachronistic use of classification recently pioneered in cultural analytics (Tello, 2021, Underwood, 2019). Probability rates, cross-classifications of the same text and, even, classification failures are reinterpreted as a way to measure the complex evolution of literary genres.

To investigate the transformation of chivalric novels into contemporary genres like fantasy, we created two historical models of literary genres: a 21st century model and an early modern "Fresnoy" model (from a 1731 catalog). The combined use of anachronistic classification aims to locate the two parallel processes of genre survival/transformation (for chivalric romance) and genre emergence/coagulation (for fantasy).

Classification was made with SVM using an R library initially developed for the classification of newspaper genres in the 19th century, *TidySupervise* (Langlais, 2021a).

The 21st century model is created using nine genre categories of the social cataloguing website *Babelio*[3]. User-generated tags have recently emerged as an important source of information in computational analysis of contemporary literature and literary reception (Walsh and Antoniak, 2021). A French counterpart to *Goodreads*, *Babelio* has a significant impact among French literary readers with nearly 1 million visits per month. In this project, we focused on a subset of generic tags that were either major acknowledged genres in contemporary literature or relevant for our ongoing projects: romance, fantasy, detective fiction, science fiction, historical novel, adventure novel, social novel, *fantastique* novel and erotica. Obviously, this is not a straightforward classification, since one novel could belong to several categories. We aimed rather to reconstruct the fuzzy space of contemporary literary genres in France with all its underlying uncertainties and overlaps.

The model was trained on 4,081 segments of 1000 words extracted from 1,346 novels. We applied a random selection of three 1000 words segments by work. This selection aims to limit over-fitting and ensure that the model will be correctly trained on generic features and not on the style of specific novels. Bootstrap evaluation of the model yield a 75% accuracy, yet with significant variations among the genres (fantasy being the highest rated with science fiction and erotica).

Preliminary results from the 21st century model have revealed significant examples of "missing links" between the early modern chivalric romances and the fantasy novels (Table 1). Of special interest is *La Mort de Roland* a 1858 rewrite of the *Chanson de Roland* by Alfred Assolant that explicitly claims to be an *epic fantasy* (*Fantaisie épique*).

Most of these works are poorly attested in literary history. In comparison, the results from Science-fiction yields much more expected and "canonic" works (especially from Jules Verne).

The early modern model (or "Fresnoy" model) has been made possible by an exceptional historical source on literary genre classification: the second volumes of *De l'usage des romans, où l'on fait voir leur utilité & leurs différens caractères* by Nicolas Lenglet du Fresnoy (first published in 1731 under the pseudonym of Gordon de Percel). It is a catalog of a large among of French, Spanish and Italian novels published since 1731 broken down by genres according to the prevalent taxonomy of the time: *Roman de chevalerie*, *Roman d'amour*, *Roman historique*, *Roman comique*, *Roman politique*, etc. Ongoing work aims to reconcile the classification of Lenglet du Fresnoy with our corpus of digitized novels. Preliminary results suggest that the generic identity in the catalogue of Fresnoy is much stronger than in the Babelio dataset, with as much as 93% accuracy on four genres (*Roman d'amour*, *Roman historique*, *Roman de chevalerie* and *Roman comique & satirique*) in our initial run. While this high accuracy may be caused by overfitting on a limited samples of novels, it seems also consistent with the significance of genre classification in the

[3] `https://www.babelio.com/`.

Title	Author	Segments classified	Proportion of segments
Les rois de mer =*The Sea Kings*	Léon Cahun	74	70
Iskender	Judith Gauthier	58	66
Les aventures du dernier Abencérage =*The Adventures of the last Abencérage*	François-René de Chateaubriand	15	66
Les aventures du capitaine Magon =*The Adventures of capitain Magon*	Léon Cahun	108	65
Histoire de Don Quichotte racontée à la jeunesse =*The Story of Don Quichotte told to the Youth*	Miguel de Cervantes	55	65
Voyage de Mademoiselle Lili autour du monde =*The Travel of Miss Lili around the world*	P.-J. Stahl	14	64
Le petit duc =*The Small Duke*	Charlotte Mary Yonge	43	63
Guillaume Tell [de Schiller], adapté pour les enfants =*Schiller's William Tell, adapted for Childrens*	Friedrich von Schuller	17	58
La mort de Roland =*The Death of Roland*	Alfred Assolant	70	58
La flèche noire =*The Black Arrow*	Robert Louis Stevenson	81	55

Table 1: Top 10 works classified as fantasy in the corpus of 19th century novel digitized by Gallica

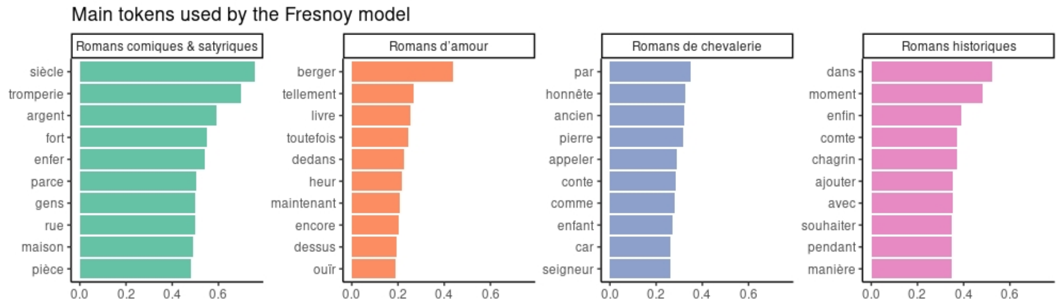

Figure 2: Main words used in the four categories of the Fresnoy model

meta-discourse about the novel in the 18th century.

A this stage, the expansion of the corpus and the re-digitization of available documents with historical models for OCR will be crucial to move beyond an exploratory phase and design a more systematic examination of the metamorphosis of literary genre.

References

Jean-Baptiste Camps, editor. *Geste: un corpus de chansons de geste, 2016-... (Version 02)*. Paris, April 2019. URL `http://doi.org/10.5281/zenodo.2630574`. textes du domaine public, développements CC-BY-SA.

Paul Fièvre. Théâtre classique, 2007. URL `http://www.theatre-classique.fr/`.

Claire Jahan and Simon Gabay. Ocr17+ - layout analysis and text recognition for 17th c. french prints, 2021. URL `https://github.com/e-ditiones/OCR17plus`.

Pierre-Carl Langlais. Classified News, Redefining the history of newspaper genre with supervised models. In *Digital Newspaper: a new Eldorado for the historians*. De Gruyter., 2021a.

Pierre-Carl Langlais. Fictions littéraires de Gallica / Literary fictions of Gallica, April 2021b. URL `https://doi.org/10.5281/zenodo.4751204`.

Franco Moretti. Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850). *Critical Inquiry*, 36(1):134–158, 2009. ISSN 0093-1896. doi: 10.1086/606125. URL `http://www.jstor.org/stable/10.1086/606125`. Publisher: The University of Chicago Press.

Carolin Odebrecht, Lou Burnard, and Christof Schöch. European literary text collection (eltec), 2021.

José Calvo Tello. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Bielefeld University Press, transcript, October 2021. ISBN 978-3-8394-5925-6. doi: 10.1515/9783839459256. URL `https://www.degruyter.com/document/doi/10.1515/9783839459256/html`. Publication Title: The Novel in the Spanish Silver Age.

Ted Underwood. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, February 2019. ISBN 978-0-226-61283-6. Google-Books-ID: fQo5uwEACAAJ.

Melanie Walsh and Maria Antoniak. The goodreads “Classics”: A computational study of readers, amazon, and crowdsourced amateur criticism. *Journal of Cultural Analytics*, 6(2), April 2021. doi: 10.22148/001c.22221. Publisher: Department of Languages, Literatures, and Cultures.