



The Collection of Eighteenth-Century French Novels 1751–1800

DATA PAPER

JULIA RÖTTGERMANN 

 ubiquity press

ABSTRACT

The French Enlightenment is a pivotal period in European intellectual and literary history, which can be studied through this dataset of French novels first published between 1751 and 1800. This collection contains 200 French novels in TEI/XML, encoded according to the ‘level-1 schema’ of the European Literary Text Collection (ELTeC), and carefully compiled to reflect the known historical publication of French Novels in that period regarding publication year, gender of author and narrative form. The dataset is connected to a bigger knowledge graph of 331,671 Resource Description Framework triples (RDF) built within the project ‘Mining and Modeling Text’ at Trier University, Germany (2019–2023).

CORRESPONDING AUTHOR:

Julia Röttgermann

Trier Center for Digital
Humanities, Trier University,
Trier, Germany

roettger@uni-trier.de

KEYWORDS:

French; literature; linked open
data; Enlightenment

TO CITE THIS ARTICLE:

Röttgermann, J. (2024). The
Collection of Eighteenth-
Century French Novels
1751–1800. *Journal of Open
Humanities Data*, 10: 31,
pp. 1–5. DOI: [https://doi.
org/10.5334/johd.201](https://doi.org/10.5334/johd.201)

(1) OVERVIEW

REPOSITORY LOCATION

Zenodo: <https://doi.org/10.5281/zenodo.10404966>

CONTEXT

Our goal was to generate a balanced dataset of French Novels first published between 1751 and 1800 that can be used to generate Resource Description Framework (RDF)¹ statements for a knowledge graph on French Enlightenment novels, but can also serve as a resource for other projects in the Digital Humanities or in the domain of Eighteenth-Century French literature. The dataset was produced by the project Mining and Modeling Text; it has been used in several project publications, including Klee & Röttgermann, 2022; Schöch et al., 2022; Röttgermann, Klee, et al., 2022; Röttgermann, Hinzmann, et al., 2022.

(2) METHOD

STEPS

The starting point was a subset of novels carefully digitized by double keying. Using this first group of novels, an OCR-model has been trained in cooperation with Christian Reul (Centre for Philology and Digitality, University of Würzburg), one of the developers of OCR4all (Reul et al., 2019, Figure 1).

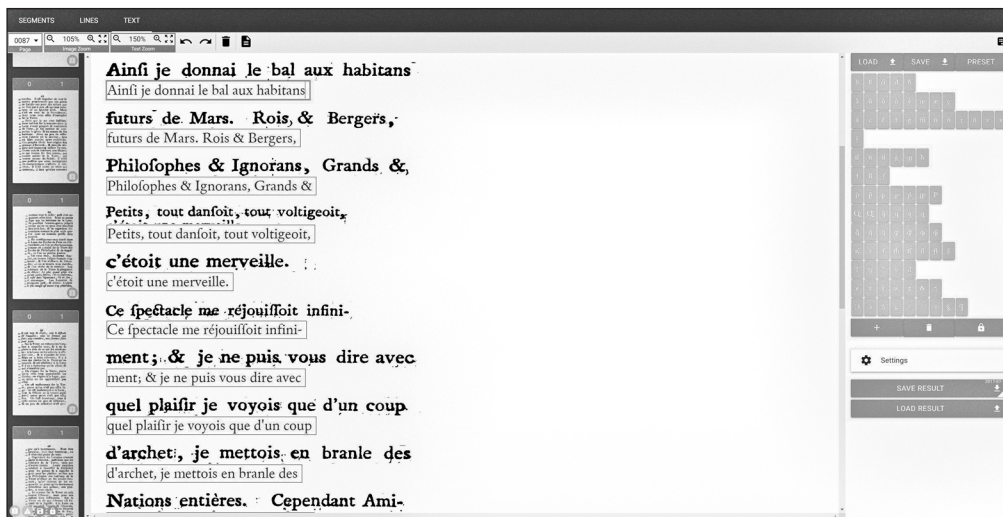


Figure 1 Training the OCR model for late Eighteenth-Century prints of French novels with OCR4all.

Applying this OCR-model for French prints of the late 18th-Century to additional scans provided for instance by Gallica² ([bnf.fr](https://gallica.bnf.fr)) and other sources (see metadata for details), a second group of novels which were not yet available in full text (or only in low quality) was produced. A third group of texts, based on existing full texts, helped us reach 200 volumes.³

SAMPLING STRATEGY

As shown in Figure 2, we used bibliographic data on the overall literary production in France 1751–1800 (Martin et al., 1977) to balance the corpus of full texts regarding the parameters gender, year of first publication and narrative form in approaching the historical distribution of these parameters in our corpus composition.

1 RDF triples are constructed by a subject-predicate-object structure via unique URIs; the statement “Voltaire is the author of *Candide*” might be represented as <https://data.mimotext.uni-trier.de/wiki/Item:Q981> <https://data.mimotext.uni-trier.de/wiki/Property:P7> <https://data.mimotext.uni-trier.de/wiki/Item:Q1022>.

2 Gallica: <https://gallica.bnf.fr> (Last accessed: 26 March 2024).

3 We converted existing text files to the uniform TEI format of the text collection. Metadata: <https://github.com/MiMoText/roman18/blob/master/metadata.tsv> (Last accessed: 26 March 2024).

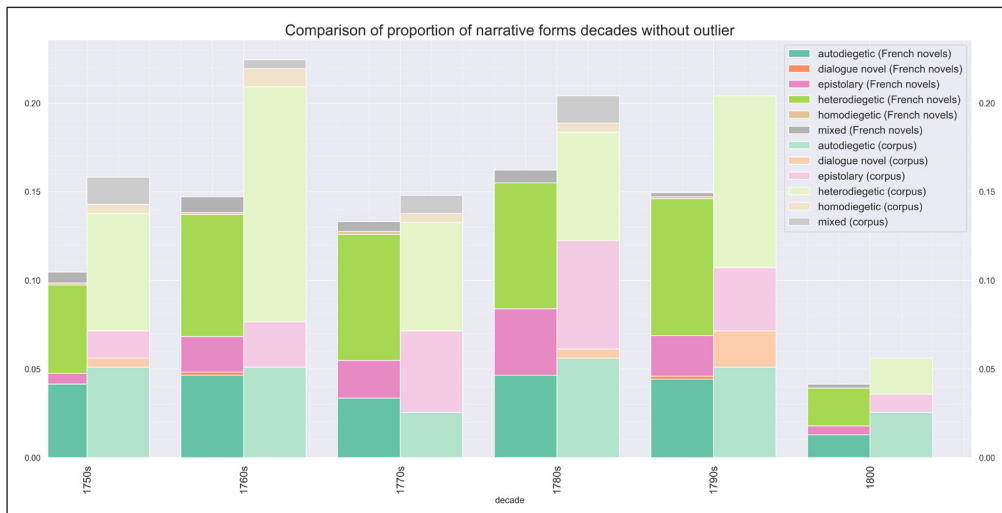


Figure 2 Narrative forms of French novels 1751–1800 (Martin et al., 1977) and in corpus metadata.

We compared the overall novel publication with the corpus data and added novels per year according to the known historical publication proportions. Regarding gender (Figure 3), we used information from Wikidata as well as a python script designed to identify gender-specific titles such as “Abbé” or “Marquis”.⁴

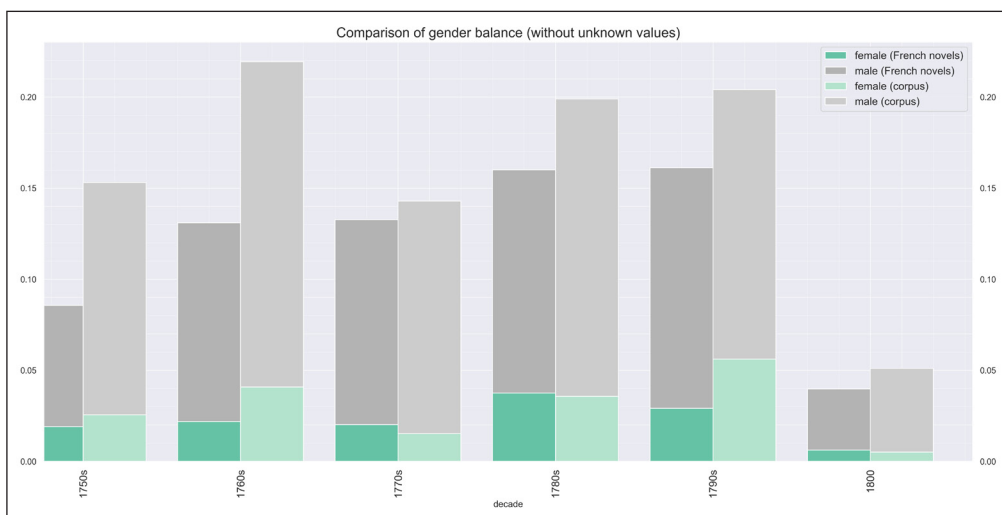


Figure 3 Gender balance in bibliographic metadata (Martin et al., 1977) and in corpus metadata.

Data regarding narrative form was derived from bibliographic metadata (Martin et al., 1977), complemented by human evaluations carried out on the full texts.

QUALITY CONTROL

Optical character recognition

The output of the OCR4all pipeline has undergone several quality controls including by a French native speaker correcting the output of OCR4all, documented by versioning control (GitHub).

Metadata

Additionally, we made sure that the data set meets the FAIR data criteria of findability, accessibility, interoperability and reusability (Röttgermann & Schöch, 2020). Every item is provided with a stable Uniform Resource Identifier (URI) (MiMoTextID) and additional authoritative data. In the process of reconciling data against entities in Wikidata, the output of OpenRefine (Huynh, 2012/2010) was manually corrected if necessary.

⁴ In instances where names did not correspond to a Wikidata entry or possess a discernible title, we employed the `gender_guesser` Python package to predict gender: https://github.com/MiMoText/balance_novels (Last accessed: 26 March 2024). Please note that the binary gender categories used here for metadata are due to data availability.

(3) DATASET DESCRIPTION

REPOSITORY NAME

Zenodo, GitHub.

OBJECT NAME

Collection de romans français du dix-huitième siècle (1751–1800)/Collection of Eighteenth-Century French Novels 1751–1800 (V1.2).

FORMAT NAMES AND VERSIONS

V1.2: 200 files in TEI/XML according to the ‘level 1’-schema of the European Literary Text Collection; TXT files in two versions: (automatically) normalized and historical spelling; controlled vocabularies used to describe metadata are documented on GitHub.⁵

CREATION DATES

2019-12-01 to 2023-12-06.

DATASET CREATORS

Julia Röttgermann (editor), Johanna Konstanciak (researcher), Christof Schöch (researcher), Julia Dudar (researcher), Henning Gebhard (researcher), Anne Klee (researcher), Sarah Ondraszek (researcher), Amélie Probst (researcher), Damir Padieu (researcher). Affiliation of all (at the time of data development): University of Trier, Trier, Germany.

LANGUAGE

French; English for metadata.

LICENSE

Public Domain.

PUBLICATION DATE

2023-12-06 (V1.2).

(4) REUSE POTENTIAL

Our data set can be used for varying language-processing tasks that use 18th-Century French language. As we provide detailed metadata, one could generate subsets of the data set, for example regarding gender, decade of publication or narrative form. One could study for example distinctive words for different decades or investigate linguistic differences between male and female authors in a diachronic perspective.

Moreover, the dataset provides a comprehensive and structured resource for analysing literary and cultural trends during the Enlightenment era, enabling researchers to gain insights into the intellectual and societal transformations of that time. As the data set is balanced according to different parameters, it can be regarded as representative of the time period.

Computational Literary Studies Methods that have already been used on the data set in the context of the project are topic modeling (Röttgermann, Klee, et al., 2022), named entity recognition (Röttgermann, Hinzmann, et al., 2022), sentiment analysis or stylometry.

Furthermore, the Linked Open data paradigm used to connect these full text resources with the knowledge graph ‘MiMoTextbase’ allows to run sophisticated SPARQL queries on these texts combining them in the graph with metadata on about 2000 French novels 1751–1800.⁶

⁵ Controlled vocabularies of the MiMoText project: <https://github.com/MiMoText/vocabularies> (Last accessed: 26 March 2024).

⁶ SPARQL-endpoint with showcase queries: <https://query.mimotext.uni-trier.de/> (Last accessed: 26 March 2024).

FUNDING INFORMATION

This project has received funding from Forschungsinitiative Rheinland-Pfalz 2019–2023. The publication was funded/supported by the German Research Foundation (DFG).

Röttgermann
*Journal of Open
Humanities Data*
DOI: 10.5334/johd.201

5


COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR CONTRIBUTIONS

Julia Röttgermann: Conceptualization, Writing – original draft.

AUTHOR AFFILIATIONS

Julia Röttgermann  orcid.org/0000-0002-1918-8117
Trier Center for Digital Humanities, Trier University, Trier, Germany

REFERENCES

- Huynh, D.** (2010). *OpenRefine*. Retrieved from <https://github.com/OpenRefine/OpenRefine> (Last accessed: 26 March 2024).
- Klee, A., & Röttgermann, J.** (2022). Nuit, correspondance, sentiment – Topic Modeling auf einem Korpus von französischen Romanen 1750–1800. *Apropos: Perspectives on Romania*, 9, 57–86. DOI: <https://doi.org/10.15460/apropos.9.1888>
- Martin, A., Mylne, V., & Frautschi, R. L.** (1977). *Bibliographie du genre romanesque français, 1751–1800*. London: Mansell.
- Reul, C., Christ, D., Hartelt, A., Balbach, N., Wehner, M., Springmann, U., Wick, C., Grundig, C., Büttner, A., & Puppe, F.** (2019). OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings. *arXiv*. DOI: <https://doi.org/10.20944/preprints201909.0101.v1>
- Röttgermann, J., Hinzmann, M., Gebhard, H., Klee, A., Konstanciak, J., Christof, S., & Steffes, M.** (2022, July 25–27). Mining and Modeling Spaces and Places for Literary History as Linked Open Data. *Digital Humanities 2022*, Tokyo, Japan. DOI: <https://doi.org/10.5281/zenodo.6948236>
- Röttgermann, J., Klee, A., Hinzmann, M., & Schöch, C.** (2022, March 7–11). Literaturgeschichtsschreibung datenbasiert und wikifiziert? *DHD2022*, Potsdam, Germany. DOI: <https://doi.org/10.5281/zenodo.6328157>
- Röttgermann, J., & Schöch, C.** (2020, May 11). FAIRe Daten in den Literaturwissenschaften? *Romanistik-Blog*. Retrieved from <https://blog.fid-romanistik.de/2020/11/05/faire-daten-in-den-literaturwissenschaften/> (Last accessed: 26 March 2024).
- Schöch, C., Hinzmann, M., Röttgermann, J., Dietz, K., & Klee, A.** (2022). Smart Modelling for Literary History. *IJHAC: International Journal of Humanities and Arts Computing*, 16(1), 78–93. DOI: <https://doi.org/10.3366/ijhac.2022.0278>

TO CITE THIS ARTICLE:

Röttgermann, J. (2024). The Collection of Eighteenth-Century French Novels 1751–1800. *Journal of Open Humanities Data*, 10: 31, pp. 1–5. DOI: <https://doi.org/10.5334/johd.201>

Submitted: 01 February 2024

Accepted: 22 March 2024

Published: 22 April 2024

COPYRIGHT:

© 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.