Journal of Open
Humanities Data

The Oupoco Database of French Sonnets from the 19th Century

DATA PAPER

FRÉDÉRIQUE MÉLANIE-BECQUET
CLAUDE GRUNSPAN
MYLÈNE MAIGNANT
CLÉMENT PLANCQ
THIERRY POIBEAU

*Author affiliations can be found in the back matter of this article

]u[ubiquity press

ABSTRACT

The Oupoco Database is a collection of 4,872 French sonnets developed in the framework of the Oupoco Project. It is mainly composed of poems from the 19th and early 20th century. The sonnets come from different sources from the Internet and from a collaboration with the Bibliothèque nationale de France. Every sonnet has a specific license (depending on the source it comes from), but the whole collection can be reused for free (under the Creative Commons Attribution 4.0 International license).

CORRESPONDING AUTHOR:
Thierry Poibeau
Lattice, CNRS & Ecole normale supérieure/PSL and Université Sorbonne nouvelle, Paris, FR
thierry.poibeau@ens.psl.eu

KEYWORDS:
French poetry; Sonnets; Poetry generation; Oulipo; Oupoco

TO CITE THIS ARTICLE:
Mélanie-Becquet, F., Grunspan, C., Maignant, M., Plancq, C., & Poibeau, T. (2022). The Oupoco Database of French Sonnets from the 19th Century. *Journal of Open Humanities Data*, 8: 25, pp. 1–5. DOI: https://doi.org/10.5334/johd.89

(1) OVERVIEW

REPOSITORY LOCATION

https://doi.org/10.5281/zenodo.5646939.

CONTEXT

The Oupoco Database is a collection of 4,872 French sonnets developed in the framework of the Oupoco Project (Poibeau *et al*., 2020). The database is mainly composed of poems from the 19th and early 20th century. We have identified 760 authors: 4,414 sonnets written by men (660 authors), 439 sonnets written by women (107 authors), which leaves 19 sonnets to which we have not been able to assign an author. The sonnets come from different sources from the Internet, or not: we especially want to thank the Bibliothèque nationale de France (BnF) (French National Library) that gave us access to a large corpus, from which we were able to extract an invaluable number of French poems. To all the sonnets is attached a specific license related to the source they come from, but all are freely available and can be reused for free (under the Creative Commons Attribution 4.0 International license).

This database has initially been developed for the Oupoco project (L'Ouvroir de littérature combinatoire),[1] which consists in producing new sonnets by recombining verses from existing ones from the French literature, following the idea put forward by Queneau in his famous conceptual book *Cent mille milliards de poèmes* (Queneau, 1961). Different scripts have been developed for the Oupoco project (to analyse the rhymes and recombine the verses), which are available on the GitHub repository linked to the project.[2] Beyond Oupoco, this database can be used for various purposes, for teaching and for research, especially in the following domains: literature studies, corpus linguistics, digital humanities, arts and technology.

(2) COLLECTION METHOD

SOURCES

The sonnets were collected from five different sources:

SOURCE	NUMBER OF SONNETS	LICENSE/COMMENTS
BnF	3,979	CC-BY-SA-NC
Wikisource	772	CC BY-SA 3.0
Web	67	Source (Blog) cited in the database
Books (anthology)	37	Manual collections of sonnets from different anthologies
Malherbe project	7	No explicit license (https://git.unicaen.fr/malherbe/corpus)

BnF directly provided us a collection of books containing poetry, in the XML Alto format.[3] We selected the books with an OCR quality score above 98% and used the BnF API to collect metadata about the books. We then identified and automatically retrieved the sonnets from these books (mainly through the pattern: two quatrains followed by two tercets). There was no exhaustive verification of the extraction process, so this corpus may contain poems that are not sonnets.[4] Metadata were then collected from the BnF, except the title of the sonnet and the page number, which were retrieved automatically. Here again, errors can probably be found, as the process was automatic with no comprehensive quality check.

Texts from Wikisource were taken as is, with no added information. Sonnets coming from the Web were manually collected at the start of the project. We made sure these sonnets are not subject to copyright.

[1] https://oupoco.org/ (last accessed: 14/10/2022).

[2] https://github.com/lattice-8094/oupoco-api (last accessed: 14/10/2022).

[3] https://www.loc.gov/standards/alto/ (last accessed: 14/10/2022).

[4] Errors can be reported to info@oupoco.org.

Anthologies have been used more recently to augment the corpus, with a specific focus on female authors, as Wikisource, for example, is highly unbalanced towards male authors). Five anthologies have been used, and the identified sonnets were also all copyright free. These four anthologies are:

- H. Blanvalet (1856). *Femmes - poëtes de la France – Anthologie*. Paris: J. Kessmann éd.
- Rachilde (1908). *Le missel de Notre-Dame des Solitudes*. E. Sansot, Paris.
- Le Comte de Saint-Jean (Mme Eugène Riom) (1892). *Les femmes poètes bretonnes*. Nantes: Société des bibliophiles bretons et de l'histoire de Bretagne.
- Alphone Séché (1908). *Les muses françaises – Anthologie des femmes poètes (1200 à 1891)*. Paris: Louis-Michaud éd. (two volumes).

The last source is the Malherbe corpus. Its contribution is marginal as only seven sonnets come from this database (that covers other kinds of French poems, most of them not being sonnets). The Malherbe database has been developed in the framework of the Malherbe project by Éliane Delente et Richard Renault. The goal of the project was to provide an overview of the diversity of French versification as well as automatic tools related to this issue. The corpus and related information can be found here: https://git.unicaen.fr/malherbe/corpus. The poems in this collection are also copyright free and although there is no license attached to the project at the time of writing, the authors confirmed that the database can be used free of charge, as long as the original repository is mentioned.

SAMPLING STRATEGY

The collection is provided as is. No sampling strategy has been used. The corpus is thus highly unbalanced, especially between authors (a few authors provided lots of sonnets, lots of other authors provided only a few of them).

QUALITY CONTROL

The process to collect the sonnets and the metadata has been highly automated, therefore errors can be found (poems that are in fact not sonnets, errors in the metadata or due to the OCR). We are however confident that these errors remain marginal, from an extensive quality check performed randomly on a sample of the data.

(3) DATASET DESCRIPTION

OBJECT NAME

oupoco.dtd and sonnets_oupoco_tei.xml

FORMAT NAMES AND VERSIONS

XML Unicode files.

CREATION DATES

The project began in 2018. The current version has been published 2022-06-29.

DATASET CREATORS

Frédérique Mélanie-Becquet: conceptualization, data curation and supervision; Claude Grunspan: data curation; Mylène Maignant: conceptualization and data curation; Clément Plancq: conceptualization; Thierry Poibeau: funding acquisition, supervision and writing.

LANGUAGE

French.

LICENSE

Creative Commons Attribution 4.0 International.

REPOSITORY NAME

Zenodo (https://doi.org/10.5281/zenodo.5646939).

PUBLICATION DATE

2022-06-29.

(4) REUSE POTENTIAL

This database can be used by anyone interested in French poetry, for teaching and for research, especially in the following domains: literature studies, corpus linguistics, digital humanities, arts and technology (for this purpose, a short video has been released to explain the Oupoco project, see Lattice and RIVA Illustrations, 2021). The corpus is large enough to be analysed with machine learning methods for stylometric studies, for example. One direct perspective would be to apply authorship attribution methods to the anonymous sonnets, so as to propose potential authors to these poems.

ACKNOWLEDGEMENTS

We want to thank the Bibliothèque nationale de France that provided us with a corpus related to French poetry, from which we were able to extract an invaluable number of French poems.

FUNDING STATEMENT

This work has received support of Translitteræ (Ecole universitaire de recherche, program “Investissements d’avenir” ANR-10-IDEX-0001-02 PSL* and ANR-17-EURE-0025) and of CNRS through the IRN (International Research Network) Cyclades. This work was also supported in part by the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Frédérique Mélanie-Becquet: conceptualization, data curation and supervision; Claude Grunspan: data curation; Mylène Maignant: conceptualization and data curation; Clément Plancq: conceptualization; Thierry Poibeau: conceptualization, funding acquisition, supervision and writing.

PUBLISHER’S NOTE

The link to the Repository Location of this dataset has been updated to include the DOI.

AUTHOR AFFILIATIONS

Frédérique Mélanie-Becquet
Lattice, CNRS & Ecole normale supérieure/PSL and Université Sorbonne nouvelle, Paris, FR

Claude Grunspan
Lattice, CNRS & Ecole normale supérieure/PSL and Université Sorbonne nouvelle, Paris, FR

Mylène Maignant
Lattice, CNRS & Ecole normale supérieure/PSL and Université Sorbonne nouvelle, Paris, FR

Clément Plancq
Lattice, CNRS & Ecole normale supérieure/PSL and Université Sorbonne nouvelle, Paris, FR

Thierry Poibeau orcid.org/0000-0003-3669-4051
Lattice, CNRS & Ecole normale supérieure/PSL and Université Sorbonne nouvelle, Paris, FR

REFERENCES

Lattice and RIVA Illustrations. (2021). Oupoco, la boîte à poésie (video). Retrieved from https://odhn.ens.psl.eu/newsroom/oupoco-la-boite-poesie, on the Observatoire des Humanités numériques (ODHN) de l'ENS-PSL web site (last accessed: 16 October 2022).

Poibeau, T., Maignant, M., Mélanie-Becquet, F., Plancq, C., Raffard, M., & **Roussel, M.** (2020). Sonnet Combinatorics with OuPoCo. *Proceedings of the the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 133–137.

Queneau, R. (1961). *Cent mille milliards de poèmes*. Paris: Gallimard.

TO CITE THIS ARTICLE:
Mélanie-Becquet, F., Grunspan, C., Maignant, M., Plancq, C., & Poibeau, T. (2022). The Oupoco Database of French Sonnets from the 19th Century. *Journal of Open Humanities Data*, 8: 25, pp. 1–5. DOI: https://doi.org/10.5334/johd.89

Published: 15 November 2022