The Corpus for Idiolectal Research (CIDRE)
Olga Seminck, Philippe Gambette, Dominique Legallois, Thierry Poibeau

To cite this version:
Olga Seminck, Philippe Gambette, Dominique Legallois, Thierry Poibeau. The Corpus for Idiolectal Research (CIDRE). Journal of Open Humanities Data, Ubiquity Press, 2021, 7, pp.15. 10.5334/johd.42 . hal-03310451

HAL Id: hal-03310451
https://hal.archives-ouvertes.fr/hal-03310451
Submitted on 30 Jul 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
The Corpus for Idiolectal Research (CIDRE)

ABSTRACT

The Corpus for Idiolectal Research (CIDRE) is a collection of fiction works from 11 prolific 19th-century French authors (4 women, 7 men; 22–62 works/author; total of 37 million words). Every work is dated with the year it was written. Using programming scripts, the works have been gathered from open source platforms, for example La Bibliothèque électronique du Québec, and stripped of paratext (text not being part of the novel, e.g. prefaces). We distribute the text files, the dating, other metadata and the programming scripts under an open source license. CIDRE is the first resource of French for the study of style and idiolect in a diachronic manner (i.e. stylochronometry) on a larger scale.

CORRESPONDING AUTHOR:
Olga Seminck
Laboratoire Lattice (Langues, Textes, Traitements informatiques, Cognition) – CNRS & ENS/PSL & Université Sorbonne nouvelle, Montrouge, France
olga.seminck@cri-paris.org

KEYWORDS:
stylochronometry; linguistics; stylometry; idiolect; diachrony; literature; fiction; French

TO CITE THIS ARTICLE:
Seminck, O., Gambette, P., Legallois, D., & Poibeau, T. (2021). The Corpus for Idiolectal Research (CIDRE). Journal of Open Humanities Data, 7: 15, pp. 1–6. DOI: https://doi.org/10.5334/johd.42
(1) OVERVIEW
An idiolect is the language of an individual and, like language in general, it is subject to change over time (Dittmar, 1996; Kuhl, 2003). However, the notion of idiolect remains an understudied topic, especially in quantitative linguistics, due to the lack of relevant large corpora (Barlow, 2010, 2013; Mollin, 2009). We thus developed the CIDRE corpus, the first corpus for the diachronic and quantitative study of idiolect in French (logo in Figure 1). Together with the EMMA corpus on 17th-century English (Petré et al., 2019), it is one of the rare quantitative resources suited to stylochronometry (see Klaussner and Vogel, 2018 and Stamou, 2008 for examples of stylochronometric studies).

With the purpose of collecting as much data per person as possible within one genre (to enable comparison), we decided to use the fiction works of prolific 19th-century writers. The advantages of this type of data are the following: fiction works tend to be long, providing us with large quantities of data; they are in the public domain, there are high quality e-books available; and the orthography of that period is very similar to today's, making the use of off-the-shelf NLP systems possible.

Using various websites distributing free epub files, we included in CIDRE, as exhaustively as possible, the fiction works of Gustave Aimard, Honoré de Balzac, Paul Féval, Henry Gréville, Daniel Lesueur, Pierre-Alexis Ponson du Terrail, George Sand, La Comtesse de Ségur, Jules Verne, Michel Zévaco and Émile Zola (see Table 1 and Figures 2 and 3 for more details). We dated the works with the year they were written in, if this information was available, and with the first year of publication otherwise. In this way, each work can be seen as a datapoint characteristic of the way the author was writing at the time.

REPOSITORY LOCATION
https://zenodo.org/record/4707812#.YK-Tai8ivs0

CONTEXT
This resource was produced as part of a research project investigating large corpora of French literature with advanced natural language processing methods.

(2) METHOD
CIDRE was produced using programming scripts in Python and manual gathering of metadata.

STEPS
1. For each author, we produce a list of fiction works to be included in the corpus (in the metadata file).
2. Hyperlinks to downloadable sources of the aforementioned works in epub format are collected.

1 Details about the exact sources can be found in the metadata of our corpus.
3. Each fiction work is manually dated using various sources. The result is stored in the metadata file, together with the source used for the dating.

4. We feed the metadata into a first Python script that downloads all epub files.

5. We applied a second Python script to the downloaded files to obtain .txt formats that are stripped of paratext (e.g. prefaces, or license declarations).

**SAMPLING STRATEGY**

To select relevant authors, we searched for authors from the 19th century, whose fiction works are available in the public domain in epub format of good quality. Once works have been preprocessed, they should be at least 100 Kb (~16,500 words) in .txt format. We removed works that have co-authors (for example some posthumous novels by Jules Verne), or authors that were known to work with ghostwriters, e.g. Alexandre Dumas (Chodorowicz, 2019), or works that we were not able to date.

**QUALITY CONTROL**

The first Python script, step1-getEBooks.py, is responsible for the correct naming of all e-books. The second, step2-convertToTei.py, removes prefaces, image descriptions and license declarations by first converting the epub file into a TEI file, using the software, then parsing through the TEI structure and selecting only the text that has been written by the author in the year of writing of the novel. Finally, a manual cleaning phase removed dedications and prefaces that remained undetected.

**(3) DATASET DESCRIPTION**

**OBJECT NAME**

CIDRE.zip

**FORMAT NAMES AND VERSIONS**

Fiction works are distributed in repositories named after the authors’ last name in .txt format. The filenames of the works always start by the year of writing, followed by _ and the title of the novel, with words separated by underscores. For example, 1886_Un_mysterieux_amour.epub.txt in the repository ‘lesueur’.

The metadata of the corpus is stored in a CSV file. The scripts to gather corpora from online libraries (e.g. Wikisource, Project Gutenberg, etc.) must be executed using Python 3. Beforehand, one needs to install the Python packages selenium, Geckodriver (from https://github.com/mozilla/geckodriver/releases), and pandoc (from https://pandoc.org/installing.html).

| AUTHOR                  | NUMBER OF WORKS | EARLIEST | LATEST |
|-------------------------|-----------------|----------|--------|
| Gustave Aimard          | 24              | 1858     | 1881   |
| Honoré de Balzac        | 59              | 1829     | 1848   |
| Paul Féval              | 23              | 1843     | 1881   |
| Henry Gréville          | 36              | 1876     | 1892   |
| Daniel Lesueur          | 31              | 1882     | 1911   |
| Pierre-Alexis Ponson du Terrail | 42 | 1852     | 1870   |
| George Sand             | 62              | 1831     | 1875   |
| La Comtesse de Ségur    | 22              | 1856     | 1871   |
| Jules Verne             | 58              | 1862     | 1905   |
| Michel Zévaco           | 29              | 1906     | 1926   |
| Émile Zola              | 35              | 1864     | 1903   |

Table 1 Summary of the content of CIDRE.

---

2 We chose the plain text format to facilitate import into different NLP tools. However, our programming scripts allow users to produce a TEI-format file from the epub downloads.
CREATION DATES
This corpus set has been created between 2020-10-01 and 2021-04-07.

DATASET CREATORS
Olga Seminck and Philippe Gambette

LANGUAGE
French

LICENSE
Corpus: public domain; metadata: Licence Creative Commons – Attribution – Partage dans les Mêmes Conditions 4.0 International; processing scripts: GPLv3 License.

REPOSITORY NAME
Zenodo and Ortolang

PUBLICATION DATE
2021-03-30
(4) REUSE POTENTIAL

Our resource can be used not only for idiolect studies, but can also serve as data in other contexts, like authorship attribution, stylistic studies, and for literature studies on the genres of realism, naturalism, adventure novels and detectives of the 19th and early 20th century. The fact that four women are included among the eleven authors of our corpus may also open some perspectives in gender studies; see Rybicki (2016) for an example on English. Moreover, the scripts can be re-used by anyone who wants to compose their own corpus of e-books.

ACKNOWLEDGEMENTS

We thank all contributors who worked on the open source projects we collected data and information from: Wikisource, Project Gutenberg, La Bibliothèque électronique du Québec, the website of Les Amis de Daniel-Lesueur, the eBalzac Project and the Gallica platform of the Bibliothèque nationale de France. We also thank the anonymous reviewers for their helpful suggestions.

FUNDING INFORMATION

This work was funded in part by the French government under the management of the Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Olga Seminck: Conceptualization, Data curation, Software, Formal analysis, Writing – original draft

Philippe Gambette: Conceptualization, Data curation, Software, Writing – original draft

Dominique Legallois: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review and editing

Thierry Poibeau: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review and editing

AUTHOR AFFILIATIONS

Olga Seminck [orcid.org/0000-0003-4617-5992]  
Laboratoire Lattice (Langues, Textes, Traitements informatiques, Cognition) – CNRS & ENS/PSL & Université Sorbonne nouvelle, Montrouge, France

Philippe Gambette [orcid.org/0000-0001-7062-0262]  
LIGM, Université Gustave Eiffel & CNRS, Champs-sur-Marne, France

Dominique Legallois  
Laboratoire Lattice (Langues, Textes, Traitements informatiques, Cognition) – CNRS & ENS/PSL & Université Sorbonne nouvelle, Montrouge, France

Thierry Poibeau [orcid.org/0000-0003-3669-4051]  
Laboratoire Lattice (Langues, Textes, Traitements informatiques, Cognition) – CNRS & ENS/PSL & Université Sorbonne nouvelle, Montrouge, France

REFERENCES

Barlow, M. (2010). Individual usage: A corpus-based study of idiolects. Cognitive Linguistics Bibliography [online]. Berlin, Boston: De Gruyter Mouton. Available at: https://www.degruyter.com/document/database/COGBIB/entry/cogbib_19/html [Last accessed 2021-06-11].

Barlow, M. (2013). Individual differences and usage-based grammar. International Journal of Corpus Linguistics, 18(4), 443–478. DOI: https://doi.org/10.1075/ijcl.18.4.01bar

Chodorowicz, S. (2019). The Stylometry of Authorial Collaboration: Alexandre Dumas and Auguste Maquet in French and English, Master’s thesis, Krakow: Jagiellonian University.
Dittmar, N. (1996). Explorations in idiolects. *Amsterdam Studies in the Theory and History of Linguistic Sciences Series*, 4, 109–128. DOI: https://doi.org/10.1075/cilt.138.10dit

Klaussner, C., & Vogel, C. (2018). Temporal predictive regression models for linguistic style analysis. *Journal of Language Modelling*, 6(1), 175–222. DOI: https://doi.org/10.15398/jlm.v6i1.177

Kuhl, J. W. (2003). The idiolect, chaos, and language custom far from equilibrium: Conversations in Morocco. PhD thesis, Athens, GA: University of Georgia.

Mollin, S. (2009). “I entirely understand” is a Blairism: The methodology of identifying idiolectal collocations. *International Journal of Corpus Linguistics*, 14(3), 367–392. DOI: https://doi.org/10.1075/ijcl.14.3.04mol

Pettré, P., Anthonissen, L., Budts, S., Manjavacas, E., Silva, E., Standing, W., & Strik, O. A. O. (2019). Early Modern Multiloquent Authors (EMMA): Designing a large-scale corpus of individuals’ languages. *ICAME journal*, 43(1), 83–122. DOI: https://doi.org/10.2478/icame-2019-0004

Rybicki, J. (2016). Vive la différence: Tracing the (authorial) gender signal by multivariate analysis of word frequencies. *Digital Scholarship in the Humanities*, 31(4), 746–61. DOI: https://doi.org/10.1093/llc/fqv023

Stamou, C. (2008). Stylochronometry: Stylistic development, sequence of composition, and relative dating. *Literary and Linguistic Computing*, 23(2), 181–199. DOI: https://doi.org/10.1093/llc/fqm029