



HAL
open science

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). European Association of Digital Humanities Conference (EADH 2021), Sep 2021, Krasnoyarsk, Russia. hal-03353520

HAL Id: hal-03353520

<https://hal.science/hal-03353520>

Submitted on 24 Sep 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

It is well known that the idiolect (the language of an individual) evolves over time. However, there is a lack of quantitative studies on this topic, due to the lack of large corpora (but see Barlow 2013; Mollin 2009; Petré et al. 2019 for a few examples). To study what is specific in an idiolect and how it evolves over a lifetime, we assembled, cleaned and dated the fiction works of 11 very prolific 19th and early 20th century French writers. This resulted in the CIDRE corpus counting 37 million words and over 400 books.

Motivation

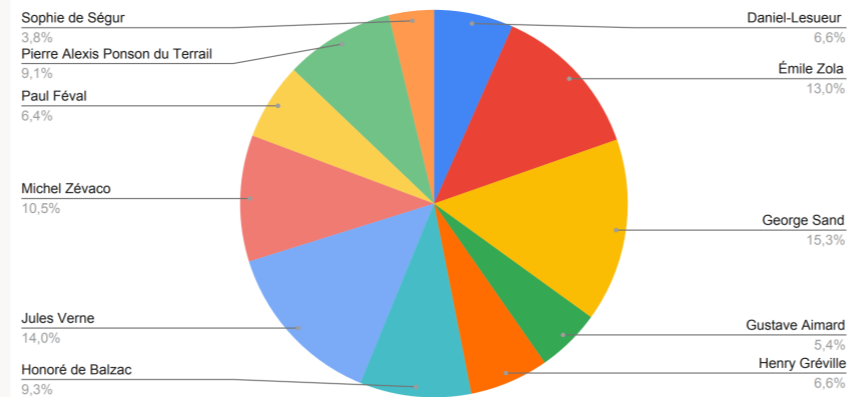
- ▶ We want to assemble a longitudinal corpus for stylistics studies, that is:
 - homogeneous (only fiction work)
 - open (contain only works in the public domain)
 - dense (covering the 19th and early 20th century with overlaps between the different authors)

Corpus Assembling

- ▶ Criteria to Select Relevant Authors:
 - Large oeuvre of fiction novels
 - E-books available as high quality e-pub files
 - No collaborative works (e.g. A. Dumas)
- ▶ Programming Scripts:
 - Automatic download of e-pub files from open source library projects (e.g. Wikisource, etc)
 - Automatic cleaning of e-pub files (remove forewords, image captions, etc)
- ▶ Manual correction

Data

36.9 Million Words



Dating of Works

Important contribution of ours:

Annotation of Books with year of writing

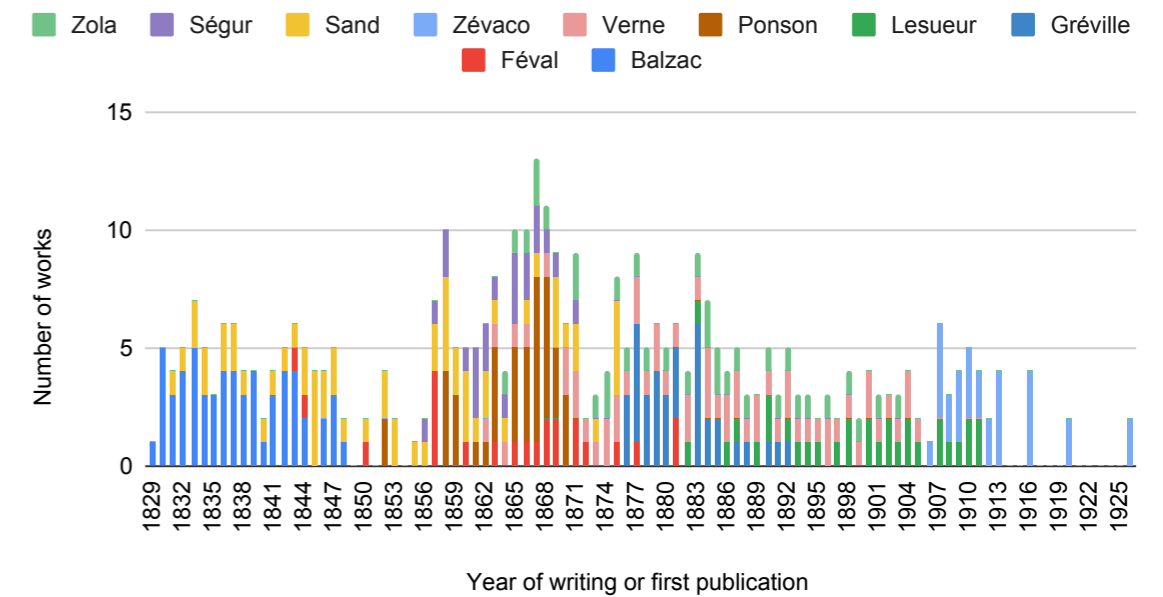
- ▶ Crucial for a diachronical study of the idiolect
- ▶ Might differ largely from:
 - First year of publication
 - Year of printing
- ▶ Provided in a metadata file with the corpus

Title	Year	Source	Comments
L'héritage de Xénie	1880	BnF	1924: Wikipedia
Cité Ménard	1880	Wikipedia, BnF	
Le moulin Frappier	1880	Wikipedia	1881: BnF
Madame de Dreux	1881	Wikipedia, BnF	
Marier sa fille	1881	BnF	
Perdue	1881	Wikipedia, BnF	

Table: Some examples from the Gréville Corpus

Dates of Works

Corpus CIDRE



Availability and Licenses

- ▶ Download: <https://github.com/oseminck/cidre/tree/v2.0>
- ▶ Texts: public domain
- ▶ Metadata: Creative Commons - Attribution-ShareAlike 4.0
- ▶ Processing scripts: GPLv3 License



References

- ▶ Barlow, M. (2013). Individual differences and usage-based grammar. *International Journal of Corpus Linguistics*, 18(4), 443–478. DOI: 10.1075/ijcl.18.4.01bar
- ▶ Mollin, Sandra. (2009). "I entirely understand" is a Blairism: The methodology of identifying idiolectal collocations. *International Journal of Corpus Linguistics*, 14(3), 367–392. DOI: 10.1075/ijcl.14.3.04mol
- ▶ Petré, Peter, Lynn Anthonissen, Sara Budts, Enrique Manjavacas, Emma-Louise Silva, William Standing, Odile AO Strik (2019). Early Modern Multiloquent Authors (EMMA): Designing a large-scale corpus of individuals' languages. *ICAME Journal* 43(1): 83-122. DOI: 10.2478/icame-2019-0004

Olga Seminck, Philippe Gambette, Dominique Legallois & Thierry Poibeau

olga.seminck@cri-paris.org, philippe.gambette@univ-eiffel.fr, dominique.legallois@sorbonne-nouvelle.fr, thierry.poibeau@ens.psl.eu

Centre national de la recherche scientifique, Université Gustave Eiffel, Université Sorbonne nouvelle, Lattice (Langues, Textes, Traitements informatiques, Cognition) - UMR 8094

Funding: ANR-19-P3IA-0001 (PRAIRIE 3IA Institute)