The Evolution of the Idiolect over the Lifetime: A Quantitative and Qualitative Study of French 19th Century Literature

Olga Seminck, Philippe Gambette, Dominique Legallois, Thierry Poibeau

▶ **To cite this version:**

Olga Seminck, Philippe Gambette, Dominique Legallois, Thierry Poibeau. The Evolution of the Idiolect over the Lifetime: A Quantitative and Qualitative Study of French 19th Century Literature. Soirée networking PRAIRIE, Sep 2022, Paris, France. 2022. hal-03784713

HAL Id: hal-03784713
https://hal.science/hal-03784713

Submitted on 23 Sep 2022

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 Evolution of the Idiolect over the Lifetime: A Quantitative and Qualitative Study of French 19th Century Literature

Abstract

We propose new methods to identify, quantify and describe the grammatical-stylistic changes that take place during the lifetime of an author. To examine the strength of the chronological signal of change, we first developed a method to calculate if a distance matrix of literary works contains a stronger chronological signal than expected by chance. 10 out of 11 corpora showed a higher than chance chronological signal. Second, we proposed a machine learning task: predicting the year in which a work was written. The accuracy and the amount of variance that is explained by the model were high for most authors we studied. After applying a feature selection algorithm, we examined the most important ones, i.e. patterns that have the greatest influence on idiolectal evolution.

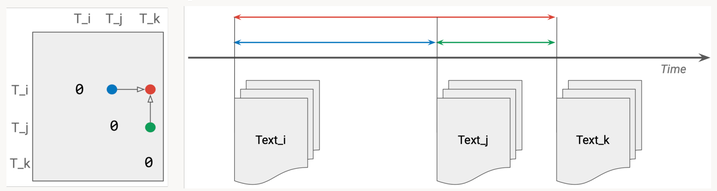
Corpus

CIDRE
Corpus for Idiolectal Research

- 37 million words
- > 400 books dated by year of writing
- 11 prolific French 19th Century Writers
- Download: `https://github.com/oseminck/cidre/tree/v2.0`

Robinsonian Score

- If there is a chronological signal, we expect that two books that are closer in time are more similar than two books further away in time.
- With stylo R, we calculated a distance matrix (δ) of the works of an author.
- We say that δ is Robinsonian if for any set of three distinct texts $text_i$, $text_j$ and $text_k$ such that $\text{date}(text_i) < \text{date}(text_j) < \text{date}(text_k)$, $\max(\delta(text_i, text_j), \delta(text_j, text_k)) \leq \delta(text_i, text_k)$.
- We calculate the rate of cells that are Robinsonian and the probability (P-value) that this rate is found by chance.
- A chronological evolution was found in 10 of the 11 corpora.

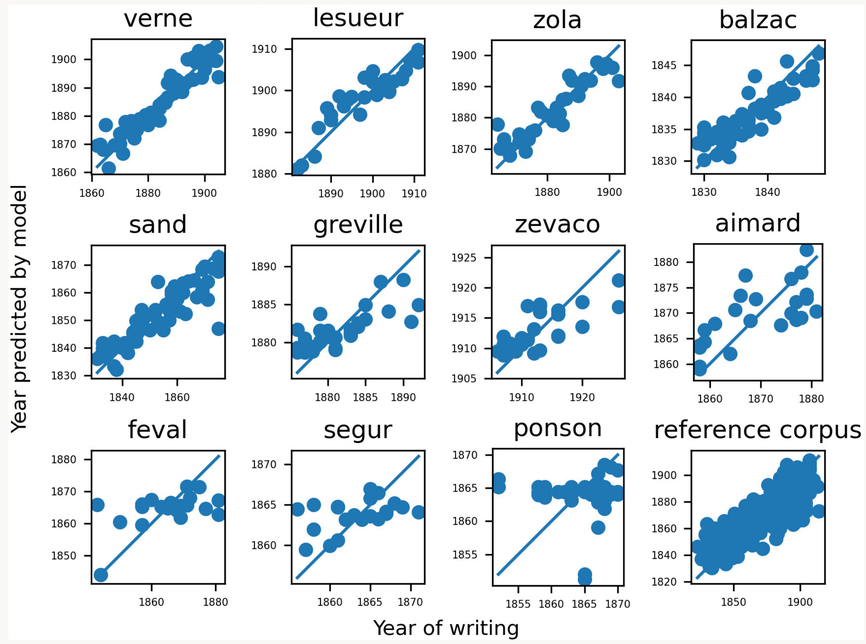


Author	Robinsonian Score	P-value
Comtesse de Ségur	0.38	0.14
Daniel Lesueur	0.41	0.00
Pierre-Alexis Ponson du Terrail	0.41	0.00
Gustave Aimard	0.42	0.01
Honoré de Balzac	0.44	0
Michel Zévaco	0.46	0
Jules Verne	0.47	0
George Sand	0.49	0
Paul Féval	0.49	0.00
Henry Gréville	0.62	0
Émile Zola	0.63	0
Reference Corpus	0.34	0

Table: Rates of number of Robinsonian cells

Predicting Year of Writing with Regression Models

- Goal: predict the year of writing of the books of an author.
- Regression model and feature selection using Lasso Lars, resulting in 10 to 61 stylistic/linguistic patterns per author.
- For example, the increasing pattern "…_DETPOSS_NC_…" in the work of Daniel-Lesueur:
  - Ah ! ma mère … ma mère … pensait Hervé, […]
    Ah ! my mother … my mother… thought Hervé, […]
  - Je suis perdue ! … Perdue ! … Ma chérie … Invente quelque chose ! … Ah ! I'm lost! … Lost! … My darling… Think of something! … Ah!
- Most models were successful.



Reference

Journal of Cultural Analytics, Vol. 7, Issue 3, 2022. CCBY-4.0
`https://doi.org/10.22148/001c.37588`

Olga Seminck, Philippe Gambette, Dominique Legallois & Thierry Poibeau

`olga.seminck@cnrs.fr, 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 UMR 8094

Funding: ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and ANR-16-IDEX-0003 (I-Site Future, programme "Cité des dames, créatrices dans la cité")

PR[AI]RIE
PaRis Artificial Intelligence Research InstitutE

ENS

PSL