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:

HAL Id: hal-03784713
https://hal.archives-ouvertes.fr/hal-03784713
Submitted on 23 Sep 2022
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

- 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 (\( \delta \)) of the works of an author.
- We say that \( \delta \) is Robinsonian if for any set of three distinct texts \( \text{text}_i, \text{text}_j, \text{text}_k \) such that date(\( \text{text}_i \)) < date(\( \text{text}_j \)) < date(\( \text{text}_k \)), \( \max(\delta(\text{text}_i, \text{text}_j), \delta(\text{text}_j, \text{text}_k)) \leq \delta(\text{text}_i, \text{text}_k) \).
- 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.

<table>
<thead>
<tr>
<th>Author</th>
<th>Robinsonian Score</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comtesse de Ségur</td>
<td>0.38</td>
<td>0.14</td>
</tr>
<tr>
<td>Daniel Lesueur</td>
<td>0.41</td>
<td>0.00</td>
</tr>
<tr>
<td>Pierre-Alexis Ponson du Terrail</td>
<td>0.41</td>
<td>0.00</td>
</tr>
<tr>
<td>Gustave Aimard</td>
<td>0.42</td>
<td>0.01</td>
</tr>
<tr>
<td>Honoré de Balzac</td>
<td>0.44</td>
<td>0.00</td>
</tr>
<tr>
<td>Michel Zévaco</td>
<td>0.46</td>
<td>0.00</td>
</tr>
<tr>
<td>Jules Verne</td>
<td>0.47</td>
<td>0.00</td>
</tr>
<tr>
<td>George Sand</td>
<td>0.49</td>
<td>0.00</td>
</tr>
<tr>
<td>Paul Féval</td>
<td>0.49</td>
<td>0.00</td>
</tr>
<tr>
<td>Émile Zola</td>
<td>0.62</td>
<td>0.00</td>
</tr>
<tr>
<td>Reference Corpus</td>
<td>0.34</td>
<td>0.00</td>
</tr>
</tbody>
</table>

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

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”)