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

CA Journal of Cultural Analytics

Seminck, Olga, et al. "The Evolution of the Idiolect over the Lifetime: A Quantitative and Qualitative Study of French 19th Century Literature." *Journal of Cultural Analytics*, vol. 7, no. 3, Sept. 2022, doi:10.22148/001c.37588.

ARTICLE

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

Olga Seminck[1], Philippe Gambette[2], Dominique Legallois[1], Thierry Poibeau[1]

[1] Laboratoire Langues, Textes, Traitements informatiques, Cognition UMR 8094, [2] Laboratoire d'Informatique Gaspard-Monge UMR 8049

Keywords: idiolect, French literature, authorship, stylometry, literature

https://doi.org/10.22148/001c.37588

Journal of Cultural Analytics

Vol. 7, Issue 3, 2022

The way in which authors express themselves is unique but changes over their lifetime. However, quantitative studies of this idiolectal evolution are rare. Using the Corpus for Idiolectal Research (CIDRE) that contains the dated works of 11 prolific 19th century French fiction writers, we propose new methods to identify, quantify and describe the grammatical-stylistic changes that take place using lexico-morphosyntactic patterns, also called motifs. To examine the strength of the chronological signal of change, we developed a method to calculate if a distance matrix of literary works contains a stronger chronological signal than expected by chance. Ten out of 11 corpora showed a higher than chance chronological signal, leading us to conclude that the evolution of the idiolect is in a mathematical sense monotonic, supporting the rectilinearity hypothesis previously put forward in the stylometric literature. The rectilinear property of the evolution of the idiolect found for most authors in CIDRE subsequently enabled us to propose a machine learning task: predicting the year in which a work was written. For the majority of the authors in our corpus, the accuracy and the amount of variance that is explained by the model were high and we discuss why the technique might fail for others. After applying a feature selection algorithm, we examined the most important features, i.e. the motifs that have the greatest influence on idiolectal evolution. We find that some of those features are stylistic and have been previously identified in qualitative literature studies. We report some remarkable stylistic constructions revealed by our algorithm to illustrate which kind of stylistic patterns can be extracted using our method.

1. Introduction

Is it true that we do not speak at 20 as we do at 60? In this article we examine if an individual's representation of a language — the idiolect — and the utterances that are its product, are fixed once and for all or not. Little research has been carried out to characterize and measure the evolution of the idiolect in an extensive manner. Our main goal was thus to develop methods in this direction that can also be applied to other longitudinal corpora. We evaluated our methods on a corpus containing the idiolects of 11 French fiction writers.

There are several reasons why it is interesting to take into account interpersonal variation over time. First, the notion of idiolect is relevant for corpus linguistics. As Heck stressed, idiolects are the primary objects of study in linguistics (in the end, we can only observe utterances that are the products of idiolects). However, most often corpora are considered as homogenous, and do not take into account the influence of individual authors on their content. It is somehow assumed that the large number of different authors in a resource will erase any individual differences. But taking these differences into account could

help better understand to what extent some features are specific to a genre, a community or, on the contrary, to an individual author or speaker. In this article, we focus on the chronological evolution of the idiolect.

Let us start by introducing some important terminology. The first term that needs to be clarified is idiolect. In Bloch's original definition, the idiolect represents: "The totality of the possible utterances of one speaker at one time in using a language to interact with one other speaker." More recently, Dittmar (111) proposed this definition: an idiolect is "the language of the individual, which because of the acquired habits and the stylistic features of the personality differs from that of other individuals and in different life phases shows, as a rule, different or differently weighted CM [communicative means]".

In addition to this, we believe that the definition of idiolect should take into account the fact that every utterance (written or oral) of an individual is part of a particular discursive practice — or, put differently, of a particular textual genre (informal conversation, tweet, philosophical essay, etc.). The idiolect should thus necessarily be considered in relation to a particular practice: it corresponds to the use by an individual of only part of the possible linguistic forms related to a discursive practice. Bloch takes this into account when he states "that a given speaker may have different idiolects at successive stages of his career, and [...] that he may have two or more different idiolects at the same time."

Another relevant notion is style. Style corresponds to linguistic forms that an observer considers as remarkable from an aesthetic point of view, in a particular discourse, compared to the discourse of others. Although the definitions of idiolect and style are intertwined, especially for those working with literary corpora, the main difference is that, for stylistic studies, a judgement has generally to be performed on the stylistic value of the linguistic phenomena under study. It should be noted that the notion of stylistic judgement is in itself highly subjective, and no clear criteria seem to be available to determine what has a stylistic value and what does not. Consequently, we focused on the evolution of idiolects, instead of styles, so as to avoid aesthetic judgements. We therefore use Bloch's definition of the idiolect that includes "the totality of possible utterances", instead of Dittmar's that focuses on stylistic features.

In this article, after the presentation of related work, we present two computational experiments to study the chronological evolution of the idiolect, followed by a qualitative analysis. We start by examining the rectilinearity hypothesis mentioned in the work of Stamou, i.e. *"the hypothesis that certain aspects of an author's writing style evolve rectilinearly over the course of an author's lifetime, hence with appropriate methods and stylistic markers, such changes ought to be detectable"*. First, we evaluate the chronological signal in corpora including the dated works of French 19th century authors by so-called Robinsonian matrices. Second, we build linear regression models for

each author studied to see whether it is possible to predict the year in which a particular novel was written by extrapolation from other works by the same author. Our regression models rely on special linguistic-stylistic patterns, called *motifs* (Legallois et al.) and are able to identify the patterns that play the greatest role in the chronological evolution of the idiolect. We discuss the stylistic value of some of these motifs in a qualitative analysis presented in section 6. The article ends with a discussion and conclusion.

2. Literature Review

Our work is directly in line with the notion of stylochronometry, a special research "niche" that studies the diachronic evolution of style. The term was coined by Forsyth and encompasses the characterization of style according to different time periods, as well as the attribution of tentative dates to literary works. Stamou reviewed a large number of studies on this topic, discussing literary works by writers such as the poet W.B. Yeats (Jaynes), or the prose of Samuel Becket (Opas), to the lyrics of the Beatles (Whissell). The review draws some important conclusions that are still valid today. The first is that even though dating methods would eventually be most useful to date works with an uncertain date of creation — such as texts by Plato, Euripides and Shakespeare — these methods should be developed, tested and evaluated using gold standard corpora (where texts can be reasonably associated with a precise date of creation so as to get a solid performance evaluation), a criterion that was already underlined by Forsyth. Despite this, there is a large literature on the topic of stylochronometry with experiments investigating only works with problematic dating (e.g. Cox and Brandwood; Wishart and Leach; T. M. Robinson; Ledger; Temple on Plato's works; Devine and Stephens; Cropp and Fick; Smith and Kelly on Euripides, and Brainerd; Derks; Jackson on Shakespeare), making it difficult to evaluate the results. Experiments in which the methods used are carefully compared to reference corpora with known dates are rather rare (however, see Can and Patton, on two modern Turkish writers). Daelemans also underlines this evaluation problem and suggests using evaluation methods from the field of Natural Language Processing (NLP). In the same vein, Craig states that "stylistic analysis needs finally to pass the same tests of rigor, repeatability, and impartiality as authorship analysis if it is to offer new knowledge".

Specifically for works on French, we came across studies using off-the-shelf methods developed for statistical textual analysis (Pincemin), for instance stylo R (Eder et al.), Lexico (Lamalle et al.), TXM (Heiden et al.), Le Trameur (Fleury and Zimina), and Hyperdeep (Vanni et al.). For example, Guaresi et al. study the evolution of style on a corpus of the annals of the congress of the French communist party from 1936 to 2018 using correspondence analysis on the vocabulary. However, a drawback of off-the-shelf methods is that they provide mainly exploratory analyses or visualizations which leave considerable room for interpretation and cannot be used directly to focus on specific stylochronometry questions or hypotheses with a rigorous evaluation

procedure. We will therefore not go into detail about this type of study but will instead discuss more focused studies that have, in our opinion, developed interesting and relevant approaches for the task.

Mollin investigated Tony Blair's idiolect. Her goal was to identify "maximizer collocations" (collocations involving adverbs such as *fully*, *entirely* or *absolutely*) that were specific to Blair's idiolect, when compared to the English language in general (comparing a three million word corpus of Tony Blair with the British National Corpus (BNC XML Edition)). Collocations were selected using the three measures: relative frequency, Mutual Information (Church et al.) and the log-likelihood measure (Dunning). A series of collocations that are typical of Blair's idiolect was identified using these three measures.

Mollin's article nicely combines quantitative and qualitative analysis and presents a clear methodology. Unfortunately, we could not find an online accessible repository of the Tony Blair Corpus, but the methods are explained to an extent that the research should be replicable. Mollin's results suggest that the notion of idiolect is indeed a relevant linguistic concept, and that there are some linguistic patterns that are highly idiosyncratic for a speaker (even though her study only included one individual).

A second researcher working on the notion of idiolect is Barlow. He studied the idiolect of five White House Press Secretaries who held this function from 1 to 4 years. For each person the author collected a corpus of approximately 200K to 1200K tokens and compared the individual frequencies of the most frequent bigrams (lexical and Part of Speech) of each press secretary against the others. He showed that individual patterns are highly recognizable and that inter-speaker variability is much larger than intra-speaker variability. Moreover, he found that the inter-speaker differences were "core aspects of language and not peripheral idiosyncrasies", meaning that they play a role in the use of function words and high frequency words, such as 'by the' and 'we have'. He also found that the speech of an individual remained remarkably stable over time, but of course, one needs to keep in mind that the maximum period for a secretary in the corpus was only four years.

Another study that concluded in favour of the staticness of the idiolect is Meyerhoff and Walker. They tried to determine to what extent the grammar of individuals is morpho-syntactically similar to that of a community. They studied the absence of the verb 'be' in a community speaking an English-based creole compared to other members of this community who had joined an urban community speaking a more 'standard' English. Their conclusions are mixed, suggesting that, despite the possibility of the idiolect evolving, conservatism can also play an important role. However, it should be noted that this study only applied to one grammatical construction in a multilingual setting, so that the reported results may be hard to generalize.

Evans wrote her PhD thesis on diachronic morpho-syntactic changes in the idiolect of Queen Elizabeth I from a sociolinguistic perspective by comparing her letters, speeches and translations (forming a corpus of 78K tokens) from before her ascension to those from the period after this event, which is often speculated to have had the greatest influence on Elizabeth's language by other scholars. Interestingly, Evans used a reference corpus — the Corpus of Early English Correspondence (Raumolin-Brunberg and Nevalainen) — and previous studies on it to identify 9 morpho-syntactic features present in this corpus and the corpus of Queen Elizabeth. The goal was to see whether the two corpora evolved in the same way. The author found that the ascension to the throne only influenced two features (the increase in the use of the royal 'we' and the decrease in the use of periphrastic superlative adjectives), and that time in general and the long education of Queen Elizabeth had a constant influence on the development of her idiolect. Evans' study provides a good example of how corpus linguistics and qualitative analysis can jointly contribute to the study of the evolution of the idiolect.

In the field of stylochronometry, the work of Klaussner and Vogel of 2015 and 2018 and Klaussner of 2017 should be mentioned. They developed regression models and evaluated them on two individual corpora of North American writers, a reference corpus and against a baseline. They used a machine learning task that aimed to predict the year of writing of a given work, using a relevant evaluation metric. Their methods play an important role in the second part of our quantitative study (Section 5) and will therefore be discussed in more detail below. However, we can already conclude that the methods proposed by Klaussner and Vogel show how quantitative machine learning methods can be used to fuel qualitative research on stylistic changes.

We have said that relevant large scale resources in this domain are scarce. We should however mention a large recent resource: the EMMA corpus (Petré et al.). It features 87 million words of prolific English 17th century writers. Various studies on the evolution of the idiolect were conducted using this resource, for example Petré and Van de Velde — although using an earlier slightly smaller version of the corpus — investigated the role of individual language users and the language community in the semantic and morphosyntactic process of grammaticalization of a specific construction: 'be going to INF'. They show how the rate of this construction was influenced before, during and after its conventionalization and observe differences between generations. Their method shows how the process of grammaticalization can be studied for individuals but also for a community at the same time. Anthonissen and Petré also show how the use of larger corpora helps to study lifespan changes affecting syntactic constructions; they demonstrate by the example of the construction 'be going to' that individual writers in the EMMA corpus can adopt and continue to participate in grammatical innovations during adulthood.

Contrary to most of the studies examined in this section, our approach takes into account all kinds of patterns (called motifs) and not only a handful of predefined and carefully selected sequences. With this more comprehensive approach, we hope to produce a more robust and reliable model of the evolution of the idiolect.

3. Corpus

For this study, we used the Corpus for Idiolectal Research (CIDRE) (Seminck et al.). This corpus features 11 French 19th and early 20th century prolific fiction writers, with a total of 37 million words. All the works have been carefully dated and the corpus includes only works of fiction, so that the problem of individuals having different idiolects related to their social situation does not interfere.

To address the question of diachronic language evolution in general in opposition to idiolectal evolution, we assembled a 'reference corpus'. This corpus contains 361 works of fiction by French authors from the same time period as the works in CIDRE, but no particular attention was paid to individual authors: they can be included in the resource if they wrote only one work. To assemble the reference corpus, we used the online tool GutenTag (Brooke et al.) that enables one to download a subcorpus from Project Gutenberg. With a semi-automatic approach to run GutenTag several times, to filter sufficiently long books and to automatically date the first edition of each work using the catalogue of the Bibliothèque nationale de France, we were able to obtain a total of 361 works of fiction in French, dated with a reasonably good precision (mean error of 1.71 years per book for books present in the CIDRE corpus). The quality of this corpus can be considered sufficient for it to be used as a reference corpus that serves to account for language change in general. Note that there is a substantial overlap between the content of the CIDRE corpus and our reference corpus (146 novels out of the 361).

4. Testing the rectilinearity hypothesis

In the introduction, we mentioned the rectilinearity hypothesis which posits that some aspects of the idiolect evolve in a linear fashion over time and that this evolution should be detectable. Importantly, the hypothesis does not say that this evolution is relevant for all linguistic features and should affect the same features for each individual. The use of some linguistic features may remain stable, but some evolution should nevertheless be observed for some others, contradicting the conservatist hypothesis (which assumes no linguistic changes). Furthermore, we add the prediction that idiolects evolve constantly and do not return to earlier stages, even if some linguistic features might.

4.1. Methods: Robinsonian Matrices

Robinsonian matrices are distance matrices that have cells whose values increase when moving away from the diagonal. They were introduced in the context of archeological deposits, to study the chronological evolution of the

Table 1. An example of a Robinsonian distance matrix: both ($text_1$, $text_2$) and ($text_2$, $text_3$) are lower than ($text_1$, $text_3$).

	$text_1$	$text_2$	$text_3$
$text_1$	0	2	4
$text_2$		0	1
$text_3$			0

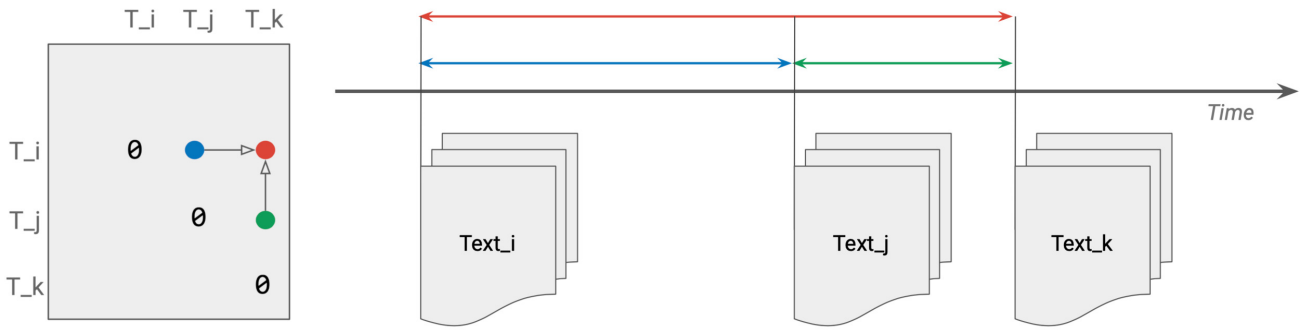

Figure 1. Illustration of the idea of Robinsonian distances between texts.

The colored dots (schema on the left) and arrows (schema on the right) represent distances. If the chronology of the text in the corpus is reflected in the measured distances, we expect that $\max(\delta(text_i, text_j), \delta(text_j, text_k)) \leq \delta(text_i, text_k)$ is true.

style of pottery fragments (W. S. Robinson). More formally, applying this concept to texts, given a matrix δ expressing the distance between novels, we say that δ is Robinsonian if for any set of three distinct texts $text_i$, $text_j$ and $text_k$ such that $\text{date}(text_i) < \text{date}(text_j) < \text{date}(text_k)$,
$\max(\delta(text_i, text_j), \delta(text_j, text_k)) \leq \delta(text_i, text_k)$.

To evaluate the rectilinearity hypothesis on a distance matrix reflecting changes in the idiolect, we measure to what extent the distance matrices corresponding to the texts of the CIDRE corpus are Robinsonian. In order to do this, we can compute the *Robinsonian score*, which we define as the percentage of triples of cells $(\delta(text_i, text_j), \delta(text_j, text_k), \delta(text_i, text_k))$, for which the inequality above is verified.

It is also possible to estimate a *p*-value, i.e. the probability that a Robinsonian score as high as the one being tested could be obtained by chance, by evaluating this score again after randomly changing the order of the texts. Getting a low *p*-value would support the rectilinearity hypothesis.

Before evaluating the rectilinearity hypothesis on the CIDRE corpus, we first used a dated corpus of Maurice Leblanc as a testbed to compare different feature representations for the texts and different ways of measuring distance. We used tokens, characters, lemmas and so-called *motifs* (Legallois et al.) as features. A motif is a sequence of lemmas and POS-tags. As function words tend to be the most relevant features of idiolectal signals (Barlow), grammatical information, i.e. function words and POS-tags, are crucial for the task. However, the tagset of our part-of-speech tagger is not fine-grained enough, losing important information for some categories. For example the difference

Table 2. Examples of unigrams and bigrams and the different types of features

"Il est fâcheux que cela traîne en longueur…"	tokens	characters	lemmas	motifs
unigrams	['Il', 'est', 'fâcheux', 'que', 'cela', 'traîne', 'en', 'longueur', '...']	['I', 'l', ' ', 'e', 's', 't', ' ', 'f', 'â', 'c', 'h', 'e', 'u', 'x', ' ', 'q', 'u', 'e', ' ', 'c', 'e', 'l', 'a', ' ', 't', 'r', 'a', 'î', 'n', 'e', ' ', 'e', 'n', ' ', 'l', 'o', 'n', 'g', 'u', 'e', 'u', 'r', '...']	['il', 'être', 'fâcheux', 'que', 'cela', 'traîner', 'en', 'longueur', '…']	['il', 'être', 'ADJ', 'que', 'cela', 'PRES', 'en', 'NC', '...']
bigrams	[('Il',"est"), ('est',"fâcheux"), ('fâcheux',"que"), ('que',"cela"), ('cela',"traîne"), ('traîne',"en"), ('en',"longueur"), ('longueur',"...")]	['Il', 'l ', ' e', 'es', 'st', 't ', ' f', 'fâ', 'âc', 'ch', 'he', 'eu', 'ux', 'x ', ' q', 'qu', 'ue', 'e ', ' c', 'ce', 'el', 'la', 'a ', ' t', 'tr', 'ra', 'aî', 'în', 'ne', 'e ', ' e', 'en', 'n ', ' l', 'lo', 'on', 'ng', 'gu', 'ue', 'eu', 'ur', 'r...']	[('Il',"être"), ('être',"fâcheux'),('fâcheux',"que'), ('que',"cela'), ('cela',"traîner'), ('traîner',"en'), ('en',"longueur'), ('longueur',"...")]	[('Il',"être'), ('être',"ADJ')),('ADJ',"que'), ('que',"cela'), ('cela',"PRES'), ('PRES',"en'), ('en',"NC'), ('NC',"...")]

between 'un' and 'le' ('a' and 'the') is ignored, and both are tagged as determiners. We therefore used the following strategy: content words were replaced with their POS-tags while function words were replaced with their lemma. This approach allowed us to keep relevant linguistic information, especially at the grammatical level. Legallois et al. proved that these motifs were effective in finding author-specific style characteristics, making it possible to identify interesting examples in corpus studies.

We compared different lengths of n-grams (unigrams to pentagrams) of tokens, characters, lemmas and motifs (see Table 2 for examples of different types of features). The texts of the corpus were represented by the top 500 features with the highest relative frequency. The different distance metrics we used come from stylo R (Eder et al.), in which we entered our corpora. Figure 2 shows the percentage of the distance matrix (calculated for the Leblanc corpus) that is Robinsonian for different feature configurations.

We therefore decided to detect the chronological signal in the corpora of CIDRE and the reference corpus using motif trigrams and the canberra metric as the default option, since trigrams are a medium size and the canberra metric performed slightly better than the others. However, the choice of motifs was motivated by the use of these features in the second series of experiments presented next in this section and by the fact that the lower part of the error bar in Figure 2 is at the highest level of all four tested features. The feature vectors in this experiment contain the scores of the 500 features with the highest relative frequencies.

4.2. Results

The scores for the different authors of the CIDRE corpus and the reference corpus can be found in Table 3. To know whether these scores are meaningful, we compared them with a distribution of random permutations of the distance matrix. For 10 000 random permutations, we calculated the percentage that obtained a Robinsonian score higher than the score of the actual distance

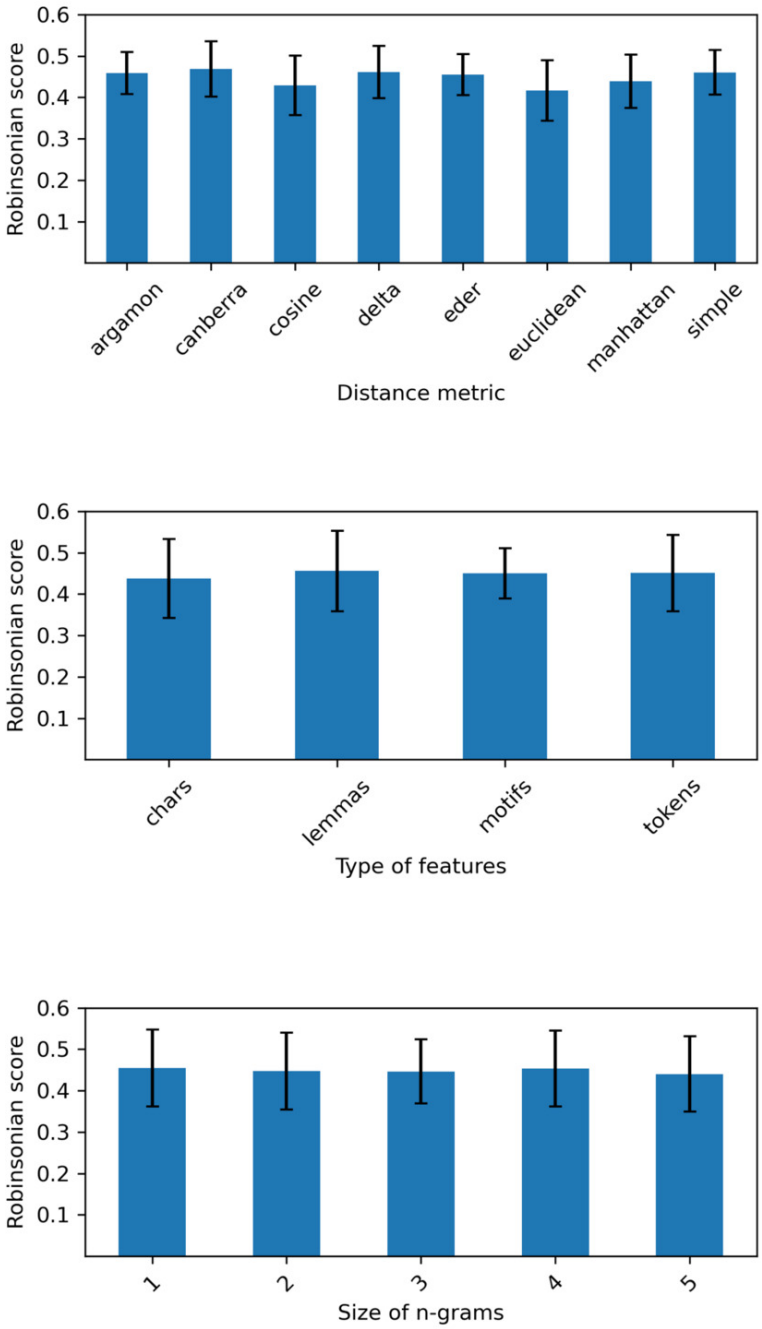

Figure 2. Robinsonian scores for different configurations of features used (eight different distance metrics, four different types of features, five different lengths of n-grams).

Each combination of distance metric, type of features and length of n-gram was tested. Error bars represent the 95% confidence interval of the mean score of all configurations that includes a given parameter. No configuration is significantly better than others. The highest result of 0.50 is obtained using quadrigrams of lemmas with the canberra metric, but this experiment does not allow us to identify a combination of features that is significantly the best to capture the chronological signal in the data.

matrix. Table 3 demonstrates that the distance matrices obtained are significantly more Robinsonian for all the authors than random permutations, except for Comtesse de Ségur.

Table 3. The Robinsonian scores for the authors in CIDRE and the reference corpus, followed by the probability of obtaining these scores by chance if the chronological signal in the data is absent.

Corpus	Robinsonian score	Probability of Robinsonian score if no chronological signal is present in the data
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

Decimal numbers come from rounding and plain zeros have not been rounded.

4.3. Discussion

First, it should be noted that the percentage of Robinsonian cells in the matrix is dependent on the number of works in the corpus. Larger corpora lower the probability of getting Robinsonianness by chance. Therefore, the score of 0.34 for the reference corpus seems low, but is actually very high for this number of works. Second, it should be kept in mind that only the absolute order of works plays a role in our method and that it does not take into account the exact difference in years between works. That is to say: if $\max(\delta(text_i, text_j), \delta(text_j, text_k)) \leq \delta(text_i, text_k)$ is false, it does not matter how much more $\max(\delta(text_i, text_j), \delta(text_j, text_k))$ is than $\delta(text_i, text_k)$.

The fact that different types of features produce similar results on the Maurice Leblanc corpus is not that unexpected regarding the literature. Stamou identified a number of stylistic markers that were of interest in many stylochronometric studies, namely: punctuation, characters, part of speech tags, most common words including function words, frequencies of selected content words, hapax and vocabulary richness. She suggested that there might not be a "single universal stylochronometer" that can apply to every corpus.

The results from our experiments show that there is a strong chronological signal in the data, except for the corpus of Comtesse de Ségur. A possible explanation for this exception could be that this corpus is too small, representing only 3.8% of the total tokens in CIDRE. Another explanation is that this corpus might be heterogeneous, as it includes children's stories, bible stories for children and fairy tales. However, in general our results are in line with the rectilinearity hypothesis: the style of an author generally evolves smoothly over time. No regression (texts stylistically similar to earlier texts) can be observed.

In the next section, we discuss our second series of experiments in which we trained a linear regression model to automatically predict the date of writing of various novels from our reference corpus.

5. Predicting Year of Writing using Linear Regression

In this section, we first examine the chronological evolution of the idiolect (and the reference corpus) by training models on the corpora of idiolects in CIDRE and then predicting the year of writing of different works using cross-validation. The hypothesis is simple: if this type of experiment is successful, the results are in favor of the rectilinearity hypothesis. In other words, the frequency of some linguistic forms increases or decreases in a linear fashion to such an extent that we can detect the year of writing. Second, we do not just want to verify if there is a chronological signal, but also if we can identify the linguistic material at the heart of this evolution. Therefore, we will present a feature selection method that identifies the features that change the most in frequency over time. These features will be used for the qualitative study in section 6. Furthermore, besides these hypotheses and goals, we also have the general objective of proposing new ways to evaluate stylometric methods. For this purpose, we will have recourse to the state of the art literature on linear regression models and verify that it can be used in the stylochronometry context.

5.1. Methods: Regression Models

Various previous studies have used regression techniques in order to date literary works. For example, Frischer used regression techniques (among other methods) to date the *Ars Poetica* of Horace. However, by today's standards, the number of features in this regression was very low so we will mainly discuss more recent work. A representative study using regression is Klaussner and Vogel work from 2018 (henceforth K&V). They used it in a machine learning task that consisted in predicting the year of writing of a work, focusing on two corpora in English: the work of Henry James and that of Mark Twain. They also used the years 1860-1919 of The Corpus of Historical American English (COHA; Davies et al.) as a reference corpus to capture the 'general' language in North America at the time, to check whether the changes detected in the work of James and Twain were shared by the community or were idiosyncratic. Four types of features were considered: character n-grams, part-of-speech tags, word stems, and lemmas with POS-tags; for each of them unigrams to quadrigrams were tested. This resulted in a total of 32 models (2 authors, 4 types of features and 4 types of n-grams). On each model, the elastic nets algorithm was applied to reduce the number of parameters. The models were evaluated using the measure of root mean squared error (RMSE), which reflects the difference (measured in years) between the prediction and the real year of writing. At this point, note that one should keep in mind that this metric is quite sensitive to outliers, as the error is squared. A baseline performance was also measured "*by using the mean of the data for the prediction of every instance*"; meaning that every work that was dated by the baseline received the same prediction (the

mean of all training instances). This would correspond to a model that has a R^2 score of 0 (Field). The best results were obtained by K&V using lemmas with POS-tags in unigrams and bigrams.

Although our experiments share many similarities with K&V, we made some different choices for our models. First, we used only motifs consisting of n-grams (unigram to pentagram, but all incorporated in the same model instead of different models as in K&V). The notion of motif is anchored in previous studies: it has been shown that they are helpful through qualitative analysis (Legallois et al.). Second, the feature section algorithm we chose is Lasso LARS (Efron et al.) with cross validation of 5 (80% training, 20% testing) and not elastic nets. We chose this algorithm because our aim was not to find the most compact model, unlike K&V, but a model that drastically reduces the number of features so that they can be inspected manually (see our qualitative study, Section 6 of this paper). Moreover, as a selection criterion of features, we require that the features be present in at least 20% of an author's texts, whereas in the work of K&V, features had to be present in all data points. This much lower threshold was chosen here because we think it is possible — at least theoretically — for a language innovation to be totally new or for some structures to entirely disappear. Also, K&V concatenated texts written in the same year into one data point by putting the texts behind each other in the same file, whereas we kept them as separate data points with the same value for the year, since we believe that this better represents the data. However, to ensure comparability with K&V, we decided to measure the RMSE and the RMSE-baseline for our experiments. We also compared our results to our reference corpus, and tried the algorithm of elastic nets,[1] as well as elastic nets cross validated.[2] In the end, however, we found that Lasso LARS cross validated performed much better on most of our corpora and that the number of features it selected was better suited for qualitative studies (the elastic nets selected either no features or thousands of features, making a qualitative study impossible). Details about the comparison of feature selection algorithms can be found in the supplementary material.[3]

5.2. Results

For every author, we measured the correlation between the actual year and the predicted year and the value of R^2 (expressed between 0 and 1), which represents the amount of variation of the data that is explained by the model (Field). The results can be found in Table 4 and Figure 3. Excellent results were obtained for Jules Verne, Émile Zola, George Sand, Henry Gréville, Daniel-Lesueur and Honoré de Balzac: the models (selected n-grams of motifs) were capable of predicting the large majority of the variation in the data. The models

1 https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html

2 https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNetCV.html#sklearn.linear_model.ElasticNetCV

3 The filename is: 'results_LassoLars_vs_ElasticNet.txt'.

Table 4. The regression experiment was very successful in explaining the variance of the corpora in gray, and considerable for the other corpora, except for Pierre Alexis Ponson du Terrail, where it was inefficient.

Author	Correlation	R^2	#β	RMSE	RMSE-b	Remarks
Jules Verne	0.94	0.89	57	3.91	11.83	
Daniel-Lesueur	0.92	0.84	14	3.39	8.46	
Émile Zola	0.92	0.83	34	4.50	11.02	
Honoré de Balzac	0.90	0.78	42	2.44	5.26	
George Sand	0.88	0.77	61	6.13	12.78	
Henry Gréville	0.78	0.55	31	2.85	4.27	Regressors in active set degenerate (1/5 folds)
Michel Zévaco	0.75	0.55	23	3.52	5.22	ConvergenceWarning: Regressors in active set degenerate (2/5 folds)
Gustave Aimard	0.70	0.49	21	5.96	8.21	
Paul Féval	0.51	0.26	17	8.72	10.14	ConvergenceWarning: Regressors in active set degenerate (3/5 folds)
Comtesse de Ségur	0.45	0.18	18	3.57	3.96	ConvergenceWarning: Regressors in active set degenerate (1/5 folds)
Pierre Alexis Ponson du Terrail	-0.04	-0.55	10	5.69	4.57	
Reference corpus	0.84	0.70	208	11.30	20.50	

explained a substantial amount of variation in the data for the authors Michel Zévaco, Gustave Aimard, la Comtesse de Ségur and Paul Féval, but less than half of it. Lastly, for Pierre Alexis Ponson du Terrail, the model was not able to explain any variance in the data, and thus the experiment was not successful at all. The same observations can also be made by comparing the evaluation metric root mean squared error (RMSE) and the baseline metric (RMSE-baseline) put forward by K&V.

It is important to mention that the modelling does not always (completely) converge for a given K-fold. This problem is mostly noticeable for Paul Féval. There seems to be a relation between the performance and this issue, but convergence cannot explain the poor performance of the model on the corpus of Pierre Alexis Ponson du Terrail: all the models on this corpus converged. We also had a look at which works were well predicted and which ones were outliers. See for example Figure 4, where it can be seen that *L'Auberge des Saules* by Daniel-Lesueur was predicted about 6 years too late, but *Comédienne* exactly at the time it was written. Other figures in the same style can be found in the supplementary material (directory *plots_regression_par_auteur*).

5.3. Discussion

The average proportion of each author in our whole corpus is 9% since we have 11 authors. Our results show that larger corpora (Sand: 15.3% of the whole corpus, Verne: 14%, Zola: 13% and Zévaco: 10.5%) perform very well and that smaller corpora obtain lower scores (Ségur: 3.8%, Aimard: 5.4%, Féval: 6.4%). This suggests that Petré and Van de Velde are right when they say that corpus size matters a lot, even if it does not explain why our worst performing corpus

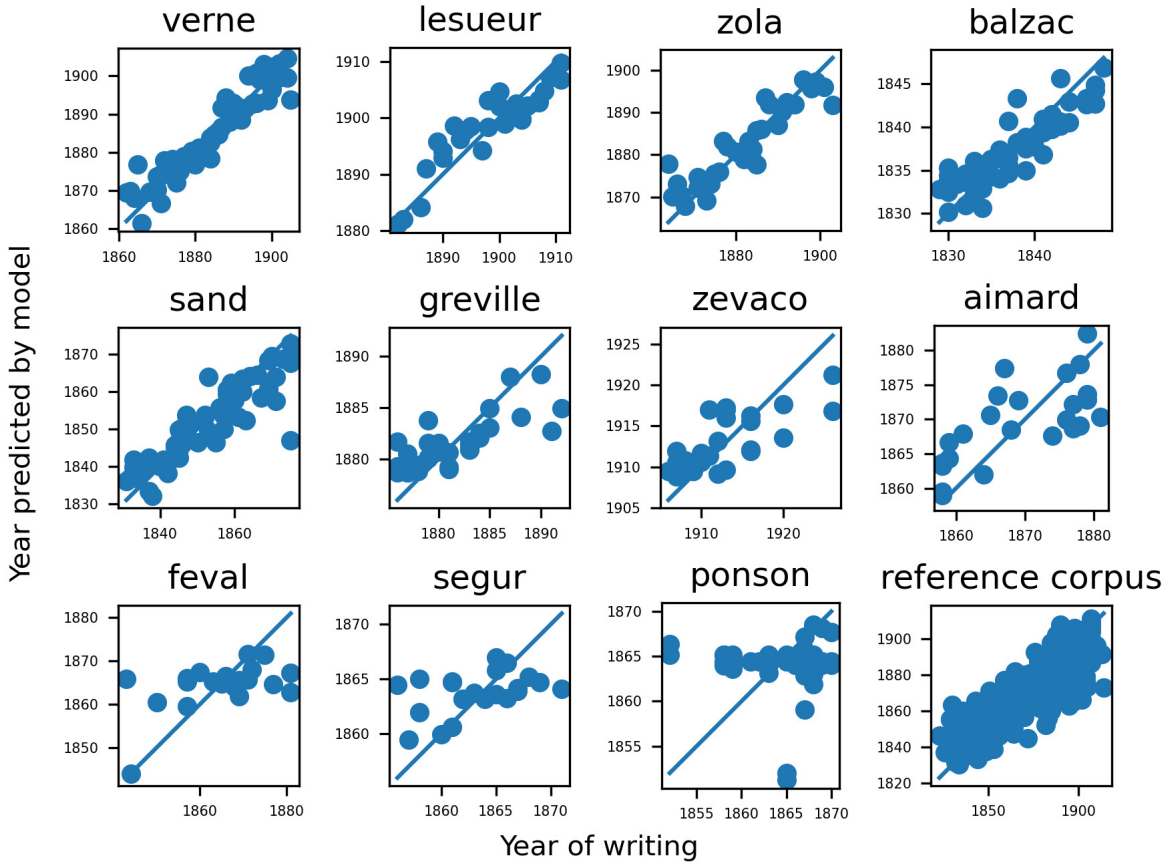

Figure 3. The results of the regression experiment for all corpora in CIDRE, sorted by performance (left to right, top to bottom).

The blue line represents a perfect correlation.

is Pierre Alexis Ponson du Terrail, which is a medium size corpus (representing 9.1% of the total size of CIDRE). However, it is possible of course that we are not aware of certain circumstances in the publication process of the different authors that might explain these results or that the dating of this corpus is of lesser quality. For example, the literary work of Ponson du Terrail is less well known than that of some other writers in CIDRE: his works were dated using information mostly from Wikipedia, which is not as reliable a source as those used for other authors. If the poorer dating of some novels by this writer is the source of the failure of our model, it means that our method is sensitive to individual data points. Indeed, going back to the previous experiment and Table 3, we see that there is a highly significant chronological signal for this corpus, which means that the approach works globally and that specific cases of failure should be further investigated.

However that may be, for most of the authors we get a very high value of R^2, which means that the chronology can explain almost all the variance of the models. This is confirmed by the fact that the value of RMSE is much lower than the RMSE-baseline value. We can thus conclude that the results on all our corpora minus one are in line with the rectilinearity hypothesis.

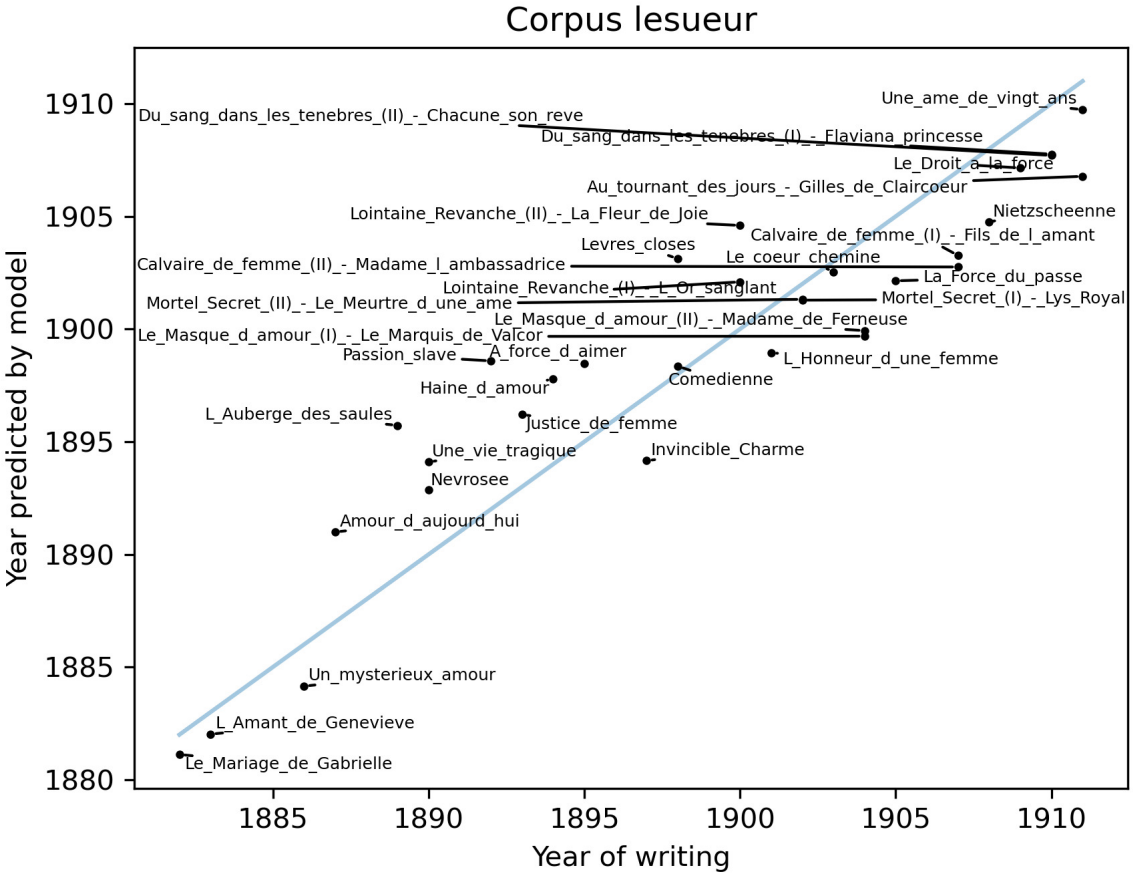

Figure 4. The result of the regression experiment for the corpus of Daniel-Lesueur with annotated data points on the scatterplot.

The blue line represents a perfect correlation.

A second interesting result that we obtained is the number of features selected per model (column #β in Table 4). For all the corpora in CIDRE the number lies between 10 and 61, which are numbers that make it possible to examine all the features of a corpus in a qualitative study. A direct comparison with K&V is difficult, as we did not work on the same corpora. However, we observe that the number of predictors per model (column #β in Table 4) has a smaller range for the different models we developed (the models of K&V range from counting 1 predictor to 315). Nevertheless, it should be kept in mind that our feature selection algorithm probably removes correlated features and that the random seed plays a role in which ones. Therefore, for a complete qualitative study of one author, it might be worthwhile repeating the regression experiment a number of times with different random seeds.

The features that characterize the evolution of an author do not necessarily have to be frequent. While some features are quite frequent, such as the increasing motif 'y' (anaphor referring to a prepositional phrase beginning with the preposition 'à') found for Zola, which obtains the relative frequency of 0.0027 at its maximum, some others, such as the decreasing motif 'autrui' (other people), have a low relative frequency, 0.0008 at its maximum. This

conclusion is similar to a finding in Koppel and Schler, who found that idiosyncratic features that play an important role in authorship attribution also tend to be of low frequency. How we explore the features of the models will be discussed in the next section.

6. A Qualitative Study of some Motifs Sensitive to Diachronic Change

In this section, we look more closely at features selected by the regression algorithm (presented in the previous section). These features, or motifs, are at the heart of idiolectal evolution and are assumed to be easily interpretable. Let's examine if this is true.

6.1. Methods: Manual Inspection

We scrutinized the motifs relevant for four authors on whom our models obtained good results: Balzac, Daniel-Lesueur, Sand and Zola (looking deeper into the motifs attached to authors for whom we obtained poor results would not make much sense). For these authors, we inspected whether the selected motifs were interpretable by looking at examples from the corpus in context, to see if they corresponded to meaningful linguistic patterns.

First, it is clear that some forms are not interpretable: in Balzac, for example, there is a decrease in the use of adverbs over time, but the adverbial category is relatively heterogeneous so that it is difficult to interpret this phenomenon. The same can be observed with the motif "NC_,_avoir" (also in Balzac) which increases; this motif is realized in sequences (*patronne, avait* - mistress, had; *frère, ont* -brother, had; *ciel, as* - heaven, have) of which, at first sight, nothing really relevant can be said. Another group of difficult motifs in Balzac is the more frequent use of the pronoun "*on*" ("on PRES", ". on", "ADJ que on"). These patterns are hard to interpret, especially since the pronoun "*on*" has a fairly wide range of referential values.

Another example from Zola concerns verbs such as *bouleverser* (to upset) and *convaincre* (to convince), which have an increasing frequency over time. But here again, there is no immediate explanation for this usage. However, we were pleased to see that most motifs are interpretable (for example, we estimated that about three quarters of the motifs retained for Zola were). In the rest of this section, we will discuss some of the (groups of) motifs that we found interesting, or that have been previously noticed by researchers of the field of stylistic analysis.

6.2. Results

A number of interpretable motifs can be considered as stylemes. For example, in Zola, there is a set of motifs organized around ". Et" (the conjunction "and" at the beginning of the sentence): ". Et le NC être" ; ". Et, dès" ; ". Et ce être" whose use increases, as shown in example 1:

> (1) Quoi donc ? Était-ce la fin ? Un souffle glacé avait couru sur le camp, anéanti de sommeil et d'angoisse. **Et ce fut** alors que Jean et Maurice reconnurent le colonel de Vineuil [...]
> (Zola, *La débâcle*)
>
> What then? Was it the end? An icy breath had run over the camp, annihilated by sleep and anguish. **And it was** then that Jean and Maurice recognized Colonel de Vineuil [...]

This use (called the revival "et") was noticed very early by stylisticians (Thibaudet; but see Bordas and Badiou-Monferrand for modern accounts). It is not considered an idiosyncrasy, since this motif was also used by Flaubert, who can be said to have been imitated by Zola (Thibaudet; Gauthier). Flaubert considered that it was "an old biblical tic which is annoying".

Another set of motifs is, meanwhile, on the decline, including "NCCOR", a tag for parts of the human body, which are used in the physical description of the characters (*épaules; tête; main; yeux*, etc. - shoulders; head; hand; eyes, etc.). It is difficult to interpret the reason for this decrease; perhaps the form, or at least its repetition, was considered a cliché by the author.

In the same way, in Sand, we also notice that motifs linked to units referring to parts of the body tend to decrease, for example "NCCOR avec NCABS": *et se jeta dans mes* ***bras avec joie****; Suzanne baissa la* ***tête avec embarras***... and threw herself into my **arms with joy**; Suzanne lowered her **head in embarrassment**. This motif associates a movement of the body with a feeling. Again, this change could be considered to be due to the avoidance of a cliché, but this is a hypothesis that will have to be verified in further analysis.

Among the positive motifs of Sand, we note these two forms (", et, comme", ", et, si ce") which share the same rhythmic pattern (see examples 2 and 3). Without going into detail, these patterns may highlight the subordinate sentence by a kind of tension (↗), while the main phase is constructed in detension (↙).

> (2) Après avoir fait quelques tours sous les galeries, il se crut assez calme pour retourner à l'atelier, **et, comme** il redescendait l'escalier des Géants, il se trouva tout à coup face à face avec le Bozza.
> (Sand, *Les Maîtres mosaïstes*)
>
> After having taken a few turns under the galleries, he believed himself calm enough to return to the workshop, **and, as** he went back down the staircase of the Giants, he found himself suddenly face to face with the Bozza.

> (3) Dès lors, j'espérais qu'elle pourrait aimer Narcisse, **et, si cet** excellent jeune homme pouvait être heureux par elle, c'était à la condition de ne plus souffrir du passé. (Sand, *Narcisse*)
>
> From then on, I hoped that she would be able to love Narcisse, **and, if this** excellent young man could be happy thanks to her, it was on the condition of not suffering from the past anymore.

For Balzac, we found the decreasing motif "tout_à_NC" which corresponds in the vast majority of cases to the adverb "tout à coup" *all of a sudden*. Again, we suspect that the decrease of this motif could be caused by the avoidance of clichés. An interesting example of increasing motifs of Balzac is the motif "dire_à_NP" *"say to Proper Name"*. When inspecting the corpus, we noticed that this phrase is used in different ways: sometimes it is inserted inside a dialogue as illustrated in example (4), often it is used to mark the transition from narration to dialogue as in (5) and vice-versa, as in (6). We consider it a stylistic means to dynamize the switches between narratives and dialogues. Often this construction is used to put a long grammatical subject after the verb and direct object (as in 4 and 6), which also creates a stylistic effect.

> (4) Il ne faut pas demander à monsieur pourquoi il vient, **dit à Castanier** une vieille portière, vous ressemblez trop à ce pauvre cher défunt.
> (Balzac, *Melmoth reconcilié*)
>
> One shouldn't ask this gentleman why he came, **said** an old doorkeeper **to Castanier**, you look too much like the poor, dear deceased.
>
> (5) Le commandant, qui l' étudiait, s'apercevant de cette insensibilité, **dit à Gérard** : Le serin n'en sait pas long.
> (Balzac, *Les Chouans*)
>
> The commander, who was studying him, and noticed this insensitivity, **said to Gérard**: The fool does not know much.
>
> (6) J'attends la réponse, **dit à Rastignac** le commissionnaire de madame de Nucingen.
> (Balzac, *Le père Goriot*)
>
> I'm waiting for an answer, **said** the commissioner of Madame de Nucingen **to Rastignac**.

Finally, for Daniel-Lesueur, it is worth mentioning the increasing motif "..._DETPOSS_NC_..." (see examples 7 and 8), by which a noun preceded by a possessive determiner in between two ellipsis punctuation marks dramatizes reported thoughts and speech by invoking a close relation *mon enfant; ma soeur; mon amie* (*my child; my sister; my friend*).

> (7) Ah ! ma mère **... ma mère ...** pensait Hervé, [...]
> (Daniel-Lesueur, *Le Masque d'Amour II - Madame de Ferneuse*)
>
> Ah ! my mother **... my mother...** thought Hervé, [...]
>
> (8) Je suis perdue ! ... Perdue ! **... Ma chérie ...** Invente quelque chose ! ... Ah ! sauve-moi !
> (Daniel-Lesueur, *Justice de femme*)
>
> I'm lost! ... Lost! **... My darling...** Think of something! ... Ah! save me!

6.3. Discussion

As already mentioned, not all the motifs identified automatically are interpretable. Many, however, are stylistic in nature without it being possible to determine whether these uses are a deliberate choice by the author, or whether they are a form of automatism. To shed light on this question, a more precise analysis involving literary expertise should be undertaken. Our analysis provides the literary scholar, the stylistician and the linguist with statistically relevant evidence of the evolution of certain forms. It is up to these specialists to show correlations between forms, to propose interpretations. This type of approach can provide an empirical basis for more theoretical research (Philippe). Our hope is to have demonstrated that our method, which combines the use of motifs and the feature selection method of Lasso LARS, identifies a large number of stylistically interesting patterns and can be a useful tool in the qualitative analysis of the evolution of the idiolect.

7. General Discussion and Future Work

7.1. Contribution of the Work

In this article we investigated the chronological evolution of the idiolect. We examined whether support could be found for the rectilinearity hypothesis which states that the evolution of the idiolect is rectilinear in time, and whether the linguistic structures at the heart of idiolectal change could be identified. Using the Corpus for Idiolectal Research (CIDRE), we developed two methods that could help reach these goals. First, we introduced the idea of evaluating to what extent the distance matrices of works of one author are robinsonian. For ten out of eleven corpora in CIDRE, we found that the Robinsonian score was significantly high, suggesting that chronology plays a crucial role in the idiolect of an author. Second, we developed linear regression methods to predict the year of writing of a work and selected linguistic features that are key in the process of idiolectal change. We found that the majority of regression models were highly successful, again supporting the rectilinearity hypothesis. Third, these models allowed us to find a number of features (in the form of *motifs*) that lent themselves to manual examination in a qualitative study, demonstrating both the usefulness of these features and the validity of our methods. We believe that the use of motifs is complementary to the use

of lemmas and tokens. As for example Brunet illustrates in his study of the vocabulary used by Zola, using lemmas allows a researcher to interpret the topics of a writer. In the present study we demonstrate that motifs, on the other hand, might give more insights in stylistic forms.

We believe that working on the concepts of idiolect and chronological change can have an impact on related research themes. Modeling the idiolect could be useful, for example for the task of automatic text dating that was included in the 2015 SemEval campaign: *Task 7: Diachronic Text Evaluation* (Popescu and Strapparava). A corpus of snippets from newspaper articles dating from 1700 until 2010 was composed and the task consisted in dating these snippets. It could be interesting to see if the idiolect plays a role and if it can enhance the classification results.

A theme for which the concept of chronological variation could be interesting is authorship attribution and authorship verification, which involves checking whether a pair of documents are written by the same person (Kestemont et al.). Nowadays, the chronology of the writing is not taken into account; only the idiolect of each author in the corpus is modeled. It is quite possible, however, that taking the date of writing into consideration would enhance the modeling. Many different features have been explored to model the idiolect of authors for this task: n-grams of words or characters (e.g. Stamatatos; Antonia et al.; Sari et al.), syntactic structures (Sundararajan and Woodard; Zhang et al.) and even discourse structure (Ding et al.) but we are not aware of models that take idiolectal variance over time into account. However, especially for writers with long careers, it could be meaningful.

In this study, we focused on methods and on the evaluation of results. We argue that the use of standard corpora, baselines and evaluation metrics could help enhance the comparability of studies in the field of stylometry and that this would help the research community gain greater insight into the robustness of the results. In our experiment on the Robinsonian matrices, we used random results as a baseline. For research questions that have not yet been addressed in the literature, this is a useful starting point, as shown in the work of Bulteau et al., who developed two algorithms to estimate the probability that a tree produced by a hierarchical clustering algorithm — for instance produced by stylo R (Eder et al.) — reflects a chronological order by chance. In our experiment using regression models, we compared our methods with those of Klaussner and Vogel from their 2018 publication, using their baseline RMSE and the standard baseline of regression models, R2 (Field).

An important contribution of this study is that it addresses questions of evaluation. We have seen that the development of off-the-shelf-packages has made it possible to shed new light on long-standing research questions. For example Schmidt-Petri et al. used the rolling-classification algorithm from stylo R (Eder et al.) to examine the contribution of Harriet Taylor Mill to the essay *On Liberty*, which is officially contributed solely to John Stuart Mill,

her husband. They found that there is stylometric evidence that she should indeed be considered a co-author of the work. However, as stylo R does not enable any statistical evaluation of the classification results, the authors had no straightforward means of interpreting their reliability and had to undertake considerable extra work to estimate the robustness of the results. We therefore think that working on the question of evaluation of stylometric methods is a topic in the field of stylometry that needs to be developed further and we hope to have made a useful contribution to it.

7.2. Future Work

The most obvious future work that should result from this study is a detailed qualitative analysis of the selected motifs in CIDRE, for which the regression models obtain good results. These studies should also contain a detailed comparison with the reference corpus in order to decide if the observed change can be interpreted as a rather general diachronic language change or an idiosyncrasy, using for example the methods that Mollin used to identify idiosyncratic collocations. This should be done, however, in collaboration with literary experts of the writers in question in order to compare the findings of the method with what is already known in the field of stylometry and stylistics. In addition to the identification of idiosyncratic motifs, collaborations with literary experts would allow us to get a more precise interpretation of the motifs of an author. Indeed, we could for example examine the role that dialogs, narratives, and descriptions play when experts provide us with theoretically and empirically motivated hypotheses on specific authors.

Another straightforward direction for future work is to repeat our experiments on other text genres, for example drama or correspondence. We are considering trying our methods on plays, for example by using the *Théâtre Classique* corpus of Fièvre. However, as theatrical works might be influenced/written by the actors in the plays, the idiolectal signal of the playwright may not be as strong as for works of fiction. Correspondence could be interesting in order to investigate idiolectal changes with respect to the addressee of the letter. We could for example use the Corpus of Early English Correspondence (Raumolin-Brunberg and Nevalainen) and the correspondence of George Sand. Another advantage of using correspondence is that dating letters might be more precise than dating works of fiction. However, it would probably result in corpora that are significantly smaller than the corpora of the authors in CIDRE. As we suspect a strong relation between corpus size and the statistical power of the experiments, the success is not guaranteed for smaller corpora. But on the other hand, the number of letters per author is probably higher than the number of books in CIDRE, which could enhance the statistical power.

A third direction for future work is to evaluate how different people influence each other with their idiolects. Evans investigated how the idiolect of Queen Elizabeth I was influenced by others. It could be interesting to develop a methodology on how influence could be established between authors or even between literary movements.

8. Conclusion

Our experiments demonstrate that there is a significant evolution of the idiolect during an author's lifetime. Our experiments also suggest that some features evolve in a rectilinear manner, steadily increasing or decreasing as the years go by. These features are sufficiently clear-cut to be used to date the year of writing of a book very accurately. We therefore conclude that we found strong support for the rectilinearity hypothesis and that the evolution of the idiolect is a relevant type of intrapersonal variation that exists alongside the strong signal of interpersonal variation. We thus dismiss the proposal that idiolects are stable over time, even though it is true that not all linguistic features evolve. A second contribution of our article is the development of new methods for which we have demonstrated the usefulness in 1) assessing the chronological signal of the idiolect in corpora and 2) identifying linguistic structures that are at the heart of this evolution. These features can in turn be used for qualitative studies with stylistic objectives.

Peer-Reviewers: Simon DeDeo, David Mimno

All scripts, supplementary materials and data used for our experiments are available in the online Harvard Dataverse directory: https://doi.org/10.7910/DVN/WCMZOK, except for the CIDRE corpus, that is freely available in the following Zenodo repository: https://doi.org/10.5281/zenodo.4707812.

Acknowledgments

We thank the two reviewers for their insightful comments and suggestions. This work has been developed in the framework of the IRN (International Research Network) Cyclades (Corpora and Computational Linguistics for Digital Humanities). 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) and reference ANR-16-IDEX-0003 (I-Site Future, programme "Cité des dames, créatrices dans la cité").

Submitted: February 02, 2022 EDT, Accepted: April 26, 2022 EDT

REFERENCES

Anthonissen, Lynn, and Peter Petré. "Grammaticalization and the Linguistic Individual: New Avenues in Lifespan Research." *Linguistics Vanguard*, vol. 5, no. s2, June 2019, doi:10.1515/lingvan-2018-0037.

Antonia, A., et al. "Language Chunking, Data Sparseness, and the Value of a Long Marker List: Explorations with Word n-Grams and Authorial Attribution." *Literary and Linguistic Computing*, vol. 29, no. 2, May 2013, pp. 147–63, doi:10.1093/llc/fqt028.

Badiou-Monferrand, Claire. "Rémanence Des Et de Relance En Français Moderne et Contemporain: Du 'Résidu' Au 'Reliquat.'" *Le Français Moderne*, vol. 88, no. 2, 2020, pp. 295–312.

Barlow, Michael. "Individual Usage: A Corpus-Based Study of Idiolects." *Proceedings of LAUD Conference*, 2010.

Bloch, Bernard. "A Set of Postulates for Phonemic Analysis." *Language*, vol. 24, no. 1, Jan. 1948, pp. 3–46, doi:10.2307/410284.

Bordas, Éric. "Et La Conjonction Resta Tensive. Sur Le et de Relance Rythmique." *Français Moderne*, vol. 73, no. 1, 2005, pp. 23–39.

Brainerd, Barron. "The Chronology of Shakespeare's Plays: A Statistical Study." *Computers and the Humanities*, vol. 14, no. 4, Dec. 1980, pp. 221–30. *Crossref*, doi:10.1007/bf02404431.

Brooke, Julian, et al. "GutenTag: An NLP-Driven Tool for Digital Humanities Research in the Project Gutenberg Corpus." *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, 2015, pp. 42–47, doi:10.3115/v1/w15-0705.

Brunet, Etienne. *Le Vocabulaire de Zola*. Slatkine, Champion, 1985.

Bulteau, Laurent, et al. "Reordering a Tree According to an Order on Its Leaves." *33rd Annual Symposium on Combinatorial Pattern Matching (CPM 2022)*, Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022. *Google Scholar*, doi:10.4230/LIPIcs.CPM.2022.24.

Can, Fazli, and Jon M. Patton. "Change of Writing Style with Time." *Computers and the Humanities*, vol. 38, no. 1, Feb. 2004, pp. 61–82, doi:10.1023/b:chum.0000009225.28847.77.

Church, Kenneth, et al. "Using Statistics in Lexical Analysis." *Lexical Acquisition: Exploiting on-Line Resources to Build a Lexicon*, Psychology Press, 1991, pp. 115–64.

Cox, D. R., and Leonard Brandwood. "On a Discriminatory Problem Connected with the Works of Plato." *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 21, no. 1, Jan. 1959, pp. 195–200. *Crossref*, doi:10.1111/j.2517-6161.1959.tb00329.x.

Craig, Hugh. "Stylistic Analysis and Authorship Studies." *A Companion to Digital Humanities*, vol. 3, 2004, pp. 233–334.

Cropp, Martin, and Gordon Fick. "Resolutions and Chronology in Euripides: The Fragmentary Tragedies." *Bulletin Supplement (University of London. Institute of Classical Studies)*, 1985, pp. iii–92.

Daelemans, Walter. "Explanation in Computational Stylometry." *Computational Linguistics and Intelligent Text Processing*, edited by Alexander Gelbukh, vol. 7817, Springer Berlin Heidelberg, 2013, pp. 451–62. *Crossref*, doi:10.1007/978-3-642-37256-8_37.

Davies, Mark, et al. "The 400 Million Word Corpus of Historical American English (1810–2009)." *English Historical Linguistics 2010: Selected Papers from the Sixteenth International Conference on English Historical Linguistics (ICEHL 16), Pécs, 23-27 August 2010*, vol. 325, John Benjamins Publishing, 2012, pp. 231–62, doi:10.1075/cilt.325.11dav.

Derks, Peter L. "Clockwork Shakespeare: The Bard Meets the Regressive Imagery Dictionary." *Empirical Studies of the Arts*, vol. 12, no. 2, July 1994, pp. 131–39. *Crossref*, doi:10.2190/h489-jh64-lq8c-l4t1.

Devine, A. M., and Laurence D. Stephens. "A New Aspect of the Evolution of the Trimeter in Euripides." *Transactions of the American Philological Association (1974-)*, vol. 111, 1981, p. 43. *Crossref*, doi:10.2307/284118.

Ding, Steven H. H., et al. "Learning Stylometric Representations for Authorship Analysis." *IEEE Transactions on Cybernetics*, vol. 49, no. 1, Jan. 2019, pp. 107–21. *Crossref*, doi:10.1109/tcyb.2017.2766189.

Dittmar, Norbert. "Explorations in 'Idiolects.'" *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, 1996, pp. 109–28.

Dunning, Ted E. "Accurate Methods for the Statistics of Surprise and Coincidence." *Computational Linguistics*, vol. 19, no. 1, 1993, pp. 61–74.

Eder, Maciej, et al. "Stylometry with R: A Package for Computational Text Analysis." *The R Journal*, vol. 8, no. 1, 2016, doi:10.32614/rj-2016-007.

Efron, Bradley, et al. "Least Angle Regression." *The Annals of Statistics*, vol. 32, no. 2, Apr. 2004, pp. 407–99, doi:10.1214/009053604000000067.

Evans, Mel. *Aspects of the Idiolect of Queen Elizabeth I: A Diachronic Study on Sociolinguistic Principles*. University of Sheffield, 2011.

Field, Andy. *Discovering Statistics Using SPSS: Book plus Code for E Version of Text*. SAGE Publications Limited, 2009.

Fièvre, Paul. "Théâtre Classique." *Université Paris-IV Sorbonne Http://Www. Theatreclassique. Fr*, 2007.

Fleury, Serge, and Maria Zimina. "Trameur: A Framework for Annotated Text Corpora Exploration." *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, 2014, pp. 57–61.

Forsyth, R. "Stylochronometry with Substrings, or: A Poet Young and Old." *Literary and Linguistic Computing*, vol. 14, no. 4, Dec. 1999, pp. 467–78. *Crossref*, doi:10.1093/llc/14.4.467.

Frischer, Bernard. *Shifting Paradigms New Approaches to Horace's Ars Poetica*. 1991.

Gauthier, E. Paul. "Zola as Imitator of Flaubert's Style." *Modern Language Notes*, vol. 75, no. 5, May 1960, p. 423, doi:10.2307/3039860.

Guaresi, Magali, et al. "Entre Rupture et Continuité, Le Discours Du PCF (1920-2020)." *Histoire & Mesure*, vol. XXXVII–1, no. 2, Dec. 2021, pp. 125–62, doi:10.4000/histoiremesure.14904.

Heck, Richard. "Idiolects." *Content and Modality: Themes from the Philosophy of Robert Stalnaker*, Oxford University Press on Demand, 2006, pp. 61–92.

Heiden, S., et al. *Manuel de TXM, Version 0.7.9*. ENS de Lyon & Université de Franche-Comté, 2018, http://textometrie.ens-lyon.fr/files/documentation/Manuel%20de%20TXM%200.7%20FR.pdf.

Jackson, MacD. P. "Pause Patterns in Shakespeare's Verse: Canon and Chronology." *Literary and Linguistic Computing*, vol. 17, no. 1, Apr. 2002, pp. 37–46. *Crossref*, doi:10.1093/llc/17.1.37.

Jaynes, Joseph T. "A Search for Trends in the Poetic Style of WB Yeats." *ALLC Journal*, vol. 1, 1980, pp. 11–18.

Kestemont, Mike, et al. "Overview of the Cross-Domain Authorship Attribution Task at PAN 2019." *CLEF (Working Notes)*, 2019.

Klaussner, Carmen. "Elements of Style Change." *University of Dublin, Ireland*, 2017.

Klaussner, Carmen, and Carl Vogel. "Stylochronometry: Timeline Prediction in Stylometric Analysis." *Research and Development in Intelligent Systems XXXII*, edited by Max Bremer and Miltos Petridis, Springer International Publishing, 2015, pp. 91–106. *Crossref*, doi:10.1007/978-3-319-25032-8_6.

---. "Temporal Predictive Regression Models for Linguistic Style Analysis." *Journal of Language Modelling*, vol. 6, no. 1, Aug. 2018, doi:10.15398/jlm.v6i1.177.

Koppel, Moshe, and Jonathan Schler. "Exploiting Stylistic Idiosyncrasies for Authorship Attribution." *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, vol. 69, 2003, pp. 72–80.

Lamalle, C., et al. *Lexico 3 Version 3.41 Février 03. Outils de Statistique Textuelle. Manuel d'Utilisation.* Laboratoire SYLED-CLA2T, Université de la Sorbonne nouvelle - Paris 3, 2003, http://www.lexi-co.com/ressources/manuel-3.41.pdf.

Ledger, Gerard R. *Re-Counting Plato a Computer Analysis of Plato's Style*. 1989.

Legallois, Dominique, et al. "The Balance Between Quantitative and Qualitative Literary Stylistics: How the Method of 'Motifs' Can Help." *The Grammar of Genres and Styles: From Discrete to Non-Discrete Units*, 2018, pp. 164–93.

Meyerhoff, Miriam, and James A. Walker. "The Persistence of Variation in Individual Grammars: Copula Absence in ?Urban Sojourners? And Their Stay-at-Home Peers, Bequia (St Vincent and the Grenadines)." *Journal of Sociolinguistics*, vol. 11, no. 3, June 2007, pp. 346–66. *Crossref*, doi:10.1111/j.1467-9841.2007.00327.x.

Mollin, Sandra. "'I Entirely Understand' Is a Blairism: The Methodology of Identifying Idiolectal Collocations." *International Journal of Corpus Linguistics*, vol. 14, no. 3, Aug. 2009, pp. 367–92, doi:10.1075/ijcl.14.3.04mol.

Opas, L. L. "A Multi-Dimensional Analysis of Style in Samuel Beckett's Prose Works." *Research in Humanities Computing 4.*, edited by S. Hocking and N. Ide, Clarendon Press., 1996.

Petré, Peter, et al. "Early Modern Multiloquent Authors (EMMA): Designing a Large-Scale Corpus of Individuals' Languages." *ICAME Journal*, vol. 43, no. 1, Mar. 2019, pp. 83–122, doi:10.2478/icame-2019-0004.

Petré, Peter, and Freek Van de Velde. "The Real-Time Dynamics of the Individual and the Community in Grammaticalization." *Language*, vol. 94, no. 4, 2018, pp. 867–901, doi:10.1353/lan.2018.0056.

Philippe, Gilles. *Pourquoi le style change-t-il?* Les Impressions Nouvelles, 2021, doi:10.14375/np.9782874498671.

Pincemin, Bénédicte. *Sept Logiciels de Textométrie*. 2018.

Popescu, Octavian, and Carlo Strapparava. "Semeval 2015, Task 7: Diachronic Text Evaluation." *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015, pp. 870–78, doi:10.18653/v1/s15-2147.

Raumolin-Brunberg, Helena, and Terttu Nevalainen. "Historical Sociolinguistics: The Corpus of Early English Correspondence." *Creating and Digitizing Language Corpora*, edited by Joan C. Beal, et al., Palgrave Macmillan UK, 2007, pp. 148–71, doi:10.1057/9780230223202_7.

Robinson, T. M. "Plato and the Computer." *Ancient Philosophy*, vol. 12, no. 2, 1992, pp. 375–82, doi:10.5840/ancientphil19921228.

Robinson, W. S. "A Method for Chronologically Ordering Archaeological Deposits." *American Antiquity*, vol. 16, no. 4, Apr. 1951, pp. 293–301, doi:10.2307/276978.

Sari, Yunita, et al. "Continuous N-Gram Representations for Authorship Attribution." *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 267–73, doi:10.18653/v1/e17-2043.

Schmidt-Petri, Christoph, et al. "Who Authored *On Liberty*? Stylometric Evidence on Harriet Taylor Mill's Contribution." *Utilitas*, vol. 34, no. 2, Dec. 2021, pp. 120–38, doi:10.1017/s0953820821000339.

Seminck, Olga, et al. "The Corpus for Idiolectal Research (CIDRE)." *Journal of Open Humanities Data*, vol. 7, 2021, p. 15, doi:10.5334/johd.42.

Smith, Joseph A., and Coleen Kelly. "Stylistic Constancy and Change across Literary Corpora: Using Measures of Lexical Richness to Date Works." *Computers and the Humanities*, vol. 36, no. 4, 2002, pp. 411–30. *Crossref*, doi:10.1023/a:1020201615753.

Stamatatos, Efstathios. "On the Robustness of Authorship Attribution Based on Character N-Gram Features." *JL & Pol'y*, vol. 21, 2012, p. 421.

Stamou, C. "Stylochronometry: Stylistic Development, Sequence of Composition, and Relative Dating." *Literary and Linguistic Computing*, vol. 23, no. 2, Oct. 2007, pp. 181–99. *Crossref*, doi:10.1093/llc/fqm029.

Sundararajan, Kalaivani, and Damon Woodard. "What Represents 'Style' in Authorship Attribution?" *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 2814–22.

Temple, J. T. "A Multivariate Synthesis of Published Platonic Stylometric Data." *Literary and Linguistic Computing*, vol. 11, no. 2, June 1996, pp. 67–75. *Crossref*, doi:10.1093/llc/11.2.67.

Thibaudet, Albert. *Gustave Flaubert*. Éditions Gallimard, 1922.

Vanni, Laurent, et al. "Hyperdeep: Deep Learning Descriptif Pour l'analyse de Données Textuelles." *JADT 2020*, 2020.

Whissell, Cynthia. "Traditional and Emotional Stylometric Analysis of the Songs of Beatles Paul McCartney and John Lennon." *Computers and the Humanities*, vol. 30, no. 3, 1996, pp. 257–65, doi:10.1007/bf00055109.

Wishart, David, and Stephen V. Leach. "A Multivariate Analysis of Platonic Prose Rhythm." *Computer Studies in the Humanities and Verbal Behavior*, vol. 3, no. 2, 1970, pp. 90–99.

XML, BNC. *The British National Corpus XML Edition DVD*. Oxford: Oxford University Press, 2007, http://www.natcorp.ox.ac.uk/docs/URG/.

Zhang, Richong, et al. "Syntax Encoding with Application in Authorship Attribution." *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2742–53, doi:10.18653/v1/d18-1294.