Beyond Idiolectometry? On Racine's Stylometric Signature

Simon Gabay

Beyond Idiolectometry? On Racine's Stylometric Signature

Simon Gabay

University of Geneva Rue des Battoirs 7, CH-1205 Genève – Suisse/ Switzerland

Abstract

If stylometry has proven to be useful for literary history, especially for distant reading approaches of texts, it still has to show its efficiency regarding close reading. Taking the example of famous French playwright Jean Racine, we propose a double analysis of his plays, both distant and close, following the double objective of controlling its newly alleged paternity on Campistron's plays (which proves to be wrong using standard methods in stylometry), and interpreting the stylometric markers used for this attribution procedure. 17th c. French having a relatively unstable spelling system, we also propose a new method for denoising, based on full linguistic annotation rather than simple lemmatisation.

Keywords

stylometry, serial stylistics, Jean Racine, Authorship attribution, French classical theatre

1. Introduction

Stylometry relies on the assumption that each person not only has a genome, but also a "stylome", *i.e.* linguistic idiosyncrasies [26] such as specific words, called *markers* by Mosteller and Wallace in their seminal study on the *Federalist papers* [40]. The two American scholars have indeed demonstrated that it is possible to distinguish texts written by Hamilton, who uses *while*, from those written by Madison, who prefers *whilst* (among many other features). The use of such a technique is however not limited to authorship attribution or document classification, and literary scholars have used it to investigate intertextuality [12] or periodisation [44].

Great efforts have been deployed to explain the inner functioning of stylometry and its effectiveness. Some have emphasised the importance of function words (prepositions, articles, conjunctions…), which have for instance the advantage to be frequent and uncorrelated to the topic of the book [29]. Others have developed extensively on the computational part (distance type, number of words…) to assess possible configurations and determine the ideal experiment conditions [18]. It seems however that the call of Walter Daelemans "to increase understanding rather than maximizing performance" [13] has not yet been fully heard, especially by literary scholars, and we are not aware of any analysis going really beyond quantitative observations. A void remains between the computational detection of features or patterns on the one hand, and the traditional stylistic analysis of texts on the other hand.

Among the numerous ways that exist to fill this gap, approximating similar concepts such as "stylometric stylome" and "stylistic signature" [55] could be productive, since they both

CHR 2021: Computational Humanities Research Conference, November 17–19, 2021, Amsterdam, The Netherlands

simon.gabay@unige.ch (S. Gabay)
https://github.com/gabays (S. Gabay)
0000-0004-1957-6448 (S. Gabay)

try, with different means, to characterise the writing of an author. In other words, do computationally detected markers have a literary value? It is indeed not clear if a stylome is only composed of idiolectal traits, or, to an extent that needs to be determined, of sylistic features with an interpretative yield.

2. Racine's case

Jean Racine (La Ferté-Milon, 1639 – Paris, 1699) is one of the most prominent French writers, so important that he has been considered a "zero point of the critical object" (*degré zéro de l'objet critique*) by critics such as Barthes [3]. He is the author of twelve plays, eleven of which are tragedies which are considered the quintessence of the genre, and a comedy: *Les Plaideurs*. This production is relatively small compared to the production of his famous contemporaries Molière and Pierre Corneille, who have both written more than thirty plays.

2.1. Problem A

New theories have recently emerged regarding the work of Racine: Dominique Labbé has postulated that he should be attributed fourteen other tragedies signed by another playwright, Jean Galbert de Campistron (Toulouse, 1656 - *ibid.*, 1723) [32, 4]. Such a claim has to be put in the broader context of D. Labbé's research on classical French theatre and the *théorie des prête-noms* ("figurehead theory"), according to which more than half of the plays (90% of the comedies) published and played in 17$^{\text{th}}$ c. France were signed by intermediaries rather than by real authors [31] – the most famous of these figureheads being Molière [33].

D. Labbé's theories have to be taken with a lot of care, since they have already been severely discarded by both solid traditional [20] and computational [10] cross-checking, but we still think that such ideas deserve a scientific answer – at the possible cost of a Streisand effect – for two reasons. First, editions of those tragedies supposedly written by Racine have already been published under his name [5], which might create confusions among readers. Second, in an age of credulity [7], it is important not to let slip without a meticulous verification hypotheses that have already been considered as conspiracy theories by some scholars [21].

2.2. Problem B

Investigating the attribution of new plays to Racine will lead to the identification of stylometric markers, which should differentiate him from other writers. It is therefore a perfect opportunity to explore the possible meaning(s) of these markers, and assess their literary value, following an approach inspired by serial stylistics. This evaluation will obviously benefit from previous works, such as Leo Spitzer's article on Racine's style [51] [1] and his idea of a *klassische Dämpfung* ("muting effect"), defined as followed:

> das oft Nüchtern-Gedämpfte, Verstandesmäßigkühle, fast Formelhafte an diesem Stil, das dann oft plötzlich und unvermutet für Augenblicke in poetisches Singen und erlebte Form übergeht, worauf aber wieder rasch ein Löschhütchen von Verstandeskühle das sich schüchtern hervorwagende lyrische Sich-Ausschwelgen des Lesers niederdämpft. [52]

[1] The article has been fully translated into French [52] and partially into English [53].

> *the frequently sober, muted quality of this style, rational, cool and formulistic, which then often, suddenly and unexpectedly, makes a transition for some moments into poetic song and form realised in experience, after which, however, an extinguisher of rational coolness quenches the shy beginnings of the reader's lyrical expansiveness.* [53]

Racine's style would be characterised by intensity variations, and especially attenuation, the trace of which can be found, according to Spitzer, in a long list of examples, such as *die Entindividualisierung durch den ubestimmten Artikel* ("the de-individualisation by means of the indefinite article"):

> Je révoque *des* lois dont j'ai plaint la rigueur. (*Phèdre* II.2)
> (*I revoke laws whose rigour I have blamed.*)

or *der distanzierend Gebrauch des Demonstrativ* ("the distancing use of the demonstrative"):

> Mais j'ai vu près de vous *ce* superbe Hippolyte. (*Phèdre* II.1)
> (*But I have seen next to you this superb Hippolytus.*)

Such stylistic stylemes (*i.e.* textual units characterising the discourse as literary in traditional stylistics[2]) are particularly interesting because they are based on stop words (articles, prepositions, adverbs, etc.), and therefore possible stylometric markers. By analysing the overlap between stylemes and markers, we should be able to evaluate the stylistic nature of the stylometric analysis.

3. Data

For this experiment, we follow a approach focused on data extracted directly from the sources, without the mediation of editions, presenting significant engineering challenges. Stylometric analysis requires an important amount of data that is hard to gather, especially for historical documents such as 17$^{\text{th}}$ c. French texts. If until now most of the research has been carried on already existing corpora [49], we have to prepare the ground for further data acquisition directly from sources that are sometimes hardly readable (*e.g.* scans of old prints or manuscripts, cf. figure 1) and require additional processing to neutralise inconsistencies (*e.g.* spelling variation or ancient glyphs such as ‹ſ›).

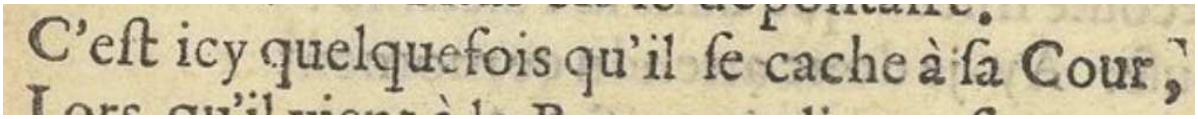

Figure 1: Racine, *Bérénice*, I.1. (https://gallica.bnf.fr/ark:/12148/bpt6k990581p/f21.item)

To cope with these problems, we follow a dedicated pipeline partially inspired by those designed for other states of language to process and clean data [11]. This latter idea is particularly crucial in our case, because not only do we want to interpret the classification itself, but also

[2]We paraphrase here Georges Molinié, who uses the jakobsonian concept of *literaturnost* ("literariness") to define the styleme as *un caractérisème de littérarité* ("a characteriseme of literariness") [38].

the stylometric markers used to produce the different clusters. Is it the personal pronoun *je* ("I")? The negative adverb *pas* ("not")? Or the verb *étoit* ("was")? The answer is particularly complicated for the 17$^{\text{th}}$ c. because:

1. Letters can be elided in some cases: *je* → *j'*.
2. Some words are homographs: negative adverb *pas* ("not") vs noun *pas* ("step").
3. Spelling varies from one occurrence to the other: *étoit* ("was") = *eſtoit*.

It is more than likely that minor variations such as these have a limited impact on the classification [16] but affect the stylometric signature: *je* and *j'* are counted as two different words, and both *pas* as one.

Table 1
Linguistic annotation

HTR	C'	eſt	icy	quelquefois	qu'	il	ſe
Normalisation	C'	est	ici	quelquefois	qu'	il	se
Lemmatisation	ce	être	ici	quelquefois	que	il	se
POS	PROdem	VERcjg	ADVgen	ADVgen	CONsub	PROper	PROper
Morphology	NOMB.=s GENRE=m	MODE=ind TEMPS=pst PERS.=3 NOMB.=s	-	-	-	PERS.=3 NOMB.=s GENRE=m CAS=n	PERS.=3 NOMB.=x GENRE=x CAS=x

The tools used for our pipeline have already been described in detail [22]. The text is extracted from old prints (cf. figure 1) with a text recognition engine, then both linguistically normalised (*i.e.* aligned with contemporary French modulo some exceptions, mainly for metric reasons) and annotated with lemma, POS and full morphology (cf. table 1). Normalisation provides a simple denoising and deals mainly with simple phenomena such as unstable spellings (*étoit* vs *eſtoit* → `était`). Linguistic annotation offers a more efficient reduction of noise: it differentiates homographs such as the noun *pas* (annotated `pas NOMcom sing. masc.`) and the adverb *pas* (`pas ADVgen`), or reconciles the elided *j'* (`je PROper P1`) with its full form *je* (also `je PROper P1`).

The corpus (cf. table 2) has been deliberately designed as heterogeneous to allow a precise exploration of stylometric markers used for the classification of our texts. If they all are plays dating from the last third of the 17$^{\text{th}}$ c., they belong to two different major genres (tragedies, comedies) and an additional minor one (heroic comedy). They are written indiscriminately in prose or in verse. Prints are produced by different *marchands-libraires* (*i.e.* publishers) and printers, from Paris and abroad (Bruxelles), to maximise the variation of spelling choices. We have prepared two plays for each playwright (Pradon and Campistron), three when there are two genres for one writer (Molière and Racine).

Texts have been corrected before being normalised and annotated. Because they have all been encoded in XML-TEI (cf. figure 2) it has been possible to keep only replies and to remove stage directions, notes, numbering of scenes and acts, etc. because we aim to study the text and not the paratext [54]. The name of places and characters, which could introduce biases[3], have also been removed.

Such a corpus being too small to provide robust results and the process to create additional data being extremely time consuming[4], we have decided to fall back on modernised versions of

[3]Two plays about the same event would be artificially overcorrelated because of similar rare words.

[4]The very poor quality of many prints forces editors to correct the entire transcription produced by the OCR engine.

Table 2
Breakdown of the primary corpus

Author	Title	Place	Publisher	Printer	Date	Form	Genre
Campistron	*Achille*	Paris	Ac. royale de musique	Ch. Ballard	1687	verse	Tragedy
Campistron	*Arminius*	Paris	Th. Guillain	Ch. Journel	1690	verse	Tragedy
Molière	*Dom Garcie de Navarre*	Bruxelles	G. De Baker	G. De Baker	1694	verse	Heroic comedy
Molière	*L'École des femmes*	Paris	L. Billaine	J. Hénault Cl. Blageart	1663	prose	Comedy
Molière	*George Dandin*	Paris	J. Ribou	Cl. Audinet	1669	prose	Comedy
Pradon	*Scipion*	Paris	J. Ribou	-	1700	verse	Tragedy
Pradon	*Statira*	Paris	J. Ribou	Cl. Blageart	1680	verse	Tragedy
Racine	*Les Plaideurs*	Paris	Cl. Barbin	Cl. Blageart	1669	verse	Comedy
Racine	*Bérénice*	Paris	J. Ribou	J.-B. I. Coignard	1676	verse	Tragedy
Racine	*Andromaque*	Paris	J. Ribou	J.-B. I. Coignard	1676	verse	Tragedy

```
<sp>
    <stage>ACHILE.</stage>
    <l n="97">
        <choice>
            <orig>Je vois avec plaifir les pertes de la <placeName>Grece</placeName>,</orig>
            <reg>Je vois avec plaisir les pertes de la <placeName>Grèce</placeName>,</reg>
        </choice>
    </l>
</sp>
```

Figure 2: TEI encoding with, in parallel, the original and the normalised (*i.e.* aligned with contemporary French) version

plays available online [19] to increase the amount of texts studied (cf. tab. 6, in the appendix). However, merging the primary and this secondary corpus remains possible at two different levels: using the normalised version of the original texts automatically produced, but also *via* the linguistic annotation, the model providing it being trained on both original and normalised transcriptions [23].

A control corpus, with a symmetrical composition to the primary corpus, but composed by 18^{th} c. French plays, has been prepared for benchmarking purposes (cf. table 5). Reproducibility of our experiments with similar results on another corpus has been thought to be an additional safety net, on top of the careful use of previous methodological studies on stylometric evaluation [16, 18] and similar experiments [10, 48].

4. Problem A: Authorship attribution

4.1. Set up

We have drawn the ascendant hierarchical clustering (henceforth AHC), using mainly two R packages, *FactoMineR* [34] and *Stylo* [17], with the following parameters:

- Distance is calculated with Burrows's delta (*i.e.* computing a manhattan distance between two z-scored vectors) [8] combined with vector-length Euclidean normalisation,

following here the conclusions of previous stylometric studies on French literature [10, 48].

- Linkage criterion follows Ward's minimum variance method (*i.e.* the pair of clusters to merge at each step is based on the optimal value of an objective function). [56]

In order to evaluate the results, two evaluation measures have been used:

- The agglomerative coefficient (henceforth AC) measures the strength of the clustering structure by calculating the mean similarity of each object with the first cluster it is merged with, normalised on the total height of the plot (*i.e.*, the height of the merger in the last step of the classification algorithm [45]). Let H be the vector of the heights at which each node i is merged with its first cluster:

$$1 - \frac{1}{n} \frac{\sum_{i=1}^{n} H_i}{max(H)}$$

  it is expressed by a number between 0 and 1, the closer to one being the better.

- Cluster purity (henceforth CP) is the average percentage of the dominant class label (the putative author) in each cluster [1]. A result below 1 (=100%) indicates that some objects (texts) were not classified correctly regarding the class label. Because our corpus is made of texts written by four authors and belong to two different genres, we expect six clusters, none of which containing two different authors or two different genres.

Four different methods to select the most relevant features have been experimented:

- Using the 100 most frequent words (henceforth MFW). We are well aware that such a number is well below the recommended average [15], but it allows us to minimise the importance of thematic/generic words and does not affect the the clustering (100, 1,000 and 3,000 MFW have been tested with equivalent results). A bootstrap consensus tree [14] (cf. figure 3) confirms this stability, no matter the number of tokens used (between 100 and 5000).

- Using function words, *i.e.* a selection of tokens excluding nouns, verbs (except auxiliary verbs), adjectives, and including preposition, articles, determiners, prepositions. Pronouns have not been kept because previous studies on the *indice pronominal* ("pronominal index") in French have shown that they are a generic rather than a stylistic feature [41, 50].

- Using (pseudo-)affixes (suffixes and prefixes), *i.e.* the first / last three characters of each word, and the first / last two characters of each word and the space preceding / following [46].

- For stop words and affixes, we have additionally applied Moisl's selection method [36] adapted for stylometry by Camps and Cafiero [10] with a 1.645 critical value (*i.e.* 95% confidence interval).

No matter what the scenario is, the main split is made according to the genre (with comedies in the upper part of the dendogram and tragedies in the lower part) and all the possible pairs are correctly classified (except for affixes without Moisl's selection). These results being

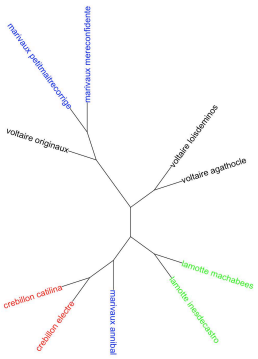

Figure 3: Bootstrap consensus tree, 100-5000 MFW

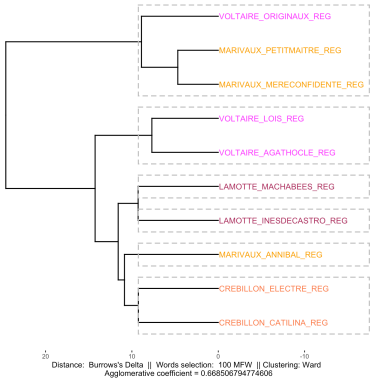

Figure 4: False clustering (cf. Lamotte's plays)

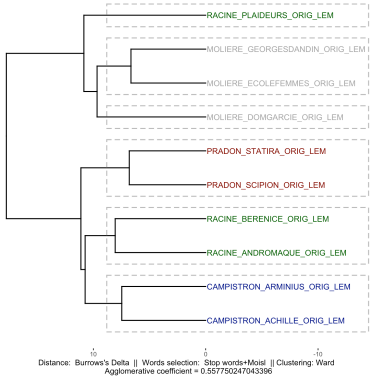

Figure 5: AHC with the best configuration

relatively clear, it has not been thought relevant to pursue with other tests *via* other features (*e.g.* character N-grams) or other methods (*e.g.* support vector machine), whose results are expected to be the same. Regarding CP, we observe minor misclassifications (cf. figure 4): only stop words (with and without Moisl's selection) and affixes (with Moisl's selection) offer 100% purity (cf. figure 5).

4.2. Stylometric Results

The same experiments have been repeated on three different versions of our corpus:

- The original version, with maximal spelling variation.
- A normalised version, the spelling of which has been aligned with contemporary French.
- An annotated version, with lemma, POS and full morphology for each token (for which we do not offer a clustering based on affixes for obvious reasons).

A detailed breakdown of the results (cf. table 3) shows that a perfect CP is achieved in many different ways, no matter the version of the corpus (cf. figure 6), and that normalisation has an ambivalent (but marginal) impact. Moisl's selection always improves the CP (if it is not at its maximum) no matter what it is combined with (stop words or affixes). MFW offer a slightly lower CP.

These results show that we are able to disentangle the authorial and the generic signal from one another, despite an unstable spelling, with maximal denoising of data *via* a complete linguistic annotation. Because it provides a unique ID for each type, impermeable to spelling variation, flexion or elision, this strategy offers, for pre-orthographic states of language, an excellent alternative to character N-grams [11]. However, with the use of a tagger, linguistic annotation introduces an additional step in the workflow, which inevitably increases noise in the data, especially when performed on unclean transcriptions. Full annotation should however be preferred to a simple lemmatisation [30], which is not precise enough and too dependent on annotation choices behind the lemmatisation model (*e.g.* nominalisation, etc.).

Regarding authorship attribution, despite variations in the results, no scenario suggests that Racine's and Campistron's plays would be written by the same playwright. When extending the size of the corpus by merging the primary and the secondary corpus, the AHC given with the best configuration produces the same classification, confirming our first results (cf. figure 7).

Version	Method	CP	AC
Original	100 MFW	0.9	0.6
	Stop words	0.9	0.5
	Stop+Moisl	1	0.55
	Affixes	1	0.45
	Affixes+Moisl	1	0.53
Normalised	100 MFW	0.9	0.54
	Stop words	1	0.52
	Stop+Moisl	1	0.52
	Affixes	0.8	0.44
	Affixes+Moisl	1	0.5
Linguistic annotation	100 MFW	0.9	0.57
	Stop words	1	0.51
	Stop+Moisl	**1**	**0.56**

Table 3: Efficiency of the clustering according to corpus version and features selection method

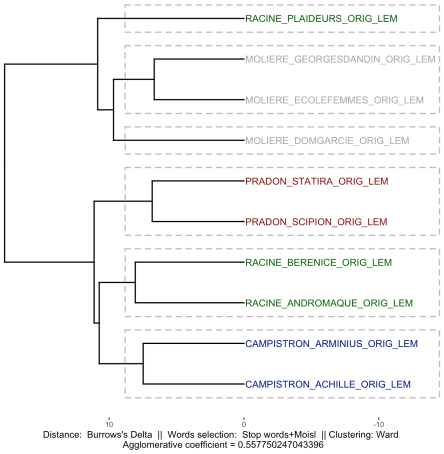

Figure 6: AHC with the best configuration

Labbé's hypothesis clearly proves to be, once again, wrong when using standard methods in stylometry.

4.3. Additional experiments

As previously explained, because our model for linguistic annotation has been trained on more than 17$^{\text{th}}$ c. prints, may they be normalised or not, it has been possible to tag the control corpus and merge it with our primary corpus. Interestingly, the results not only validate the previous ones, but prolong them. Using the same configuration (Moisl+stop words on linguistic annotation), we are now able to disentangle genre, authors but also centuries. Looking at the AHC (cf. figure 9), we see a first split according to the genre (comedies are in the upper part, tragedies in the lower part of the tree), then centuries (with 17$^{\text{th}}$ c. texts in the upper sub-parts and 18$^{\text{th}}$ c. texts in the lower sub-parts), and finally authors.

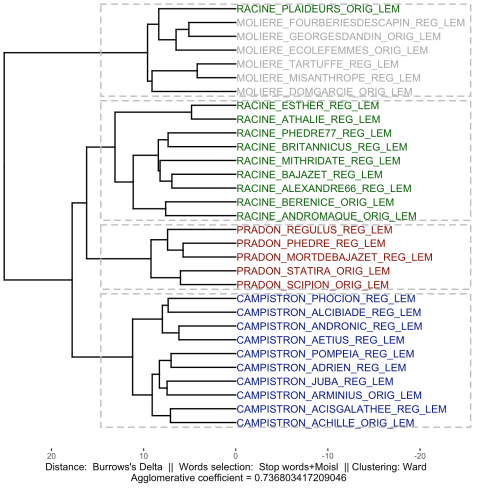

Figure 7: AHC with the best configuration (primary + secondary corpora)

A fairly reliable PCA (cf. figure 8, 45% of the total information is retained) shows similar results with comedies on the left and tragedies on the right, 18$^{\text{th}}$ c. texts on the upper part and 17$^{\text{th}}$ c. texts on the lower part, but emphasises some limits of our clustering, especially for three texts which are loosely attributed to a cluster (Marivaux's tragedy as a green square, Molière's heroic comedy as a turquoise circle, and, to a lesser extent, Racine's comedy as a purple square). It is extremely interesting to note that these three texts are all a play of the other genre than the speciality of the writer: this tension between personal and generic traits could be interpreted as a limited ability to mimic the characteristics of an other genre that one's speciality, literary "cross-dressing" showing here its limits. Tragedies of a comic writer would be, in a way, less tragic, and comedies of a tragedian less comic.

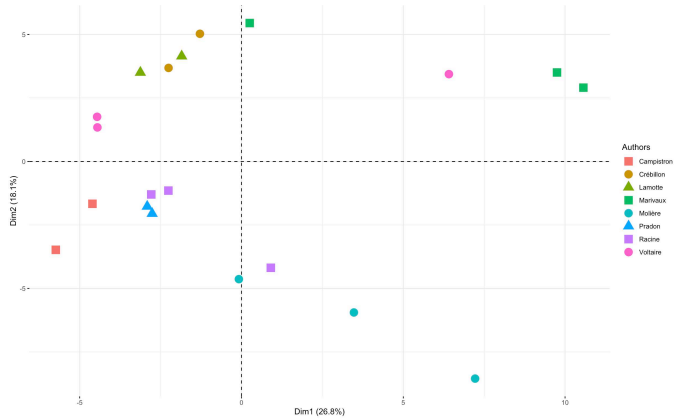

Figure 8: PCA with both corpora (Moisl+stopwords)

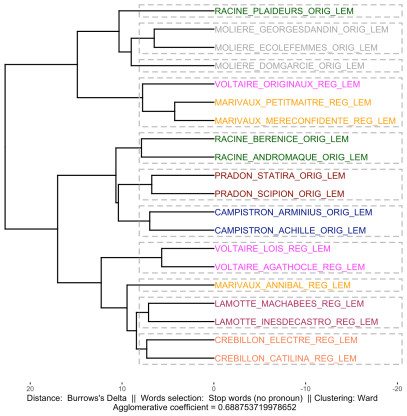

Figure 9: AHC with both corpora

This capacity to cross-dress seems however to vary from one author to another according to the PCA: some playwrights show a better homogeneity of the authorial signal despite genre variation, such as Racine, whose works are less spread on the graph than Voltaire's. A t-SNE (cf. figure 10) – *i.e.* a visualisation of high-dimensional data in a two-dimensional space prioritising short distances rather than long ones [35] – can highlight such a phenomenon by clustering together all the plays of a single author no matter the genre, revealing the inner homogeneity of apparently scattered works. Thus, if Marivaux' or Voltaire's plays are clearly divided by genre, it is not the case for Molière and Racine, whose stylometric signature seems more dense.

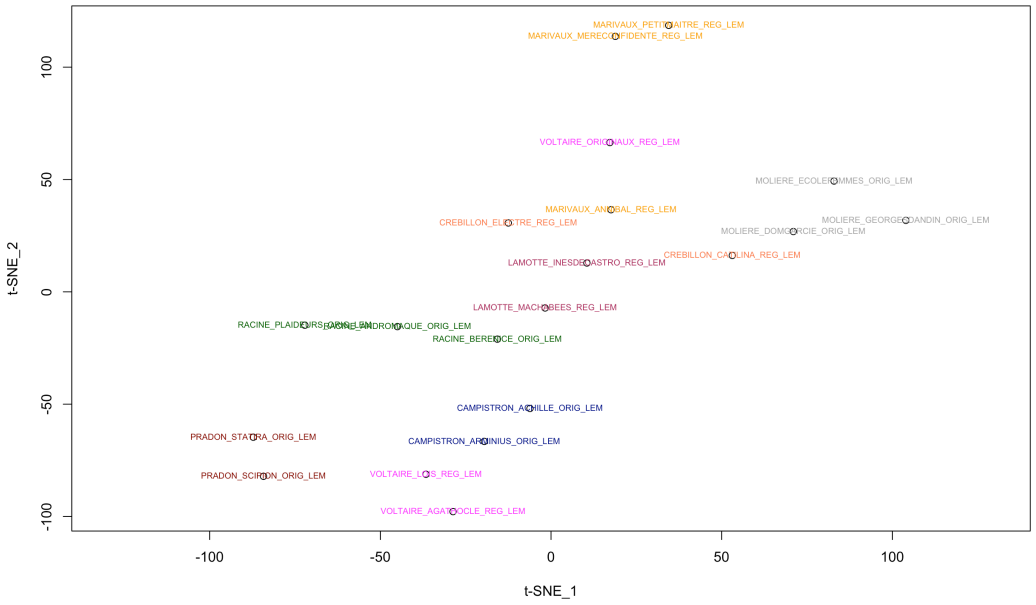

Figure 10: t-SNE (Moisl+stopwords, intial dims=2, perplexity=3)

5. Problem B: On style

5.1. Stylistic interpretation of stylometric markers

Now, another question arises: what are the tokens behind these clusters? The splits we have presented in our AHC are indeed based on tokens in texts, and we can hypothesise that those tokens reflect a specific trait, a stylome, of the author. To do so, we need to compute the link between a token and a cluster, which can be done with a v test (a test value).

$$v - test = \frac{\bar{x}_q - \bar{x}}{\sigma}$$

In the following equation, $\bar{x}_q$ is the average of variable X for the individuals (tokens) for a category q (clusters), $\bar{x}$ the average of the variable X accross all categories, and σ is the square root of the variance (*i.e.* the standard deviation) [28].

Token	V-test			
	17th c. (without Moisl)	17th c. (with Moisl)	17th c. (with Moisl) extended	17th c.+18th c. (with Moisl)
même (indefinite adjective, singular)	2.612539	2.574263	4.635370	-
tant (adverb)	2.534095	2.556951	2.361631	3.282359
encore (adverb)	2.413010	2.442929	2.096351	2.268252
après (preposition)	2.332788	2.356990	-	2.946071
enfin (adverb)	2.332233	2.339742	-	2.655457
si (adverb)	2.022235	-	2.411819	-
quel (relative determiner, masculine singular)	1.978633	2.000917	2.672914	-
contre (preposition)	-2.069312	-	-	-
jusque (preposition)	-	-	2.272614	2.245277
où (relative pronoun)	-	-	-	2.215313

Table 4
Selection of tokens attributed to the cluster of Racine's tragedies and their respective v test with four configurations, all on the stop words: with and without Moisl's selection on the primary corpus only, with Moisl's selection on the primary corpus merged with the secondary one, and with Moisl's selection on the primary and the control corpus merged.

With our best configuration on the primary corpus (cf. the second column of table 4), the v test defines six tokens as typical of Racine's tragedies, but these results are not absolute: test values are computed in contrast to the rest of the corpus, and change with the latter. However, the results of the v test for the same plays, but with another configuration (without Moisl's selection) or a larger corpus (*e.g.* the 17th c. and 18th c. texts merged), are similar, proving a relative stability, which would deserve further research.

The interpretation of these markers is partially facilitated by Spitzer's study: the intensive adverb *tant* ("so much"), analysed together with *si* ("so"), is identified as characteristic of Racine's muting effect.

> *An das distanzierende Demonstrativ können wir das beteuernde si und tant anschließen. Ein* si *[…] ruft ja den Gesprächspartner zum Zeugen an, man sollte also*

> *auf eine besonders 'warme' Wirkung schließen. Die gegenteilige Wirkung scheint nun bei Racine herauszukommen: das* si *hat etwas Kühl-Abgeschwächtes* [52]
>
> To the distancing demonstrative we can add the affirming *si* and *tant*. A *si* [...] calls the interlocutor to witness, so we should conclude that it has a particularly 'warm' effect. The opposite effect now seems to come out of Racine: the *si* has something cool and weakened [our translation]

A typical example with *tant* would be the following:

> Astyanax, d'Hector jeune et malheureux fils,
> Reste de *tant* de rois sous Troie ensevelis. (*Andromaque* I.1)
> (*Astyanax, Hector's young and unfortunate son,*
> *Remainder of so many kings buried under Troy.*)

The notion of plenty introduced by *tant* is immediately counterbalanced by the idea that this profusion has disappeared: so many kings are dead.

Stylometric markers seem to point in another direction: the unfolding of the narration, altering the flow of the story with a similar "muting effect". In that sense, the best example is *encore* ("again"):

> Où suis-je ? Qu'ai-je fait ? Que dois-je faire *encore* ? (*Andromaque* V.1)
> (*Where am I? What have I done? What do I still have to do?*)

Rather than adding new peripeteias, it conveys a sensation of lingering, of endless repetition without clear direction highlighted in our example by the interrogation.

The use of the indefinite adjective *même*, used in pronominal locution such as *lui-même* ("himself") or *nous-mêmes* ("ourselves"), creates a similar effect of circularity, but within the sentence itself, with a redundancy of pronouns provoking a loop in the narration.

> Mais moi-*même*, seigneur, que faut-il que je croie ? (*Bérénice* III.2)
> (*As for myself, my lord, what must I believe?*)

The polyptoton (*moi-même*/"myself"-*je*/"I") does strengthen the affirmation of the self (*me, myself and I*) but is used to emphasises doubt, accentuated in this very example by an interrogation.

Finally, the two tokens *après* ("after") and *enfin* ("finally") both have this "cooling" effect Spitzer talks about: the analepsis offers a reversed perspective on the story, looking at it from its end and therefore preventing any potential suspense:

> *Enfin*, *après* un siège aussi cruel que lent,
> Il dompta les mutins, reste pâle et sanglant
> Des flammes, de la faim, des fureurs intestines, (*Bérénice* I.4)
> (*Finally, after the siege as cruel as slow,*
> *He tamed the rebels, pale and bloody remainders*
> *of flames, hunger and intestine feuds,*)

The suddenness with which Racine wraps up the story (*enfin*/"finally") contrasts with its supposed length (*lent*/"slow"), and all the adventures seem to be concealed. In the light of this last example and the previous ones, we can conclude that if stylometric markers are stylistically relevant, their interpretation is far from being straightforward, and a careful examination of occurrences remains compulsory to avoid misinterpretations.

5.2. Stylometry, style and idiolect

For a few decades, it has been accepted that among all the available criteria, statistical repetition and deviation are not sufficient to identify a styleme [39], because this latter is not hyletic [43] but has a fully heuristic status [37]: a word is not *per se* poetic, but used poetically, and it is this usage that defines its literariness. Stylometric markers are nothing more than idiolectal traits with a potential aesthetic value that awaits to be deciphered.

The too loose definition of style proposed by Herrmann, Schöch & van Dalen-Oskam, "a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively" [27], is therefore unsatisfactory because it does not disentangle idiolectometry from stylistics – two related, yet substantially different approaches to the text, and potentially any other work of art. As G. Philippe explains: *l'idiolecte c'est le style sans la signification, et le style, l'idiolecte en tant qu'il peut faire l'objet d'une interprétation* ("the idiolect is style without signification, and style is the idiolect in so far as it can be interpreted") [42].

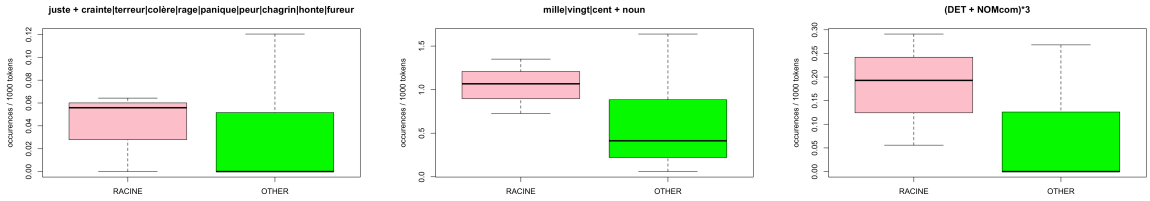

Figure 11: Three stylemes identified by Stpitzer, whose statistical over-representation in Racine's plays (compared with the rest of the primary corpus plus our control corpus) is confirmed.

The example of Racine demonstrates that with a minimal definition of style as "a property of texts constituted by an ensemble of formal features *with an interpretative yield* which can be observed quantitatively or qualitatively", stylometry, if carefully used, can potentially contribute to the identification of stylistic signatures. This identification would however not be complete *via* this only mean and needs to be combined with other approaches, such as textometry, to be fully captured. Textual motifs, which combine several words in a (semi-)rigid order [25], remain for instance a blind spot of a purely stylometric research despite their importance to describe Racine's style (cf. figure 11). It is the same for more syntactic studies, looking at sentences [24].

If stylometric markers are not perfect and constitute only a portion of the stylistic features of a writer, they also have their virtues. They could be of great help for an old, important and complex challenge of stylistics: the contextualisation of stylemes. What, from a given author, belongs to his time? or his school? Studying carefully markers, we can observe that the genre, as previously mentioned and exposed elsewhere [9], but also the date of writing do play an important role in the clustering, and even that the generic and the diachronic signals prevail upon the authorial one. In that sense, because a sub-cluster (the author) inherits from characteristics of the previous ones (the genre and the period), *tant*, *encore* or *enfin* are not only Racinian stylemes, but also classical[5] and tragic features.

[5] Used in the French sense to designate 17$^{\text{th}}$ c. literature

6. Conclusion

We can now, with certainty, remove any doubt about the paternity of Racine on Campistron's plays: the latter is not a figurehead of the former, despite Labbé's claims. Such a result is guaranteed by a battery of tests which all confirm our classification, but a careful study of their respective accuracy highlights the efficiency of performing an HCA on a linguistically annotated corpus rather than the raw text. Such a method not only offers a perfect CP, but also disambiguates homographs and corrects polymorphism due to spelling variations or elisions, which is of high interest when one needs to interpret both the classification itself and the words behind this classification.

These words can be identified with a standard v test, which computes the link between a given token and a cluster. Despite a certain volatility of the results, relatively dependent of the corpus used, stylometric markers do have a clear interpretative yield, which confirm Spitzer's idea of a muting effect in Racine's plays, but also prolong this idea by new examples. If traditional and stylometric stylemes concur, they are however of a slightly different nature because the latter characterise (mathematically) Racine's tragedies and cannot be transverse, *i.e.* shared to a significant extent with another writer, genre or period.

Such results show that stylometry does not recognise only idiolects, and can contribute to stylistic surveys at various levels, starting with the close reading of the text by identifying the stylemes that make it special. However, because of its comparative nature, stylometry does not limit itself to authors and does identify other broader clusters related to the genre or the period. Doing so, it answers Barthes' wish to *dépasser la notion d'idiolecte (primitivement retenue comme point de départ) et à voir dans toute écriture, fût-elle apparemment très individuelle, le fragment d'un sociolecte ou langage de groupe* ("to go beyond the notion of idiolect (originally retained as a starting point) and to see in all writing, however apparently very individual, the fragment of a sociolect or group language") [2]. Stylometry naturally articulates individual traits to global ones and contradicts S. Vaudrey-Luigi's affirmation, according to whom *ce n'est peut-être pas tant un style d'auteur que l'on reconnait qu'un style d'époque* ("it is not the style of an author that we recognise, but the style of a period") [55]: it might very well be both of them, one hidden under the other. Just like J. Scherer explained with the construction of plays [47], behind a frame made of strict rules, we see the apparition of individual traits, in the background, probably until the advent of romanticism and *le sacre de l'écrivain* [6].

The identification of these traits is however still problematic, because stylometric results remain specific to the primary corpus, whereas a proper "stylome", like the "genome" it has been named after, should be absolute. Indeed, if the definition of a stylistic signature that is relative to a given context is sufficient for authorship attribution, it remains of a limited interest for stylistic studies. The solution to this problem is still unclear to us, but clearly passes by a different approach to corpora, which need to be less homogeneous, and more representative of the production of the time, to offer more precise results.

Data and scripts

Supplementary materials (doc+code) are available on zenodo: https://doi.org/10.5281/zenodo.5526586.

Acknowledgments

The final version of this article would not have been possible without the help of J.-B. Camps, Fl. Cafiero, Th. Clérice, K. Abiven, G. Forestier and our reviewers. Thank you also to Éléonore, *directrice du "projet suisse"*.

References

[1] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. "A framework for projected clustering of high dimensional data streams". In: *Proceedings of the Thirtieth international conference on Very large data bases - Volume 30*. Vldb '04. Toronto, Canada: VLDB Endowment, 2004, pp. 852–863.

[2] R. Barthes. *Œuvres complètes: 1968-1971*. Paris: Éditions du Seuil, 2002.

[3] R. Barthes. *Sur Racine*. Pierres vives. Paris: Éditions du Seuil, 1965.

[4] J.-C. Basson and D. Labbé. "De précieux manuscrits". In: *Actes des 15es Journées internationales d'Analyse statistique des Données Textuelles*. 15es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2020). Toulouse, France, 2020.

[5] J.-C. Basson and D. Labbé, eds. *Jean Racine. Aétius, Juba, Tachmas. Tragédies inédites transcrites et présentées par Jean-Charles Basson et Dominique Labbé*. Montréal: Monière-Wollank Editeurs, 2015. URL: https://hal.archives-ouvertes.fr/hal-01165969.

[6] P. Benichou. *Le Sacre de l'Écrivain, 1750-1780. Essai sur l'avènement d'un pouvoir spirituel laïque dans la France moderne*. Paris: Joseph Corti, 1973.

[7] G. Bronner. *La Démocratie des crédules*. Paris: Presses Universitaires de France, 2013.

[8] J. Burrows. "'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship". In: *Literary and Linguistic Computing* 17.3 (2002), pp. 267–287. DOI: 10.1093/llc/17.3.267.

[9] F. Cafiero, J.-B. Camps, S. Gabay, and M. Puren. "La naissance du style: auteur vs genre aux XVIIe et XIXe siècles". In: *Humanistica 2020 - Archives du colloque*. Bordeaux, France: Humanistica, 2020. URL: https://hal.archives-ouvertes.fr/hal-02577853.

[10] J.-B. Camps and F. Cafiero. "Why Molière most likely did write his plays". In: *Science Advances* 5.1 (2019). URL: https://advances.sciencemag.org/content/5/11/eaax5489.

[11] J.-B. Camps, T. Clérice, and A. Pinche. "Stylometry for Noisy Medieval Data: Evaluating Paul Meyer's Hagiographic Hypothesis". In: *Digital Scholarship in the Humanities* 36 (2021). URL: http://arxiv.org/abs/2012.03845.

[12] M. Choiński, M. Eder, and J. Rybicki. "Harper Lee and Other People: A Stylometric Diagnosis". In: *Mississippi Quarterly* 70.3 (2017), pp. 355–374. DOI: 10.1353/mss.2017.0022. URL: https://muse.jhu.edu/article/747862.

[13] W. Daelemans. "Explanation in Computational Stylometry". In: *Computational Linguistics and Intelligent Text Processing*. Ed. by A. Gelbukh. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2013, pp. 451–462. DOI: 10.1007/978-3-642-37256-8_37.

[14] M. Eder. “Computational stylistics and Biblical translation : how reliable can a dendrogram be ?” In: *The Translator and the Computer*. Wrocław: Wyższa Szkoła Filologiczna we Wrocławiu, 2012, pp. 155–170. URL: http://docplayer.pl/949875-The-translator-and-the-computer.html.

[15] M. Eder. “Does size matter? Authorship attribution, small samples, big problem”. In: *Digital Scholarship in the Humanities* 30.2 (2015), pp. 167–182. DOI: 10.1093/llc/fqt066.

[16] M. Eder. “Mind your corpus: systematic errors in authorship attribution”. In: *Literary and Linguistic Computing* 28.4 (2013), pp. 603–614. DOI: 10.1093/llc/fqt039.

[17] M. Eder, J. Rybicki, and M. Kestemont. “Stylometry with R: A Package for Computational Text Analysis”. In: *The R Journal* 8.1 (2016), pp. 107–121. URL: https://journal.r-project.org/archive/2016/RJ-2016-007/index.html.

[18] S. Evert, T. Proisl, F. Jannidis, I. Reger, S. Pielström, C. Schöch, and T. Vitt. “Understanding and explaining Delta measures for authorship attribution”. In: *Digital Scholarship in the Humanities* 32 (suppl_2 2017), pp. ii4–ii16. DOI: 10.1093/llc/fqx023.

[19] P. Fièvre. *Théâtre classique*. 2007. URL: http://www.theatre-classique.fr.

[20] G. Forestier. *Molière auteur des œuvres de Molière*. 2011. URL: http://moliere-corneille.huma-num.fr.

[21] G. Forestier. “Révéler la vérité cachée : le cas Molière comme symptôme du fonctionnement et des enjeux de la pensée hypercritique de la Renaissance à aujourd’hui”. In: *La Vérité. Congrès annuel de l’IUF*. Toulouse, France, 2013. URL: https://hal.archives-ouvertes.fr/hal-01888357.

[22] S. Gabay, A. Bartz, and Y. Deguin. “CORPUS17: a philological corpus for 17th c. French”. In: *Proceedings of the 2nd International Digital Tools & Uses Congress (DTUC ’20)*. Hammamet, Tunisia, 2020. DOI: 10.1145/3423603.3424002.

[23] S. Gabay, T. Clérice, J.-B. Camps, J.-B. Tanguy, and M. Gille-Levenson. “Standardizing linguistic data: method and tools for annotating (pre-orthographic) French”. In: *Proceedings of the 2nd International Digital Tools & Uses Congress (DTUC ’20)*. Hammamet, Tunisia, 2020. DOI: 10.1145/3423603.3423996.

[24] R. Garrette. *La Phrase de Racine : étude stylistique et stylométrique*. Champs du signe: sémantique, rhétorique, poétique. Toulouse: Presses universitaires du Mirail, 1995. 331 p. URL: http://data.rero.ch/01-2209268/html.

[25] L. Gonon, V. Goossens, O. Kraif, I. Novakova, and J. Sorba. “Motifs textuels spécifiques au genre policier et à la littérature ’blanche’”. In: *SHS Web of Conferences* 46 (2018), p. 06007. DOI: 10.1051/shsconf/20184606007.

[26] H. v. Halteren, H. Baayen, F. Tweedie, M. Haverkort, and A. Neijt. “New Machine Learning Methods Demonstrate the Existence of a Human Stylome”. In: *Journal of Quantitative Linguistics* 12.1 (2005), pp. 65–77. DOI: 10.1080/09296170500055350.

[27] J. B. Herrmann, C. Schöch, and K. van Dalen-Oskam. “Revisiting Style, a Key Concept in Literary Studies”. In: *Journal of Literary Theory* 9.1 (2015). DOI: 10.1515/jlt-2015-0003.

[28] F. Husson, S. Lè, and J. Pagès. *Exploratory Multivariate Analysis by Example Using R*. Computer Science and Data Analysis Series. Boca Raton London New York: Chapman and Hall/CRC, 2017.

[29] M. Kestemont. “Function Words in Authorship Attribution. From Black Magic to Theory?” In: *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*. Gothenburg, Sweden: Association for Computational Linguistics, 2014, pp. 59–66. DOI: 10.3115/v1/W14-0908.

[30] C. Labbé and D. Labbé. “Inter-Textual Distance and Authorship Attribution. Corneille and Molière”. In: *Journal of Quantitative Linguistics* 8.3 (2001), pp. 213–231. URL: https://halshs.archives-ouvertes.fr/halshs-00139671.

[31] D. Labbé. “Comédiens et écrivains au XVIIe siècle. À la redécouverte des frères Corneille”. In: *Séminaire de stylistique française*. Cologne, Germany, 2011. URL: https://halshs.archives-ouvertes.fr/halshs-00657083.

[32] D. Labbé. “Jean Racine, plume de l’ombre ?” In: *Séminaire Linguistique du français moderne*. Neuchâtel, Switzerland, 2017. URL: https://hal.archives-ouvertes.fr/hal-01480917.

[33] D. Labbé. *Si deux et deux sont quatre, Molière n’a pas écrit Dom Juan...: Essais - documents*. Paris: Max Milo Editions, 2009.

[34] S. Lê, J. Josse, and F. Husson. “FactoMineR: A Package for Multivariate Analysis”. In: *Journal of Statistical Software* 25.1 (2008), pp. 1–18. DOI: 10.18637/jss.v025.i01.

[35] L. v. d. Maaten and G. Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[36] H. Moisl. “Finding the Minimum Document Length for Reliable Clustering of Multi-Document Natural Language Corpora”. In: *Journal of Quantitative Linguistics* 18.1 (2011), pp. 23–52. DOI: 10.1080/09296174.2011.533588.

[37] G. Molinié. “Sémiostylistique : à propos de Proust”. In: *Versants: revue suisse des littératures romanes* 18 (1990), pp. 21–30. DOI: 10.5169/seals-259858.

[38] G. Molinié and A. Viala. *Approches de la réception*. Perspectives littéraires. Paris: Presses Universitaires de France, 1993. DOI: 10.3917/puf.molin.1993.01.

[39] J. Molino. “Pour une théorie sémiologique du style”. In: *Qu’est-ce que le style ?* Paris: Presses Universitaires de France, 1994, pp. 213–261.

[40] F. Mosteller and D. L. Wallace. “Inference in an Authorship Problem”. In: *Journal of the American Statistical Association* 58.302 (1963), pp. 275–309. DOI: 10.2307/2283270.

[41] C. Muller. “Les ‘pronoms de dialogue’: interprétation stylistique d’une statistique de mots grammaticaux”. In: *Langue française et linguistique quantitative*. Travaux de linguistique quantitative. Genève: Slatkine, 1979, pp. 117–124.

[42] G. Philippe. “Traitement stylistique et traitement idiolectal des singularités langagières”. In: *Cahiers de praxématique* 44 (2005), pp. 77–92. DOI: 10.4000/praxematique.1659.

[43] F. Rastier. *Sémantique interprétative*. Formes sémiotiques. Paris: Presses Universitaires de France, 2009.

[44] S. Rebora and M. Salgaro. “Is ‘Late Style’ measurable? A stylometric analysis of Johann Wolfgang Goethe’s, Robert Musil’s, and Franz Kafka’s late works”. In: *Elephant & Castle: laboratorio dell’immaginario* 18 (2018), pp. 4–39. URL: https://www.dlls.univr.it/?ent=pubbdip%5C&id=988359.

[45] P. J. Rousseeuw. “A visual display for hierarchical classification”. In: *Data Analysis and Informatics* 4 (1986), pp. 743–748.

[46] U. Sapkota, S. Bethard, M. Montes, and T. Solorio. “Not All Character N-grams Are Created Equal: A Study in Authorship Attribution”. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2015, pp. 93–102. DOI: 10.3115/v1/N15-1010.

[47] J. Scherer. *La Dramaturgie classique en France*. 1 vols. Paris: Nizet, 1950. 488 pp.

[48] C. Schöch. “Fine-tuning Stylometric Tools: Investigating Authorship and Genre in French Classical Theater”. In: *DH2013 conference - Book of abstracts*. Lincoln (NE), 2013. URL: http://dh2013.unl.edu/schedule-and-events/program/.

[49] C. Schöch. “Zeta für die kontrastive Analyse literarischer Texte: Theorie, Implementierung, Fallstudie”. In: *Quantitative Ansätze in den Literatur- und Geisteswissenschaften: Systematische und historische Perspektiven*. Berlin: de Gruyter, 2018, pp. 77–94. DOI: 10.1515/9783110523300.

[50] M. K. Sjöblom. “L’indice pronominal est-il encore d’actualité ?” In: *Lexicometrica* 5 (2004). URL: http://lexicometrica.univ-paris3.fr/article/numero5.htm.

[51] L. Spitzer. “Die klassische Dämpfung in Racines Stil”. In: *Archivum romanicum* 12 (1928), pp. 361–472. URL: http://digitale.bnc.roma.sbn.it/tecadigitale/giornale/TO00176940/1928/unico/00000379.

[52] L. Spitzer. *Etudes de style*. Trans. by E. Kaufholz. Bibliothèque des idées Gallimard. Paris: Gallimard, 1970.

[53] L. Spitzer. “The Muting Effect of Classical Style in Racine (1928)”. In: *Racine: Modern Judgements*. Trans. by R. C. Knight. Modern Judgements. London: Macmillan Education UK, 1969, pp. 117–131. DOI: 10.1007/978-1-349-15297-1_9.

[54] J.-M. Thomasseau. “Pour une analyse du para-texte théâtral : quelques éléments du para-texte hugolien”. In: *Littérature* 53.1 (1984), pp. 79–103. DOI: 10.3406/litt.1984.2218.

[55] S. Vaudrey-Luigi. “De la signature stylistique à la reconnaissance d’un style d’auteur”. In: *Le francais aujourd’hui* n°175.4 (2011), pp. 37–46. URL: https://www.cairn.info/revue-le-francais-aujourd-hui-2011-4-page-37.htm.

[56] J. H. Ward. “Hierarchical Grouping to Optimize an Objective Function”. In: *Journal of the American Statistical Association* 58.301 (1963), pp. 236–244. DOI: 10.1080/01621459.1963.10500845.

A. Appendix

Table 5
Breakdown of the control corpus

Author	Title	Place	Publisher	Printer	Date	Form	Genre
Crébillon	*Catilina*	Paris	Praut fils	-	1748	verse	Tragedy
Crébillon	*Électre*	Paris	Pierre Ribou	-	1708	verse	Tragedy
La Motte	*Inés de Castro*	Paris	Grégoire Dupuis François Flahault	-	1723	verse	Tragedy
La Motte	*Les Machabées*	Paris	-	-	1721	verse	Tragedy
Marivaux	*Annibal*	Paris	Noël Pissot	-	1727	verse	Tragedy
Marivaux	*La Mère confidente*	Paris	Praut fils	-	1735	prose	Comedy
Marivaux	*Le Petit Maître corrigé*	-	-	-	1734	prose	Comedy
Voltaire	*Agathocle*	-	-	-	1779	verse	Tragedy
Voltaire	*Lois de Minos*	Paris	Valade	-	1773	verse	Tragedy
Voltaire	*Les Originaux*	-	-	-	1732	prose	Comedy

Table 6
Breakdown of the secondary corpus

Author	Title	Place	Publisher	Printer	Date	Form	Genre
Campistron	*Acis et Galatée*	-	Paris	-	1690	verse	Pastorale héroïque
Campistron	*Adrien*	E. Lucas	Paris	-	1683	verse	Tragedy
Campistron	*Aétius*	-	-	-	1685	verse	Tragedy
Campistron	*Alcibiade*	J. Garrel	Amsterdam	-	1685	verse	Tragedy
Campistron	*Andronic*	J. Garrel	Amsterdam	-	1685	verse	Tragedy
Campistron	*Juba, roi de Mauritanie*	NA	NA	NA	1685	verse	Tragedy
Campistron	*Phocion*	-	-	-	1695	verse	Tragedy
Campistron	*Pompéia*	Paris	Compagnie des libraires	-	1750	verse	Tragedy
Molière	*Fourberie de Scapin*	P. Le Monnier	Paris	-	1671	prose	Comedy
Molière	*Le Misanthrope*	J. Ribou	Paris	-	1667	verse	Comedy
Molière	*Tartuffe*	J. Ribou	Paris	-	1669	verse	Comedy
Pradon	*Tamerlan*	J. Ribou	Paris	-	1676	verse	Tragedy
Pradon	*Phèdre et Hippolyte*	Th. Amaulry	Paris	-	1677	verse	Tragedy
Pradon	*Régulus*	Th. Guillain	Paris	-	1687	verse	Tragedy
Racine	*Alexandre le Grand*	Th. Girard	Paris	-	1666	verse	Tragedy
Racine	*Athalie*	D. Thierry	Paris	-	1691	verse	Tragedy
Racine	*Bajazet*	P. Le Monnier	Paris	-	1672	verse	Tragedy
Racine	*Britannicus*	Cl. Barbin	Paris	-	1670	verse	Tragedy
Racine	*Esther*	D. Thierry	Paris	-	1689	verse	Tragedy
Racine	*Mithridate*	Cl. Barbin	Paris	-	1673	verse	Tragedy
Racine	*Phèdre*	Cl. Barbin	Paris	-	1677	verse	Tragedy