

A *Gueuloir* of One's Own: Computing the Acoustic Signature of Flaubert

Simon Gabay¹ , Florian Cafiero^{2,3} , and Jean-Luc Falcone¹ 

¹ Université de Genève, Geneva, Switzerland

² EPITA, Paris, France

³ Centre Jean Mabillon, École nationale des chartes - PSL, Paris, France

Abstract

Can the traces of an ear-driven literary practice be detected in prose? This paper addresses the question through the case of Gustave Flaubert, whose *gueuloir*, his habit of declaiming sentences aloud and revising them by ear, has long been assumed to shape his writing. We compare Flaubert's novels against those of three contemporaries (Aimard, Chevalier, Ponson du Terrail). We first show that phonetic transcriptions consistently outperform orthographic versions of the texts for authorship attribution, suggesting that French spelling obscures part of the authorial signal and that phonemes are more reliable stylometric features than graphemes. Building on this finding, we examine different dimensions of the acoustic signal to characterise each author's signature. At the phonemic level, Flaubert displays a higher cluster density, a lower vowel-to-consonant ratio, and a reduced bigram entropy. Rhythmically, his writing is also distinctive, with sparser stresses, longer unstressed runs, and more predictable local sequences, though rhythm alone provides a weaker basis for attribution. These results demonstrate that the *gueuloir* is characterised by a consonantal texture and a monotonous flow. More broadly, they make a case for phonetic-based stylometry as a method for studying acoustic phenomena in prose composition, especially for French.

Keywords: acoustics, stylometry, computational stylistics, French literature

1 Introduction

One of the most enduring yet contested principles of modern linguistics is the arbitrariness of the sign: the claim that the bond between a word's acoustic form and its meaning is conventional rather than motivated [36]. Between these two dimensions, literary criticism has overwhelmingly privileged the signified, leaving the acoustic properties of the signifier comparatively neglected. This neglect is historically surprising: since at least the sixteenth century, the *jugement de l'oreille* ("judgement of the ear") has shaped linguistic choices in French [49]. More broadly, literary production has long been guided by euphonic and prosodic considerations, as evidenced by devices such as alliteration and related sound-based figures. Sound, nevertheless, remains understudied in both traditional and computational stylistics, especially for prose.

Recent work in medieval song studies has proposed integrating sound into literary analysis through a "three-tiered model of description that distinguishes between the meaning layer, the linguistic sound layer, and the melodic sound layer" [25], with the goal of incorporating "the imaginings and modelling of sound into the analysis in a methodically controlled manner" [50]. We take the linguistic sound layer as our point of departure, arguing that it can be operationalised to yield new stylistic insights through a signifier-based approach. Our test case is Gustave Flaubert (1821–1880), an author known for his acute attention to the sound of his prose, whose texts we examine for significant acoustic differences from those of his contemporaries.

Simon Gabay, Florian Cafiero, and Jean-Luc Falcone. "A *Gueuloir* of One's Own: Computing the Acoustic Signature of Flaubert." In: *Computational Cultural Science Workshop 2026*, ed. by Suzanne Mpouli, Jean-Baptiste Camps, Florian Cafiero, and Chahan Vidal-Gorène. Vol. 6. Anthology of Computers and the Humanities. 2026, 32–45. <https://doi.org/10.63744/txBiSXmISDgd>.

© 2026 by the authors. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

Flaubert is famously associated with the *gueuloir*, literally a “shouting room”, whereby he tested the musicality of his prose by reading it aloud [16]. However, the extent to which this practice produces perceptible stylistic effects was questioned by some of his contemporaries, as illustrated by the remark Gautier made to the Goncourt brothers:

Très souvent, son rythme, à Flaubert, n’est que pour lui seul, il nous échappe. Il se gueule ça à lui-même. Vous savez, il a des gueuloirs, de ses phrases qui lui semblent très harmoniques ; mais il faudrait lire comme lui, pour avoir l’effet de ses gueuloirs. “Very often, Flaubert’s rhythm is meant for him alone; it escapes us. He shouts it out to himself. You know, he has his ‘shouting rooms’, phrases that strike him as highly harmonious; but one would have to read the way he does to experience the effect of his shouting rooms.”

— Théophile Gautier, quoted by Edmond and Jules de Goncourt, *Journal*, 3 March 1862 [23].

This study draws on stylometry and computational analysis of style to investigate the acoustic dimension of literary style in general, and of Flaubert’s *gueuloir* in particular. We argue that his stylistic analysis can benefit from experiments conducted on the phonetic transcriptions of his texts rather than on their orthographic versions. As this approach is new, we first assess, through authorship attribution, whether phonetic transcriptions capture an authorial signal. Since an acoustic signature depends not only on phonemic patterning but also on rhythmic structuring, we then supplement our phonemic analyses with tests based on accentual transcriptions of the novels.

2 State of the art

Research on the most relevant features for stylometry has attracted considerable efforts and led to a wide range of solutions. Character *n*-grams [28], rhymes [30], function words [29], pseudo-affixes [46], syntactic patterns [33], and combinations of lexicological and morpho-syntactic data [4; 17; 27] have all been proposed.

Among them, acoustic properties have received attention through metre and accentual patterns in poetry [8; 11; 39; 40; 41; 42], though such features have not yet been adopted consistently even for verse [1; 5; 48], and no applications have been proposed for French prose. Regarding phonemes, Deng and Allahverdyan [13] have shown that the rank-frequency curve of phonemes in English texts reveals an author-specific “signature” in phoneme usage that cannot be reduced to shared vocabulary or genre. Despite this finding, very few authorship attribution studies have relied on phonetics. Khomytska and Teslyuk [31] have applied their “multifactor method” to phonemes, and forensic linguists have exploited phonetic-prosodic clues for author profiling [24; 26].

In the Humanities, attempts to exploit sound remain rare [38], beyond the work of Haug [25], Stock [50], and our recent stylometric study of the *Illuminations* [18]. Phonostylistics, which draws on phonetics, sociolinguistics, and psycholinguistics [34], focuses primarily on oral variation (social, regional, etc.) rather than on written literary texts. From a different theoretical standpoint, Henri Meschonnic forcefully argued that rhythm should be considered “a linguistic level like the lexicon or syntax” [37, p. 363], but his approach remains resolutely qualitative and text-specific.

The scarcity of studies is probably explained by the difficulty of producing a hermeneutic yield from an acoustic analysis of language [12], and the technical difficulty of conducting such an analysis on large corpora. Regarding Flaubert’s work, the few existing approaches to the *gueuloir* thus remain limited in scope: Fried’s close reading [16] remains entirely qualitative, and Richaudeau’s study of Flaubert’s prose style [44] is restricted to samples of 192 sentences per novel.

Because no author uses the phonetic alphabet, we need to convert orthographic texts into phonetic transcriptions. While text-to-speech synthesis remains problematic, the grapheme-to-

phoneme conversion itself is relatively manageable [9], though neural network-based approaches for French remain limited by scarce training data [32].

3 Data

Author	Title	Date	Tokens
Gustave Aimard (1818–1883)	<i>Les Trappeurs de l’Arkansas</i>	1857	90
	<i>L’Eclaireur</i>	1859	128
	<i>Le Guaranis</i>	1864	77
	<i>Les Bohêmes de la mer</i>	1865	60
	<i>Ourson Tête de fer</i>	1875	62
	<i>L’Olonnais</i>	1876	106
	<i>Le Fils du soleil</i>	1879	56
	<i>Les Bandits de l’Arizona</i>	1882	72
			=651
Henri-Émile Chevalier (1828–1879)	<i>La Capitaine</i>	1862	50
	<i>La Tête plate</i>	1862	64
	<i>Les Nez percés</i>	1862	61
	<i>Poignet d’acier, ou les Chippiouais</i>	1863	51
	<i>Peaux rouges et peaux blanches</i>	1864	59
	<i>L’Enfer et le paradis de l’autre monde</i>	1865	44
	<i>Jacques Cartier</i>	1868	62
	<i>Le Chasseur noir</i>	1877	56
	<i>La Fille du pirate</i>	1878	53
	<i>Le Gibet</i>	1878	60
			=560
Gustave Flaubert (1821–1880)	<i>L’Education sentimentale (1st version)</i>	1845	105
	<i>Madame Bovary</i>	1857	120
	<i>Salammbô</i>	1862	105
	<i>L’Education sentimentale (2nd version)</i>	1869	163
	<i>La tentation de Saint-Antoine</i>	1874	43
	<i>Bouvard et Pécuchet</i>	1881	94
			=630
Pierre-Alexis de Ponson du Terrail (1829–1871)	<i>La Baronne trépassée</i>	1852	69
	<i>L’Héritage mystérieux</i>	1857	189
	<i>Les Chevaliers du clair de lune</i>	1859	139
	<i>Le Chambrion</i>	1865	40
	<i>L’Orgue de Barbarie: les héros de la vie privée</i>	1869	70
	<i>Les Aventures du capitaine La Palisse</i>	1880	76
			=583

Table 1: Corpus per author, with publication date and size (in k tokens).

Our corpus of novels (Table 1) draws on data distributed by ANR Chapitres [2]. All selected authors are novelists active during the same period as Flaubert and born within a decade of him: Gustave Aimard (1818–1883), a Parisian *feuilletonniste* specialising in adventure fiction; Henri-

Émile Chevalier (1828–1879), a French writer who spent a decade in exile in America; and Pierre-Alexis de Ponson du Terrail (1829–1871), best known for his *Rocamboles* series.

A key selection criterion was the exclusion of authors known to have relied on ghostwriters, a widespread practice in nineteenth-century French literary production [15]. On Flaubert’s side, we extend the corpus to include the posthumously published first version of *L’Éducation sentimentale* and the unfinished *Bouvard et Pécuchet*.

4 Methods

4.1 Textual representations and transcription pipeline

4.1.1 Phonetic transcription

All texts are transcribed into a phonetic alphabet that uses one symbol per phoneme. We use a modification of X-SAMPA [51], replacing symbols encoded as two characters to single-character variants. We design a lightweight phonetic transcription process drawing on *Morphalou 3.1* [3], an extensive French lexicon containing the inflected forms of 159,271 lemmas along with their phonetic transcription, loaded into an embedded SQLite database. The lexicon is complemented with an *ad hoc* dictionary for missing terms, and texts are preprocessed with *Pie extended* [10] for tokenisation and part-of-speech annotation, using models specialised for modern French [19]. All proper nouns are removed to avoid interference with the authorial signal.

Each token is first looked up in the *ad hoc* dictionary; if not found, a lookup in *Morphalou* retrieves all possible transcriptions, with part-of-speech information used to resolve ambiguities. We then add basic French *liaisons*, in which a normally silent final consonant is pronounced when the following word begins with a vowel (e.g. *Les Illuminations* → lezilyminasjõ for [lezilyminasjõ], with z representing the *liaison* consonant). Given the complexity of French liaison, we focus only on mandatory liaisons so as not to introduce erroneous ones, applying them between token pairs satisfying conditions relating to grammatical category and the absence of aspirated h.

Orthographic	Nous étions à l’Étude, quand le Proviseur entra, suivi d’un nouveau habillé en bourgeois et d’un garçon de classe qui portait un grand pupitre.
Phonetic	nuzetjõaletyd@kāl@pR0viz9RātRasHividūnuvoabijeãbuRZwae dūgaRsõd@klas@kipORtEũgRāpypitR@

Table 2: Example of a phonetic transcription (from *Madame Bovary*).

4.1.2 Rhythmic transcription

We complement the phonemic analysis by conducting a comparable experiment using rhythm-oriented transcriptions, automatically generated through a rule-based procedure following the model of Plecháč and Birnbaum [42] as implemented by Nagy et al. [39]. Accentual stress (marked S) falls on the final pronounced syllable of each lexical word, excluding final mute -e. Weak syllables (marked w) include function words and clitics. These patterns are concatenated per sentence with | marking sentence boundaries.

Orthographic	Il connut la mélancolie des paquebots, les froids réveils sous la tente, l’étourdissement des paysages et des ruines, l’amertume des sympathies interrompues.
Rhythmic	wwSwwwwSwSwSwSwSwwwwSwSwSwSwSwwwwS

Table 3: Example of a rhythmic transcription (from *L’Éducation sentimentale*).

4.2 Analytical features

4.2.1 Calibration for stylometric classification

A supervised classification analysis is conducted using a Support Vector Machine (SVM) classifier implemented in *MegaStyl* [20], a version of *SuperStyl* [7] enhanced for phonetic data. We systematically explore several dimensions of the classification setup: feature representations (character trigrams, function words, affixes, and combinations), sampling strategies, dimensionality reduction via PCA, penalised vs. non-penalised SVM, and two validation protocols. We investigate the influence of sample length on classification performance, with orthographic sample sizes ranging from 500 to 3,000 tokens and phonetic sample sizes from 5,000 to 10,000 phonemes.

4.2.2 Phonemic metrics

Rather than concentrating on individual prominent trigrams, we interpret results through metrics that capture the structural properties and distributional patterns of trigram usage for each author. We evaluate four phonemic metrics:

- *Hiatus avoidance ratio*: the proportion of vowel pairs avoided to ensure phonetic fluidity, calculated as $1 - \frac{\text{hiatus count}}{\text{vowel count}}$.
- *Vowel/consonant ratio*: the number of vowels divided by the number of consonants.
- *Consonant cluster density*: the frequency of consecutive consonants (clusters of two or more), normalised per 1,000 phonemes.
- *Bigram entropy*: Shannon entropy [47] on bigram frequencies, quantifying the unpredictability of consecutive phoneme pairs.

4.2.3 Rhythmic metrics

Five complementary metrics are computed across each author’s corpus:

- *Local rhythm regularity*: the consistency of per-token syllable counts (the sum of *w* and *S*) within local windows of 10 consecutive tokens. For each sliding window, we compute the variance of these per-token syllable counts; the final score is the mean of these variances across the text.
- *S/w ratio*: the overall proportion of stressed syllables relative to total syllables.
- *Average length of weak-syllable groups*: the mean number of consecutive weak syllables (*w*) between two stressed syllables (*S*).
- *Variability of weak-syllable group lengths*: the standard deviation of the lengths of these weak groups across all sentences, indicating the degree of local rhythmic irregularity.
- *Bigram pattern entropy*: Shannon entropy applied to consecutive pairs of stress markers, quantifying the predictability of rhythmic sequences.

4.3 Statistical testing

Given the small number of observations per author and no assumption of normality, we use the Wilcoxon rank-sum test [35; 52] for pairwise comparisons between Flaubert and each of the remaining authors. This non-parametric test assesses whether one distribution is stochastically greater or lesser than another by ranking pooled observations. All tests are one-sided, reflecting directional hypotheses, and use a confidence level of 0.95.

4.4 Dependence between variables

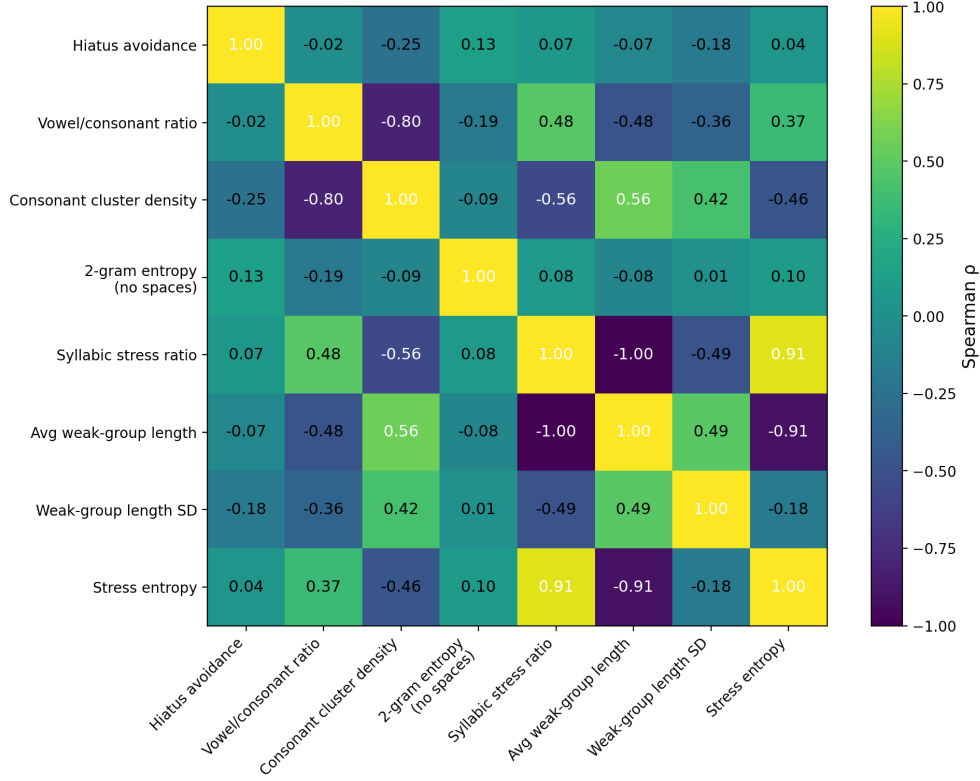


Figure 1: Spearman correlation matrix of phonemic and rhythmic metrics computed on the corpus. The structure reveals partially distinct segmental and prosodic dimensions, alongside highly collinear rhythmic descriptors and more independent entropy-based measures.

Our variables are not fully independent of one another. The correlation matrix (Figure 1) reveals that features cluster into partially distinct blocks. Phonetic descriptors such as vowel/consonant ratio and consonant cluster density are strongly opposed, reflecting a coherent segmental dimension. Rhythmic metrics exhibit a high degree of internal collinearity, with near-deterministic relationships between stress ratio, weak-group length, and related measures, pointing to a reduced effective dimensionality. Entropy-based descriptors, by contrast, display weaker and more diffuse correlations with the other variables, suggesting that they capture complementary aspects of the signal. Overall, the selected features do not collapse into a single dimension but organise into a limited number of structured yet partially independent components.

5 Results

5.1 Stylometric performance of phonetic transcriptions

We assess whether phonetic transcriptions capture each author’s stylometric signal, using SVM classifiers and comparing performance against the orthographic baseline on the same corpus. Our results (Table 4) are consistent with those of a similar study conducted on different corpora [14], showing that operating on phonetic transcriptions provides a consistent boost to classification performance. Orthographic versions already yield strong results, but phoneme-based classifications outperform them, yielding a gain of approximately 3 percentage points overall, with a macro-averaged F1 (hereafter macro F1) of 0.99 versus 0.96. This improvement may plausibly be attributed to the complexity of the French orthographic system, which introduces several sources of noise: silent letters, multiple grapheme combinations for a single phoneme, and phenomena

partially invisible in writing, such as *liaisons*. By working directly on phonemes, the classifier bypasses these irregularities and picks up the stylometric signal more cleanly. These results support the use of phonetic transcriptions for the analyses that follow.

Table 4: Best-performing SVM classifiers using orthographic and phonetic transcriptions of the corpus, with character 3-grams, class weights, and 10-fold cross-validation held constant.

(a) Orthographic transcription, 2,500 tokens/sample, PCA dimensionality reduction, linear Kernel. **(b)** Phonetic transcription, 8,500 characters/sample, downsampling, linear Kernel.

	Precision	Recall	F1-score	Support
Gustave Aimard	0.94	0.98	0.96	257
Émile Chevalier	0.92	0.94	0.93	218
Gustave Flaubert	0.97	0.97	0.97	251
Pierre Ponson du Terrail	1.00	0.92	0.96	231
accuracy			0.96	957
macro average	0.96	0.95	0.96	957
weighted average	0.96	0.96	0.96	957

	Precision	Recall	F1-score	Support
Gustave Aimard	0.98	0.99	0.98	253
Émile Chevalier	0.97	0.98	0.97	212
Gustave Flaubert	1.00	0.99	0.99	246
Pierre Ponson du Terrail	1.00	0.99	0.99	217
accuracy			0.99	928
macro average	0.99	0.99	0.99	928
weighted average	0.99	0.99	0.99	928

5.2 Phonemic profile

Table 5 reports four phonemic metrics computed across the novels attributed to each author. Three metrics reveal a statistically distinctive profile for Flaubert. His consonant cluster density (162.54 occurrences per 1,000 characters) is the highest in the corpus and significantly elevated relative to the non-Flaubert subcorpus ($W = 131$, $p < 0.001$). His vowel-to-consonant ratio (0.9572) is markedly lower than that of the other authors ($W = 30$, $p = 0.014$). His phoneme bigram entropy (8.488 bits) is the lowest across all sampled authors ($W = 30$, $p = 0.014$). Hiatus avoidance ratios, by contrast, are remarkably stable across all four authors (0.9953–0.9956), most likely reflecting the structural disfavouring of hiatus in French, leaving little room for individual variation.

Table 5: Comparison of key euphonic metrics across authors. Bold values indicate metrics where Flaubert shows a distinctive profile.

Author	Hiatus avoidance ratio	Vowel/consonant ratio	Consonant cluster density (per 1k chars)	Bigram entropy (bits)
Gustave Flaubert	0.9954	0.9572	162.54	8.488
Gustave Aimard	0.9954	0.9715	156.71	8.5075
Émile Chevalier	0.9956	0.9673	155.61	8.53
Pierre Ponson du Terrail	0.9953	0.9958	149.73	8.4951

The three significant metrics converge on a coherent phonetic profile. Flaubert’s elevated consonant cluster density suggests a systematic preference for a denser, more textured sound fabric than his contemporaries. This is corroborated by his lower vowel-to-consonant ratio, indicating a relative predominance of consonants over vowels. His lower bigram entropy, reflecting a more constrained and recurrent distribution of consecutive phoneme pairs, may be interpreted as a signature of acoustic control at the sublexical level – a pattern consistent with the hypothesis that the *gueuloir* produces measurable and distinctive stylistic effects.

5.3 Rhythmic profile

5.3.1 Classification

Table 6 reports the performance of the best SVM classifier trained on rhythmic features. Results are moderate overall, with a macro F1 of 0.76 on our corpus, suggesting that rhythmic patterns alone do not constitute a fully reliable stylometric signal. Flaubert achieves an F1 of 0.83, the highest score, closely followed by Aimard (0.82), and well above the macro average. The relatively strong classifier performance on Flaubert’s prose compared to Chevalier (0.66) or Ponson du Terrail (0.71) is consistent with the hypothesis that the *gueuloir* leaves a rhythmic imprint: a writer who listens to his sentences as he writes them is more likely to leave a rhythmic imprint that a model can learn. The moderate overall results may therefore reflect not a failure of the method but a meaningful finding: rhythm discriminates where it has been cultivated.

Table 6: Best SVM model based on rhythmic transcriptions. 3,000 tokens/sample, downsampling, RBF kernel, 9-grams.

	Precision	Recall	F1-score	Support
Gustave Aimard	0.84	0.81	0.82	196
Émile Chevalier	0.63	0.70	0.66	164
Gustave Flaubert	0.85	0.81	0.83	190
Pierre Ponson du Terrail	0.70	0.71	0.71	175
Accuracy			0.76	725
Macro average	0.76	0.76	0.76	725
Weighted average	0.76	0.76	0.76	725

5.3.2 Exploration

Table 7 presents rhythmic metrics aggregated at the author level. Flaubert stands out on two dimensions: he has the lowest S/w ratio (0.3625) and the longest average weak-group length (1.761). Both differences are statistically significant: his S/w ratio is significantly lower than that of the other authors ($W = 30$, $p = 0.014$), and his average weak-group length is significantly greater ($W = 114$, $p = 0.014$). Flaubert’s prose thus distributes its stresses more sparingly, with longer unstressed intervals, than that of his contemporaries. His bigram pattern entropy (1.803) is also the lowest ($W = 18$, $p = 0.0017$), indicating that his rhythmic sequences are more uniform and predictable.

Table 7: Rhythmic metrics across authors.

Author	Local rhythm regularity	S/w ratio	Avg weak-group length	Group length variability	Bigram pattern entropy
Flaubert-Gustave	0.5558	0.3625	1.761	1.251	1.803
Aimard-Gustave	0.5833	0.3664	1.301	1.299	1.823
Chevalier-H-Emile	0.5819	0.3707	1.698	1.267	1.827
Ponson-du-Terrail-Pierre	0.5956	0.3790	1.639	1.224	1.834

Taken together, these metrics suggest that Flaubert’s prose is rhythmically distinctive in a specific and coherent way: stresses arrive less frequently, at wider and more regular intervals, producing a cadence that is both more measured and more predictable at the local level. The low bigram entropy is consistent with the famous monotony of Flaubert’s style, a controlled rhythmic regularity compatible with deliberate revision.

6 Discussion

Our classification experiments show that phonetic transcriptions are more discriminative than their orthographic versions for authorship attribution: authorial style appears to be carried in part by acoustic patterns that spelling only imperfectly captures. Nevertheless, all three representations (orthographic, phonetic and rhythmic) capture, to varying degrees, the stylometric signal of each

author, supporting their use as data for further experiments aimed at exploring each author’s style in greater detail.

Phonemic metrics all converge on a profile marked by a stronger consonantal texture, where consonants tend to group into clusters. Rhythm, by contrast, is less powerful as a classification basis overall, even though Flaubert remains relatively distinctive within that weaker space: stresses are sparser and unstressed sequences are longer. Across both dimensions, rhythmic and phonemic, the low bigram entropy suggests that his style is locally more predictable than that of his contemporaries. The *gueuloir* may have acted by imposing a smooth rhythmic template and by sharpening the fine-grained acoustic texture of prose through repeated adjustments to phoneme sequences, vowel-consonant balance, and sentence cadence.

Our findings partially align with stylistic effects already identified in Flaubert’s prose. The observation by Proust [43] regarding the *défilement continu, monotone* (“continuous, monotonous flow”) of Flaubert’s style is consistent with the low phonemic and rhythmic entropy and long sequences of unstressed syllables we observe in his prose (*il connut la mélancolie des paquebots, les froids réveils sous la tente, l’étourdissement des paysages et des ruines, l’amertume des sympathies interrompues*). Similarly, Zola [53] notes that *il avait besoin d’un certain nombre de r, pour donner du roulement à la période* (“he needed a certain number of r’s to give a rolling effect to his sentence”); in our data, consonant clusters frequently centre on r (*un petit étendard de toile verdâtre, enrichi d’inscriptions en lettres d’or*). However, these clusters rarely form alliterations and vary across occurrences, lending support to Gothot-Mersch [22, p. 249], who emphasises that [*Flaubert*] *est tellement sensible à ces ‘échos’ involontaires* (“Flaubert is so sensitive to these involuntary ‘echoes’”).

7 Conclusion

In 1982, Henri Meschonnic [37] argued that rhythm should be understood not as a technical property of verse but as a structural principle of discourse, one that informs prose no less than poetry. Our study takes this claim seriously: approaching rhythm in an extended sense that encompasses both the accentual and phonemic dimensions of the text, we show that the linguistic sound layer carries genuine stylistic information. In most of our analyses, Flaubert occupies a distinctive position, which suggests that sound is not merely ornamental in his novels but part of his style.

Although our results are compatible with the presence of computationally detectable traces of the *gueuloir*, the nature of those traces raises a broader methodological question. Phoneme frequencies used by the SVM classifier are difficult to interpret: Legallois et al. [33], working at the grammatical rather than acoustic level, already argued that discrete frequency-based descriptors support a stylistics of identification but resist characterisation. Knowing that Flaubert uses a given trigram more or less than his contemporaries (e.g. the 3-gram [pje], which occurs in words as diverse as *pied*, *expier*, *copiai*, and *pompier*) does not, in itself, tell us much about his prose, and risks encouraging the kind of overinterpretation that Delbouille [12] warned against, reading global distributions as evidence of conscious acoustic choices. What yields more interpretable results are targeted tests, such as the patterning of sequences, scores computed locally (e.g. for 1,000 tokens), or metrics computed at a macro level but on abstract categories such as consonants versus vowels. They not only separate Flaubert from his contemporaries but do so in a way that connects more directly to the acoustic experience of reading his prose. This closer connection, in turn, makes the question of intentionality more tractable.

A writer who revises by ear is more likely to optimise abstract features at a local level (density of consonants, regularity of stress placement, repetitions or mere “echoes”, etc.) than to target specific phonemes or sequences directly at a global level (the 3-gram [pje]). These are adjustments that can plausibly be heard and consciously made, unlike global distributional tendencies, which may reflect deeper stylistic automatisms of the kind Fried identifies in Flaubert’s relationship to his

own *gueuloir* [16]. More broadly, this suggests that computational stylistics should move beyond frequency-based representations towards phonetic-based or phonetic-derived models sensitive to order and context, an agenda that applies to verse no less than prose, and that would bring computational approaches closer to the kind of hermeneutic yield that acoustic analysis of literature has long struggled to produce [12].

Limitations. Several limitations should be noted. The temporal span covered varies considerably across authors, which may introduce diachronic stylistic drift and partly confound authorial distinctiveness with career-stage effects [6; 21; 45]. The phonetic transcription remains an approximation, particularly for optional liaisons and diachronically variable realisations. The experimental design remains deliberately constrained: rhythmic features are operationalised independently of higher-level syntactic structure. More fundamentally, the *gueuloir*’s effect is observed only indirectly: we identify textual patterns compatible with oral revision, but the causal mechanism cannot be demonstrated from published texts alone.

Acknowledgements

Many thanks to Isabelle Pitteloud for her valuable comments, as well as for the time she took to discuss and suggest insightful directions. Errors remain our own.

This work has received support (to author 2) under the Major Research Programme of PSL Research University “CultureLab” launched by PSL Research University and implemented by ANR with the reference ANR-10-IDEX-0001.

References

- [1] Alqurashi, Lama, Sharoff, Serge, Watson, Janet, and Blakesley, Jacob. “BERT-based Classical Arabic Poetry Authorship Attribution”. In: *Proceedings of the 31st International Conference on Computational Linguistics*, ed. by Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert. Abu Dhabi, UAE: Association for Computational Linguistics, 2025, pp. 6105–6119. URL: <https://aclanthology.org/2025.coling-main.409>.
- [2] ANRChapitres. “ANRChapitres / 2000 romans 19e 20e: Corpus Chapitres”. Version v1.0.0. Dec. 2022. DOI: 10.5281/zenodo.7446728.
- [3] ATILF. “Morphalou”. ORTOLANG (Open Resources and Tools for Language). 2023. URL: <https://hdl.handle.net/11403/morphalou/v3.1>.
- [4] Cafiero, Florian and Camps, Jean-Baptiste. “Why Molière Most Likely Did Write His Plays”. In: *Science Advances* 5, no. 11 (2019). DOI: 10.1126/sciadv.aax5489.
- [5] Cafiero, Florian, Camps, Jean-Baptiste, and Gabay, Simon. “Louise Labé: une créature de papier?” In: *Humanistica 2023*, ed. by Simon Gabay. Geneva, Switzerland: Association francophone des humanités numériques, 2023. URL: <https://hal.science/hal-04090284v1>.
- [6] Cafiero, Florian, Ing, Lucence, Gabay, Simon, and Clérice, Thibault. “‘I Am Too Old For This Style!’ A Stylometric Benchmark of Age Effect on Authorship Attribution”. In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Esch-sur-Alzette, Luxembourg: Anthology of Computers and the Humanities, 2025, pp. 1249–1261. DOI: 10.63744/By09x5ZX3yWX.
- [7] Camps, Jean-Baptiste and Cafiero, Florian. “SUPERvised STYLometry (SuperStyl)”. Version 1.0. Nov. 2024. DOI: 10.5281/zenodo.14069799.

- [8] Camps, Jean-Baptiste, Cafiero, Florian, Chaumet-Riffaud, Philippe, Conceicao, Damien, Godreau, Ulysse, Guidi, Émilie, Lionnet-Rollin, Alexandre, Moins, Théo, Nistor, Pierre-Alexandre, and Salvati, Benedetta. “Style in Eight Syllables: Metric Annotation and Stylometry of Chrétien de Troyes and Contemporaries”. In: *Digital Humanities Benelux 2025 Conference*. Amsterdam, The Netherlands, 2025, pp. 1–12. DOI: 10.5281/zenodo.15600182.
- [9] Catach, Nina. “The Automatic Phoneticization of the French Language”. In: *Computers and the Humanities* 20 (1986), pp. 159–166. DOI: 10.1007/BF02404455.
- [10] Clérice, Thibault. “Pie Extended”. Version 0.0.40. May 2022. DOI: 10.5281/zenodo.6534764.
- [11] Corbara, Silvia, Moreo, Alejandro, and Sebastiani, Fabrizio. “Syllabic Quantity Patterns as Rhythmic Features for Latin Authorship Attribution”. In: *Journal of the Association for Information Science and Technology* 74, no. 1 (2023), pp. 128–141. DOI: 10.1002/asi.24660.
- [12] Delbouille, Paul. *Poésie et sonorité: la critique contemporaine devant le pouvoir suggestif des sons*. Paris: Les Belles Lettres, 1961. URL: <https://hdl.handle.net/2268.1/12533>.
- [13] Deng, Weibing and Allahverdyan, Armen E. “Stochastic Model for Phonemes Uncovers an Author-Dependency of their Usage”. In: *PLOS ONE* 11, no. 4 (2016). DOI: 10.1371/journal.pone.0152561.
- [14] Falcone, Jean-Luc and Gabay, Simon. “Low Phonographicity, High Stakes. On the Value of Phonetic Transcription for Stylometry in French”. In: *Workshop on Computational Methods in the Humanities 2026*. Lausanne, Switzerland, 2026.
- [15] Fillaire, Bernard. *Alexandre Dumas et associés*. Paris: Bartillat, 2002.
- [16] Fried, Michael. *Flaubert’s “Gueuloir.” On “Madame Bovary” and “Salammbô”*. New Haven and London: Yale University Press, 2012.
- [17] Gabay, Simon. “Beyond Idiolectometry? On Racine’s Stylometric Signature”. In: *Proceedings of the Conference on Computational Humanities Research 2021*, ed. by Maud Ehrmann, Folger Karsdorp, Melvin Wevers, Tara Lee Andrews, Manuel Burghardt, Mike Kestemont, Enrique Manjavacas, Michael Piotrowski, and Joris van Zundert. Amsterdam, The Netherlands, 2021, pp. 359–376. URL: <https://hal.science/hal-03402994>.
- [18] Gabay, Simon. “Rien de Nouveau chez Rimbaud: sur l’attribution des “Illuminations””. In: *Humanités numériques* 11 (2025). DOI: 10.4000/1498n.
- [19] Gabay, Simon, Clérice, Thibault, Camps, Jean-Baptiste, Tanguy, Jean-Baptiste, and Gille-Levenson, Matthias. “Standardizing Linguistic Data: Method and Tools for Annotating (Pre-Orthographic) French”. In: *Proceedings of the 2nd International Digital Tools & Uses Congress (DTUC ’20)*, ed. by E. Reyes, G. Kembellec, F. Siala-Kallel, L. Sfaxi, M. Ghenima, and I. Saleh. Hammamet, Tunisia: Association for Computing Machinery, Oct. 2020, pp. 1–7. DOI: 10.1145/3423603.3423996.
- [20] Gabay, Simon and Falcone, Jean-Luc. “Megastyl”. Oct. 2025. DOI: 10.5281/zenodo.19461873.
- [21] Gómez-Adorno, Helena, Posadas-Durán, Juan-Pablo, Ríos-Toledo, Germán, Sidorov, Grigori, and Sierra, Gerardo. “Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts”. In: *Computación y Sistemas* 22, no. 1 (2018). DOI: 10.13053/cys-22-1-2882.
- [22] Gothot-Mersch, Claudine. *La Genèse de “Madame Bovary”*. Paris: José Corti, 1966.

- [23] Guinot, Jean-Benoît. “Gueuloir”. In: *Dictionnaire Flaubert*. Paris: CNRS Éditions, 2010, pp. 307–308. URL: <https://shs.cairn.info/dictionnaire-flaubert--9782271069283-page-307>.
- [24] Harris, Michael J., Gries, Stefan Th., and Miglio, Viola G. “Prosody and Its Application to Forensic Linguistics”. In: *Linguistic Evidence in Security, Law and Intelligence* 2, no. 2 (2014). DOI: 10.5195/lesli.2014.12.
- [25] Haug, Andreas. “Musikalische Lyrik im Mittelalter”. In: *Musikalische Lyrik*. Lilienthal: Laaber-Verlag, 2004, pp. 59–129.
- [26] Hollien, Harry. *The Acoustics of Crime: The New Science of Forensic Phonetics*. New York: Springer, 2013. DOI: 10.1007/978-1-4899-0673-1.
- [27] Juola, Patrick. “The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions”. In: *Digital Scholarship in the Humanities* 30, no. suppl. 1 (2015), pp. i100–i113. DOI: 10.1093/llc/fqv040.
- [28] Keselj, Vlado, Peng, Fuchun, Cercone, Nick, and Thomas, Calvin. “N-Gram-Based Author Profiles For Authorship Attribution”. In: *Proceedings of the Pacific Association for Computational Linguistics*. Halifax, Canada: Pacific Association for Computational Linguistics, 2003, pp. 255–264. URL: <http://www.cs.dal.ca/~vlado/papers/pacling03.pdf>.
- [29] Kestemont, Mike. “Function Words in Authorship Attribution. From Black Magic to Theory?” In: *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, ed. by Anna Feldman, Anna Kazantseva, and Stan Szpakowicz. Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, pp. 59–66. DOI: 10.3115/v1/W14-0908.
- [30] Kestemont, Mike, Daelemans, Walter, and Sandra, Dominiek. “Robust Rhymes? The Stability of Authorial Style in Medieval Narratives”. In: *Journal of Quantitative Linguistics* 19, no. 1 (2012), pp. 54–76. DOI: 10.1080/09296174.2012.638796.
- [31] Khomytska, Iryna and Teslyuk, Vasyl. “The Multifactor Method Applied for Authorship Attribution on the Phonological Level”. In: *Computational Linguistics and Intelligent Systems*, ed. by Vasyl Lytvyn, Victoria Vysotska, Thierry Hamon, Natalia Grabar, Natalia Sharonova, Olga Cherednichenko, and Olga Kanishcheva. Lviv, Ukraine, 2020, pp. 189–198. URL: <https://ceur-ws.org/Vol-2604/paper14.pdf>.
- [32] Lee, Hoyeon, Jang, Hyeeun, Kim, Jonghwan, and Kim, Jaemin. “A Two-Step Approach for Data-Efficient French Pronunciation Learning”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 19096–19103. DOI: 10.18653/v1/2024.emnlp-main.1064.
- [33] Legallois, Dominique, Charnois, Thierry, and Larjavaara, Meri. “The Balance Between Quantitative and Qualitative Literary Stylistics: How the Method of ‘Motifs’ Can Help”. In: *The Grammar of Genres and Styles*, ed. by Dominique Legallois, Thierry Charnois, and Meri Larjavaara. Berlin-Boston: Walter de Gruyter Mouton, 2018, pp. 164–193. DOI: 10.1515/9783110595864-008.
- [34] Léon, Pierre. “Nouveau regard sur la phonostylistique”. In: *La Linguistique* 45, no. 1 (2009), pp. 159–169. DOI: 10.3917/ling.451.0159.
- [35] Mann, H. B. and Whitney, D. R. “On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other”. In: *The Annals of Mathematical Statistics* 18, no. 1 (1947), pp. 50–60. URL: <http://www.jstor.org/stable/2236101>.

- [36] Mathieu, Cécile. “L’arbitraire saussurien: résistances et résolution”. In: *La linguistique* 54, no. 1 (2018), pp. 21–38. DOI: 10.3917/ling.541.0021.
- [37] Meschonnic, Henri. *Critique du rythme. Anthropologie historique du langage*. 1st edition: 1982. Lagrasse: Éditions Verdier, 2009.
- [38] Mousavishirazi, Seyed Jamal. “Structure sonore d’un texte littéraire”. In: *Corela - Cognition, représentation, langage HS-30* (2020). DOI: 10.4000/corela.10402.
- [39] Nagy, Ben, ŠeĽa, Artjoms, De Sisto, Mirella, and Plecháč, Petr. “Metronome: Tracing Variation in Poetic Meters via Local Sequence Alignment”. In: *Computational Humanities Research 1* (2025), e1. DOI: 10.1017/chr.2025.1.
- [40] Nagy, Benjamin. “Metre as a Stylometric Feature in Latin Hexameter Poetry”. In: *Digital Scholarship in the Humanities* 36, no. 4 (2021), pp. 999–1012. DOI: 10.1093/llc/fqaa043.
- [41] Plecháč, Petr and Birnbaum, David J. “Assessing the Reliability of Stress as a Feature of Authorship Attribution in Syllabic and Accentual Syllabic Verse”. In: *Quantitative Approaches to Versification*, ed. by Petr Plecháč, Barry P. Scherr, Tatyana Skulacheva, Helena Bermúdez-Sabel, and Robert Kolár. Prague: Institute of Czech Literature of the Czech Academy of Sciences, 2019, pp. 201–10.
- [42] Plecháč, Petr and Birnbaum, David J. “Rhythm-Based Authorship Recognition in Syllabic and Accentual-Syllabic Verse”. In: *Literatura: Teoría, Historia, Crítica* 25, no. 2 (2023), pp. 184–190. DOI: 10.15446/lthc.v25n2.108502.
- [43] Proust, Marcel. “À propos du «style» de Flaubert”. In: *La Nouvelle Revue Française* 76 (1920).
- [44] Richaudeau, François. “Gustave Flaubert ou l’écriture pour être gueulée”. In: *Communication & Langages*, no. 63 (1985), pp. 13–29. URL: <https://doi.org/10.3406/colan.1985.1657>.
- [45] Ríos-Toledo, Germán, Posadas-Durán, Juan Pablo Francisco, Sidorov, Grigori, and Castro-Sánchez, Noé Alejandro. “Detection of Changes in Literary Writing Style Using N-Grams as Style Markers and Supervised Machine Learning”. In: *PLOS ONE* 17, no. 7 (2022), e0267590. DOI: 10.1371/journal.pone.0267590.
- [46] Sapkota, Upendra, Bethard, Steven, Montes, Manuel, and Solorio, Thamar. “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*, ed. by Rada Mihalcea, Joyce Chai, and Anoop Sarkar. Denver, USA: Association for Computational Linguistics, 2015, pp. 93–102. DOI: 10.3115/v1/N15-1010.
- [47] Shannon, Claude E. “A Mathematical Theory of Communication”. In: *The Bell System Technical Journal* 27, no. 3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.
- [48] Skorinkin, Daniil and Orekhov, Boris. “Hacking Stylometry with Multiple Voices: Imaginary Writers Can Override Authorial Signal in Delta”. In: *Digital Scholarship in the Humanities* 38, no. 3 (2023), pp. 1247–1266. DOI: 10.1093/llc/fqad012.
- [49] Steuckardt, Agnès. “Les Dictionnaires du XVIIIe siècle et l’oreille”. In: *Le Jugement de l’oreille*, ed. by Agnès Steuckardt and Mathilde Thorel. Paris: Honoré Champion, 2017, pp. 211–227. URL: <https://hal.science/hal-01944763>.

- [50] Stock, Markus. “Triôs, triên, trisô : Klangspiele bei Wernher von Teufen und Gottfried von Neifen”. In: *Beiträge zur Geschichte der deutschen Sprache und Literatur* 138, no. 3 (2016), pp. 365–389. DOI: 10.1515/bgs1-2016-0029.
- [51] Wells, John C. “Computer-Coding the IPA: a Proposed Extension of SAMPA”. Revised version. London, 2000. URL: <https://www.phon.ucl.ac.uk/home/sampa/ipasam-x.pdf>.
- [52] Wilcoxon, Frank. “Individual Comparisons by Ranking Methods”. In: *Biometrics Bulletin* 1, no. 6 (1945), pp. 80–83. DOI: 10.2307/3001968.
- [53] Zola, Émile. *Les Romanciers naturalistes : Balzac, Stendhal, Gustave Flaubert, Edmond et Jules de Goncourt, Alphonse Daudet, les romanciers contemporains*. Paris: Charpentier, 1881. URL: <https://gallica.bnf.fr/ark:/12148/bpt6k215334g>.