

Low Phonographicity, High Stakes. On the Value of Phonetic Transcription for Stylometry in French

Jean-Luc Falcone

Université de Genève, Suisse
jean-luc.falcone@unige.ch

Simon Gabay

Université de Genève, Suisse
simon.gabay@unige.ch

Abstract

French orthography is a poor guide to the sound of a text: letters can carry several simultaneous values, and phonetic phenomena such as *liaison* are only partially, if at all, visible in writing. Character-based stylometric features computed on standard spelling therefore risk conflating genuine authorial choices with artifacts of a notation that writers do not fully control. We test whether phonetic transcription offers a more reliable basis for authorship attribution in French, comparing classifiers trained on orthographic text against classifiers trained on a phoneme-level transcription of the same corpus. As a case study, we use the novels of Gustave Flaubert, an author associated with the *gueuloir*, his practice of testing sentences by reading them aloud, set against three contemporaries (Aimard, Chevalier, Ponson du Terrail). Phonetic transcription consistently outperforms orthography for authorship classification, raising the macro F1-score from 0.96 to 0.99. This confirms that part of the stylometric signal in French is masked, rather than carried, by spelling, and argues for phonetic-based representations wherever a language’s orthography is markedly opaque.

1 Introduction

Letters in French can assume up to seven distinct values (basic, zero, positional, etc. cf. [Blanche-Benveniste and Chervel 1969](#)) or several simultaneous functions (phonological, morphological, sometimes purely distinctive, cf. [Honvault 1995](#)), which poses a major challenge for any computational analysis of the language. For example, despite an identical etymology (< lat. *FUNDUS*), orthography distinguishes *fonds* (“collection”) from *fond* (“bottom”) through an *s* that is neither a plural marker nor a verbal ending, serves no function beyond differentiating two homophones, and, like the final *d*, is not even pronounced. Only about 80% of French graphemes correspond to phono-

grams ([Catach, 1990](#)), and homographs are not always homophones (*films*, /fis/ for “son” but /fil/ for “threads”). French also features *liaison*: a normally silent final consonant is realised before a following word beginning with a vowel or mute *h* (*les amis* /le_z_ami/; *un grand homme* /œ̃ɡʁɑ̃_t_ɔm/), obligatory in some syntactic contexts, optional in others, prohibited elsewhere.

Under such conditions, what are we measuring when we count letters or character *n*-grams, as stylometry routinely does? By relying on orthography, we ground the analysis on data whose authority does not lie entirely with the author: writers control the meaning and the intended sound of their text, but the notation of both partly escapes them, and it is precisely on this notation that computation operates.

We propose instead to work on phonetic transcription, a notation that minimises this noise: it collapses spelling variants realised identically ([f] for <ph> and <f>), drops unpronounced letters, while retaining traces of some morphemes ([ɑ̃] for the adverbial suffix in *finale*ment). Because it represents an actual possible realisation of the text, the form it takes when read aloud, it should provide more robust stylometric data than spelling, at least for text recent enough for pronunciation to be reliably reconstructed ([Viémond, 2016](#)).

To test this empirically, we turn to a case where an ear-driven mode of composition has long been suspected: Gustave Flaubert (1821–1880) and his *gueuloir*, literally a “shouting room”, where he tested the musicality of his prose aloud ([Fried, 2012](#)). We compare classifiers trained on Flaubert’s novels and on those of three contemporaries, using orthographic and phonetic representations of the same texts, to ask whether moving from spelling to sound improves authorship attribution.

2 State of the art

Stylometry has explored many feature types – character n -grams (Keselj et al., 2003), function words (Kestemont, 2014), morpho-syntactic patterns (Gabay, 2021) – almost always computed on orthographic text. Acoustic properties remain marginal, mostly confined to metre in poetry (Plecháč and Birnbaum, 2023; Nagy et al., 2025), with no applications to prose. Deng and Allahverdyan (Deng and Allahverdyan, 2016) showed phoneme rank-frequency curves carry an author-specific signature in English, yet very few attribution studies exploit phonetics (Khomytska and Teslyuk, 2020), partly because converting orthographic French into phonetic transcription at scale remains technically demanding (Catach, 1986; Lee et al., 2024). Existing studies of the *gueuloir* itself remain qualitative and small-scale (Fried, 2012; Richaudeau, 1985).

3 Data

Our corpus (Table 3) draws on novels distributed by ANRChapitres (2022), comparing Flaubert to three novelists active in the same period and born within a decade of him: Gustave Aimard, Henri-Émile Chevalier, and Pierre-Alexis de Ponson du Terrail. Authors known to have relied on ghostwriters were excluded. For Flaubert, the corpus includes the posthumous first version of *L’Éducation sentimentale* and the unfinished *Bouvard et Pécuchet*.

4 Methods

All texts are transcribed into a one-symbol-per-phoneme alphabet, a modification of X-SAMPA (Wells, 2000), using *Morphalou 3.1* (ATILF, 2023), a lexicon of 159,271 lemmas with their phonetic transcriptions, complemented by an *ad hoc* dictionary and *Pie extended* (Clérice, 2022) for tokenisation and part-of-speech tagging. Each token is looked up and disambiguated by part of speech; basic mandatory *liaisons* are then added. Proper nouns are removed to avoid interference with the authorial signal (Table 1 gives an example).

A supervised classification is conducted with an SVM classifier implemented in *MegaStyl* (Gabay and Falcone, 2025), an extension of *SuperStyl* (Camps and Cafiero, 2024) for phonetic data. We systematically vary feature representations (character trigrams, function words, affixes, and combinations), sampling strategies, dimensionality reduction, and validation protocol, and we test the effect

Orthographic	<i>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.</i>
Phonetic	nuzetjôaletyd@kâl@pRoviz9RâtRasHiv idûnuvoabi jeâbuRZwaedûgaRsöd@klas@ kipORtEûgRâpypitR@l

Table 1: Example of a phonetic transcription (from *Madame Bovary*).

of sample length, with orthographic samples of 500–3,000 tokens and phonetic samples of 5,000–10,000 phonemes, always comparing phonetic transcriptions against the orthographic baseline under matched conditions.

5 Results

Table 2 shows that phonetic transcription provides a consistent boost over orthography for authorship classification. Orthographic transcriptions already perform well, but phoneme-based classification outperforms them by roughly 3 percentage points, reaching a macro F1-score of 0.99 against 0.96 for the orthographic baseline. This improvement is best explained by the sources of noise identified above: silent letters, several graphemes mapping to a single phoneme, and phenomena partly invisible in writing, such as *liaisons*. Working directly on phonemes removes these irregularities and lets the underlying stylometric signal emerge more cleanly.

6 Discussion

The central finding is methodological: phonetic representation is more discriminative than orthographic representation for authorship attribution in French, confirming that character-level features computed on spelling partly measure orthographic convention rather than authorial choice. Grounding the analysis on phonetic transcription instead brings the data closer to a level over which the author has fuller control.

The benefit likely scales with the opacity of a language’s orthographic system (Seymour et al., 2003): shallow orthographies, such as Italian or German, may show limited gains, while opaque ones, such as French or English, stand to benefit substantially. For such languages, phonetic transcription is not a mere reformatting: it grants access to a layer of stylistic organisation otherwise measurable only indirectly. Flaubert illustrates the point rather than exhausting it: an author whose relationship to sound is already well documented

(a) Orthographic transcription, 2,500 words/sample, downsampling, linear kernel.

Author	Precision	Recall	F1-score	Support
Gustave Aimard	0.92	0.96	0.94	380
Emile H. Chevalier	0.88	0.89	0.89	361
Gustave Flaubert	0.94	0.93	0.94	372
Pierre Ponson du Terrail	0.94	0.89	0.91	366
accuracy			0.92	1479
macro avg	0.92	0.92	0.92	1479
weighted avg	0.92	0.92	0.92	1479

(b) Phonetic transcription, 8,500 characters/sample, downsampling, linear kernel.

	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

Table 2: Comparison of the two SVM, both with class weights, downsampling, k -fold=10.

offers a favourable test, but the underlying argument concerns French orthography in general, not this author in particular.

7 Conclusion

Our results show that a phonetic transcription pipeline, applied at corpus scale, substantially improves authorship classification relative to orthographic text. This supports a broader claim: for orthographically opaque languages like French, digital stylistics has something to gain by working on phonetic or phonetic-derived representations, since spelling both adds elements without phonemic realisation and conceals phenomena such as *liaison*. This opens a path for revisiting authorship and style questions in French prose through the lens of sound rather than script.

Limitations. The temporal span covered varies across authors, which may introduce diachronic drift alongside authorial distinctiveness (Gómez-Adorno et al., 2018; Ríos-Toledo et al., 2022; Cafiero et al., 2025). The phonetic transcription remains an approximation, particularly for optional liaisons and diachronically variable realisations, and the choice of Flaubert as illustration means the gain observed here should still be confirmed on authors with no particular reputation for ear-driven composition.

References

ANRChapitres. 2022. [Anrchapitres / 2000 romans 19e 20e: Corpus chapitres](#).

ATILF. 2023. [Morphalou](#). ORTOLANG (Open Resources and Tools for Language).

Claire Blanche-Benveniste and André Chervel. 1969. *L’Orthographe*. Maspero, Paris.

Florian Cafiero, Lucence Ing, Simon Gabay, and Thibault Clérice. 2025. [‘I Am Too Old For This](#)

[Style!’ A Stylometric Benchmark of Age Effect on Authorship Attribution](#). In Taylor Arnold, Margherita Fantoli, and Ruben Ros, editors, *Computational Humanities Research 2025*, volume 3, pages 1249–1261. Anthology of Computers and the Humanities, Esch-sur-Alzette, Luxembourg.

Jean-Baptiste Camps and Florian Cafiero. 2024. [SUPERvised STYLometry \(SuperStyl\)](#).

Nina Catach. 1986. [The Automatic Phoneticization of the French Language](#). *Computers and the Humanities*, 20:159–166.

Nina Catach. 1990. *L’orthographe française: traité théorique et pratique avec des travaux d’application et leurs corrigés*. Nathan, Paris.

Thibault Clérice. 2022. [Pie Extended](#).

Weibing Deng and Armen E. Allahverdyan. 2016. [Stochastic Model for Phonemes Uncovers an Author-Dependency of their Usage](#). *PLOS ONE*, 11(4).

Michael Fried. 2012. *Flaubert’s “Gueuloir.” On “Madame Bovary” and “Salammbô”*. Yale University Press, New Haven and London.

Simon Gabay. 2021. [Beyond Idiolectometry? On Racine’s Stylometric Signature](#). In *Proceedings of the Conference on Computational Humanities Research 2021*, pages 359–376, Amsterdam, The Netherlands.

Simon Gabay and Jean-Luc Falcone. 2025. [Megastyl](#).

Helena Gómez-Adorno, Juan-Pablo Posadas-Durán, Germán Ríos-Toledo, Grigori Sidorov, and Gerardo Sierra. 2018. [Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts](#). *Computación y Sistemas*, 22(1).

Renée Honvault. 1995. [Statut linguistique et gestion de la variation graphique](#). *Langue française*, 108:10–17.

Vlado Keselj, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. [N-Gram-Based Author Profiles For Authorship Attribution](#). In *Proceedings of the Pacific Association for Computational Linguistics*, pages 255–264, Halifax, Canada. Pacific Association for Computational Linguistics.

- Mike Kestemont. 2014. [Function Words in Authorship Attribution. From Black Magic to Theory?](#) In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 59–66, Gothenburg, Sweden. Association for Computational Linguistics.
- Iryna Khomytska and Vasyl Teslyuk. 2020. [The Multi-factor Method Applied for Authorship Attribution on the Phonological Level.](#) In *Computational Linguistics and Intelligent Systems*, pages 189–198, Lviv, Ukraine.
- Hoyeon Lee, Hyeun Jang, Jonghwan Kim, and Jaemin Kim. 2024. [A Two-Step Approach for Data-Efficient French Pronunciation Learning.](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19096–19103, Miami, Florida, USA. Association for Computational Linguistics.
- Ben Nagy, Artjoms Šēla, Mirella De Sisto, and Petr Plecháč. 2025. [Metronome: Tracing Variation in Poetic Meters via Local Sequence Alignment.](#) *Computational Humanities Research*, 1:e1.
- Petr Plecháč and David J. Birnbaum. 2023. [Rhythm-Based Authorship Recognition in Syllabic and Accentual-Syllabic Verse.](#) *Literatura: Teoría, Historia, Crítica*, 25(2):184–190.
- François Richaudeau. 1985. [Gustave Flaubert ou l’écriture pour être gueulée.](#) *Communication & Languages*, pages 13–29.
- Germán Ríos-Toledo, Juan Pablo Francisco Posadas-Durán, Grigori Sidorov, and Noé Alejandro Castro-Sánchez. 2022. [Detection of Changes in Literary Writing Style Using N-Grams as Style Markers and Supervised Machine Learning.](#) *PLOS ONE*, 17(7):e0267590.
- Philip H. K. Seymour, Mikko Aro, Jane M. Erskine, and collaboration with COST Action A8 network. 2003. [Foundation Literacy Acquisition in European Orthographies.](#) *British Journal of Psychology*, 94(2):143–174.
- Marc Viémon. 2016. [L’Apprentissage de la prononciation française par les espagnols aux XVIe, XVIIe et XVIIIe siècles.](#) Ph.D. thesis, Universidad de Sevilla, Sevilla.
- John C. Wells. 2000. [Computer-Coding the IPA: a Proposed Extension of SAMPA.](#) Revised version.

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