More Sound, More Soundness? Improving Authorship Attribution with Phonemes

Simon Gabay[1], Florian Cafiero[2], and Jean-Luc Falcone[1]

[1] Université de Genève, Geneva, Switzerland
[2] École nationale des chartes | Paris Sciences Lettres, Paris, France

Abstract

This paper assesses whether turning written French poetry into a speech-oriented representation can improve the performance of authorship attribution methods. To this end, we develop a phonetic transcription system to automatically convert poems from six authors – including the disputed *Illuminations* of Rimbaud – into phonetic transcriptions, and adapt existing tools to ingest and process phonetic data. The output of this grapheme-to-phoneme task is then enriched with minimal prosodic cues, namely the creation of synthetic tokens based on punctuation and the addition of basic French *liaisons*. Using the same trigram features and classifier across all representations, we observe that moving from orthographic to phonetic transcriptions with a modest prosodic enrichment raises the F-score from 0.89 to 0.95, while reducing inter-author confusion. These results suggest that even lightweight speech-based features, produced with reproducible rules and open tools, can meaningfully enhance stylometric analysis of French verse and warrant further study for contested texts.

Keywords: Stylometry, Phonetics, Stylistics, Authorship attribution

1 Introduction

Saussurean linguistics established the arbitrariness of the linguistic sign as a foundational principle: there is no inherent link between a word's meaning and its visual and acoustic realization [29]. Yet while researchers often emphasize the primacy of the signified, which carries semantic content, the signifier also holds considerable importance: since at least the 16th century, the *jugement de l'oreille* ("judgment of the ear") has, for instance, played a significant role in shaping linguistic choices in French [35]. This importance of the signifier is especially pronounced in literature, and even more so in poetry, where one encounters eye rhymes and numerous stylistic figures based on sound effects such as alliteration. Its relevance increases further in the computational treatment of texts, for it is primarily this material aspect of language to which the computer has access.

The fact that letters can assume up to seven distinct values (basic value, zero value, position value, etc.) [2] or several simultaneous functions (phonological, morphological, sometimes distinctive) [19] therefore poses a major challenge for any computational analysis of the language. For example, despite their identical Latin etymology (lat. < FUNDUS), French artificially distinguishes *fonds* (eng. "fonds") from *fond* (eng. "bottom") through an *s* that is neither a plural marker (*enfant<u>s</u>*, eng. "children") nor a verbal morpheme (*tu mange<u>s</u>*, eng. "you eat"): it serves no purpose other than to differentiate two homophones. Moreover, this *s* (like the final *d*) is not even pronounced: only about 80% of French graphemes correspond to phonograms [6]. Under such conditions, what exactly are we measuring when we count letters or character *n*-grams? By relying on the orthographic form of a text, we ground the detection of an idiolect (and potentially a

Simon Gabay, Florian Cafiero, and Jean-Luc Falcone. "More Sound, More Soundness? Improving Authorship Attribution with Phonemes." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 1121–1131. https://doi.org/10.63744/JsYBzks5UCwg.

style) on data whose authority does not lie entirely with the author. While the writer is responsible for the meaning and the intended sound of the text, the notation of both dimensions partly escapes his control – and yet it is precisely on this notation that computation operates.

To address this issue, we propose using a notation that minimizes ambiguity and reduces noise from orthographic artifacts: phonetic transcription. It reduces variation (\f\ for both ‹ph› and ‹f›, deletion of the unpronounced ‹d› in *fond*…), retains traces of certain morphemes (\mã\ for the adverbial suffix as in *finalement*), and thus appears to be a promising approach, at least for more recent texts, since older language stages have a phonic realization that is more difficult to reconstruct [37]. Although such transcription inevitably removes some visual information from the text, it should yield more robust data and enable us to address questions that have recently emerged in literary studies.

In recent years, scholars of medieval song have sought to "open a path to the sound of language" by integrating "the imaginings and modeling of sound into the analysis in a methodically controlled manner" [36]. This integration rests on a "three-tiered model of description, that distinguishes between the meaning layer (*Bedeutungsschicht*), the linguistic sound layer (*sprachlicher Klangschicht*), and the melodic sound layer (*melodischer Klangschicht*)" [17]. Our working hypothesis is that the study of the linguistic sound layer, whose significance deserves renewed attention, can be operationalized through phonetic transcription and may lead to improved stylometric performance.

To explore this hypothesis, we turn to a concrete case study: Rimbaud's *Illuminations* [33], whose authorship has been the subject of renewed debate following E. Breuil's proposal [3]. We replicate a recent stylometric study on this corpus [14] to compare the effectiveness of analyses conducted on orthographic texts with those based on automatically produced phonetic transcriptions. For this project, we design a new lightweight system for generating phonetic transcriptions.

2 Case Study: Rimbaud's *Illuminations*

Breuil's theory[1] is simple: the author of the *Illuminations* would be Germain Nouveau, and Rimbaud would have only served as the copyist of the poems. To persuade his readers, Breuil conducts a thorough historical-literary investigation, drawing on biographical facts (Rimbaud, often short of money, was a copyist to earn a living), philological evidence (the manuscript, although in Rimbaud's handwriting, contains copying errors), and literary traits (for example, the association of summer with unhappiness, typical of Nouveau's poetry, rather than with happiness as in Rimbaud's).

Clearly invalidating such a seductive demonstration is particularly problematic for several reasons, notably three. First, the literary qualities of Germain Nouveau, which are in no way in doubt, do not allow us to exclude *a priori* that he might be the author of these poems: this is not a question of attributing the *Illuminations* to an unknown author. Second, although questionable, Breuil's observations make sense: for example, *baou* is indeed an Occitan word that more plausibly belongs to the lexicon of the Varois Nouveau than to that of the Ardennais Rimbaud. Third, the recurrent association between the two poets by readers as discerning as Aragon and Breton implies a certain poetic proximity between our two *poètes maudits*. One could nevertheless object that proximity is not identity, that Rimbaud could very well have borrowed a word heard from a friend he often met, and that Nouveau's talent in no way proves that he could be the author of the poems in question.

Unfortunately, E. Breuil did not conduct any stylometric analysis to substantiate his hypothesis, which would have been useful given that a recent stylometric investigation on this case appears to refute his theory [14]. The following study aims to revisit this issue using alternative features in

[1] This idea is not entirely new: Jacques Lovichi [28] attempted to demonstrate that Nouveau contributed to the *Illuminations*.

order to assess the robustness of this first computational approach.

3 State of the art

Given the computational nature of stylometry, much research has focused on algorithmic aspects such as distance functions and clustering methods, etc. [12; 31]. However, the nature of the data being analysed is equally important. Research on stylometric features has therefore received considerable attention, resulting in a variety of more or less effective solutions. Proposed features include function words [23], *n*-grams [22], affixes [34], syntactic patterns (also called *motifs*) [27], rhymes [24], a combination of lexicological and morpho-syntactic data [4; 13; 21], among others.

Among stylometric features, phonemes have drawn special interest because they sit below the word level and are less tied to topic. Deng et al. [9] model how often phonemes occur by examining the rank–frequency curve. They show that this curve is well captured by a Dirichlet model. In their setup, the model has a single *shape parameter* β that controls how even or spiky the distribution is: larger β implies more uniform phoneme usage, whereas smaller β implies that a few phonemes dominate. Fitting β to 48 English novels, they report two patterns: (i) the estimated β values cluster by author (books by the same writer tend to have similar β); and (ii) when comparing entire phoneme profiles (full vectors of phoneme frequencies), within-author distances are consistently smaller than between-author distances—even after removing all words shared across the texts, indicating that the effect is not driven by overlapping vocabulary or genre. Works by Iryna Khomytska and colleagues [25] have built upon their "multifactor method" of authorship attribution to use it with phonemes.

Adjacent disciplines have also informed research on phonetic features. Because they often work with audio recordings, forensic linguists and phoneticians [18] have exploited phonetic-prosodic clues, but primarily for author profiling rather than direct authorship attribution. For instance, Harris et al. [16] combine vowel-duration metrics, pitch movement and intensity patterns in mixed-effects and random-forest models to distinguish bilingual from monolingual Spanish speakers and discriminate English, Spanish, and Portuguese with 94 % accuracy.

While the use of phonemes in computational stylistics is already uncommon, attempts to engage with the sound of language are even rarer within the Humanities, except perhaps from a pedagogical perspective [30]. We have already mentioned the work of Haug [17] and Stock [36], as well as our own recent stylometric study on the *Illuminations* [14], which highlights the importance of the phonic dimension of character trigrams in Rimbaud's poetry. This scarcity of studies is likely due both to the difficulty of deriving a clear hermeneutic yield from phonetic analyses of language and to the technical challenges of performing such analyses on large corpora without computational tools.

Since no author writes using the phonetic alphabet, orthographic texts must first be converted into phonetic transcriptions. This task has long been the subject of numerous studies [10; 11]. Although text-to-speech synthesis poses significant challenges for machines, the transcription process itself is relatively straightforward, as a limited set of rules is generally sufficient [7]. Recently, neural network-based approaches attempting to combine grapheme-to-phoneme and lexical post-processing (particularly for *liaisons*) have demonstrated their effectiveness; however, the scarcity of training data still prevents the development of fully operational systems [26].

4 Data

First, we compile a training corpus comprising texts from the two putative authors, Rimbaud and Nouveau, along with several additional writers serving as distractors. We select four other prominent poets active during the same period (the latter third of the 19th century) and in the same place (Paris), since both time and location may carry linguistic traces detectable in statistical analyses.

Two of them–Paul Verlaine (1844–1896) and Charles Cros (1842–1888)–are poetically close to Rimbaud and Nouveau and belong to the circle of the *poètes maudits*. The other two–François Coppée (1842–1908) and Léon Dierx (1838–1912)–stand somewhat further apart stylistically. Most of the poems date from the 1860s and 1870s and are drawn from the following collections:

- Arthur Rimbaud: *Les Premiers vers* (circa 1870-1871) as published by P. Berrichon in 1912, some of the early poems rejected as an appendix by Berrichon but retained in the 1895 edition of *Poésies*, and finally *Les Vers nouveaux*;
- Germain Nouveau: *Les Valentines et autres vers* (1922 edition, but whose proofs date from the 1880s) and the first part of *Les Poésies d'Humilis et vers inédits* (1924), i.e. only Les *Poésies d'Humilis* (probably written before 1880);
- Paul Verlaine: *Les Poèmes saturniens* (1866), *Les Fêtes galantes* (1891 edition, dating from 1869), *La Bonne Chanson* (1891 edition, dated 1870), and *Romances sans paroles* (1891 edition, dated 1874);
- François Coppée: *Poèmes divers* (1885 edition, dated 1869) and *Poèmes modernes* (1885 edition, dated 1869);
- Léon Dierx: *Poèmes et poésies* (1861);
- Charles Cros: *Le Coffret de Santal* (1879).

5 Methods

5.1 Features

To recreate the "the linguistic sound layer", all texts are transcribed into a phonetic alphabet using a one-to-one correspondence between symbols and phonemes, enabling them to be processed as ordinary characters. We use a modification of X-SAMPA, the extension of the Speech Assessment Methods Phonetic Alphabet (SAMPA) notation [38]. SAMPA provides a keyboard-compatible encoding for the full set of IPA symbols covered by the 1993 IPA Chart, including diacritics and tone marks, and was proposed as a standard means of transmitting IPA-transcribed data. In our system, symbols originally encoded as two characters (\o~\) were replaced with single-character variants (\õ\).

We designed a novel phonetic transcription pipeline leveraging data from *Morphalou 3.1* data [1], a comprehensive French lexicon containing the inflected forms of 159 271 lemmas together with their phonetic transcriptions. To ensure efficient performance, the complete lexicon was loaded into an embedded SQLite database, enabling rapid queries. The resource was supplemented with an *ad hoc* dictionary, to account for a few missing entries, such as the poetic spelling *encor* (*encore*, en. "again", "still"), and to exclude irrelevant tokens such as Roman numerals (*iii*, *VI*, etc.). Before phonetic transcription itself, the texts are processed with *Pie extended* [8] and models specialized for modern French [15], which ensures accurate tokenization and part-of-speech annotation. At this stage, all proper nouns, may they be toponyms (*Angoulême*, *Rome*…) or anthroponyms (*Paul*…), are removed.

The texts may undergo three types of processing, which can be applied selectively or in combination to produce five distinct versions, as detailed in Table 1.

Phonetic transcription. This grapheme-to-phoneme task is applied independently to each token to generate a phonetic transcription. The following protocol is followed while performing this task:

1. each token is first checked against the manual dictionary;

2. if not found, a lookup is performed in Morphalou to retrieve all possible transcriptions:
    - if more than one possible transcription is available, the part-of-speech tag is used to select the most plausible form – for instance the verb *est* (\E\, en. “is”) vs. the noun *est* (\Est\, en. “east”);
    - if no match is found, the word is ignored but recorded in a log file for subsequent inspection.

Synthetic tokens. Simili-prosodic units are created by concatenating the phonetic transcriptions of successive tokens into a single synthetic token, except at verse boundaries or where punctuation occurs. A blank space is then inserted as an artificial proxy for prosodic pauses, intended to approximate the natural grouping of words in speech.

French *liaisons*. We incorporate basic French *liaisons* (linking consonants), applied depending on the scenario either at the end of the first token in a pair (e.g., *Les Illuminations* → \lez ilyminasjÕ\ with \z\ representing the *liaison*) or directly within synthetic tokens (e.g., *Les Illuminations* → \lezilyminasjÕ\). The system of French *liaisons* is notoriously complex, even for native speakers, as it varies considerably across periods, regions, and registers. Consequently, although most *liaisons* occur in poetry, we choose to implement only the so-called “mandatory *liaisons*” to avoid introducing prosodically or stylistically implausible forms.

A *liaison* is introduced between any pair of words that meets the following conditions:

- the first word ends with a consonant letter, but a vowel phoneme;
- the second word begins with a vowel phoneme;
- there is no punctuation between the words, and they belong to the same verse;
- at least one of the following condition is met:
    - the first word is a determiner, a preposition of a pronoun;
    - the second word is the postpositive pronoun *-y* or *-en*;
    - both words form a pair noun-adjective or adjective-noun pair;
    - the first word is an adverb and the second is a preposition;
- the second word does not start with an aspirated *h*[2].

Version	Text
Orthographic	aux abattoirs, dans les cirques,
Phonetic	o abatwaR dã le siRk@
Phonetic-synthetic	oabatwaR dãlesiRk@
Phonetic with *liaisons*	oz abatwaR dã le siRk@
Phonetic-synthetic, with *liaisons*	ozabatwaR dãlesiRk@

Table 1: Different processings of short extract from the poem *Après le déluge* in the *Illuminations*.

[2] In French, the initial *h* is aspirated if the word has a non-Latin or non-Greek etymology, such as those of Germanic (*harangue*) or Arabic (*hasard*) origin. A comprehensive list of such words was used as a lookup table after the lemmatization with Morphalou 3.1.

5.2 Classification

A supervised classification analysis is performed using a Support Vector Machine (SVM) classifier. To this end, we employ MegaStyl, an enhanced version of SuperStyl [5], specifically designed for the processing of phonetic data. Because the X-SAMPA phonetic alphabet incorporates symbols that are commonly interpreted as punctuation marks (for instance, "@" representing the *schwa*), we modify both the normalisation and tokenisation components of SuperStyl's original module. In our implementation, the normalisation step can be safely omitted when the input data have already been standardised–as is the case with phonetic transcription–while tokenisation operates straightforwardly on whitespace-delimited units.

While certain aspects of the training configuration remain fixed (namely, Z-score and L2 normalisation), we evaluate the influence of several additional parameters. Specifically, we investigate various sampling strategies (including over- and undersampling, as well as Tomek links), the use of penalised versus unpenalised models, dimensionality reduction through PCA, both linear and non-linear kernels, and two validation schemes: standard k-fold cross-validation (k = 10) and group-based cross-validation. Furthermore, we compare the four aforementioned feature types (original or synthetic tokens, with or without *liaisons*, etc.).

We also experiment with samples of varying lengths. Since synthetic tokens are considerably longer than the average word length in French, words cannot serve as a reliable sampling unit; we therefore operate at the character level. We generate samples ranging from 5 000 to 10 000 characters in length, with increments of 250. This functionality has also been integrated into MegaStyl, our modified version of SuperStyl.

6 Results

Empirical evaluation indicates that the best-performing configurations consistently involve a penalised model combined with downsampling (Tab. 2). The use of non-linear kernels occasionally provides further improvements, and dimensionality reduction via PCA can also enhance performance. Sample sizes vary considerably across configurations, ranging from 5,000 to 9,000 characters. While phonetic transcription does not improve upon the orthographic baseline (89%), the inclusion of prosodic information consistently enhances performance, yielding gains of 1 to 6 percentage points.

Among the various configurations, the use of synthetic tokens improves the SVM accuracy, particularly when combined with *liaisons* (F-score: 94%). Our results indicate that *liaisons* contribute the largest performance gain and, notably, achieve even higher accuracy when applied without synthetic tokens (95%). Among all the writers, Dierx and Nouveau are consistently the best recognised, which is noteworthy given that Nouveau is one of the two putative authors of the *Illuminations*. In contrast, Rimbaud's poetry consistently proves the most difficult to identify, regardless of the experimental setup.

We employ the optimal configuration to perform a rolling analysis with a step size of 100 characters, resulting in a total of 271 overlapping text segments. None of these segments is attributed to Nouveau, who can therefore be excluded as a potential author of the *Illuminations*. However, only 28% of the segments are attributed to Rimbaud, while the remaining 72% are assigned to Verlaine.

7 Discussion

The Rimbaud case. The fact that a large majority of the *Illuminations* is attributed to Verlaine is unexpected, though not entirely surprising, as it corroborates—albeit with much greater strength—previous findings obtained through traditional stylometric methods [14]. A plausible explanation for this attribution is not that Verlaine authored the *Illuminations*, but rather that several of Rimbaud's poems used to train the SVM were copied by Verlaine [32]. This is, for instance, the case

	Precision	Recall	F-score	Support
Coppée	0.83	0.56	0.67	9
Cros	0.83	1.00	0.91	10
Dierx	1.00	1.00	1.00	10
Nouveau	1.00	1.00	1.00	9
Rimbaud	0.78	0.78	0.78	9
Verlaine	0.89	1.00	0.94	8
Accuracy			0.89	55
Macro avg	0.89	0.89	0.88	55
weighted avg	0.89	0.89	0.88	55

(a) Results of the SVM on the orthographic version of the text, 7 750-characters sample, k=10, penalised model, downsampling and a linear kernel.

	Precision	Recall	F-score	Support
Coppée	1.00	0.44	0.62	9
Cros	0.91	1.00	0.95	10
Dierx	1.00	1.00	1.00	10
Nouveau	1.00	1.00	1.00	10
Rimbaud	0.67	0.89	0.76	9
Verlaine	0.89	1.00	0.94	8
Accuracy			0.89	56
Macro avg	0.91	0.89	0.88	56
weighted avg	0.91	0.89	0.88	56

(b) Results of the SVM on the phonetic version of the text, 6 000-characters sample, k=10, penalised model, downsampling and a sigmoid kernel.

	Precision	Recall	F-score	Support
Coppée	1.00	0.67	0.80	9
Cros	1.00	1.00	1.00	10
Dierx	1.00	1.00	1.00	10
Nouveau	1.00	1.00	1.00	9
Rimbaud	0.73	0.89	0.80	9
Verlaine	0.75	0.86	0.80	7
Accuracy			0.91	54
Macro avg	0.91	0.90	0.90	54
weighted avg	0.92	0.91	0.91	54

(c) Results of the SVM on the phonetic version of the text with synthetic tokens, 5 000-characters sample, k=10, penalised model, downsampling and a linear kernel.

	Precision	Recall	F-score	Support
Coppée	0.83	1.00	0.91	5
Cros	0.86	1.00	0.92	6
Dierx	1.00	1.00	1.00	6
Nouveau	1.00	1.00	1.00	5
Rimbaud	1.00	0.60	0.75	5
Verlaine	1.00	1.00	1.00	5
Accuracy			0.94	32
Macro avg	0.95	0.93	0.93	32
weighted avg	0.95	0.94	0.93	32

(d) Results of the SVM on the phonetic version of the text with synthetic tokens and *liaisons*, 6 000-characters sample, k=10, penalised model, PCA reduction, downsampling and a linear kernel.

	Precision	Recall	F-score	Support
Coppée	0.85	1.00	0.92	11
Cros	1.00	1.00	1.00	7
Dierx	1.00	1.00	1.00	7
Nouveau	1.00	1.00	1.00	6
Rimbaud	1.00	0.80	0.89	5
Verlaine	1.00	0.83	0.91	6
Accuracy			0.95	42
Macro avg	0.97	0.94	0.95	42
weighted avg	0.96	0.95	0.95	42

(e) Results of the SVM on the phonetic version of the text with *liaisons*, 9 000-characters sample, k=10, penalised model, downsampling and a linear kernel.

Table 2: Best-performing SVM results for each feature set. Details of all parameters for each experiment are given in the table caption.

with *Les Effarés*, for which our base edition uses the version copied by Verlaine[3]. As a result, these texts may carry traces of his authorial signal [20, p. 109], which could bias the model's attribution.

The phonetic case. This study set out to determine whether bringing the text closer to the phonetic signal—through automatic phonetic transcription and minimal prosodic enrichment—rather than using its orthographic form could enhance stylometric attribution in late-nineteenth-century French poetry. Three main findings emerge.

First, replacing the orthographic signal with an enriched phonetic representation systematically improves the decision margin of the SVM classifier. Classification accuracy rises from 0.89 to 0.95 under comparable conditions. This finding suggests that a substantial portion of the confusion observed in orthodox stylometric analysis originates from orthographic artefacts that obscure the genuine distribution of linguistic units under an author's control.

Second, grouping consecutive phoneme strings into pseudo-prosodic units and, more importantly, incorporating the most frequent liaisons yielded a significant performance gain across all poets—including Rimbaud, who appears to be more difficult to predict. This overall improvement indicates that rhythm-sensitive features such as token concatenation and liaison consonants capture higher-level phonotactic regularities. However, the fact that synthetic tokens perform less well than liaison-based representations suggests that excessive concatenation may reduce classification accuracy, possibly due to confounding factors such as reduced punctuation density or, when relevant, verse length.

Third, the phonetic pipeline markedly reduced between-author confusion without inflating within-author variance. This result aligns with previous evidence that phoneme-level representations capture author-specific preferences that persist even after lexical pruning.

Limitations. This study presents several methodological limitations. First, out-of-vocabulary tokens—most notably neologisms and archaisms typical of *fin-de-siècle* symbolism—were excluded from processing, which may have introduced biases in the delineation of class boundaries. Second, *liaison* modeling was restricted to so-called "mandatory" cases, while other *liaisons* were disregarded, although such phenomena may contribute to author-specific phonostylistic patterns. The results for the *Illuminations* should therefore be interpreted with great caution, especially given the surprising strength of Verlaine's signal detected in Rimbaud's text, which must absolutely be confirmed.

Despite these limitations, the incremental yet consistent gains achieved through the use of phonetic features with prosodic enrichment suggest that addressing the "linguistic sound layer" [17] represents a genuinely promising avenue not only for literary studies, but also for future stylometric research.

Code and data

The code and data supporting this study are openly accessible at: `https://gitlab.unige.ch/dh/fonetik` and `https://gitlab.unige.ch/dh/megastyl`.

References

[1] ATILF. "Morphalou". ORTOLANG (Open Resources and TOols for LANGuage). 2023. URL: `https://hdl.handle.net/11403/morphalou/v3.1`.

[3] In our version *grogne* (eng. "growl") rather than "chante" (eng. "sing"), "Et que leur chemise tremblote" (eng. "And may their shirt tremble") rather than *Et que leur lange blanc tremblote* (eng. "And may their white swaddling cloth tremble"), etc.

[2] Blanche-Benveniste, Claire and Chervel, André. *L'Orthographe*. Paris: Maspero, 1969.

[3] Breuil, Eddie. *Du Nouveau chez Rimbaud*. Paris: Honoré Champion, 2014.

[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] Camps, Jean-Baptiste and Cafiero, Florian. "SUPERvised STYLometry (SuperStyl)". Version 1.0. Nov. 2024. DOI: 10.5281/zenodo.14069799.

[6] Catach, Nina. *L'orthographe française: traité théorique et pratique avec des travaux d'application et leurs corrigés*. Paris: Nathan, 1990. URL: https://archive.org/details/lorthographefran0000cata.

[7] Catach, Nina. "The Automatic Phoneticization of the French Language". In: *Computers and the Humanities 20* (1986), pp. 159–166. DOI: 10.1007/BF02404455.

[8] Clérice, Thibault. "Pie Extended". Version 0.0.40. May 2022. DOI: 10.5281/zenodo.6534764.

[9] 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.

[10] Derouault, A.-M. and Merialdo, B. "Probabilistic Grammar for Phonetic to French Transcription". In: *ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 10. Tampa, USA, 1985, pp. 1577–1580. DOI: 10.1109/ICASSP.1985.1168078.

[11] Divay, M. and Guyomard, M. "Grapheme-to-Phoneme Transcription for French". In: *ICASSP '77. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 2. 1977, pp. 575–578. DOI: 10.1109/ICASSP.1977.1170351.

[12] Evert, Stefan, Proisl, Thomas, Jannidis, Fotis, Reger, Isabella, Pielström, Steffen, Schöch, Christof, and Vitt, Thorsten. "Understanding and Explaining Delta Measures for Authorship Attribution". In: *Digital Scholarship in the Humanities 32*, no. suppl_2 (June 2017), pp. ii4–ii16. DOI: 10.1093/llc/fqx023.

[13] Gabay, Simon. "Beyond Idiolectometry? On Racine's Stylometric Signature". In: *Proceedings of the Conference on Computational Humanities Research 2021*, ed. by Maud Ehrmann, Folgert Karsdorp, Melvin Wevers, Tara Lee Andrews, Manuel Burghardt, Mike Kestemont, Enrique Manjavacas, Michael Piotrowski, and Joris van Zundert. Amsterdam, Netherlands, Nov. 2021, pp. 359–376. URL: https://hal.science/hal-03402994.

[14] Gabay, Simon. "Rien de Nouveau chez Rimbaud : sur l'attribution des Illuminations". In: *Revue Humanités Numériques 11* (2025). DOI: 10.4000/1498n.

[15] 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, Oct. 2020, pp. 1–7. DOI: 10.1145/3423603.3423996.

[16] 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).

[17] Haug, Andreas. "Musikalische Lyrik im Mittelalter". In: *Musikalische Lyrik*. Lilienthal: Laaber-Verlag, 2004, pp. 59–129.

[18] Hollien, Harry. *The Acoustics of Crime: The New Science of Forensic Phonetics*. Springer Science & Business Media, 2013. DOI: 10.1007/978-1-4899-0673-1.

[19] Honvault, Renée. “Statut linguistique et gestion de la variation graphique”. In: *Langue française 108* (1995), pp. 10–17. URL: https://www.persee.fr/doc/lfr_0023-8368_1995_num_108_1_5312.

[20] Ischi, Stéphane. “Rimbaud et Corbière se sont-ils lus?” In: *Romantisme 151* (1 2011), pp. 101–112. DOI: 10.3917/rom.151.0101.

[21] 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.

[22] Keselj, Vlado, Peng, Fuchun, Cercone, Nick, and Thomas, Calvin. “N-gram based author profiles for authorship attribution”. In: *In Proceedings of the Pacific Association for Computational Linguistics*. Halifax, Canada, 2003, pp. 255–264. URL: http://www.cs.dal.ca/~vlado/papers/pacling03.pdf.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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 Legallois, Charnois, and Larjavaara. De Gruyter Mouton, 2018, pp. 164–193. DOI: 10.1515/9783110595864-008.

[28] Lovichi, Jacques. “Germain Nouveau et les Illuminations”. In: *Rimbaud vivant 6* (1975), pp. 17–25.

[29] Mathieu, Cécile. “L’arbitraire saussurien: résistances et résolution”. In: *La linguistique 54* (1 2018), pp. 21–38. DOI: 10.3917/ling.541.0021.

[30] Mousavishirazi, Seyed Jamal. “Structure sonore d’un texte littéraire”. In: *Corela - Cognition, représentation, langage HS-30* (2020). DOI: 10.4000/corela.10402.

[31] Nagy, Ben. “Bootstrap Distance Imposters: High precision authorship verification with improved interpretability”. In: *Proceedings of the Computational Humanities Research Conference 2024*, ed. by Wouter Haverals, Marijn Koolen, and Laure Thompson. Aarhus, Denmark, 2024, pp. 482–493. URL: https://ceur-ws.org/Vol-3834/paper61.pdf.

[32] Pierrot, Roger. “Verlaine copiste de Rimbaud: les enseignements du manuscrit Barthou à la Bibliothèque Nationale”. In: *Revue d’histoire littéraire de la France 87* (2 1987), pp. 213–220. DOI: 10.3917/rhlf.g1987.87n2.0213.

[33] Rimbaud, Arthur. *Les Illuminations*. Paris: Publications de la Vogue, 1886. URL: https://gallica.bnf.fr/ark:/12148/btv1b8610832j.

[34] 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, Colorado: Association for Computational Linguistics, 2015, pp. 93–102. DOI: 10.3115/v1/N15-1010.

[35] 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.

[36] Stock, Markus. “Triôs, triên, trisô : Klangspiele bei Wernher von Teufen und Gottfried von Neifen”. In: *Beiträge zur Geschichte der deutschen Sprache und Literatur 138* (3 2016), pp. 365–389. DOI: 10.1515/bgsl-2016-0029.

[37] Viémon, Marc. *L’Apprentissage de la prononciation française par les espagnols aux XVIe, XVIIe et XVIIIe siècles*. PhD thesis. Sevilla: Universidad de Sevilla, 2016. URL: http://hdl.handle.net/11441/40533.

[38] 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.