



Științele limbajului

**Roxana Patraș
Loredana Cuzmici**
(editors)

**Tools, Methods, and Solutions
for the Exploration of Romanian Corpora**

INSTITUTUL EUROPEAN

Roxana Patraș, Loredana Cuzmici (editors), *Tools, Methods, and Solutions for the Exploration of Romanian Corpora*

© 2024 Institutul European Iași, pentru prezenta ediție

INSTITUTUL EUROPEAN

Iași, Str. Bălușescu nr. 2, et. I, O P 6, CP 1309

euroedit@hotmail.com.; www.euroinst.ro

Descrierea CIP a Bibliotecii Naționale a României

Tools, methods, and solutions for the exploration of Romanian corpora /

ed.: Roxana Patraș, Loredana Cuzmici. - Iași : Institutul European, 2024

Conține bibliografie

ISBN 978-606-24-0395-9

I. Patraș, Roxana (ed.)

II. Cuzmici, Loredana (ed.)

811.135.1

All rights reserved. No part of this work covered by the copyrights hereon may be reproduced or copied in any form or by any means – graphic, electronic, or mechanical, including photocopying, recording, taping – without written permission of publisher.

PRINTED IN ROMANIA

Roxana Patraş, Loredana Cuzmici
(editors)

TOOLS, METHODS, AND SOLUTIONS FOR THE EXPLORATION OF ROMANIAN CORPORA

INSTITUTUL EUROPEAN
2024

This work was supported by a grant of the Romanian Ministry of Education and Research, CNCS–UEFISCDI, project number PN-III-P1-1.1-TE-2019-0127, within PNCDI III.

The revision and editing of this collection was performed by BolderWords translation & revision and was supported by the Altissia Chair in Digital Cultures and Ethics held by Chris Tanasescu at Université catholique de Louvain.

Table of Contents

INTRODUCTORY NOTES	7
Tactics: your DH is not my DH Roxana PATRAȘ	9
CREATING TOOLS FOR ROMANIAN CORPORA	23
RELATE: a modern processing platform for Romanian language Vasile PĂIȘ, Radu ION, Andrei-Marius AVRAM, Maria MITROFAN, Dan TUFÎȘ	25
Tools and Resources for Digitizing Historical Romanian Documents Victoria BOBICEV, Tudor BUMBU, Ludmila MALAHOV, Alexandru COLESNICOV, Svetlana COJOCARU, Liudmila BURTSEVA, Cătălina MĂRÂNDUC	53
Digitalization of Romanian Lexicography Elena Isabelle TAMBA	85
ADAPTING TOOLS FOR ROMANIAN CORPORA	97
Postdigital Creativity: modeling poetry translation with multiplexes, neural networks, and large language models Raluca TANASESCU, Chris TANASESCU	99
Challenges of Digitization and Digitalization for the Study of Writing: from archival survey to process analysis Georgeta CÎȘLARU	125
Towards a Digital Corpus-Based Method for Assessing Language Level in EFL Student Writing: a case study on Romanian undergraduate literary analyses Alexandru ORAVIȚAN, Mădălina CHITEZ	139
DIGITIZING ROMANIAN HERITAGE	151
A Computer-Assisted Analysis of Eminescu's <i>Doină</i> Ioana GALLERON	153
Mapping Eastern European Political Discourse and Inequalities through Public Monuments: a digital cartography crowdsourcing project Voica PUȘCAȘIU	175
The Title of the Feuilleton Novel: a model of semantic annotation Lucreția PASCARIU	193
Exercises in Literary Geography: where hajduks roam in the second half of the nineteenth century Alexandra OLTEANU	227

<i>DIGITUS DEI EST HIC!</i>	251
Proxy Structures in Computational Analyses of Text Corpora	253
Laura PRICOP	
The Many Faces of Virtual: charting the family resemblance network for conceptual understanding	267
Matei Alexandru STOENESCU	
CARTOGRAPHIES OF THE ROMANIAN NOVEL: THEN AND NOW	297
The Recontextualization of a Genre	299
Loredana CUZMICI	
Notes on the Contributors	321

A Computer-Assisted Analysis of Eminescu's *Doină*

Ioana GALLERON

In a paper published in 2023, Patraş, Galleron and Patraş pleaded for a digital scholarly edition of Mihai Eminescu's work, using *Doină* as a case study for defining preferred editorial principles and technical choices. As shown repeatedly over time, digital editions have the advantage of being accessible and interoperable; also, to the difference to the paper editions, the one provided for *Doină* appears easier to follow.¹ In addition to this rationale, the present article aims to consolidate the argument by providing examples of how such an endeavor might facilitate the interpretation of literary texts. Thanks to a large array of tools and methods, digital editions bring to the fore patterns that are unseen or difficult to perceive, thereby contributing to the necessary and constant reappraisal of the literary patrimony. Without claiming—in the name of a dubious “scientific approach”—to replace the experience of reading for pleasure, such methods enable the testing of intuitions and the discovery of unexpected features; in short, they enable an increase in the body of knowledge and the esthetic emotion that are gained through literary works of art.

Without going into detail about the choice of the poem, or about the technical and philological choices made for *Doină*, it is necessary to summarize the ideas

¹ When Nicolae Iorga writes about Perpessicius's work, which he views as exemplary, he explains: “De altfel, matematica adnotatorului va putea fi [...] foarte cu greu urmarită. Eu unul n-o pot face ușor: uneori mă pierd. Era mai bine, orice proporții ar fi luat ediția, să se dea și mai des, cum se făgăduise, integralitatea versiunilor. ‘Etaj’, ‘subsol’, ca ‘terminologie arhitectonică’, îngreueie rezultatul acestei admirabile osteneli. Fără a mai vorbi de semne, a căror amintire nu poate urmări pe cititor” (*In addition, editor's mathematics of annotation will be... very difficult to follow. I cannot do it easily: sometimes I get lost. It would have been better, whatever the volume of the book, to offer more frequently, in accordance to what has been promised, the full versions. ‘Main text’, ‘footnotes’, all this ‘architectural terminology’ weighs down on this admirable endeavor. Not to mention the various signs, whose signification the reader may forget.*) (Iorga 1939–1940, 152–153).

[All translations from Romanian to English belong to this chapter's author].

developed in Patraş, Galleron and Patraş (2023) to gain a better understanding of the discussion that follows. Written between 1878² and 1882, *Doină* exists in 10 versions, which have been assigned the letters A through J.³ They have been transcribed with a semi-diplomatic protocol, then encoded following the parallel segmentation method. The result can be visualized with Versioning Machine or any other tool supporting such annotation.⁴ All verses from the ten versions (258 lines in total) have been compiled in a single TEI-xml document, presenting almost 800 variation points, whether phonetic, graphic, lexical, grammatical or otherwise. The sheer number of variants explains why, in print editions, scholars concentrate on so-called “significant variations”—that is, those that carry a high semantic load and offer an interesting point of view into the main choices made during the creative process. One of the main advantages of digital methods is to allow taking into account all variations, whether big or small. This leads to new insights into Eminescu’s creative process of *Doină*. In what follows, an in-depth study of the thematic patterns, then of the melodic outline of the poem and finally of the

² The emergence of *Doină* seems to have been precipitated by Eminescu’s reaction to the decisions of the Congress of Berlin in June-July 1878, following the end of the war between The Russian Empire and The Ottoman Empire. Several amendments were made to the Treaty of San Stefano. In addition to granting Romanian independence—and even rewarding the country, by expanding its territory to include two thirds of Dobruja—the new Treaty of Berlin (as it is now known) also ratified the definitive loss of Bessarabia, much to Eminescu’s wrath. Furthermore, the treaty required the Ottoman Empire, as well as the various newly formed nations, to recognize several civil rights for the Jews, which was a measure bitterly debated by the intellectual circles of those times (Eminescu included). Commenting the ‘Jewish issue’ as well as the poem’s reference to *railways* (“*drum de fier*”), Călinescu contends that the starting point of the poem is the Strousberg scandal (Călinescu 1969(1932), 276). Indeed, in the aftermath of the Romanian War of Independence, the Romanian government negotiated with German shareholders in order to buy back the costly railway which had been built in part by the consortium directed by the Jewish entrepreneur Bethel Henry Strousberg. However, by the time Eminescu put the first verses of *Doină* on paper, this scandal was already quite old, which suggests that its contribution to the conception of Eminescu’s text might be less important than Călinescu argues.

³ As indicated in Patraş, Galleron and Patraş 2023, the designation of the versions follows Perpessicius for A and B. We distinguish two separate versions corresponding to Perpessicius’s C; as a consequence, we propose 10 versions instead of 9, and the designating letters have been modified accordingly after C.

⁴ The annotated file can be downloaded from <https://nakala.fr/10.34847/nkl.ae9214uz>, DOI : [10.34847/nkl.ae9214uz].

toponyms used in the text, will be conducted on the basis of these 800 variants, perused with digital data and methods.

Thematic patterns

In an interesting attempt to analyze *Doină* by adopting an interdisciplinary approach that, on the one hand, takes into account the belated diagnosis of Eminescu's bipolar disorder and, on the other hand, incorporates what the authors call "verbal information," Pîrvu *et al.* (2016) affirm that:

the first part of the poem ("From Tisa to Dniester's tide/All Romanians to me cried,/That they could no longer dwell/Amidst the foreign swell... The foreigner is everywhere [...]") may have been written also in 1882 (late summer, early autumn), whereas the second part ("He who loves the foes about/ [...] / Rise, O'Stephen, from the ground [...]") will have been written during the hypomanic episode in late 1882 winter, early 1883 spring. The date for *Doina* [*sic*] should consequently be given as 1882–1883. (373)

While such a hypothesis is clearly ill-informed, ignoring the wealth of information uncovered by Perpessicius about the duration of the poem's creation process, Pîrvu *et al.* affirmation serves to draw the reader's attention to the rather multidimensional nature of the poem. *Doină* mixes several literary motifs that convey quite different messages. The use of the digital methods can help unravelling the various stages through which *Doină* has developed, from its first outline in 1878–1879 (version A) to the most recent known versions (I and J).

A count of the variants per version (annotated with the <rdg> tag in TEI) allows to establish both a rate of differentiation and a rate of similarity. The former is calculated as the number of cases in which a version brings a variant of any nature (spelling, grammar, addition, deletion, etc.) that does not appear in any of the other versions. From this point of view, the most original version of *Doină* would be F, with 63 own variant lines; the most "common ground" version turns out to be E, with only 33 own variant lines. It should be noted that the text published in *Convorbiri*

literare (version J) is not one of the most “common ground” versions: with 56 own variant lines, it is the second most distinct version. Most points of differentiation relate to spelling and punctuation, as Perpessicius observed; however, to the difference to Perpessicius, I consider these aspects to be not so neutral, since they contribute to revealing the dialectal Moldavian dimension of Eminescu’s scriptural practice.

The similarity rate is more complex, requiring a pairwise calculation of versions. Given the exploratory nature of the present study, as well as the relatively limited number of variants, I calculated this rate based on XQuery extractions;⁵ in the event of a larger-scale analysis (one involving a much larger number of poems and variants), the testing of other methods is necessary, such as a stemmatology program written in R,⁶ or the automation of extractions with Python. For the present study, the collected values were entered into an Excel table with the different variants in double entry (where variants appear in both rows and columns), and a heat map has been generated (*Figure 1*⁷).

Several lessons can be learned from this visual representation. First, the heat map confirms that *Doină* evolved in a non-linear manner. Versions A and B are produced shortly after each other, and therefore share a great number of similarities. With versions C and D, Eminescu tries out what Perpessicius calls a new “pattern.” It should be noted that version E—which constitutes a first stable version of the poem, written neatly and maybe intended for publication in the “Familia” journal—resembles version C more than it does version D; also, to some extent, version E is a return to the earliest attempts (A and B). Versions F and H explore themes not yet addressed in the previous versions; in between these two versions one finds version G, which largely returns to the thematic patterns of E. Version I revisits certain choices made in E and G, while also coinciding in many ways with A and B. Its

⁵ Annotation as well as the XQuery extractions have been carried out using oXygen, <https://www.oxygenxml.com/>.

⁶ For instance: <https://cran.r-project.org/web/packages/stemmatology/vignettes/stemm-atology.pdf>.

⁷ The graph is to be read from left to right. To give but an example, the second value in the first line means that in 181 cases version B matches version A. The lowest number, 127, indicates that version H meets version F only in 127 variation points. Since each version necessarily matches itself, the meeting points between A and A, B and B, etc. are in deep green. The total number of variation points (that is, the total number of <app>) recorded in the TEI-xml file is 261.

colors in the graph tend to confirm Perpessicius's assumption, according to which this version I is the one Eminescu has read in Iași just before falling ill;⁸ it can therefore be considered as a quasi-definitive version. However, it should be noted that the similarity rate of J with the previous versions (including I) is decreasing: contrary to what Perpessicius states, the orthographic differences are not trivial; rather, they distinguish the published version not only from the "copy from Iași" (if indeed this is what Eminescu shared with his listeners) but also from the other versions written by Eminescu.

With respect to the rate of similarity, it is worth emphasizing the special character of the H version. Perpessicius describes it as a last exercise before the definitive pattern. It is strange, however, to place this version after E and F, with which version H has less in common than with D, especially since the (rather imprecise) dating proposed by Perpessicius ultimately associates it with 1880 rather than with 1881–1882:

For dating we are guided by a draft of verses 37–48 that appears in Ms. 2264, f° 378, on the back of a proof form, in the vicinity of and with the same writing as a note, for the newspaper, which heralds the decoration of Macedonski with *Bene merenti*, class I. (Perpessicius 1944, 16)

Alexandru Macedonski (1854–1920) having been decorated in 1880, the argument suggests that H was drafted before G. Therefore, either the fragment on folio 378 has nothing to do with version H and is instead linked to versions C and D, or the folio fragment is indeed associated with version H (in keeping with Perpessicius's claim), and H would therefore be considerably older than F and G, probably written before *Doină*'s first stable version E. All this leads to the idea of *Doină* having two great periods of creation. The first one precedes 1881–1882; the poet experiments with several thematic patterns, and expresses anger and a certain feeling of helplessness. The second period comes after 1881–1882, during which the

⁸ According to Călinescu, Eminescu read *Doină* in front of the "Junimea" members only, and did not make it to read it the following day, during the gathering in front of the monument dedicated to the famous Moldavian king of the Middle Ages, Ștefan cel Mare [Stephen the Great] (see Călinescu 1969 (1932), 276 et sq.).

poem matures on the formal level, while the tone becomes more moderate and the overall orientation more pragmatic.

The alignment of the variants allows us to observe that, before 1881, Eminescu develops two literary motifs that are significantly reduced starting with version G. First, the poet delights in listing the people of various foreign nations: Tatars and Kalmyks, Gagauz people, Hungarians and even—in version F—Bulgarians and Greeks. Of these, only the Russians (“*muscalii*”), which were introduced in version C, remain in the later versions (I and J). The other nationalities are designated as “foreigners” or “enemies,” nouns which preserve the polemical aspect of the text but reduce the contentious nature of the poem. Also, the elimination of the other demonyms allows the poem to focus on the Russians; after all, they were the main beneficiaries of the Berlin Treaty of 1878 through the annexation of Bessarabia, a main catalyst for the writing of *Doină*.

The two periods of creation are also clearly distinct with regards to the use of the curses. These are much more abundant in versions A, B, C and D. Appeals to interventions from the natural world, whose forces lead the homeland to cleanse itself of impure bodies, are present in all four of the versions written before 1881–1882. First, Eminescu references various diseases—such as the plague or the cholera; see also “*partea boalei*” (any form of disease) in H—that are likely to attack foreign swarms that attack helpless Romania. Second, he evokes natural catastrophes (blizzards, floods) that are capable of sweeping such swarms away, and he goes on to develop the image of a symbolic dismemberment by various animals, with dogs and crows mentioned in the final versions. These appeals to primitive—even chthonian (as of version C, the “desert” is invoked)—forces are all the more significant because the poem engages in a kind of alternation between them and invocations of Stephen the Great. The image of the king’s resurrection is absent from versions C and D, which contain the largest number of curses. Under the influence of anger, Eminescu seems to adopt a form of magical thinking, according to which solutions to problems come from non-human forces. In this first period, the poem expresses feelings, rather than reflecting calls for action.

After 1881–1882, the profile of *Doină* changes. With the exception of version H, whose chronological location is debatable (as discussed above), the poem opts for

a decidedly apotheotic ending, calling for the return of Stephen the Great. The image of the return of the king appears from the very beginning of the creative process (i.e. in version A), but it wanes at certain moments, as we have already seen. Moreover, in versions A and B, Stephen the Great is primarily the lead organizer of a campaign to scorch the ground in front of the invaders, rather than a revivalist figure. In the toned-down, more optimistic context of the later versions, the same literary motif of the king's return symbolizes Romanians' capacity to act of their own accord, by mobilizing their proven virtues.⁹ Magical thinking, which calls on supernatural forces, does not completely disappear in the later versions of the poem (forests, animals, and even the core spirit of the nation continue to be invoked), but it no longer plays a central role. While nature seems to protect the suffering Romanians in the early versions, the situation is somewhat reversed in the later versions, in which it is nature that predominantly suffers and asks for protection. In this respect, the digital juxtaposition of the different versions shows that the development of the nature motif, which increases from version A to version F, is somewhat slower from version G to version J. Eminescu eliminates the allusion to the spruce leaves, and reduces the redundancy between "forests" (păduri) and "groves" (codri), present in versions C and D. A description of the negative impacts of the country's enemies thus replaces the evocation of an invasion ("the Tartar hordes are coming"), which brings to mind a medieval era and is therefore effective on the affective level but less cognitively pertinent for Romanian readers in the 1880s. Similarly, the poet develops the motif of a Romanian illness ("moving backwards as a crayfish," etc.): foreigners are not themselves harmful (as in the A to D variants) but are dangerous because of the effects they produce. This argument is meant to give an ethical basis to the xenophobic appeal.

To sum up, before 1881–1881 *Doină* expresses a terrible but also self-destructive reaction to the amputation of Romania of some of its territory. After 1881–1882, the poem offers rather a kind of diagnosis of the national problems, while outlining an action plan. We are referring here to an ethical and esthetic call to

⁹ Stephen the Great ("Ștefane, Măria Ta") is not the only tutelary figure invoked: manuscript A shows that Eminescu also explored other historical or mythical characters, such as Matias Corvin and Romanian peasants (e.g. "Ioan", "Romanian brothers").

action, rather than to a political discourse, which is more developed in Eminescu's journalistic activity from the same period (Vatamaniuc 1996, Mocanu 2020). The poem is meant to revitalize the past and to reinvigorate the threatened nation, by reviving the ghost of Stephen the Great and the national desire for self-determination. This evolution may testify to a precocious depressive episode (to borrow the medical terminology used by Pîrvu *et al.*), which might have taken place in 1879–1880, followed by a manic episode in the second part of the creative period (1881–1882) and a subsequent low, which led to Eminescu's admission to the Caritas sanatorium (Neghina and Neghina 2011).

The melodic evolution of *Doină*

The preceding analysis of the thematic patterns in *Doină* calls into attention the evolution of the melodic outline of the poem, which seems to have interfered, at least in certain versions and to some extent, with the cognitive acceptance of the images created. The most conclusive evidence of this is to be found in version C, where Eminescu seems to focus on the musical dimension of the poem, sometimes to the detriment of the text's semantic dimension. In this version, one sees caterpillars eating “spruce flowers”, and ravens that “carry flesh on thorns/ And bones through stalks.” This is because the rhyming words for “caterpillars” (omizi) are not numerous, and because the verb “to leaf out” (“a înfrunzi”), which could have been an acceptable alternative, has already been used in verse 26. One of the few options left is to rhyme “omizi” with “molizi” (spruce), evoking in the process a little-known “spruce flower” (which, incidentally, seems not to be threatened by predators, at least when compared to pulp and needles). In version D, Eminescu replaces “flower” with “leaf”, rendering the image more acceptable; also, he moves the flowering to the previous verse, in connection with elms and cornels (*cornus mas*¹⁰). The alliteration in “z” (“frunze” [leaves], “molizi” [spruce], “omizi” [caterpillars]) therefore constitutes the main driving force in the writing of these lines, somewhat to the

¹⁰ The Latin name has been kept here because a satisfactory English-language translation of the Romanian word for “corn” has not been found.

detriment of the sense. After 1882, Eminescu focuses more on the image of depleting domestic resources, while the alliteration becomes less perceptible or disappears altogether. The poem's themes and melodic outlines come together, achieving those formal qualities of poetry that are easy to follow, pleasant to listen to, and familiar to most Romanian readers.

A larger study of the melodic outline of *Doină* confirms this conclusion. In order to conduct such research, a phonetic transcription of the 10 versions has been realized. For pragmatic reasons, this task has not been conducted in strict accordance with the recommendations of the international phonetic alphabet. While ChatGPT 3.5 (which was used to generate a first transcription) and the Phonemizer Python script¹¹ (which has been a second option) use several characters from this alphabet to indicate specificities of Romanian phonology (such as /tʃ/ for the palatal sound present in “nici,” or the short i sound (/ɪ/) at the end of the same word), counting the annotated phonemes with Python proved to be more complicated, insofar as the scripts identify two elements (rather than one) in the first case (/t/ and /ʃ/) and add an extra grapheme /ɪ/ in the second case. Such errors can probably be corrected with the use of a more specialized dictionary; however, the aim here was not to achieve transcription accuracy, but to count the occurrences of phonetic characteristics playing a distinctive role in the evolution of the melodic outline of the poem. Such features include, among others, semi-closed vowels, semi-vowels (/j/ in particular), affricative and fricative consonants (especially /tʃ/, /dʒ/ and /ʃ/). The decision was therefore made to produce a customized phonetic transcription, using graphemes such as /ǧ/ (instead of /dʒ/), /j/ (instead of /ɪ/), /ç/ (instead of /tʃ/ or /č/), and /wa/ for the diphthong “oa.” While such workarounds imply a certain amount of post-processing (e.g. to discriminate between the real “j”—as in “jidă”—and the /j/ indicating a semi-vowel), they allow the phonetic transcription to be carried out easily and efficiently. In total, 30 distinct phonemes were identified, during the transcription process, as playing a major role in the creation of the melodic outline of the poem, and were subsequently counted through a simple character tally in Python.

¹¹ <https://bootphon.github.io/phonemizer/>.

In order to understand the results of the phonetic count, we need to appreciate two additional aspects that must be taken into account. First, the lack of precision in Eminescu's use of diacritics had to be addressed. The transcription described in Patraș, Galleron and Patraș (2023) attempted to follow Eminescu's manuscript as closely as possible; diacritics have not been added if they were missing in the manuscript, nor the various signs corrected, that Eminescu used to indicate the phonetic value of a given phoneme. However, such scrupulous respect for the letter of the text was not possible within the framework of phonetic transcription; as Vladimir Streinu¹² has shown, the editor's reluctance to engage his or her responsibility with regards to the phonetic value of a glyph leads to a kind of falsification of Eminescu's poetry, creating nonexistent and non-aesthetic pronunciations. For the sake of consistency, all instances of "a" have been rendered as "ă" in the appropriate contexts (e.g. "sătmar" instead of "satmar"; "străinătate" instead of "străinatate" or "strainatate"), even if this meant erasing the difference between the versions in which the well-known verse "poor, poor country" is spelled in standard Romanian ("săracă țară săracă") and the versions that feature the spelling "saracă țară, saracă!", with a more pronounced Moldavian accent.

¹² "Atât de stăpânit de superstiția rimei s-a dovedit unul dintre [editori], [...], că intervine în două feluri: fie crezând că acordă mai muzical cuvintele neacordate deplin de poet, fie că îngăduie forme curioase la un cuvânt, care s-ar găsi în manuscris, pentru a salva... sonorizarea cu celălalt cuvânt. În amândouă cazurile, se ajunge la modificări flagrante față de textul eminescian" [One of the [editors] proved to be so possessed by the superstition of rhyme, ... that he intervened in two ways: on the one hand, through giving a more musical tune to the words not fully tuned by the poet, on the other hand through allowing curious word forms, which are supposedly to be found in the manuscript, in order to save... the voicing with the other word. In both cases, these are blatant changes compared to the Eminescu's text]. Streinu also adds: "poetul tindea în grafia lui, chiar cea mai îngrijită, la uniformizarea semnelor pentru ă, â, sau î și i; de cele mai multe ori, același semn al lui ă indică și pe â sau î, semn care adesea devine o linie și nu o dată un simplu punct. [...] Dar aparența manuscriselor, din cauza uniformității către care tind acele câteva semne grafice, nu e însăși realitatea lor, iar cât despre rimă, e de presupus că Eminescu, în rafinamentul auzului său, cuprindea și asonanța cu posibilitățile ei moderne, mai subtile decât ale rimei clasice." [Even when he used his neatest handwriting, the poet tended to employ the same signs for ă and â, or î and i; most of the time, the same diacritic as for ă also indicates â or î, becoming frequently a line and even a simple dot... But the appearance of the manuscripts, with the uniformity towards which these few graphic signs tend, is not their true reality, and as for the rhyme, it can be assumed that Eminescu, in the refinement of his hearing, also considered assonance with its modern and more subtle possibilities than the classical rhyme] (Streinu 1939, 103-105).

The second aspect to note in order to understand the results of the phonetic count is that transcribed versions do not take into account text fragments that appear in certain versions of the poem but not in others. These cuts have an obvious impact on the total number of vowels, semi-vowels, diphthongs etc. counted, since some occurrences may be present in the deleted parts but are absent or very limited in the remaining lines. One may therefore consider that some phonetic characteristics observed and interpreted in what follows have been given outsized importance. However, one must bear in mind that Eminescu himself considered the lines not taken into account as being less important for he was trying to express. On the melodic side, although the poet experimented with the same phonemes, words and patterns throughout the creation of *Doină*, some of them had clearly been given greater prominence in the building of the final melodic outline.

The different versions contain between 824 (version E) and 1216 (version F) phonemes. An analysis of the weight¹³ of each phoneme in the total number of lines phonetically transcribed indicates that some sounds are indeed particular to certain versions. As hypothesized, such phonemes include the semi-vowel /j/, the affricative /tʃ/ and the diphthongs “oa” and “ea.”¹⁴ Other discriminants are “a”, “i”, “j”, “m”, “n”, “s”, “u” and “v,” with the two nasals (“m” and “n”) and the fricative “v” being the most unexpected in this list. The presence of these three sounds is quite limited in certain versions (“m” in version A, “n” in version D, “v” in version E), an aspect that has not been observed so far since, when listening to a poem, one has more difficulty detecting such “missing” phonemes than identifying over-represented sounds. On the whole, vowels, semi-vowels and diphthongs tend to be the main discriminants between versions, suggesting that Eminescu was especially interested in experimenting with assonance, as opposed to alliteration, in *Doină*.

¹³ The weight has been determined as follows. I divided each phoneme count by the total number of phonemes in the version to calculate its percentage of contribution to the whole. After determining an average for each phoneme, across all versions, and a standard deviation (calculated with Pearson’s formula), I noted the position of each phoneme in the data series. No value is situated beyond 2.36 standard deviation points, with only 14 values beyond -2/ +2 standard deviation points. All this indicates a fairly narrow distribution of data, and no real outliers; at the same time, this distribution supports the idea that some phonetic features are quite distinct in certain versions.

¹⁴ The semi-closed vowel “ă” and the affricative “ț” did not meet the expectations regarding their discriminant power.

How has this work been carried out, and what can be said about the final product in terms of its melodic outline? In order to answer these questions, we have prepared a table that shows the distribution of the phonetic features of each version, and we have normalized and fitted the table for a principal component analysis (PCA), using a SKLearn script.¹⁵ The results can be seen in *Figure 2*, generated with Matplotlib.¹⁶

The different versions of *Doină* can be divided into two main groups, with C, D and H very distinct on the y-axis. Significant differentiation can also be observed on the x-axis, between F, on the one hand, and C and E, on the other. On the whole, it is worthwhile noting that the distribution does not follow an arc that moves from the earliest to the latest versions: as observed when we analyzed the thematic patterns, Eminescu seems to have experimented with several possibilities, abandoning and returning to some of them, before reaching a certain stability of the melodic outline. In this regard, it is significant that, on the graph, version A is quite separate from the other versions: whereas some of its phonetic traits serve as the foundation for most of the subsequent versions, significant divergences can be noted as early as in version B. The relative weight of phonemes across versions (see note 15) indicates that version B reflects an increase in “u” sounds and a significant reduction of “j” sounds. Version C takes a sharp turn from the melodic outline of B, especially through an increase in the number of /j/ semivowels; this number remains very high in version H and quite high in version D, especially considering that this feature is very limited in almost all of the other versions. Version D incorporates several “i” and “r” sounds, signaling an interest in increasing the musicality of the text (“i” being a vowel, and “r” being a liquid consonant); it is also in this version that Eminescu explores the alliteration of “z” sounds, as seen earlier. Another significant change in phonetic choices occurs with version E. Rather than introducing new features, the poet reduces the phonetic diversity; this can be better understood by looking at the heat map in *Figure 3*, based on absolute values for each phoneme.

¹⁵ <https://scikit-learn.org/stable/index.html>.

¹⁶ <https://matplotlib.org/>.

Figure 3 is somewhat misleading, since it suggests that later versions contain a larger number of phonetic specificities than the earlier ones and does not reflect all of the observations made in the previous paragraph. This misleading effect is predominantly related to the difference in size between the versions, with versions F through J increasing considerably in the total number of phonemes, as opposed to the absolute values from versions A to E; indeed, the size of the poems—which accounts for almost 50% of the total variance—is the first component retained in the PCA. This visualization nonetheless allows us to see the extent to which version E represents a decrease when phonetically compared with traits from version D (especially), and explains why version F marks a new development in the musical conception of the poem. In terms of relative values, version F does not contain much innovation (apart from a noticeable increase of “ț” sounds), but the increase in the word count changes its melodic outline. Despite certain local variations, such as the increase in the relative number of “a” sounds in version G or the similar evolution of “l,” “î” and “n” sounds in versions I and J, version F becomes the foundation for the melodic outline of subsequent versions.

Once again, version H is distinctive because it clearly differs, in both absolute and relative values, from versions F and G, on the one hand, and from versions I and J, on the other. Version H can be named “a study in /tʃ/” (e.g. “cine ne-au adus muscalii,” “cine ne-au adus jidanii,” “cine ne-au adus pe greci”), interspersed with a considerable amount of “ea” sounds (as seen in the examples quoted above). With respect to the first feature, this version is among those that most resemble version A, an argument which supports the suspicion that this version was elaborated during the first, rather than the second, period of composition of *Doină*. The melodic outline of version H is very distinct from that of versions G or I, and from that of almost all the other versions; it thus testifies to a chaotic creative moment, during which Eminescu explores different patterns, both figurative and melodic, before arriving at a more harmonious form of poetry in the period leading up to the 1883 recital in Iași.

Overall, the phonetic evolution of *Doină* shifts from “signature” versions—characterized by a surplus of one or several phonemes—to non-outstanding versions, in which all of the traits discussed above (vowels, semi-vowels, palatal fricatives, etc.) are carefully combined to give the impression of a steady flow of the laments,

of the imprecations about the foreigners and of the appeal to patriotism. As he works on toning down the chauvinistic dimension of the poem, Eminescu endeavors to soften its melodic outline. This softening can be seen as a trick used to inhibit critical reflection among readers or listeners who were—or, unfortunately, are—already inclined to adhere to the often heinous nationalist discourse espoused by the poet.¹⁷ Adopting a more charitable view, we might instead appreciate this phonetic softening as an indication of how the poet was distancing himself from the extremely xenophobic comments he had initially made; as already noticed by scholars, Eminescu was rather ambivalent in his positions: “for almost every violent anti-Semitic statement from one article, another less violent one, which reevaluates and contradicts the former, can be found in another article or, sometimes, in the same article” (Butaru 2012, 119).¹⁸ At any rate, the digital exploration of *Doină*’s phonetic dimension draws attention to the interrelated roles and evolution of musicality and ideology.

Drawing a Romanian map

Among *Doină*’s most distinctive features are the place names used in the opening line: “Nistru” and “Tisa.” Yet few readers are able to quote other pairs of toponyms identified in the poem, despite the fact that these other place names are much more numerous (e.g. Sătmă, Boian, Turnu, Dorohoi). In fact, much like with the literary motif of the nature, the evolution of the 10 versions reflects an increase in the number of place names identified, followed by a stagnation. *Doină* is not the first poem in which pairs of toponyms introduced by “from XX to YY” appear; this

¹⁷ To a certain extent, the role that the melodic outline played in the success of nationalist discourse is suggested by Lucian Tion (2021), who comments on the xenophobic drive of a poem titled *Scrisoarea a treia* (The Third Epistle): “The Bulgar and the Greek filth have been joined by ‘the crook-faced Gagauz’ [...], and, of course, by the omnipresent Jew, who figures relentlessly in Eminescu’s political writings as the primordial threat to Romanians everywhere, however, nowhere as melodious as in this example of poetic racist slur” (Tion 2021, 76).

¹⁸ Butaru’s explanation for this consistently contradictory position is that Eminescu’s hate concerns not the foreign national *per se* but the foreigner as a vector of liberalism, carrier of a form of universalism and non-manual work ill-suited to an agrarian country. For other considerations about Eminescu’s antisemitism and chauvinism, see Chiorean (2023) and Jicu (2012).

syntactic structure had already been used in *Bogdan Dragoș* and in *La arme* (*Pick up the weapons*), which Perpessicius referred to as “collateral” poems to *Doină*. However, Eminescu’s text is certainly the one in which the phrase has almost an obsessive presence, which raises the question of its significance. Does Eminescu choose place names at random—for their euphonic virtues, for instance? Or does he make conscious choices that are semantically loaded? The first pair of toponyms suggests a rather deliberate selection, since the two rivers represent frontiers that are not only natural but also symbolic: for Eminescu, they designate the breadth of the Romanian territory, despite part of it having been confiscated by the Russian Empire. However, it is less clear what message he had hoped to send by using the other toponyms, some of which are placed at the very heart—versus on the margins—of Romania (Aiud, Brașov, Boian).

In order to better understand the meaning conveyed by these toponyms, I have projected their coordinates onto a map. The results can be seen in *Figure 4*, where green lines are used to connect the different pairs.

Though imprecise, this projection allows us to observe that the lines joining the pairing points cross out Romanian territories that were under foreign administration during the creation of *Doină*. The poem therefore very subtly expresses a sentiment of unacceptable loss, while simultaneously reintegrating the places into the national space. Rather than indicating frontiers and bringing the idea of a “round Romania” (known as “România dodoloață”) to the fore, *Doină* therefore strives to visualize an amputated country. The idea of a poetical crossing out of lost territories is partly supported by a similar appeal that Eminescu made in one of his articles, which was written during the same period as *Doină*’s creation, and in which he invites the reader to “draw[...] a line from Hotin on the Dniepr to Prut.”¹⁹ Further evidence is to be found in the fact that toponym pairs privilege those geographical landmarks that optimize coverage of the homeland territory. This is probably why phrases such as “from Brașov to Aiud” (or Banat/Arad) and “from Olt to Criș”—two of the shortest diagonals among those drawn by the poem on the imaginary map of the Romanian-

¹⁹ “Trăgând o linie de la Hotin din Nistru până la Prut” (Eminescu *qtd. in* Pleșca 2014, 39). The author does not give the exact references of the quotation, vaguely referring to a paper published in the conservative journal *Timpul* in January or February 1878.

speaking countries—are eliminated in the later versions. Overall, the accumulation of geographical points, while no doubt also informed by euphonic concerns, helps to establish, more and more clearly and from a version to the next, the image of an amputated Romania, adding to the apocalyptic view of a country that loses its vitality because of the foreigners.

Conclusion

The analysis of the literary motifs, phonemes and toponyms, as they appear and disappear from one version to another, allows us to better understand the creation process of *Doină*. Two main periods of elaboration have been observed, one during which the poem functions especially as a curse, and a second one where it primarily functions as a patriotic appeal. Eminescu's vision is perpetually evolving, and the three aspects analyzed in the previous sections support Perpessicius's intuition that the published version of *Doină* is not the final one. Had the poet not been affected by the mental illness that ultimately led to his death, we would probably have had to annotate additional versions of the poem, and maybe compare them with a stable, if not definitive, version published by Eminescu himself.

Much remains to be done in terms of using digital tools to exploit *Doină* more fully. A coding of the variants as (ortho)graphic, lexical, topographic, and so on, as called upon in Patraș, Galleron and Patraș (2023), still remains to be done: such coding would most certainly help to refine our understanding of the literary motifs, ideology and melodic outline of the text. Moreover, an analysis of the quantitative evolution of each part of speech might support (or contradict) some of the observations made above. This paper is, however, a good example of what might be accomplished if we had at our disposal a complete digital collection of Eminescu's texts—or at least a digital collection of those texts that were published during his lifetime.

The present paper is not the first attempt to read Eminescu's poetry with digital tools. As compared to previous attempts to describe Eminescu's style from a quantitative point of view (Briciu, 2019) or to apply a sentiment analysis to his work (Lupea, 2021), the specificity of this article lies both in the decision to focus on a

single text and in the exploitation of a carefully curated digital scholarly edition. It argues for an approach to Eminescu's work, and the literary corpus more generally, that is anchored in a more qualitative analysis, to the detriment of a distant reading. The digital humanities developed, in the 2000^s, with the goal of continuing, amplifying and simplifying traditional literary criticism and philology; in recent years, however, the discipline has strayed greatly from these initial promises, especially in the context of a rapid development of natural language processing and large language models. These recent achievements undoubtedly have much to offer the exploration of cultural heritage, in all of its forms. However, making sense of this heritage still requires human intervention in the preparation and exploration of data. The present contribution hopes to have been a reminder of this truth.

Annex 1: Figures

	A	B	C	D	E	F	G	H	I	J
A	261	181	162	159	187	146	165	148	172	167
B		261	151	154	187	150	171	141	173	166
C			261	177	183	129	155	150	154	150
D				261	161	145	167	160	166	159
E					261	161	182	150	180	172
F						261	162	127	154	145
G							261	168	180	171
H								261	154	147
I									261	200
J										261

Figure 1: Similarity between versions

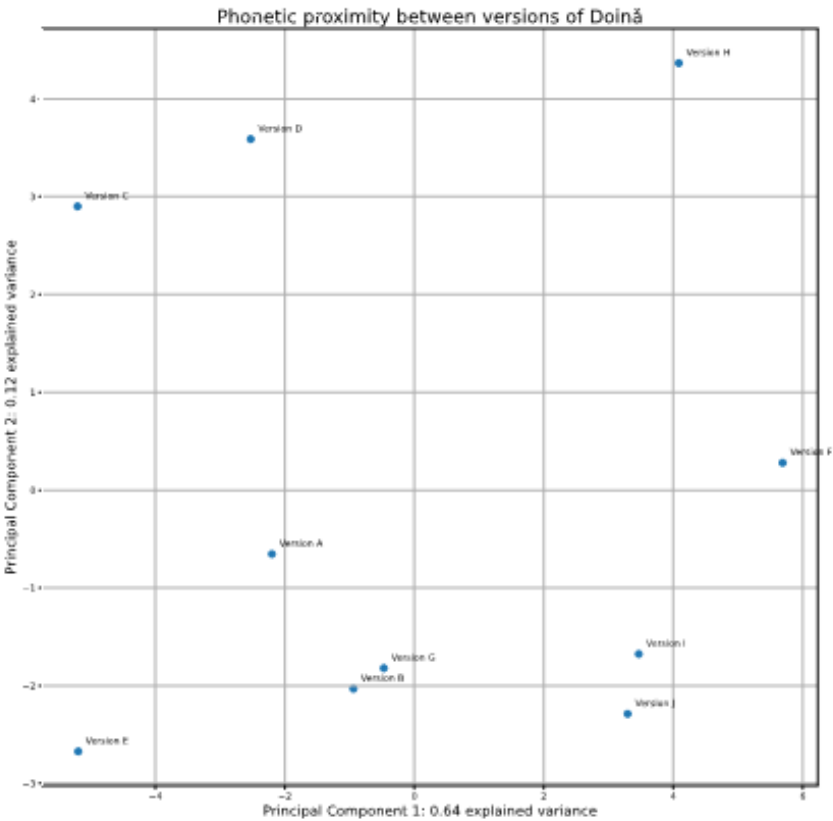


Figure 2: Principal component analysis of the phonetic transcriptions of versions of *Doină*

	Version A	Version B	Version C	Version D	Version E	Version F	Version G	Version H	Version I	Version J	Average
a	90	F	91	F	88	128	F	F	F	124	109,1
ă	F	F	45	F	47	65	F	F	F	F	53,5
b	F	F	F	F	4	F	F	12	F	F	7,1
c	F	25	F	48	23	F	F	56	F	F	37
ç	F	F	F	F	F	F	5	22	F	F	10
d	F	F	25	28	F	53	53	F	F	F	42,6
e	F	F	47	F	45	83	F	83	F	F	64,7
ea	7	F	F	7	F	F	F	24	F	F	12,9
f	F	3	F	7	2	8	3	F	F	F	4,9
g	F	F	F	F	F	F	F	F	F	0	1,4
ğ	F	F	2	4	F	F	F	F	9	10	6,6
h	2	F	3	3	F	3	F	7	F	7	4,8
i	F	F	17	17	F	F	F	F	31	31	23,4
î	F	F	F	74	48	72	55	71	F	F	63,2
/i/	37	F	F	F	33	53	F	58	53	F	45,8
j	4	F	F	F	F	F	F	F	F	F	1,2
l	F	F	34	F	34	52	F	F	51	51	41,5
m	26	F	28	F	29	47	F	46	F	F	37,1
n	F	F	68	69	65	F	F	103	101	F	84,8
o	F	F	27	F	F	46	F	32	F	F	39,6
p	F	F	F	F	F	29	F	27	F	F	22,8
r	61	F	F	F	56	F	F	F	81	F	71,3
ş	F	F	F	27	18	27	18	16	F	F	22,1
s	F	38	F	F	38	54	F	60	F	F	46,8
t	F	F	37	45	F	83	F	F	80	F	65,2
ţ	F	F	F	5	F	13	F	F	F	F	8,5
u	53	F	51	F	F	74	F	F	74	F	64,3
v	F	F	F	F	5	17	F	F	F	F	12
w	F	F	8	5	F	F	F	F	14	F	11,1
z	F	6	F	F	F	6	F	7	F	F	4,9
Total phonemes	947	989	834	942	824	1216	1014	1162	1143	1131	

Figure 3: Phoneme distribution in each version of *Doină*Figure 4: Localization of toponyms in *Doină*

References

- Briciu, Anamaria. 2019. "Quantitative Analysis of Style in Mihai Eminescu's Poetry." *Studia Univ. Babes Bolyai, Informatica* 64 (2). DOI: [10.24193/subbi.2019.2.06].
- Butaru, Lucian T. 2012. "Eminescu and the Pattern of Romanian Antisemitism." *Studia UBB Europaea* 57 (3): 117–128.
- Călinescu, George. 1969 [1932]. *Viața lui Mihai Eminescu*. In *Opere*, vol. 11. București: Editura pentru literatură.
- Chiorean, Maria. 2023. "Racialized Modernity in Late-Nineteenth-Century Romanian Literature." *Metacritic Journal for Comparative Studies and Theory* 9 (1). DOI: [10.24193/mjst.2023.15.02].
- Iorga, Nicolae. 1981. "Eminescu în și din cea mai noua ediție." In N. Iorga, *Eminescu*, ediție îngrijită, studiu introductiv, note și bibliografie de Nicolae Liu, 150–178. Iași: Editura Junimea.
- Iorga, Nicolae. 1981. "Note istorice asupra editării operei poetice a lui M. Eminescu." In N. Iorga, *Eminescu*, ediție îngrijită, studiu introductiv, note și bibliografie de Nicolae Liu, 124–129. Iași: Editura Junimea.
- Jicu, Adrian. 2012. "Antimodernii Eminescu și Caragiale." *Philologica Jassyensia* 8 (2 (16)): 65–75.
- Lupea, Mihaela, Anamaria Briciu, și Elena Bostenaru. 2021. "Emotion-based Hierarchical Clustering of Romanian Poetry." *Studies in Informatics and Control* 30 (1): 109–118.
- Mocanu, Mihaela. 2020. "Eminescu's Journalistic Works – Editing Approaches and Reading Patterns." *Philobiblon* 25 (1). DOI: [10.26424/philobib.2020.25.1.03].
- Neghina, Raul and Adriana Maria Neghina. 2011. "Medical Controversies and Dilemmas in Discussions about the Illness and Death of Mihai Eminescu (1850–1889), Romania's National Poet." *Medical Problems of Performing Artists* 26 (1): 44–50.
- Patraș, Roxana, Ioana Galleron, and Antonio Patraș. 2023. "Către o ediție digitală a operei lui Mihai Eminescu: alinierea în TEI-xml a versiunilor Doinei și vizualizarea cu Versioning Machine." *Philologica Jassyensia* 19 (1 (37)): 109–123.
- Perpessicius. 1944. "[Note și comentarii la] *Doină*." In M. Eminescu, *Opere III. Poezii tipărite în timpul vieții. Note și variante: de la Doina la Kamadeva*, ediție critică îngrijită de Perpessicius. București: Fundația Regele Mihai I.
- Pîrvu, Bogdan C., Mihai A. Stroe, and Rodica Nagy. 2016. "Towards a Linguistic Patho-Chronology: Dating Poetry Using a Language-and-Mind Approach (Case Study: Mihai Eminescu)." *Revue Roumaine de Linguistique* 61 (4): 361–377.
- Pleşca, Nae-Simion. 2014. "Basarabia în Publicistica și Poezia lui Eminescu." *Philologia* 56 (mai-august): 39–43.
- Streinu, Vladimir. 1989. "O ediție Eminescu." In V. Streinu, *Eminescu*, ediție îngrijită, prefată, note și indice de nume de Mihai Dragan, 95–115. Iași: Editura Junimea.

- Tion, Lucian. 2021. "The Use of Language in the Making of Romanian Nationalism from Dimitrie Cantemir and Mihai Eminescu to the Iron Guard and National Communism." *Acta Univ. Sapientiae, European and Regional Studies* 20: 63–92. DOI: [10.2478/auseur-2021-0012].
- Vatamaniuc, Dimitrie. 1996. *Publicistica lui Eminescu, 1877–1883, 1888–1889*. București: Minerva.