STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LXIV**, Number 2, 2019
DOI: 10.24193/subbi.2019.2.06

QUANTITATIVE ANALYSIS OF STYLE IN MIHAI EMINESCU'S POETRY

ANAMARIA BRICIU

Abstract. Quantitative stylistic methods aim to express certain aspects of a text in numeric form, thus allowing the introduction of fast, powerful and accurate computational approaches for analysis. While in the case of literature, the validity and usefulness of such studies is highly controversial, one cannot deny the opportunities brought forward by computational methods: first, the exploration of large sets of documents in search of patterns otherwise difficult to discover by human readers; second, the possibility of opening up new perspectives by uncovering latent features of texts. In this study, we investigate the poetic work of one of the most important Romanian poets, Mihai Eminescu, through a variety of quantitative methods addressing lexical, morphological, semantic and emotional aspects of text. We propose a comparison between the results of the computational approach and established interpretations of Eminescu's work in order to assess the viability of computational methods in poetic style studies.

1. Introduction

Computational studies of literary style have enriched the domain of literary criticism by quickly and efficiently analyzing large text corpora, and presenting readers with useful representations and visualizations. Poetry, in particular, has been the subject of a number of recent articles that explore perspectives such as poetic style, computational aesthetics or means of expressions regarding certain topics. The majority of studies, however, explore English and American literature.

In the present paper, we propose a quantitative analysis of style for one of the most famous Romanian poets, Mihai Eminescu, with elements of novelty in the study of the unique relationship between well-defined, unambiguous

Received by the editors: November 15, 2019.

2010 *Mathematics Subject Classification.* 68T10, 68T50.

1998 *CR Categories and Descriptors.* I.2.7 [**Computing Methodologies**]: Artificial Intelligence – *Natural Language Processing*; I.7.m [**Document and Text Processing**]: Miscellaneous.

Key words and phrases. text processing, quantitative analysis, poetry.

statistics and subjective, nuanced interpretations of poems in question. Having a large body of works available for reference with respect to Mihai Eminescu's writing style, we consider this analysis a worthwhile starting point for the investigation of the utility of computational methods in analyzing complex Romanian literature.

Word count methods, or quantitative measures of writing style are seen as features on the surface structure of the literary text that create certain aesthetic effects that provoke a reaction from the reader [13]. They have been widely contested by literary experts for being overly simplistic, artificial, unsubtle and incapable of generating any meaningful results. While it is certainly true that statistical methods lack in nuance, and hardly have the power of capturing artistic expression in its multifaceted form, they must be interpreted as guiding tools in informed literary endeavors rather than methods to search for a ground truth [14]. Moreover, they can be viewed as methods to take advantage of the large amount of digital literary texts available, thus enabling new modes of "reading" and analysis that summarize distinguishing features of very large corpora, something difficult to do in the case of traditional reading. Some experts even argue that computational approaches could bring added rigor and a degree of objectivity to the open question of interpretation and generalization [8]. John Burrows, in a novel study that used stylometry to analyze Jane Austen's works, famously stated that "...it is a truth not generally acknowledged that, in most discussions of works of English fiction, we proceed as if a third, two-fifths, a half of our material were not really there" [3, p. 1]. Consequently, the present computational analysis of style is not aimed at seeking absolute truths about Mihai Eminescu's work or simple, concise answers regarding his style - definitive solutions to the stated problem, as in most algorithmic approaches - but to prove that statistical methods can open up interesting, two-way debates between literary scholars and computer scientists by exploring and describing a representative corpus of Romanian poetry in an efficient and concise way.

This paper is structured as follows. Section 2 provides a brief overview of related works in the field. Section 3 includes a short description of Mihai Eminescu's writing style and the characteristics of each of his creation phases, while Section 4 details the methodology of the study, including information about the dataset used, the resources and tools employed, as well as the challenges met. Section 5 presents and discusses the results obtained, while the last section highlights conclusions and outlines directions for future work.

2. Related work

Quantitative studies of literature have a long history: first modern methods were pioneered in the 1850s, and, since then, the field has known periods of intense activity and evolution, especially in the last decades, when computational approaches were introduced. This includes broad topics, from simple information extraction from literary works to character trait identification and affect modeling in narratives. As far as computational poetry is concerned, there are a series of overlapping research directions: analysis of poetic style, creation of visualization tools, and poetry generation. We will cover the existing works in the first two cases as they relate to the present work.

2.1. **Computational analysis of poetic style.** SPARSAR [5] is a comprehensive system for automatic analysis of poetry style that makes use of computational linguistic methods. The developed system outputs syntactic, semantic and structural information about a poem, as well as affect and phonetic models, ultimately summarizing this data in seven complex indices that allow visual comparison between multiple works. Comparisons between linguistic styles of different poets are made in [10, 11, 22]. Kao and Jurafsky, for instance, examine elements of poetic craft such as imagery, sound devices, emotive language, and diction features to analyze differences between contemporary professional and amateur Imagist poets [10]. Interesting results have also been obtained when poems have been translated into a vector space [11],[22]. Kaplan and Blei represent a poem through a vector of stylistic features that include ortographic, syntactic and phonemic measures, while Zhang and Gao use the sum of the word embedding vectors for the most frequent terms in a poem as its representation. Such vector representations can further be used to either attempt classification, generally with the aim of distinguishing authors [12] or clustering [22].

2.2. **Poetry visualization.** Tools for text visualization can be extremely useful, especially in literary works, where some aspects of the work may not be immediately evident, but might emerge in carefully chosen visual representations.

The tool in [17], for example, offers a wide range of visuals like assessing unique, informative words in each poem, places, time periods, figures of speech, and sentiment expressed in each term and verse.

There are also works that propose similar tasks to ours, namely visualizations of author style through certain periods of time [10, 5]. In the same sense, there are studies that focus on Mihai Eminescu's writing, but they either address different aspects of his style (such as specific semantic units [20]) or target other types of writing, such as journalistic articles [4].

3. Mihai Eminescu's work: Brief Theoretical Overview

Mihai Eminescu was a Romantic poet, novelist, and journalist. He is generally regarded as the most famous and influential Romanian poet, and considered the first modern poet in Romanian literature. His work is unique through the ways of artistic expression that do not necessarily conform to rigid norms, but explore a vast space between the communication of social and political messages and the intrinsic, reflexive state of individual reality transposed in art.

These means of artistic expression are extensively analyzed by literary critics, with entire books dedicated to overviews of Eminescu's poetic style [7] and to his use of language [9]. In this study, we will address poetic style analysis from a computational perspective, investigating quantitative measures of language. In particular, we will study the ability of computational methods to synthesize important facets of the stages in Mihai Eminescu's creation.

3.1. **Linguistic style.** The appeal of Mihai Eminescu's poetic language resides in its novelty and naturalness, which blends folkloric and familiar forms with expressions of high-order, intellectualised language. His entire poetic work is, in actuality, a vast composite model that in many ways transcends literary genres. L. Galdi, in his analysis of Eminescu's poetic style, argues that "a meditation written by Eminescu means more than any other meditation of the era; it comes so far from any epigonism [...] even a teen love poem like "De-aș avea" has personal touches that, from an affect perspective, would be looked for uselessly in some of Alecsandri's works that served as model" [7].

3.2. **Phases of creation.** Most literary critics separate Eminescu's work in three phases of creation, or three poetical and ontological visions. In fact, these phases can be interpreted as three types of realities proposed by the author, each with its own specific themes, motifs and means for poetic expression. We will focus on the three main time periods of 1866-1870 (Phase 1), 1870-1876 (Phase 2) and 1877 and later (Phase 3) [18]. This temporal separation, however, is not a rigid one. Some experts argue in favor of some intermediary, transitional stages [18, p. 433], and contextualize certain poems within a different phase that the date of its creation would recommend it for [18, p. 451].

In this study, we will assess quantitative measures of style in conjunction with the three main stages of Mihai Eminescu's poetic expression, making the membership of a poem to a specific phase group a crisp one, but take into consideration the classification in [18], overwriting automatic assignment of phase based on poem year where it is needed.

3.2.1. *Phase 1.* Chronologically speaking, the first phase refers to poems written between 1866 and 1870. In the majority of the poems written during this phase, a strong influence of the forty-eighters poets can be observed (e.g. Vasile Alecsandri, Ion Heliade Rădulescu, Dimitrie Bolintineanu), both in terms of topic and poem genre. Eminescu writes odes, satires and folklore-inspired poems where he makes use of both folkloric and highly intellectual poetic means, integrating archaic terms with neologisms seamlessly. Moreover, a predilection for comparisons and longer, ornery epithets is observed. This is meaningful in contrast to later works in which the dominant figure of speech is the metaphor, and epithets are simplified in favor of creating sharper, more emotional-heavy visual images.

3.2.2. *Phase 2.* Literary critics approximate the second phase of creation to range from 1870 to 1876. During this stage, the admiring poetic tone from the previous phase is abandoned, the author negating the fabulous, transcendental motifs such as the music and heart of the universe in favor of a more disillusioned, realistic perspective. The present becomes empty of meaning and essence, and there are only two ways in which the poet can battle this realization: first, the return to the idyllic time of childhood, characterized by a series of obsessively referenced nature related motifs (e.g. forest) and the use of verbs in imperfect tense which reflect his nostalgia; and second, a rebellion against the senseless universe. The vocabulary employed to outline such themes is a somber one - there are frequent references to shadows, darkness, blackness, the void, demonic sides, sadness, detachment and alienation, death, and oneiric colors like deep green, navy blue or off-white.

In this phase, the author creates something called compensatory universes, picturesque and pristine poetic worlds that alleviate the pain of living in the real one. There are four ways in which these illusory universes can be created: dreams, love, poetic art and history. Overall, the second phase is characterized by more abstract, suggestive language, in contrast to the rational discourse in the first phase. There are a number of poems in which emotional states are expressed with higher frequency than in the first, works that seem to be deeply personal, and invoke a series of sentiments and emotions: negativity, anger, disgust, sadness, anticipation, love.

3.2.3. *Phase 3.* The last phase of Eminescu's work is defined by a conscious approach to creation in which he abandons the romantic vision. Verses become simpler, poems lack figures of speech but involve more complex structure. At this stage, the main subject of the poems becomes the human condition. This is conveyed through the introduction of terms on both sides of the

"thought/action" semantic pair and temporal references with a similar contrast ("always","infinite" vs. "suddenly"), in the syntax of the use of verbal tenses that suggest an undetermined time (imperfect tense) and verb times that accentuate the pain of the present (indicative present).

4. Methodology

4.1. **Data.** For the data considered, we collected 339 poems from an available online source[1]. Of these, only works excluded were those that could not be definitively associated with a publication year. The number of poems published in each year can be observed in Table 1.

Number of poems per year							
Year	# poems	Year	# poems	Year	# poems	Year	# poems
1866	10	**1872**	14	**1878**	27	**1884**	2
1867	8	**1873**	25	**1879**	36	**1885**	1
1868	4	**1874**	16	**1880**	18	**1886**	2
1869	19	**1875**	5	**1881**	14	**1887**	3
1870	9	**1876**	62	**1882**	12		
1871	12	**1877**	11	**1883**	29		

Table 1. Number of poems per year

4.2. **Tools and Resources.** The tools and resources used in this study will be described in this section.

4.2.1. *RoEmoLex.* RoEmoLex (Romanian Emotion Lexicon) [2, 15, 16] is a resource developed for text-based emotion detection in Romanian language and it contains 9177 terms annotated with eight primary emotions (*Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust*) and two polarity tags (*Positivity, Negativity*). Moreover, each term has part-of-speech information associated. Of the 9177 terms in the lexicon, we take into account only 8486, eliminating multiple word idioms from consideration due to the difficulty of matching highly irregular poetic language to template expressions. We use the RoEmoLex database to compute emotion features for each poem.

4.2.2. *RoWordNet.* RoWordNet (Romanian WordNet) [21, 6] is a semantic network for the Romanian language that mimics Princeton WordNet, a large English lexical database. The basic unit in WordNet is a synset, which expresses a unique concept and contains a number (a "set") of synonym words known as literals. All synsets have additional properties such as part-of-speech,

[1]https://ro.wikisource.org/wiki/Autor:Mihai_Eminescu

definition, conceptual category and relationship information with regard to other synsets (i.e. semantic relations like hypernymy ("is-a"), meronymy ("is-part"), antonymy etc). We use the RoWordNet python API[2] to search for a term's number of senses and relation to other synsets (hypernymy hierarchy).

4.2.3. *NLP Cube.* NLP-Cube [1] is an open source natural language processing framework that supports tasks such as sentence segmentation, tokenization, part-of-speech tagging, lemmatization and dependency parsing for a variety of languages. We use the NLP-Cube python API[3] with a Romanian language model for the following tasks: sentence segmentation, tokenization and part-of-speech tagging.

4.2.4. *UAIC Romanian Noun Phrase Chunker.* The UAIC Romanian Noun Phrase Chunker [19] is a complex tool which recursively detects and annotates noun phrase chunks for Romanian text. Noun phrase (NP) chunking is defined as a partial parsing task that generates an output of the nominal groups in a text, i.e. the units for which the principal word (head) is a noun. We use the web service[4] provided by the UAIC Natural Language Processing Group for our noun phrase related features: average length of noun phrases, number of noun phrases in a poem, types of noun phrases in a poem.

4.3. **Features.** We propose a number of features distributed across semantic, lexical, syntactic and affective perspectives as follows.

(1) **Number of tokens in poem.** Represents the total number of tokens in a poem. This includes stopwords[5] but excludes punctuation.
(2) **Average word length.** Represents the average length of a word in character units. Approximates vocabulary complexity on the simple assumption that longer words are more difficult ones.
(3) **Type-token ratio.** A metric conventionally used as proxy for vocabulary richness [10], defined as the number of unique words in a text divided by the number of all words in the text.
(4) **Hapax Legomena.** Some researchers argue that the type-token ratio is not a sufficient metric for assessing vocabulary richness, and introduce features that count hapax legomena [5]. A hapax legomenon is a word that occurs only once within a context, either in the written record of an entire language, in the works of an author, or in a single text[6]. It speaks to the importance of rare words in a corpus, and

[2]https://github.com/dumitrescustefan/RoWordNet
[3]https://github.com/adobe/NLP-Cube
[4]http://nlptools.info.uaic.ro/WebNpChunkerRo/NpChunkerRoWS?wsdl
[5]very common words in a given language, usually connectives
[6]https://en.wikipedia.org/wiki/Hapax_legomenon

can be used to measure vocabulary growth across different stages of creation [5].

(5) **Part-of-speech density.** The frequencies of parts of speech reflect a poet's mode of discourse [11]. These measures are defined as the number of terms belonging to a certain part-of-speech category divided by the total number of words in the poem. In this study we considered the following categories: nouns, adjectives, verbs, adverbs, and pronouns. For pronouns, we have also investigated the ratio of first, second, and third person pronouns with respect to the total number of terms in the pronoun category for greater specificity. As for verbs, we look at the use of different tenses (present, past, imperfect).

(6) **Abstract and concrete ratios.** Poetry is a type of text dense in imagery and suggestion. It makes sense, then, to define features that assess these aspects [10]. The abstract to concrete feature value is defined as the ratio between the number of abstract concepts and the number of terms referring to concrete terms. To compute this value, we traverse the RoWordNet hypernym hierarchy of each word in a poem [12]. If a hypernym with a semantic category of 'Abstract' is found, then the number of abstract concepts is incremented. Conversely, if we find a hypernym with semantic class of 'Physical' or 'Object', we increment the concreteness count. The ratio is computed by dividing these two values.

(7) **Valence and emotion ratio.** For the two valences and eight emotions for which tags exist in RoEmoLex, we have counted the number of words in each emotion and valence class in each poem and divided it by the length (in words) of the poem.

(8) **Epithets: Noun phrase information.** With the help of the UAIC Romanian Noun Phrase chunker, we compute the number of noun phrases, their average length and the count of different part-of-speech associations that make up a noun phrase for each poem.

4.4. **Archaic terms and unusual word forms.** Mihai Eminescu's poetic language is unique in that it combines literary language with popular one seamlessly. In a given poem, both neologisms and archaic terms can be found - the latter a significant challenge for natural language processing tools developed within the context of contemporary language. The main issue concerns word spellings determined by out of date language rules ("î" instead of "â" inside a word), which, more often than not, are not recognized while processing.

What is more, the author takes many liberties in prefixing and suffixing nouns, adjectives and verbs to suit the desired poetic construction [9]. These

unusual inflectional forms also pose a problem for the processing tools in the sense of lemmatization and part-of-speech tagging. Lastly, we note that some hapax legomena encountered in the text might not be true entities of this kind, due to interchangeable spellings. There may be situations where the tool used was unable to count the different versions of the same word - though in this case, it could be argued that the intentionality in choosing a particular form on the poet's part should not be ignored.

4.5. **Sentence segmentation.** In analysing a corpus of poetry, one must answer the question of the basic unit to be studied: a verse, a sentence or a stanza, taking into consideration the way the poet intended to delineate meaning, which is not always clearly deductible. Additionally challenging is the fact that there may be ambiguous sentence delimitation in such artistic constructs.

4.6. **Word order.** Syntactical parsing of poem content faces the challenge of unusual word order - inversions, repetitions and different constitutive arrangements for elements of a sentence from those in common language are frequent, employed as artistic devices for building musicality of verses, maintaining rhyme and meter, and nuancing the emotional tone of the poem. While they certainly count for aesthetic effect, these specificities of poetic language might lead to incorrect results in syntactic parsing.

4.7. **Word Disambiguation.** In working with tools like RoWordNet or RoEmoLex, which distinguish between word senses, one would wish also identify sense in poem terms for accurate results. However, simple to implement methods like the Lesk algorithm do not yield satisfying results, since the decision is based on surrounding word context. In poems, it is difficult to assess the window size for this context and to obtain any matches, for reasons previously detailed: unusual structural expressions in poetic language and uncommon word forms.

5. Results and Discussion

Stylistically-wise, Mihai Eminescu's work can be separated in three phases of creation, albeit in a slightly fuzzy manner [18]. We have covered theoretical aspects of these stages in Section 3.2. In the following sections, we will present results for a quantitative analysis of the corpus in relation with these three phases of creation.

5.1. **Vocabulary.** We have approached the task of vocabulary examination by taking into account only content words (nouns, adjectives, adverbs, verbs). Table 2 presents the number of poems and the total number of content tokens discovered in each phase.

	# poems	# tokens
Phase 1	44	5843
Phase 2	158	40669
Phase 3	137	20919

TABLE 2. Number of poems and content tokens in creation phases

It can be observed that Phase 1 contains the least poems and smallest number of tokens, while Phase 2 contains the most poems and tokens. It is interesting to note that while the difference in number of poems between Phase 2 and Phase 3 is relatively small, the number of tokens is almost double in Phase 2. This is due to the higher number of long poems (over 2000 tokens) in Phase 2 (9 poems) versus Phase 3 (2 poems).

As far as frequent words are concerned, we take the percentage represented by the total count of a given word in a phase with regards to the total number of tokens in that phase. We find a more frequent occurrence of the words *suflet/soul*, *floare/flower*,*inimă/heart*, *dor/longing*, *amor/love*, *dulce/sweet* and *alb* in the first phase of creation, which references the high number of joyous love poems written in this stage. Conversely, *umbră/shadow* and *negru/black* have a slightly larger presence in the second phase, suggesting darker emotional states and emotionally heavier subjects. For the third phase, we draw attention to the terms *trece/to pass*, *ști/to know* and *lung/long*. In this phase, Mihai Eminescu's poems address the passing of time and the place of a meditative man within a hostile historical context, concepts built upon terms like "to pass", "to know", "long". Lastly, there are terms which we interpret as motifs, such as *lună/moon*, *cer/sky*, *val/wave*, *stea/star*. They appear with the same density throughout the years, under different meanings.

Equally interesting can prove to be the analysis of vocabulary richness. Inspired by the work in [5], we choose three measures to examine this aspect: VR1, equal to the mean type-token ratio for each phase, HA1, the percentage of hapax legomena in a phase with respect to the total number of unique content tokens in that phase, and HA2, the number of hapax legomena unique to the phase in question (i.e. not present as hapax legomena in other phases) divided by the total number of hapax legomena in the phase.

As it can be observed in Table 3, the vocabulary richness measures have similar average scores throughout the three phases: the mean type-token ratio is at around 75% in all stages, while the percentage of hapax legomena with

	VR1	HA1	HA2
Phase 1	0.74	0.67	0.70
Phase 2	0.73	0.61	0.84
Phase 3	0.75	0.63	0.75

TABLE 3. Vocabulary richness measures for each phase

respect to the total number of unique content tokens varies between 60% and 70%. Lowest values are recorded for Phase 2, which is explained by the length of the poems written during these years, considerably greater than in the other phases. However, it must be noted that the second hapax legomena measure (HA2) - the percentage of terms that appear only once in a phase, and only in that phase - is highest for Phase 2. This points to the exploration of the different poetic visions at this creation stage. While in Phase 3, Eminescu's work shows maturity and is more a concentrated synthesis of his ideas, the second phase feels more like a search for answers. The compensatory universes he creates might all have the same aim, but they vary in vocabulary and tone depending on founding idea (dreams, love, history, artistic expression).

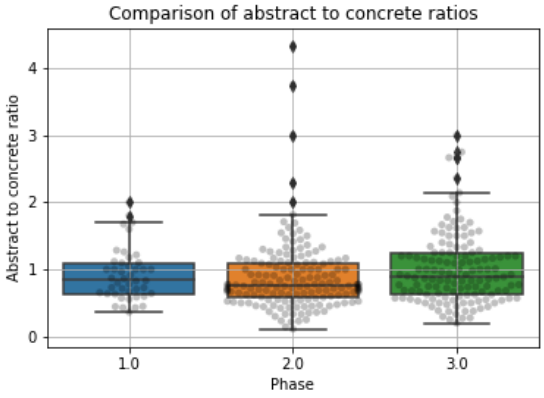

FIGURE 1. Comparison of abstract to concrete ratios in different phases of creation

As for the semantic measure of abstract to concrete concept ratio, we found very small differences between phases, as can be seen in Figure 1. While the median ratio is slightly higher for the third stage, it is a too small difference to draw any definitive conclusions in support of the theoretical observation that third-phase poems approach more abstract topics like *thought versus action*, *human condition* or *the role of the divine*. Abstract terms such as *dream*,

longing, *glory*, *sigh*, *self*, *future*, *flight* are recurrent concepts throughout the whole corpus, with the poet building a number of different meanings around them. Alternatively, concrete references such as *gold*, *wind*, *butterfly*, *sun*, *face*, *hand*, *eye* are symbols and elements of imagery that no phase lacks.

5.2. **Morphology and syntax.** With respect to morphology and syntax, we look at the average part-of-speech densities per phase. For pronouns, we also compare ratio of first, second and third person pronouns with respect to all pronouns, while for verbs, we examine the use of different tenses.

	1st person	2nd person	3rd person
Phase 1	0.17	0.15	0.68
Phase 2	0.23	0.19	0.58
Phase 3	0.25	0.15	0.60

TABLE 4. Average ratios of first, second and third person pronouns per phase

As far as part-of-speech density is concerned, the average ratio of nouns, adjectives, verbs, adverbs and pronouns is almost identical in each phase. Small differences can only be observed in the case of nouns (average of 0.28 in Phase 1 versus 0.25 in Phase 2 and 0.24 in Phase 3), which can be interpreted as a product of the high number of epithets that is present in earlier works. As for pronouns, the number and frequency of third person forms far surpasses that of first and second person uses, even though there is a high number of love poems where the relationship between the poet and the loved one is built on a "I"-"you" linguistic relation [9]. As it can be seen in Table 4, the distribution of percentages differs slightly in the first phase, which can be attributed to the more personal type of poetry of the later phases, as opposed to an overwhelming amount of scenery descriptions in the earlier works.

For verbs, results match theoretical observations as well, but in a similarly subtle manner. Overall, verbs in the present tense are the most frequent (60%), followed by forms in past tense (9-15%) and imperfect (3-7%), in that order. The highest average ratio of imperfect forms can be found in Phase 2 (7%), which suggests the idea of undetermined time, that of the idyllic, compensatory worlds Eminescu constructs in this phase, while past tense can be found most often in the first phase (15%), which may be explained by the fact that this stage of creation contains a number of odes and history-themed poems.

Therefore, in the case of morphological and syntactical features, results are often less conclusive, and more challenging to interpret. However, some subtleties of the author's writing can be translated even in a quantitative

analysis, as shown by the differences between first and third person pronouns in early and later phases, and the values obtained in the case of verbs at imperfect tense.

5.3. **Emotions.** Analysis of emotions in Mihai Eminescu's work with respect to phases yields no clear emotional profile for a creation stage, with mean percentages of emotion almost identical throughout. However, there are some interesting results with regards to the maximum value for percentage of emotion in each phase, as it is shown in Table 5.

	Anger	Anticipation	Fear	Disgust	Joy	Sadness
Phase 1	0.08	0.17	0.13	0.04	0.14	0.14
Phase 2	**0.17**	0.18	**0.17**	**0.09**	0.16	**0.22**
Phase 3	0.13	**0.30**	0.21	0.08	**0.23**	0.18

TABLE 5. Maximum value for percentage of emotion in each phase

The poem with highest percentage of Anger is *"Venin și farmec"* (Phase 2), while *"E ceasul cel de taină"* (Phase 3) has both the highest percentage of Anticipation and of Fear. As for Disgust and Sadness, the highest percentages can be found in Phase 2, in the short poem *"De ce mă-ndrept ș-acum"*, and *"Îngere palid"*, respectively. Finally, the maximum value for the emotion Joy is obtained for a Christmas poem in Phase 3, named *"Colinde, colinde"*.

In our analysis, emotion and valence as measured by frequency of occurrence do not seem to be distinguishing features for the three phases of creation in Mihai Eminescu's work. Therefore, we propose extending the investigation into this aspect in future works by considering trajectories of emotion in poems and other finer-grained emotional features.

5.4. **Stylistic devices.** As far as figures of speech like epithets are concerned, on a simple count of their occurrence, we find that the phase with most such stylistic devices is Phase 2. However, this result should be interpreted in the context of poem length at this stage, given that the number of figures of speech is proportional to the length of the poem. A more informative measure may be the average length of noun phrases, which is highest in the first phase and recording a decrease in later stages, from 2.3 words to under 2. This is in line with the observed presence of long, ornery epithets in Phase 1, and a simplification of poetic expression at stages of maturity.

Finally, we examine the five most common types of noun phrases in each phase, shown in Table 6. They are the same in each phase, but the hierarchy differs slightly. For instance, the most frequent composite phrase in each phase is made up of two nouns (N+N), but a noun followed by a pronoun (N+P) is second most common only in Phase 1.

	N+N	N+P	N+ADJ	ADJ+N	N+ADP+N	ART+N
Phase 1	207	159	156	60	122	97
Phase 2	1370	889	1097	391	916	673
Phase 3	691	383	472	178	480	332

TABLE 6. Most common types of noun phrases in each phase

What is interesting to notice is that there is a high number of nouns followed by adjectives (N+ADJ) in all phases, and, if we also consider the reversed order (ADJ+N), types which are closest to a formal definition of epithets, the total surpasses any other values for Phase 1 and 2. The exception is Phase 3, where a phrase consisting of two nouns is still more common than one with an adjective and a noun. There is also a considerable amount of nouns linked by adpositions (N+ADP+N), especially in later phases. While using adjectives in proximity to nouns is a device used for characterization and intrinsic poetic expression, the second type, which incorporates adpositions, has a more functional role in building and structuring imagery.

6. Conclusions

In this paper, we have presented a computational analysis of a representative Romanian poetry corpus, the poetic work of Mihai Eminescu. We have examined a series of features addressing vocabulary richness, language complexity and emotional content in each phase of artistic creation. Results show that for a series of measures, theoretical observations from the literary criticism field find correspondence in quantitative analysis, indicating that this approach, and visualization of results, in particular, could be used as support tool for interpretation.

However, while computational analysis of literary works and especially poetry is a promising research direction, there are a few challenges to be considered. First, the choice of features must be informed by the domain to be truly relevant, and so must interpretations. The latter, in particular, require knowledge from both the literary linguistics domain, and a familiarity with the considered author's body of work. This is the reason we mark the task of making visualization of results public for future work; we consider that by inviting debate from informed readers of poetry, new, interesting interpretations could arise, or, on the contrary, features taken into account could draw attention to aspects not considered before.

Finally, the original contribution of this paper to the field was an in-depth quantitative analysis of the works of Mihai Eminescu and a close examination of the relationship between the phases of the author's artistic expression

and quantifiable aspects of language. In particular, the emotional and valence features we defined are unique to this study, and while results in this regard were not very conclusive with respect to considered task, we propose two research directions to further examine this exciting aspect: first, the standalone exploration of emotion in the corpus, and second, finer-grained measures that capture the nuances of expression throughout a poem, such as trajectories of emotion in a text.

References

[1] T. Boroș, S. D. Dumitrescu, and R. Burtica. NLP-cube: End-to-end raw text processing with neural networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 171–179, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[2] A. Briciu and M. Lupea. RoEmolex - a Romanian Emotion Lexicon. *Studia Universitatis Babeș-Bolyai Informatica*, 62(2):45–56, 2017.

[3] J. F. Burrows. *Computation into criticism: A study of Jane Austen's novels and an experiment in method*. Clarendon Press, 1987.

[4] M. Dascalu, D. Gîfu, and S. Trausan-Matu. What makes your writing style unique? Significant differences between two famous Romanian orators. In *International Conference on Computational Collective Intelligence*, pages 143–152. Springer International Publishing, 2016.

[5] R. Delmonte. Computing poetry style. In *Proceeding of Emotion and Sentiment in Social and Expressive Media: Approaches and perspectives from AI (ESSEM 2013), CEUR Workshop*, pages 148–155, 2013.

[6] S. D. Dumitrescu, A. M. Avram, L. Morogan, and S.-A. Toma. RoWordNet–a Python API for the Romanian WordNet. In *2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–6. IEEE, 2018.

[7] L. Gáldi. *Stilul poetic al lui Mihai Eminescu*. Editura Academiei Republicii Populare Române, 1964.

[8] A. Hammond, J. Brooke, and G. Hirst. A tale of two cultures: Bringing literary analysis and computational linguistics together. In *Proceedings of the Workshop on Computational Linguistics for Literature*, pages 1–8, 2013.

[9] D. Irimia. *Limbajul poetic eminescian*. Junimea, 1979.

[10] J. T. Kao and D. Jurafsky. A computational analysis of poetic style. In *LiLT (Linguistic Issues in Language Technology)*, volume 12, pages 1–33, 2015.

[11] D. M. Kaplan and D. M. Blei. A computational approach to style in American poetry. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 553–558. IEEE, 2007.

[12] V. Kesarwani. *Automatic Poetry Classification Using Natural Language Processing*. PhD thesis, University of Ottawa, 2018.

[13] M. Kestemont and L. Herman. Can machines read (literature)? *Umanistica Digitale*, 3(5), 2019.

[14] M. G. Kirschenbaum. The remaking of reading: Data mining and the digital humanities. In *The National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation, Baltimore, MD*, 2007.

[15] M. Lupea and A. Briciu. Formal Concept Analysis of a Romanian Emotion Lexicon. In *13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 111–118, 2017.
[16] M. Lupea and A. Briciu. Studying emotions in Romanian words using Formal Concept Analysis. *Computer Speech & Language*, 57:128 – 145, 2019.
[17] L. Meneses and R. Furuta. Visualizing poetry: Tools for critical analysis. *The Journal of the Initiative for Digital Humanities, Media, and Culture*, 3(1):1–14.
[18] I. E. Petrescu. *Studii eminesciene*. Casa Cărții de Știință, 2007.
[19] R. Simionescu. Romanian deep noun phrase chunking using Graphical Grammar Studio. In *Proceedings of the 8th International Conference Linguistic Resources And Tools For Processing Of The Romanian Language*, pages 135–143, 2012.
[20] D. Tatar, M. Lupea, E. Kapetanios, and G. Altmann. Hreb-like analysis of Eminescu's poems. *Glottometrics*, 28:37–56, 2014.
[21] D. Tufiș and V. Barbu Mititelu. *The Lexical Ontology for Romanian*, volume 48 of *Text, Speech and Language Technology*, pages 491–504. Springer, 2014.
[22] L. Zhang and J. Gao. A comparative study to understanding about poetics based on natural language processing. *Open Journal of Modern Linguistics*, 7:229–237, 2017.

Babeş-Bolyai University, Department of Computer Science, 1 M. Kogălniceanu Street, 400084 Cluj-Napoca, Romania

Email address: anamaria.briciu@cs.ubbcluj.ro