Language Resources and Evaluation
https://doi.org/10.1007/s10579-023-09689-6

ORIGINAL PAPER

A corpus of Persian literary text

Shahab Raji[1] · **Malihe Alikhani**[2] · **Gerard de Melo**[3] · **Matthew Stone**[4]

Accepted: 22 August 2023

Abstract

Persian poetry has profoundly affected all periods of Persian literature and the literature of other countries as well. It is a fundamental vehicle for expressing Persian culture and political opinion. This paper presents a corpus of Persian literary text mainly focusing on poetry, covering the ninth to twenty-first century annotated for century and style, with additional partial annotation of rhetorical figures. Our resource is the largest and the most diverse corpus available in Persian literary text, with a particularly broad temporal scope. This allows us to conduct several computational experiments to analyze poetic styles, authors and time periods, as well as context shifts over time, for which we rely both on supervised models and on Persian poetry-specific heuristics. The corpus, the tools, and experiments described in this paper can be used not only for digital humanities studies of Persian literature but also for processing Persian texts in general, as well as in other broader cross-linguistic applications.

Keywords Corpus compilation · Classification · Literary text · Rhetorical figures · Data cleaning

Shahab Raji and Malihe Alikhani have contributed equally to this work.

✉ Shahab Raji
sraji@ea.com

✉ Malihe Alikhani
m.alikhani@northeastern.edu

Gerard de Melo
gdm@demelo.org

Matthew Stone
mdstone@cs.rutgers.edu

1 Electronic Arts, Redwood Shores, California, USA

2 Khoury School of Computer Science, Northeastern University, Boston, Massachusetts, USA

3 Hasso Plattner Institute, University of Potsdam, Potsdam, Germany

4 Department of Computer Science, Rutgers University, Piscataway, New Jersey, USA

Published online: 23 November 2023

1 Introduction

1.1 Motivation

Compiling corpora for low-resource languages is a valuable undertaking for preservation, education, knowledge acquisition, and monitoring demographic and political processes. The focus of this work is on presenting a large machine-readable corpus of Persian literary text suitable for studying a variety of NLP problems in Persian, including lexical semantics, authorship prediction, style classification, and computational approaches for studying rhetorical figures and metaphor. Further, since the corpus covers the majority of available Persian poems across over fourteenth centuries, the corpus facilitates computational studies to track meaning shifts over time.

Poetry explores the space of imagination beyond linguistic interpretation and pragmatics (Kadkani, 1943; Atashi, 2004), yet brings distinctive insights (Hobbs, 1990). Persian poetry, in particular, has not only played a profound role in shaping Persian culture, politics, and literature, but has a pervasive influence on world literature (Mohaqeqi et al., 2014; Tusi, 2013). Even in the United States, *Rumi*, *Khayyam*, and *Hafez* are among the best-selling poets.

1.2 Previous work

While there has been substantial computational research on poetry in English (Lau et al., 2018; Greene et al., 2010; Genzel et al., 2010; Hayward, 1996), Chinese (Zhang et al., 2017; Liu et al., 2019), and German (Baumann et al., 2018), among others, Persian poetry has not been widely studied in computational linguistics. The works that do exist point to the need for more comprehensive and systematic resources: For instance, Asgari and Chappelier (2013), Asgari et al. (2013) apply topic modeling to Persian poetry but work with a collection of works by 30 poets but only in one style, *Ghazal*. The corpus presented here covers those datasets as well as other styles. Other studies (Malmasi & Dras, 2015; Seraji et al., 2012; Khashabi et al., 2021) focus solely on contemporary Farsi language.

1.3 Contributions

We introduce a corpus of Persian literary text that encompasses poems from the ninth to twenty-first centuries as well as the two main collections of myths and stories *Gulistan* and *Panchatantra*. Gulistan Saadi includes a mix of poetry and prose. These collections are essential for designing basic models for processing literary text, such as spell checkers and temporal and structural analyses. Table 1 provides essential statistics on the corpus.

In addition, in order to study the symbolism and rhetorical figures in Persian, we annotate 4192 lines of poetry with six rhetorical figures. We explain these figures and their similarities to their English counterparts. We present detailed statistics about the corpus, and, in a series of computational experiments, we introduce a baseline

Table 1 Our corpus coverage

Item	Count
Poets	112
Collections[a]	614
Poems and prose	57, 980
Vocabulary size	260, 180
Poetry	
Lines[b]	1, 422, 501
Tokens	9, 144, 037
Prose	
Tokens	228, 501

[a] A *collection* (also known as, *Diwan*) is a compilation of several poems by a poet that is collected in one volume
[b] Number of lines are only counted for poetry

for classifying authors and styles. Finally, we present a study on semantic shifts in different eras of Persian poetry.

While our experiments focus on poetry, some collections consist of a mix of prose and poetry, and keeping only the poetry would lead to incomplete collections. Thus, including prose is useful for completeness and may be beneficial for certain kinds of temporal and structural analyses. Also, we made deliberate choices in selecting the most prominent styles and poets for the classification experiments. Our criteria are based on factors such as popularity, as well as differences in content and intent among the selected poets. By taking this approach, we aim to present a balanced representation that can be applicable to cross-linguistic studies while avoiding excessive information specific to Persian poetry.

1.4 Outline

The rest of the paper is organized as follows. In Sect. 2, we explain the data collection and normalization process used to compile a clean, well-organized corpus of Persian literary text from the ninth to twenty-first century. In Sect. 3, we describe the human annotation process used to endow our corpus with annotations for century, style, author, and a rich set of rhetorical figures. To demonstrate the merits of the corpus, in Sect. 4, we present a series of computational analyses that offer new techniques to investigate literary developments over time and present an open-source library for style classification. Additionally, we describe a series of classification experiments to show the distinction between the authors and different periods.

2 Corpus compilation

We had to overcome several challenges to make this corpus possible. The first of these is the limited availability of relevant source texts on the web. To complete some of

the collections and annotations, we had to obtain access to a diverse set of resources. Second, cleaning and normalizing online text in Farsi requires correcting for various keyboard layouts, including Arabic. We release Python code with this submission for cleaning and normalizing diverse forms of Persian text,[1] as most existing libraries are designed only for modern text. Third, annotating literary text is expensive and requires expert annotators. We have annotated part of the corpus with key rhetorical figures such as metaphor with high reliability following carefully designed protocols.

2.1 Crawling

To compile a comprehensive corpus with poems and stories spanning from the ninth to twenty-first centuries, we had to crawl multiple sources and request access to online teaching material to put together pieces and make collections complete. Our corpus was collected by crawling several Persian literary websites.[2] The released version of our corpus does not include poems from the twenty-first century due to copyright concerns. However, we include modern poetry in our experimental analyses in Sect. 4. We have obtained all required permissions to release the rest of the corpus publicly.

2.2 Linguistic challenges

Persian is an Indo-European language with a comparably rich morphology, conventionally written in Arabic script. The language poses a number of special challenges. Orthographic variability results from the frequent omission of vowel diacritics, alternative encoding of characters, and diverse shapes for affixes. Additional challenges include morphological complexities in the inflectional paradigms for nouns, verbs and adjectives, which involve multiple stems for verbs, and irregularities in nouns and verbs borrowed from Arabic. The language allows free word order, with a default of SOV.

2.3 Normalization and cleaning process

To ensure the quality of the data, we relied on both automated means and manual corrections. An important aspect of this is orthographic normalization. For this task, first we tokenized each line using space and punctuation characters as separators. We developed a tool to normalize Persian text in accordance with the list of undesired forms proposed by the Academy of Persian Language and Literature.[3] According to these, the use of certain letters and letter combinations, imported from general Arabic and Western usage, is deemed incorrect. We have replaced such instances with the standard form of Persian letters.

Arabic characters that are represented with alternative Unicode codepoints have been replaced with their correct Persian form. Such cases typically are the result of

[1] Code: https://github.com/pithysr/persian-poetry.

[2] The material was collected from https://ganjoor.net, http://shereno.com/, http://shamlou.org/.

[3] http://apll.ir/.

using an Arabic keyboard layout. For instance, *Arabic Letter kaf* (U+0643) is replaced with *Arabic letter keheh* (U+06A9) and *Arabic letter high Hamza Waw* (U+0676) is replaced with *Arabic letter Waw with Hamze above* (U+0624).

Left-to-right and Right-to-left control characters are removed, since the script is always right to left. We did not remove *Zero width non-joiner* (U+200C), since it is commonly used in Persian for inflected adjectives and nouns, as well as for morphological changes in verbs to show tense, aspect, and mood of the verb.

We are distributing the normalized version of the data (Raji et al., 2023). While normalization tools for Farsi are readily accessible, our approach prioritizes ensuring a high level of quality of the data instead of relying on such tools. Considering the limitations of most spell-checkers that are tailored to contemporary Farsi text, these tools may not adequately handle older prose or poetry.

Moreover, while the large size of the corpus makes manual proofreading prohibitively costly, we have corrected instances of misspellings as we came across them while working with the data. After cleaning, the vocabulary size is reduced by around 3000.

3 Corpus annotation

The corpus includes the poet and century as metadata. We annotated each poem based on an inventory of styles and temporal periods. Additionally, we partially annotated the corpus with six rhetorical figures commonly used in Persian poetry.

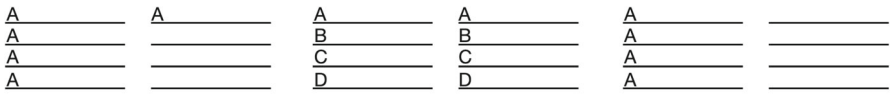
3.1 Style, author, and temporal annotations

3.1.1 Background on Persian poetry styles

Classical style Classical Persian poetry is classified into conventional styles in part based on different metrical and structural features. The prosody of classical Persian verse is based on the line, called a *bayt*, which consists of half lines that are metrically identical (Tousi, 1974; Tabatabai, 2001; Shamisa, 1999; Perry, 2011). Figures 2 and 3 from Sect. 3 show examples of two lines of a poem. The half-lines are usually written side by side.

Different styles of classic poetry are usually distinguishable by examining the rhymes of the words in the half-lines. Some of the styles have a similar pattern of rhymes, in which case the content and topic of the poem may be used to classify them.

Figure 1 illustrates the differences and similarities in the four most prevalent styles in Persian poetry. Note that Persian has a right-to-left script. The letters at the end of each half-line indicate the position of rhyming words or phrases, which signal different styles. As shown, in *Masnavi* each pair of half-lines have their own rhyme, while in *Ghazal* the rhymes occur at the end of each line. Hence, these two styles can be distinguished just by the positions of their rhymes. In contrast, the position of the rhyming words in the two styles of *ghazal* and *qaside* are the same. However, a *qaside* could be used by religious poets as an educational poem, whereas *ghazals*



(a) *Ghazal* and *Qaside* (b) *Masnavi* (c) *Qet'e*

Fig. 1 Structure of rhyming words in different classical style. The letters indicate the position of the rhyming words

(literally: love-songs) are much shorter poems, adopted by mystic poets and Sufis as a medium for the expression of love for the divine. From the fourteenth century CE, Persian poets became more interested in ghazals, and the qasida form declined.

Modern poetry Poetry remained a prominent form of literature in Iran through the twenty-first century. The modern style does not possess the features of the traditional styles and has different topics, content, and goals. Poems are not confined to the two half-lines format and the rhetorical figures, themes, metrics, and prosody are different, making this style easily distinguishable from classical poetry.

3.1.2 Annotation process

We manually labeled the collections with their styles. For this annotation, each collection was labelled by two annotators. The annotators and adjudicators are all adult native Farsi speakers, expert linguists, and completed at least two undergraduate classes in Persian literature and poetry.

Most of the collections have different chapters with different styles and the most prominent collections were annotated with corresponding styles. While our paper does not delve into the intricacies of Persian poetic meters that define fine-grained styles and genres, we have focused on "Masnavi", "Ghazal", "Qaside", and "Modern" styles for our annotations. The inter-rater agreement for style is almost perfect ($\kappa = 0.96$). We study this data in further detail in Sect. 4.

We also included the name of the poet and centuries following classifications proposed in the humanities (Safa, 1993; Browne, 1999; Rypka, 2013). This information was taken from the metadata available in the original sources. Table 2 shows the distribution of poets across different centuries.

3.2 Rhetorical figure annotations

Persian poetry uses rich symbolism (Seyed-Gohrab, 2011), and from the very beginning has extensively relied on a diverse inventory of rhetorical figures (Seyed-Gohrab, 2011; Arberry, 2008; Lewis, 2014; Meisami, 2014). Traditional figures of Persian rhetoric, when being analyzed in their stylistic function and expressive potential, are useful tools for describing the poet's style and imagery. While these devices and techniques are studied in several domains (Tom & Eves, 2012; Bush, 2012; Fengjie et al., 2016; García et al., 2018) in other languages, this is the first large resource available to study them in Persian.

Table 2 Distribution of number of poets, collections, and poems over time

Century	#Authors	#Collections	#Poems	#Tokens
9th	10	8	1325	209, 739
10th	12	15	1092	292, 016
11th	5	98	3566	1, 136, 056
12th	13	78	11, 174	2, 174, 417
13th	3	102	11, 473	1, 877, 271
14th	1	24	1812	231, 679
15th	2	15	3224	258, 479
16th	12	27	7410	498, 199
17th	3	12	3450	730, 444
18th	2	13	324	32, 737
19th	4	29	2236	483, 770
20th	2	46	6133	739, 205
21st[a]	40	147	4400	708, 526

[a]Note that twenty-first century is not released with the corpus due to copyright concerns

We chose the collection of *ghazals* by *Hafez*, which consists of 4192 lines of poems. We choose to work with this collection since the poems are among the most complex *ghazals* yet also the most popular in Iran, Afghanistan, and Tajikistan. This collection was chosen because of its rich symbolism and the important role of Hafez in Persian literature and culture. The analysis in Asgari et al. (2013) is on the same poetry style.

3.2.1 Background on rhetorical figures

There are many rhetorical figures in Persian. Here, we briefly explain the six rhetorical figures used in this study. The decision to use these figures is based on their frequency in Persian poetry and their similarity to figures used in English. Note, however, that the English counterparts may have some differences with the forms as used in Persian.

Iham The term *Iham* literally means creating doubt. or making one suppose. It refers to a deliberate use of lexical ambiguity, whereby the poet employs a word with two different meanings and arranges the context surrounding it in a way that one meaning is more immediate and the other remote, yet both can make sense.

As an example, the word "پرده" (curtain) refers to two concepts; 1. the penetralia, the most intimate part of the house, and 2. two semitones in traditional Persian music. This rhetorical figure is designed in a way that its direct meaning is the first thing that comes to the mind but really the remote meaning is intended.

Majaz (metonymy) *Majaz* refers to when only the related meaning of the word is intended. For example, when a part of a whole is used instead of the whole, such as when "سر" (head) is used instead of *person*.

Esteara (metaphor) *Esteara* refers to a form of metaphor based on similarities of with the intended subject, where the subject is removed, for example, شیر (lion) in certain contexts means a brave warrior. Or آتش (fire) can, metaphorically, describe the sorrow of losing the lover. In example (1), *musk deer* is a metaphor for the poet's lover.

(1) یا رب آن آهو-ی مشکین به ختن باز رسان
O God that deer-CONJ musk to Khotan return- IMP
O God, guide the musk deer back to Khotan.

Kenaya This refers to a particular kind of *esteara*, which we annotate separately due to its particular prominence, including in everyday language. A *kenaya* is a form of allusion employed when being indirect is deemed more polite, appropriate, or preferable for other reasons. In example (1) earlier, *Khotan*, the name of an ancient city near Kashgar, China, is used as a *kenaya* for the hometown of the poet. In many cases, kenayas are phrases. Example (2) below means "having intentions to do something". Literal meanings of Examples (3) and (4) are to "draw on the water" and "to give to the wind". As a *kenaya*, they mean "to do something in vain" and "to waste", respectively.

(2) سر چیزی داشتن
head thing have-GER

(3) نقش بر آب زدن
sketch on water hit-GER

(4) به باد دادن
to wind give-GER

Tashbih (simile) In most of the cases, such similes in Persian poetry come with an explicit marker, such as "x like y" or "x as y". As in English, when the marker is removed, *tashbih* is very similar to *esteara* (metaphor). The top most frequent unigrams that are used with this rhetorical figure in *Hafez* poetry are: زلف (hair), جام جم (love), and دل (heart).

Jenas This device involves the use of words that are (or appear to be) derived from a common root, as in example (6), ناز (pronounced *nāz*) and نیاز(pronounced *nyāz*). There are two kinds of *jenas*: (a) when the two words are homonyms, and (b) when they are very similar in writing or pronunciation.

The examples below include pairs of words in a line that are similar in spelling or pronunciations. These words can appear in any order or anywhere within a line. Incidentally, in example (6), when the two words are used together رنگ : تعلق (*goblet of the king*) is a metaphor for the universe.

(5) ناز - نیاز
need - coquetry

(6) جام - جم
king - goblet

دوا-ی دردِ عاشق را کسی کو سهل پندارد
remedy-CONJ pain-CONJ lover CONJ who 3SG-CONJ easy think-3SG
The one who thought that the remedy of the lover's pain is easy

(1)

ز فکر آنان که در تدبیر درمانند درمانند
of idea those CONJ in plan remedy-be-3PL distress-INTR
Of remedy those who are in thought, from thought distressed are

Fig. 2 Example of a line with *jenas* rhetorical figure (1) in the Hafez *ghazal* collection. The highlighted words are homonyms, the first being a noun meaning "remedy" and the second a verb meaning "distress"

(7) پیمانه - پیمان
promise - chalice

A complete list of identified jenas is released along with the corpus. Figure 2 shows an example of *jenas* where two homonyms are used with different meanings.

3.2.2 Annotation process

For these annotations, the words or phrases corresponding to the rhetorical figures in each line of poem are labelled by two annotators.[4] The annotators and adjudicators underwent a substantial period of training on the relevant linguistic devices before conducting the annotation. In our annotation protocol, we ask the annotators to label the six rhetorical figures described above. We prepared an annotation platform for the annotators where they could choose among the defined rhetorical devices and explain their observations. Each line of a poem was annotated with up to six rhetorical figures. Figure 3 shows three rhetorical figures used in one line of poetry. In general, spans of text are annotated, and some figures require marking multiple spans.

Some of these figures are purely morphological (e.g., *jenas*), or the words and phrase do not have a second meaning (e.g., *tashbih*). For others, the annotation also provides the literal meaning of each word or phrase when needed.

For *tashbih*, the annotators marked the word that is described by another word in the poem. For example, in Fig. 3, "belonging" is described as being similar to "a paint" that covers whatever is underneath it. This relationship is labelled as تعلق:رنگ (belonging:color) in the annotations. For *jenas*, the annotators provide the pair of words in the poem that create the rhetorical figure. The poem in Fig. 2 is annotated as درمانند–درمانند (remedy–distress) for *jenas*. The words creating the rhetorical figure are highlighted in both poems.

Iham, *kenaya*, *esteara*, and *majaz* are more involved. For each word or phrase, the annotators provide the concept or concepts that is described by that word or phrase. Since the concept is not in the poem, it can be described in different ways. Hence, an adjudication process was used to select a consistent description after the initial annotation process. For *iham*, which are inherently ambiguous, the annotation does not indicate which meaning may be the intended one.

[4] The research has been approved by our institutional review board for human subject studies. The annotators were paid a rate of $20/h.

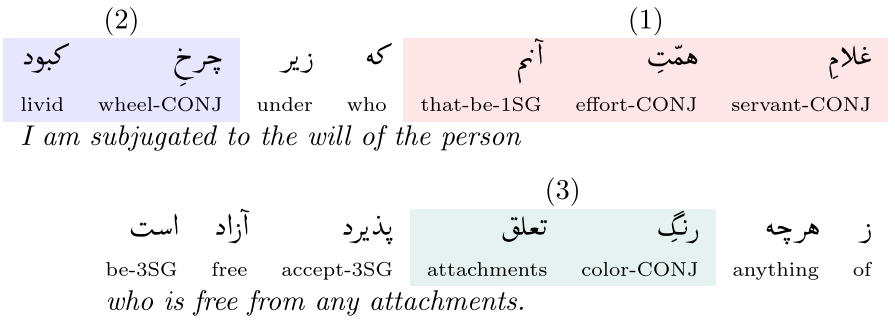


Fig. 3 An example of a line (two half lines) with *kenaya* (1), *esteara* (2) and *tashbih* (3) rhetorical figures in Hafez *ghazal* collection

Table 3 Distribution of different rhetorical figures in *Hafez* poetry

Rhetorical figure	Frequency (%)
Kenaya	19.9
Tashbih	13.3
Esteara	12.1
Iham	7.8
Jenas	5.7
Majaz	4.0

3.2.3 Annotation results

We chose the collection of *ghazals* by *Hafez* and present annotations for 4192 lines of poetry. The inter-annotator agreement was measured at the line level across a test set of 500 lines and resulted in a Cohen's κ score of 0.78 (average across the rhetorical figures), which indicates strong agreement between annotators. Table 3 shows what fraction of lines contain each of the rhetorical figures. We observe that *kenaya* is particularly common in the annotated data.

4 Experiments and analysis

In this section, we study how NLP techniques can be used automatically for informed explorations of Persian text at a variety of different analytical levels. We also provide detailed information about the distribution of texts over time, word counts, the average length of lines in poems in classic and modern texts, and more. However, we will focus our analysis on Persian poetry in this paper.

4.1 Style classification

In this section, we present several baselines for style prediction in Persian poetry.

4.1.1 Rule-based approach

We have implemented an open source tool that recognizes classical styles of Persian poetry, except for *ghazal* and *qaside*, using a rule-based algorithm based on formal features. As shown in Fig. 1, only these two styles of classical Persian poetry remain that cannot be distinguished using rules. Hence, the algorithm first looks at the positions of rhyming words and attempts to predict the style of the poem using simple rules.

4.1.2 Supervised learning

To distinguish between *ghazal* and *qaside*, we train a supervised model. We compiled a dataset of 1100 poems, 595 of which are *ghazals*, based on the most notable poets in each style,[5] We established a train–test split at a 80–20% ratio. As models, we consider a convolutional neural network (CNN) (Kim, 2014) as well as a linear SVM model. The CNN model consists of a convolutional layer and a fully-connected layer to predict the label. A dropout rate of 0.5 is applied to the convolutional layer. The CNN model obtains an accuracy of 74%, while the SVM sentence classifier obtains an accuracy of 95%. The lower accuracy of the former stems from the small size of the training data.

4.1.3 Modern style classification

The *modern* style merits special consideration. The rhetorical figures, themes, metrics and prosody are different, making this style easily distinguishable from classical poetry. We also observed a considerable difference in the length of poems compared to classical styles. However, the difference between half-lines and lines is not as obvious as in classical Persian poems.

We found that poems of the twenty-first century can easily be distinguished from others by the two models. The same SVM model using Bag-of-words and CNN model using a word2vec model as input both attain an accuracy of 89% at distinguishing modern poetry from classic poetry. The word2vec embedding model was trained on our corpus.

4.2 Poet and century classification

In what follows, we describe baseline experiments for predicting the century and the poet of poems from different historic periods.

4.2.1 Models

To further study the differences and commonalities between poems in different centuries and the style of authors, we ran a linear SVM model with Bag-of-Word features using a train–test split of 85–15%. Another CNN model uses word2vec (trained on

[5] Parvin Etesami, Saadi, Farrokhi, Onsori and Naser Khosro for *qaside* and Saadi, Hafez, Rumi for *ghazal*. Saadi has collections with both styles.

Table 4 F_1-scores for authorship classification for well-known poets

Poet	SVM	CNN[a]
Khayyam	0.53	0.45
Saadi	0.85	0.72
Hafez	0.89	0.87
Rumi	0.85	0.65
Ferdousi	0.99	0.97
Nezami	0.91	0.78
Overall	0.87	0.77

[a]The unbalanced size of test sets as well as the unbalanced length of poems decrease the performance of CNN model

Table 5 Top 10 words with highest drifts in meaning over time

Word	Absolute drift
Wine	0.185
Beware	0.165
Message	0.151
King	0.131
Mirror	0.112
Bazaar	0.104
Hunt	0.087
God	0.086
Vivid	0.085
Prophet	0.080

wine in particular has several metaphorical meanings. We can see that the dynamic embedding model can capture how context shifts over time has influenced the contextual interpretation of this word from an alcoholic drink to a metaphor for martyrdom

our corpus) as input, and consists of two parallel CNN layers with 50 filters each and kernel sizes of 4 and 10, a max-pooling layer, and two dense layers for predicting the labels. As in our previous model, a dropout rate of 0.5 is applied to the hidden layer.

4.2.2 Results

The results in Table 4 show that using different inputs and the unbalanced size of the test sets for each class significantly affect the CNN model. A t-test indicates statistically significant results with $p < 0.05$ and $t < -70.8$.

The confusion matrix for our temporal classification in Fig. 4 reveals the similarity of poems in the tenth to thirteenth centuries, as well as in the eighteenth to twentieth century.

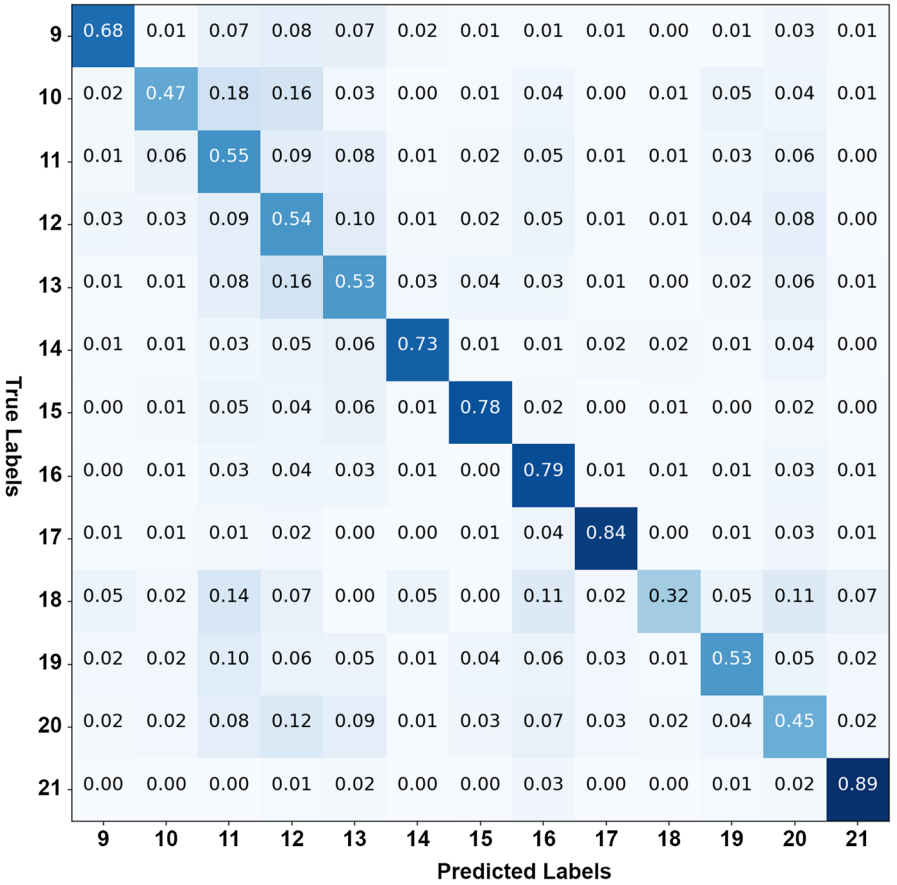


Fig. 4 The confusion matrix for century classification. The results are largely better for the sixteenth, seventeenth, and twenty-first century because cleaner data is available. Whereas, the results for some time periods are not good due to the small size of the dataset or similarity between the author styles

4.3 Tracking changes in context

In addition to the usage of the words as rhetorical figures, another meaningful study is to assess context shifts of words.

4.3.1 Algorithm

To observe how the contexts of words have changed over time, we applied dynamic Bernoulli embedding for language evolution (Rudolph & Blei, 2018) on our data. The method was originally devised to study language evolution and meaning shifts. We adapt the method to instead study changes in the contexts of words over time. Other methods that study context changes (Mihalcea & Nastase, 2012; Hamilton et al., 2016) propose algorithms for aligning embeddings that are trained separately on data in each time slice. However, such algorithms are highly sensitive to the size of the data that is

به یکی جرعه می کـازار کسـش در پی نیست
be-3SG-NEG follow in who-CONN that-annoyance wine drop one of
For sake of one draught of wine wherein is the injury of none

زحمتی می‌کشم از مردم نادان که نپرس
question-2SG-IMP-NEG that ignorant people-CONJ from tolerate trouble-INDF
From the ignorant man, such torment for suffer that asks none

Fig. 5 Example of a line (two half lines) of a poem by Hafez with *wine* referring to the alcohol prohibition in the Islamic era. *Wine* here is used an a symbol of all of the prohibited activities

used in each slice. Since the time slices in our corpus are very heterogeneous in size, such algorithms are not good candidates for our analysis.

4.3.2 Results

As we can observe in Table 5, *wine*, *beware*, *message*, *king*, and *mirror* exhibit the highest drifts. The given numbers represent the absolute total drift of the word vector, assessed in terms of the Euclidean distance between the words' embeddings between the first and the final time slices (Rudolph & Blei, 2018).

The change of context for some of these words such as *wine* and *king* in the history of Persian poetry have been the subject of previous studies (Kadkani, 1943; Sharifnasab, 2004; Shabestary, 2008). The results of our analysis accord well with the explanations and observations made in these works. For instance, neighboring words of *wine* have changed from *lover*, *beloved*, *dance*, and *happy* in Sufi poetry to *martyr*, *blood*, *country*, and *war*. One sub-cluster of this includes words such as *forbidden* and *affectation* that refer to alcohol prohibition in the Islamic era. Figure 5 presents an example of this case.

This has also been the subject of poetry critics (Pourjavadi, 2008). The context of *king* changed from the tenth and eleventh centuries (mostly influenced by Ferdowsi, Hafez, and Rumi), from *land*, *war*, *horse* to *gift*, *heart*, *love*, *poverty*. The complete list of words together with the result of the analysis and code is attached with this submission.[6]

5 Conclusion

Preservation, revitalization, and documentation purposes call for the availability of computational resources and methodologies for low-resource languages. We take a step forward with this by furthering research not just for modern Farsi, but also middle and old Persian, by introducing a large standardized and machine-readable corpus of Persian literary text that is annotated for century and style. We have additionally annotated Hafez's *ghazals* with critical rhetorical figures such as *metaphor*. Our computational experiments provide insights into how Persian poetic language has evolved.

[6] https://github.com/pithysr/persian-poetry.

Additionally, our investigations suggest the effectiveness of supervised and unsupervised techniques in studying poems and poetic styles.

By expanding the range of languages traditionally studied by computational linguistics, low-resource languages often represent a test-bed for validating current methods and techniques. Our resource can contribute to research on metaphor, lexical semantics, text generation, and entailment in addition to cross-linguistic studies. Although these have been studied in a number of poems over the years in the linguistics and Persian literature departments, such studies have never had the tools and resources to consider such a wide coverage corpus while taking advantage of NLP techniques.

Funding Open access funding provided by Northeastern University Library

References

Arberry, A. J. (2008). *Persian poems: An anthology of verse translations. Everyman's library, no. 996*. Yassavoli Publications. Retrieved from https://books.google.com/books?id=-yInMc3wWg4C.

Asgari, E., & Chappelier, J.-C. (2013). Linguistic Resources and Topic Models for the Analysis of Persian Poems. In *Proceedings of the workshop on computational linguistics for literature* (pp. 23–31).

Asgari, E., Ghassemi, M., & Finlayson, M. A. (2013, December). Confirming the themes and interpretive unity of Ghazal poetry using topic models. In *Neural Information Processing Systems (NIPS) Workshop for Topic Models*.

Atashi, M. (2004). *Ahmad Shamlou: A critical analysis*. Amitis.

Baumann, T., Hussein, H., & Meyer-Sickendiek, B. (2018). Style detection for free verse poetry from text and speech. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1929–1940). Association for Computational Linguistics. https://www.aclweb.org/anthology/C18-1164.

Browne, E. G. (1999). Literary history of Persia. Routledge. Retrieved from https://www.routledge.com/A-Literary-History-of-Persia-4-Volume-Set/Browne/p/book/9780700704064.

Bush, L. R. (2012). *More than words: Rhetorical devices in American political cartoons*. University of South Florida.

Fengjie, L., Jia, R., & Yingying, Z. (2016). Analysis of the rhetorical devices in Obama's public speeches. *International Journal of Language and Linguistics, 4*(4), 141–146.

García, J. R., Montanero, M., Lucero, M., Cañedo, I., & Sánchez, E. (2018). Comparing rhetorical devices in history textbooks and teacher's lessons: Implications for the development of academic language skills. *Linguistics and Education, 47*, 16–26.

Genzel, D., Uszkoreit, J., & Och, F. (2010). "Poetic" statistical machine translation: Rhyme and meter. In *Proceedings of the 2010 onference on Empirical Methods in Natural Language Processing*. pp. (158–166). Association for Computational Linguistics.

Greene, E., Bodrumlu, T., & Knight, K. (2010). Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 524–533).

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). *Diachronic word embeddings reveal statistical laws of semantic change*. arXiv preprint arXiv:1605.09096.

Hayward, M. (1996). Analysis of a corpus of poetry by a connectionist model of poetic meter. *Poetics, 24*(1), 1–11.

Hobbs, J. R. (1990). *Literature and cognition* (Vol. 21). Center for the Study of Language (CSLI).
Kadkani, S. (1943). *Poetry and imagination*. Neel Publication.
Khashabi, D., Cohan, A., Shakeri, S., Hosseini, P., Pezeshkpour, P., Alikhani, M., & Yaghoobzadeh, Y. (2021). Parsinlu: a suite of language understanding challenges for Persian. *Transactions of the Association for Computational Linguistics, 9*, 1147–1162.
Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1746–1751). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1181. https://www.aclweb.org/anthology/D14-1181.
Lau, J. H., Cohn, T., Baldwin, T., Brooke, J., & Hammond, A. (2018). Deep-speare: A joint neural model of poetic language, meter and rhyme. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (Vol. 1: Long Papers, pp. 1948–1958). Association for Computational Linguistics. https://www.aclweb.org/anthology/P18-1181.
Lewis, F. D. (2014). *Rumi-past and present, east and west: The life, teachings, and poetry of Jalâl al-Din Rumi*. Oneworld Publications.
Liu, Z., Fu, Z., Cao, J., de Melo, G., Tam, Y.-C., Niu, C., & Zhou, J. (2019). Rhetorically controlled encoder-decoder for modern Chinese poetry generation. In *Proceedings of ACL 2019* (pp. 1992–2001). Association for Computational Linguistics. https://www.aclweb.org/anthology/P19-1192.
Malmasi, S., & Dras, M., et al. (2015). Automatic language identification for Persian and Dari texts. In *Proceedings of PACLING* (pp. 59–64).
Meisami, J. S. (2014). *Medieval Persian court poetry* (Vol. 804). Princeton University Press.
Mihalcea, R., & Nastase, V. (2012). Word epoch disambiguation: Finding how words change over time. In *Proceedings of the 50th annual meeting of the association for computational linguistics* (Vol. 2: Short Papers, pp. 259–263).
Mohaqeqi, A., Faramarzi, P., & Mohaqeqi, J. (2014). Studying the identity of Iranian classical effective literature and contemporary impressible literature. *International Journal of Applied Linguistics and English Literature, 3*(6), 145–151.
Perry, J. R. (2011). Grammaire du persan contemporain. *Iranian Studies, 44*(2), 273–275. https://doi.org/10.1080/00210862.2011.542038
Pourjavadi, N. (2008). *The wine of love: Researches on the meaning of wine in Persian mystical poetry*. Karname.
Raji, S., Alikhani, M., de Melo, G., & Stone, M. (2023). A corpus of Persian literary text. *Zenodo*. https://doi.org/10.5281/zenodo.7923612
Rudolph, M., & Blei, D. (2018) Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web conference on World Wide Web* (pp. 1003–1011). International World Wide Web Conferences Steering Committee.
Rypka, J. (2013). *History of Iranian literature*. Springer.
Safa, Z. (1993). *History of literature in Iran*. Ferdows Publication.
Seraji, M., Megyesi, B., & Nivre, J. (2012). A basic language resource kit for Persian. In *Eight international conference on language resources and evaluation (LREC 2012), 23–25 May 2012, Istanbul, Turkey* (pp. 2245–2252). European Language Resources Association.
Seyed-Gohrab, A. A. (2011). *Metaphor and imagery in Persian poetry* (Vol. 6). Brill.
Shabestary, M. (2008). *The secret rose garden*. Forgotten Books.
Shamisa, C. (1999). *Style of poetry*. Mitra Press.
Sharifnasab, M. (2004). On "Wine" in Mathnavi, Based on Ibn-e Farezs "Xamriyye". *Literary Text Research, 7*(19), 190–204. https://doi.org/10.22054/ltr.2004.6269
Tabatabai, A. (2001). *Persian language etymology*. Bokhara Magazine.
Tom, G., & Eves, A. (2012). *The use of rhetorical devices in advertising*. Cross Currents: Cultures, Communities, Technologies.
Tousi, M. A. (1974). *The affixes in Persian language*. Gohar.
Tusi, B. (2013). *Expansion of Iranian literature and culture in the world*. Tehran University Center of Islamic Studies-McGill University.
Zhang, J., Feng, Y., Wang, D., Wang, Y., Abel, A., Zhang, S., & Zhang, A. (2017) Flexible and creative Chinese poetry generation using neural memory. In *Proceedings of the 55th annual meeting of the association for computational linguistics* (Vol. 1: Long Papers, pp. 1364–1373). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1125. https://www.aclweb.org/anthology/P17-1125.