

# Profiling Poets from Spanish Sonnets

Alysa Meng  
University of Washington  
menga@uw.edu

## Abstract

*Every poem contains traces of its invisible author. This project investigates how machine learning techniques (SVM, Bi-LSTM, BERT) can capture authorial metadata from Spanish sonnets. By analyzing how models interact with poetic features, we explore how computational models produce information about authorship, including time period, gender and region through a series of text classification tasks. While each model architecture captures poetic aspects in a different way, they all struggle to perform well on less frequent classes in an uneven dataset. We examine these biases within poetic data, which is reflected in the accuracies, confusion matrices and macro F1 scores. Examining model weights expose patterns that different architectures capture in each task. This work contributes to broader discussions on cultural representation in Hispanic literature and the limits of artificial intelligence in textual studies.*

## 1. Introduction

The question of authorship has long been a subject of scholarly investigation across disciplines such as literature, history, and philology. Traditionally, authorship attribution concerned itself with identifying the most likely author of a given text, particularly in cases where authorship was disputed, unknown, or pseudonymous. Over the past several decades, this area of inquiry has been significantly advanced by computational methods that analyze stylistic features of texts to draw inferences about their origins and gained popularity in the latter half of the 20th century, coinciding with the advent of computational linguistics and the increasing availability of digital corpora.

These stylometric approaches, focuses on quantifying linguistic style through features such as word frequency distributions, syntactic patterns, and semantic tendencies. Among the foundational tools in this domain are Burrows's Delta, Craig's Zeta, and a range of distance-based metrics designed to capture stylistic deviation across texts [11]. These metrics are often employed to construct stylistic fin-

gerprints that can be compared across a corpus to support claims of common or divergent authorship.

The integration of machine learning into stylometric practice has reshaped practices in authorship attribution. The traditional feature engineering approach is now increasingly supplemented by data-driven approaches that leverage large-scale text representations and classification models. This transition aligns with the broader emergence of the digital humanities, a field that interrogates how digital tools and computational methodologies can augment and transform traditional humanistic inquiry.

Beyond attribution, stylometric methods have found extensive application in textual forensics and historical linguistics research, allowing scholars to infer not only the likely author of a text but also aspects of the author's socio-cultural, temporal, and educational background [7]. This broader conception of authorship background inference concerned with extracting latent metadata about an author from stylistic features has been valuable in contexts where direct biographical or contextual information is limited or inaccessible.

With the recent popularity of large language models (LLMs), more recent studies have explored the problem of distinguishing human-authored poetry from that produced by LLMs, with authorship inference methods employed to assess the stylistic alignment of generated texts with established poetic traditions or specific groups of authors [10]. This application illustrates a shift focus from identifying known human authors to tracing stylistic influences embedded within the outputs of generative models.

These developments call for a reevaluation of the tools serving as the foundation for authorship inference, specifically in the context of evolving AI-driven systems that influence how we interpret literature.

Specifically, we ask the following:

*To what extent can machine learning approaches help organize, interpret, and derive patterns from Spanish poetry?*

To answer the question, this project focuses on three biographical profiling tasks:

1. Predicting the author’s country of origin,
2. Predicting the author’s gender, and
3. Predicting the historical period in which the poem was composed.

## 2. Related Work

While support vector machines (SVMs) and principal component analysis (PCA) have dominated feature-based stylometric approaches, recent scholarship has begun to explore neural network architectures, such as long short-term memory networks (LSTMs), for capturing syntactic and semantic features relevant to stylistic inference [15]. These approaches increasingly focus on integrating linguistic cues into authorship attribution models, with some early investigations employing transformer-based architectures such as BERT [6]. However, much of this work remains concentrated on English-language prose, typically restricted to texts produced within the last few centuries.

While non-English texts have received some attention, such efforts remain limited in scope and number. For instance, Hernández [8] applied stylometric analysis to explore stylistic similarities among three poets from Spain’s Golden Age, employing the `stylo` package in R [5]. Plecháč et al. [13] conducted a pilot study assessing the role of poetic versification in SVM-based authorship classification, with an emphasis on feature interpretability across Czech, German, Spanish, and English poetry.

Recent advancements have extended authorship attribution to languages and literary forms beyond Western prose. For example, BERT-based ensemble models have been applied to Arabic poetry [1] and to Japanese literary texts [9], reflecting a growing interest in the cross-linguistic applicability of transformer models for stylistic inference.

Despite these developments, the intersection of Hispanic poetry with state-of-the-art computational methods remains underexplored. This project addresses this gap by examining how machine learning models can be used to classify and infer characteristics about authors based on Spanish sonnets.

Poetry in Spanish presents a particularly challenging case for biographical inference due to its rich historical depth, regional diversity, and formal complexity. Spanish is a gendered language, making it different from English. These aspects potentially limit the effectiveness of language models trained on more uniform and contemporary datasets with methods derived from English tasks. The historical variation in vocabulary, orthography, syntax adds complexity to the classification process, making it an exciting environment for assessing the limits and capabilities of current computational approaches to literary analysis.

## 3. Methodology

Preliminary data analysis and visualization to first understand the structure of the dataset and its contents was done in Excel. Data processing was done with Python’s `pandas` package. Then we compare 3 different deep learning approaches for classification, implemented using `pytorch` libraries (e.g., `torch.nn.Module`). Training and evaluation for SVM and LSTM models were done locally. The BERT-architecture model was fine-tuned on Google Colab to take advantage of GPU speedup.

## 4. Dataset

This study uses the Diachronic Spanish Sonnet Corpus (DISCO), a richly annotated dataset comprising 4,530 Spanish sonnets spanning five centuries [14]. Annotations were initially generated with computational techniques, and then verified by domain experts. The authorial metadata labels are used for the supervised classification tasks, and the per-poem rhyme and meter information are used as SVM features.

For the 1,237 authors with associated metadata, The composition of DISCO is skewed towards male poets, Spanish authors, and 19th century works as summarized in the Appendix (5, 6, 7).

As there may be ambiguity in mapping authors to a particular century, we take the floor of the `normdate` attribute, which already has standardized the birth date format for the authors. The analogous breakdown for poems and their sources are further discussed in [14].

Furthermore, the poems in DISCO are specifically *sonetos*, 14-verse poems where each verse is 11 syllables. The poem into 4 stanzas: 2 quatrains and 2 tercets. Each quatrain has 4 verses, and each tercet has 3 verses. The rhyme scheme is structured, the quatrains and tercets following the same pattern, respectively. The stressed or unstressed nature of each syllable in a verse is known as its meter. An example is provided in the Appendix.

### 4.1. Data Pipeline

For each poem, we obtain its text from the `.txt` file and join it with its associated TEI annotations for rhyme and meter in the `.xml` format. Afterwards, data enrichment is done by joining the authorial metadata with the poem content and prosodic features. As we are interested in overall patterns across gender, countries, and periods, we remove poems by the 20th century author and by authors in a country group with less than 10 total authors. We then sanitize the poems and drop those that do not have rhyme or meter information, as there were some minor inconsistencies in DISCO’s TEI annotations. This also includes stripping punctuation and extraneous whitespace, and setting all characters to lowercase in the SVM-based and LSTM-based

models. The casing and punctuation is maintained in the BERT-based model since a cased transformer is used for fine-tuning. A visualization of this process is provided in Figure 1.

To split the data into train and test sets for each classification task, we use a stratified sampling approach to maintain the ratios of the category labels and produce a 90-10 train-test split across the poems. After data processing, each classification task had roughly 4050 poems in its training set and 450 poems in its test set. Finally, 10% of the training set was kept aside to form a validation set using the stratified splitting technique.

## 4.2. Support Vector Machine (SVM)

SVMs are a widely used, relatively interpretable machine learning model suitable for high-dimensional text classification tasks so we implement an SVM classifier as a baseline. For features, we use the top 2000 unigrams and top 2000 bigrams using TF-IDF weights, respectively. The TF-IDF weights correspond to the product of the a term frequency and its inverse document frequency.

For a term  $t$  and document  $d$  in the set of documents  $D$ ,

$$\text{tf}(t, d) = \frac{\# \text{ times } t \text{ appears in } d}{\sum_{\text{term } t' \in d} \# \text{ times } t' \text{ appears in } d}$$

$$\text{idf}(t, D) = \frac{|D|}{\# \text{ of documents in } D \text{ that contain } t}$$

To incorporate prosodic aspects of poetry, the meter and rhyme annotations from the Text Encoding Initiative (TEI) files are sequentially encoded. The n-gram and prosodic features are concatenated to get a vectorized representation for a poem.

## 4.3. Long Short-Term Memory (LSTM)

To improve upon the baseline, we evaluate a Bidirectional Long Short-Term Memory (Bi-LSTM) neural network with an attention layer before the classification layer. Bi-LSTMs are more effective at capturing sequential dependencies in text and can incorporate both forward and backward context when processing a sentence. The expectation is that Bi-LSTM models will more effectively capture underlying syntactic and rhythmic patterns that distinguish authors, especially across time and region. We use pre-trained GloVe word embeddings from the Spanish Billion Words Corpus and Embeddings project [2, 12] with a vocabulary size of 10,000 words.

## 4.4. Transformer

Futhermore, this study fine-tunes BETO [3], a Spanish-language BERT model pre-trained on a large Spanish corpus, for each classification task. BETO’s transformer architecture allows for context-sensitive word embeddings and

attention mechanisms that have proven effective in various downstream NLP tasks. This model is adapted to the three classification tasks, with particular attention to fine-tuning strategies suitable for limited, domain-specific training data such as historical poetry.

## 4.5. Evaluation

Models are evaluated using standard classification metrics, including accuracy, macro F1 score, and confusion matrices. Beyond performance metrics, this study also emphasizes model interpretability. Feature importance analysis is conducted for the SVM classifiers, and attention weight visualization is used to investigate which aspects of the poems most strongly influence the Bi-LSTM and BETO predictions.

## 5. Experiments

The code repository for data processing and model evaluation can be found at <https://github.com/kpwce/profiling-spanish-poets>. For SVM-based and LSTM-based models, grid search was done for learning rates (1e-2, 1e-3, 1e-4) and weight decay (0, 1e-5, 1e-4, 1e-3); Adam was chosen as the optimizer. For the BERT-based model, we tried various dropout rates (0.2, 0.3, 0.4, 0.5) and batch sizes (8, 16, 32).

Model	Gender	Country	Period
MostFreq	0.933	0.613	0.694
SVM	0.883	0.574	0.764
LSTM	0.930	0.605	0.836
BERT	0.928	0.656	0.858

Table 1. Model Accuracies (MostFreq denotes a classifier that always chooses the most frequent label)

Model	Gender	Country	Period
SVM	0.597	0.303	0.570
LSTM	0.686	0.278	0.683
BERT	0.582	0.248	0.737

Table 2. Model Macro F1 Scores

### 5.1. SVM

For all three classification tasks, the model consistently predicted the majority class for all elements in the train and test sets, even after hyperparameter selection. This often gave the highest accuracy, but at the cost of always getting the majority class correct and every other class incorrect.

To mitigate this issue, we implemented a weighted loss function to penalize the model more for an incorrect prediction for a less frequent class by a normalized inverse proportion. Given a class  $c$  in classes  $C$ ,





$$\text{class-weight}(c) = \frac{\left( \frac{\# \text{ samples in train set}}{\# \text{ samples in train set that have label } c} \right)}{|C|}$$

With learning rate 0.005 and weight decay  $1e-4$ , we have the 88.3% accuracy for the gender prediction, 57.4% for birth country prediction, 76.4% accuracy for the century prediction as shown in Table 1. The century classifier was the only model out of the 3 that improved beyond the accuracy evaluated on assigning everything to the majority class (19th century), which was comparatively 69.4%. The accuracy for always predicting the majority for gender (male) was 93.3% and for country (Spain) was 61.3%.

## 5.2. LSTM

Unlike the SVM-based classifiers, the Bi-LSTM architecture with no class weights did not consistently predicted majority class each time for all three tasks, but there continued to be a large emphasis on assigning the more frequent labels correctly. The tradeoff between classifying majority classes correctly for higher accuracy versus getting a lower, but balanced performance across classes was sensitive to model hyperparameters. Higher weight decay (e.g.,  $1e-2$ ) led to the SVM majority classification case, while lower weight decay led the model to overfit on the training set. To mitigate these issues, we also use the class weights calculation in the cross entropy loss.

The final version of the LSTM-based models used an embedding dimension of size 300 to align with the GloVe embedding dimensions and a hidden dimension for the Bi-LSTM of size 64 (128 divided by 2 for the bi-directional architecture).

Each of the models had a different best-performing learning rate and weight decay configuration after hyperparameter tuning using the validation set, as shown in Table 3 alongside their respective test accuracies after training for 5 epochs.

Task	Learning rate	Weight decay
Gender	$1e-3$	$1e-4$
Country	$1e-2$	$1e-5$
Century	$1e-2$	$1e-5$

Table 3. LSTM final model hyperparameters

On a high level, the LSTM-based models followed the same pattern as the SVM-based models in the confusion matrix. For all three classification tasks, the LSTM-based models had a higher overall accuracy than the SVM-based models.

## 5.3. Transformer

Like the other models, the BERT-based model had a tendency to predict the most frequent classes, so class weights were used in the cross entropy loss. Training was done for 5 epochs with a 0.3 dropout rate and a batch size of 32. The results are shown in Table 1. While the accuracy is high for country classification, we observe a low macro F1 score, finding that the model learns to classifying a limited subset of the classes with high accuracy rather than all the classes.

In regards to accuracy, the fine-tuned versions of BETO performs better than the SVM-based models and both the birth country and century models. However, Figure 2 displays lower accuracies in many cases for less frequent classes in the gender and birth country tasks.

## 6. Discussion

The weighted SVM results demonstrate more balanced classification output between the most common classes and the other classes, but ultimately prefer the majority classes. The confusion matrix shows that the model incorrectly predicts male the majority of the time when a poetry was written by a female poet. This is similar for Spain in the birth country classifier and the 19th century in the period classifier.

Taking the magnitude of the learned feature weights, we determine which SVM features are important for gender, country, and century classification. Top features for gender are in Table 4. For birth country, we list each country’s most important positive feature to understand what would push towards classification for a particular country (Spanish term, English literal translation): Philippines (misterio, mystery), Cuba (primero, first), Spain (prado, meadow), Portugal (jardín, garden), Mexico (soberbio, proud), Chile (hay un, there’s a), Argentina (los que, those that), Venezuela (bronce, bronze), Puerto Rico (y mi, and my), Costa Rica (lluvia, rain), Colombia (cadáver, cadaver). With the century classification, we have the following: 15th and earlier (eterno, eternal), 16th (guerra, war), 17th (voz y, voice and), 18th (de él, of him), and 19th (soneto, sonnet). The n-gram features were generally ranked higher in importance than the prosodic features. This may be due to the rigid structure of a Spanish sonnet that is shared across the corpus of sonnets. Furthermore, the limited vocabulary derived from the training data resulted in inability to fully represent the words in a poem in the test set.

Upon analyzing the LSTM-based model attention layer weights, a similar on focus towards noun phrases can be seen in Figure 4 for a period classification test example with a 16th century true label, which can be compared to the SVM-based 16th century model weights in Figure 3.

Furthermore, we find that birth country nor century perfectly corresponds to poetic style. For instance, the

reina	desotras	flores	fresca	rosa			
-2.05E-01	Removed	-7.53E-01	8.68E-02	1.50E-01			
primero	honor	de	abril	y	de	este	prado
2.68E-01	-1.63E-01	-2.74E-02	9.03E-01	8.58E-01	-2.74E-02	-1.89E-01	1.22E+00
así	te	privilegie	el	cierzo	helado		
9.67E-01	-2.46E-01	Removed	7.72E-02	Removed	1.68E-01		
y	respete	la	helada	rigurosa			
8.58E-01	Removed	-5.74E-01	2.10E-02	Removed			

Figure 3. SVM 16th-century weights for first quatrain of Luis Martín de la Plaza’s “Fresca rosa”

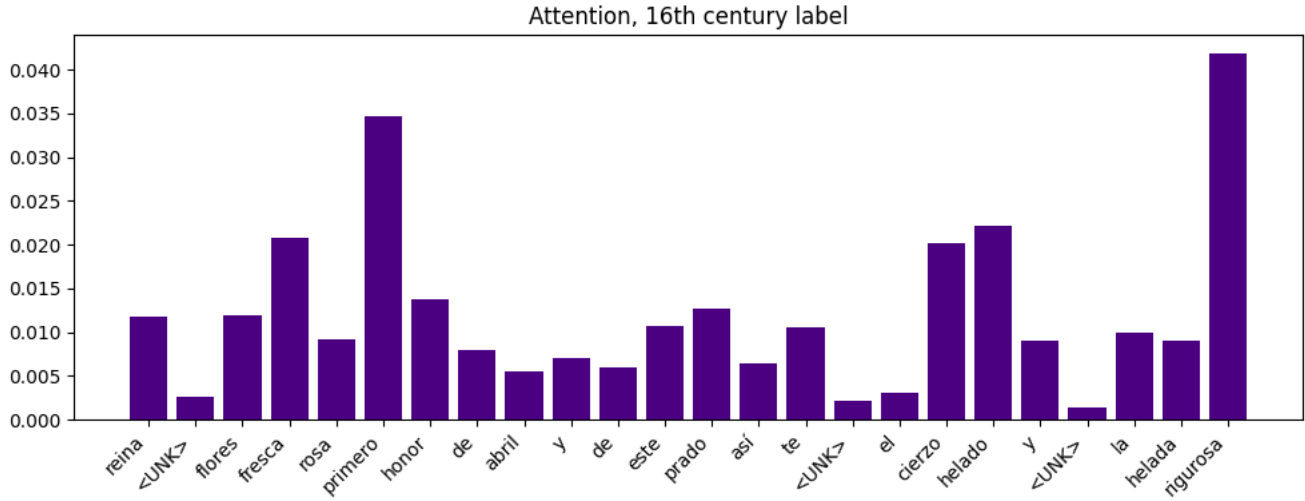


Figure 4. LSTM attention for first quatrain of Luis Martín de la Plaza’s “Fresca rosa”

Term	Male	Female
valiente (brave)	-	+
señor (sir, lord)	+	-
error (error)	-	+
herido (wounded)	-	+
del amor (of love)	+	-

Table 4. Gender feature importance. Male and female had the same ranking for the top 5 but inverted in sign, ordered most to least significant.

LSTM-based country classification model misclassified an Argentinian-born poet as Spanish, but the poet had lived in Spain for an extended period of time so it would be reasonable to find lexical traces associated with a dialect from Spain in his poetry. With continuous periods of time being mapped to discrete categories, there are some poets born immediately before the turn of a century that were classified into the later century for all the time period classifiers.

The transformer-based models demonstrate some ability to capture rhythmic features and syntactic structure. Us-

ing BertViz [16] to visualize the attention mechanism for the different layers of the model heads, we discover that later layers (8-11) appear to encode higher-level relationships that may correspond to relationships within and between verses. This preliminary observation should be further investigated to better understand the linguistic meaning, if such interpretation can be extracted, from these layers.

## 7. Limitations

Poetry is complex, and the three approaches investigated do not take into account every poetic aspect. Other features to consider include marking more linguistic traits, such as the presence of a synalepha, or tagging named entities, such as places and people. Poetry also has visual cues that standard language modeling approaches do not account for. Punctuation usually matters, even though it was stripped by design for the SVM-based and LSTM-based approaches. All three approaches remove indent and whitespace, but in some cases do meaningfully contribute to the interpretation of a poem.

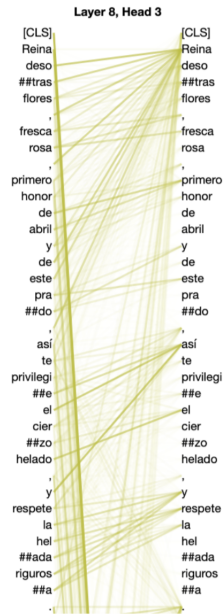


Figure 5. An attention head layer from the first quatrain of Luis Martín de la Plaza’s “Fresca rosa”

As each model produces a different representation of a poem, additional work on ensemble models between BERT-based, SVM-based and LSTM-based classifiers. Combining these distinct methods that represent the same poem in different ways can improve overall performance.

The dataset composition is not representative of Hispanic poetry as a whole. The unbalanced nature of the biographical traits posed a challenge for training the models. An alternative approach of taking the class with the minimum of examples was considered to sample a subset of data that was balanced across classes, but this would have limited the proportion of the already small dataset that we were working with. In the future, we hope that more diverse datasets are made accessible to train models that optimize across all classes rather than the majority. From the converse perspective, we hope to see improved computational methods for handling small, complex and unbalanced humanities datasets.

## References

- [1] Lama Alqurashi, Serge Sharoff, Janet Watson, and Jacob Blakesley. BERT-based classical Arabic poetry authorship attribution. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6105–6119, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics. 2
- [2] Cristian Cardellino. Spanish Billion Words Corpus and Embeddings, August 2019. 3
- [3] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*, 2020. 3
- [4] G. de la Vega and J.F. Alcina. *Poesía completa*. Austral : Literatura. Espasa Calpe, 1989. 8
- [5] Maciej Eder, Jan Rybicki, and Mike Kestemont. Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, 8(1):107–121, 2016. 2
- [6] Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. BertAA : BERT fine-tuning for authorship attribution. In Pushpak Bhattacharyya, Dipti Misra Sharma, and Rajeev Sangal, editors, *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, Indian Institute of Technology Patna, Patna, India, Dec. 2020. NLP Association of India (NLP AI). 2
- [7] D. C. Greetham. Textual forensics. *PMLA/Publications of the Modern Language Association of America*, 111(1):32–51, 1996. 1
- [8] Laura Hernández Lorenzo. Digital stylistics applied to golden age spanish poetry is fernando de herrera really a transitional poet between renaissance and baroque? 2024. 2
- [9] Taisei Kanda, Mingzhe Jin, and Wataru Zaitzu. Integrated ensemble of bert- and features-based models for authorship attribution in japanese literary works, 2025. 2
- [10] Tharindu Kumarage and Huan Liu. Neural authorship attribution: Stylometric analysis on large language models. In *2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pages 51–54. IEEE, 2023. 1
- [11] François Dominic Laramée. Introduction to stylometry with python. *Program. Hist.*, (7), Apr. 2018. 1
- [12] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 3
- [13] Petr Plecháč, Klemens Bobenhausen, and Benjamin Hammerich. Versification and authorship attribution. a pilot study on czech, german, spanish, and english poetry. *Studia Metrica et Poetica*, 5(2):29–54, 2018. 2
- [14] Pablo Ruiz Fabo, Helena Bermúdez Sabel, and Clara Isabel Martínez Cantón. pruizf/disco: Version 5.0, 2023. 2
- [15] Jacques Savoy. *Advanced Models for Stylometric Applications*, pages 153–187. Springer International Publishing, Cham, 2020. 2
- [16] Jesse Vig. Visualizing attention in transformer-based language representation models, 2019. 6

## Appendix

### DISCO author biographical data distributions

gender	count
Male	1155
Female	81

Table 5. Author gender distribution

country-birth	count
Argentina	29
Bolivia	3
Brasil	2
Chile	13
Colombia	23
Costa Rica	11
Cuba	173
Ecuador	9
España	832
Filipinas	11
Haití	1
Honduras	6
Italia	2
México	35
Nicaragua	2
Panamá	4
Paraguay	2
Perú	9
Portugal	15
Puerto Rico	15
República Dominicana	3
Uruguay	4
Venezuela	10
Unspecified	22

Table 6. Author birth country distribution

century	count
15th or earlier	10
16th	170
17th	309
18th	52
19th	694
20th	1

Table 7. Author period distribution

En tanto que de rosa y azucena  
se muestra la color en vuestro gesto,  
y que vuestro mirar ardiente, honesto,  
enciende al corazón y lo refrena;

y en tanto que el cabello, que en la vena  
del oro se escogió, con vuelo presto,  
por el hermoso cuello blanco, enhiesto,  
el viento mueve, esparce y desordena;

coged de vuestra alegre primavera  
el dulce fruto, antes que el tiempo airado  
cubra de nieve la hermosa cumbre.

Marchitará la rosa el viento helado,  
todo lo mudará la edad ligera,  
por no hacer mudanza en su costumbre.

We observe 14 lines, divided into 2 quatrains and 2 tercets. To analyze the rhyme, we find that the 2 quatrains have consonant rhymes in the pattern “ABBA ABBA” because verses 1, 4, 5, and 8 end in “na” while verses 2, 3, 5, and 6 end in “to”. Similarly, the 2 tercets that follow have consonant rhymes in the pattern “CDE DCE”. Thus, the rhyme scheme of the poem is “ABBA ABBA CDE DCE”. We map each letter to a unique integer for computation.

The meter captures the rhythm of the poem, so to analyze it, we need to analyze which syllables are stressed. Consider the first line, which flows as follows (stressed syllables denoted in capital letters: “en TAN-to que de RO-sa y a-zu-CE-na”. While the meter may vary between verse depending on the sonnet, each verse should have a stressed syllable in the 10th position. We can encode this stress pattern with binary variables since each syllable is stressed or unstressed for computation.

### How to analyze a Spanish sonnet

SONETO XXIII by Garcilaso de la Vega from [4]