View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by idUS. Depósito de Investigación Universidad de Sevilla

TOWARDS A PHILOLOGICAL METRIC THROUGH A TOPOLOGICAL DATA ANALYSIS APPROACH

Eduardo Paluzo-Hidalgo
Department of Applied Mathematics I
School of Computer Engineering
University of Seville
Seville, Spain
epaluzo@us.es

Rocio Gonzalez-Diaz
Department of Applied Mathematics I
School of Computer Engineering
University of Seville
Seville, Spain
rogodi@us.es

Miguel A. Gutiérrez-Naranjo
Department of Computer Science and Artificial Intelligence
School of Computer Engineering
University of Seville
Seville, Spain
magutier@us.es

January 14, 2020

ABSTRACT

The canon of the baroque Spanish literature has been thoroughly studied with philological techniques. The major representatives of the poetry of this epoch are Francisco de Quevedo and Luis de Góngora y Argote. They are commonly classified by the literary experts in two different streams: Quevedo belongs to the *Conceptismo* and Góngora to the *Culteranismo*. Besides, traditionally, even if Quevedo is considered the most representative of the *Conceptismo*, Lope de Vega is also considered to be, at least, closely related to this literary trend. In this paper, we use Topological Data Analysis techniques to provide a first approach to a metric distance between the literary style of these poets. As a consequence, we reach results that are under the literary experts' criteria, locating the literary style of Lope de Vega, closer to the one of Quevedo than to the one of Góngora.

Keywords Philological metric · Spanish Golden Age Poets · Word embedding · Topological data analysis · Spanish literature

1 Introduction

Topology is the branch of Mathematics which deals with proximity relations and continuous deformations in abstract spaces. Recently, many researchers have paid attention to it due to the increasing amount of data available and the need for in-depth analysis of these datasets to extract useful properties. The application of topological tools to the study of these data is known as Topological Data Analysis (TDA), and this research line has achieved a long list of successes in recent years (see, e.g., [1], [2] or [3], among many others). In this paper, we focus our attention on applying such TDA techniques to study and effectively compute some kind of nearness in philological studies.

Until now, most of the methods used in comparison studies in philology are essentially qualitative. The comparison among writers, periods or, in general, literary works is often based on stylistic observations that cannot be quantified. Several quantitative methods based on statistical analysis have been applied in the past (see [4]) but their use is still controversial [5].

Our approach, based on TDA techniques, is completely different from previous ones. Instead of using statistical methods, whose aim is to summarize the information of the literary work in a numerical description, our procedure is

based on the spatial shape of the data after embedding it in a high-dimensional metric space. Broadly speaking, our work starts by representing a literary work as a could of points. The process of making such representation word by word is called *word embedding*.

Among the most popular systems for word embedding, the `word2vec` [6], GloVe [7] or FastText [8] systems can be cited. Along this paper, the `word2vec` system with its `skipgram` variation will be used for obtaining such multidimensional representation of literary works.

The embedding techniques mentioned above try to find a representation of the literary work as a high-dimensional point cloud in such a way that the semantic proximity is kept. The latter is one of the key points of this paper. Another of the key points is the use of TDA techniques to measure the nearness between different point clouds representing different literary works.

In computer sciences, there are many different ways to measure the distance among two point clouds [9], but most of them are merely based on some kind of statistical resume of the point cloud and not on its *shape*.

In this paper, the shape of a point cloud representing a literary work is captured by using a TDA technique known as persistence diagrams, which is based on deep and well-known concepts of algebraic topology such as simplicial complexes, homology groups and filtrations. A measure between persistence diagrams, namely the bottleneck distance, provides a way to quantify the nearness among two different persistence diagrams and hence, a way to quantify the nearness among two different literary works.

As far as we know, very few papers are exploring similar research lines [10, 11] that the proximity between literary works is measured using TDA techniques. In order to illustrate the potential of such techniques, we provide a case study on the comparison of the literary works of two poets who are representatives of the two main stylistic trends of the Spanish Golden Age: Luis de Góngora and Francisco de Quevedo. We also consider a third poet, called Lope de Vega, whose literary works belong to the same stylistic trend as those of Francisco de Quevedo.

Literary experts agree that the styles of Lope de Vega and Francisco de Quevedo are *close* (they belong to the same literary trend, the so-called *Conceptismo*), but both are *far* from the style of Luis de Góngora, which corresponds to a different literary trend called *Culteranismo* [12]. The application of TDA techniques for measuring the nearness of such Spanish poets quantitatively confirms that the styles of Lope de Vega and Francisco de Quevedo are close to each other and yet both styles are far from the style of Luis de Góngora.

The paper is organized as follows: In Section 2, some preliminary notions about word embedding and TDA techniques are provided. The procedure applied to compare two different literature styles is described in Section 3. In Section 4, the specific comparison between the literary works of the three poets mentioned above is thoroughly described. Finally, in Section 5, conclusions and future work are given.

2 Background

In this section we recall some basics related to the techniques used along the paper. Firstly, word embedding methodology is briefly introduced. Later, the relevant tools from TDA used in our approach will be described.

2.1 Word embedding

Word embedding is the collective name of a set of methods for representing words from natural languages as points (or vectors) in a real-valued multi-dimensional space. The common feature of such methods is that words with similar meanings take close representation. Such representation methods are on the basis of some of the big successes of deep learning applied to natural language processing (see, for example, [13] or [14]). Next, we recall some basic definitions related to this methodology.

Definition 1 (corpus) *Given a finite alphabet Σ, the set of all possible words is $\Sigma^* = \{x_1x_2\cdots x_n \mid x_i \in \Sigma\}$. A corpus is a finite collection of writings composed with these words, denoted by C. The vocabulary, V, of a corpus C is the set of all the words that appear in C. Finally, given $d \geq 1$, a word embedding is a function $E : V \to \mathbb{R}^d$.*

The word embedding process used along this paper is the `word2vec`[1], specifically its modified version called `skipgram` [17]. It is based on a neural network architecture with one hidden layer where the input is a corpus and the output is a probability distribution. It is trained with a corpus to detect similarities in words based on their relative distance in a writing. Such distances are the base of their representation in an n-dimensional space.

[1]The model we used is the one implemented in the python library `gensim` which is based on [15, 16].

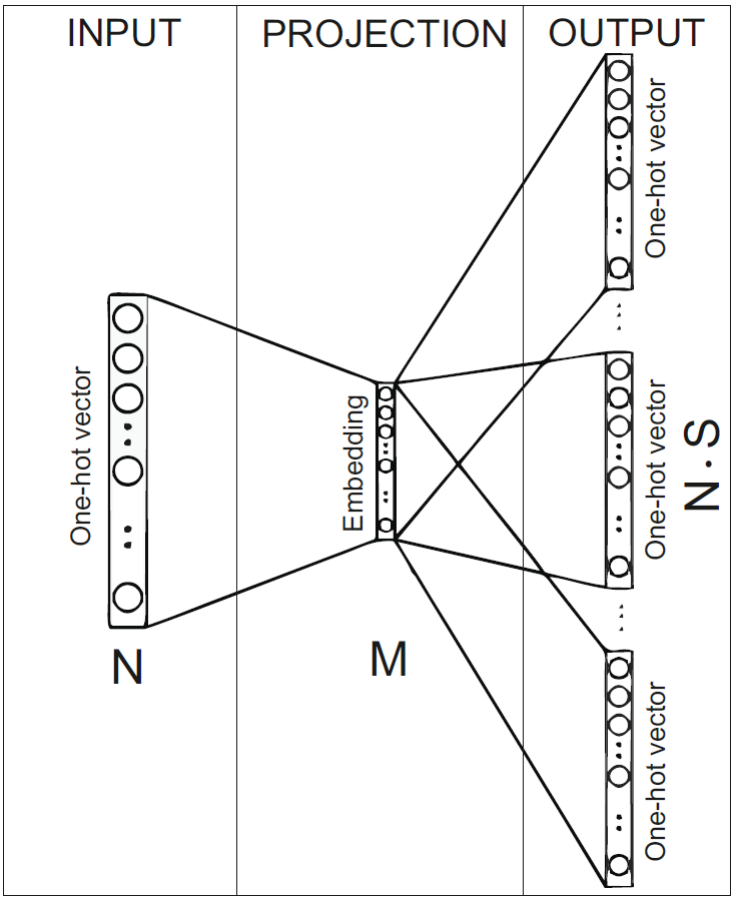


Figure 1: The `skipgram` neural network architecture. The input layer has as many neurons as the length of the one-hot vector that encode the words of the corpus, i.e., the number of words that compose the vocabulary of the corpus, N in this case. The size of the projection layer is equal to the dimension in which we want to embed the corpus, M. Finally, the output layer has $N \cdot S$ neurons where S is the size of the window, i.e., the number of surrounding words that the model tries to predict. This image is inspired in the image of the `skipgram` model in [18].

The two main models of the `word2vec` techniques are called CBOW (Continuous Bag of Words) and `skipgram`. A detailed description of such models is out of the scope of this paper. Roughly speaking, the neural network is trained by using a corpus, where the context of a word is considered as a *window* around a target word. In this way, in the `skipgram` model each word of the input is processed by a log-linear classifier with continuous projection layer, trying to predict the previous and the following words in a sentence. In this kind of neural network architecture, the input is a one-hot vector representing a word of the corpus. Then, the weights of the hidden layer are the high dimensional representation of the words, and the output is a prediction of the surrounding words. More specifically, it is a log-linear classifier with continuous projection layer following the architecture shown and explained in Figure 1.

2.2 Topological data analysis

The field of computational topology and, specifically, topological data analysis were born as a combination of topics in geometry, topology, and algorithms. In this section, some of their basic concepts are recalled. For a a detailed presentation of this field, [19, 20] are recommended.

As we will mention below, we are interested in how a space is connected taking into account, somehow, the distribution of a point cloud in the space. Considering this aim we will recall, firstly, homology, and lately, persistent homology which are fundamental TDA tools. The information obtained when computing persistent homology is usually encapsulated as a persistence barcode. Finally, the bottleneck distance will be shown as the main distance to compare persistence barcodes.

The class of the spaces where we define homology groups are the class of simplicial complexes which is a space built from line, segments, triangles, and so on for higher dimensions. These components are called simplices.

Definition 2 (n-simplex) *Let $\{v_0, \dots, v_n\}$ be a set of geometrically independent points in $\mathbb{R}^d$. The n-simplex σ spanned by $v_0, \dots, v_n$ is defined as the set of all points $x \in \mathbb{R}^d$ such that $x = \sum_{i=0}^{n} t_i v_i$, where $t_i \in \mathbb{R}$ when $0 \leq i \leq n$, and $\sum_{i=0}^{n} t_i = 1$. Besides, $v_0, \dots, v_n$ are called the vertices of σ, the number n is called the dimension of σ, and any simplex spanned by a subset of $\{v_0, \dots, v_n\}$ is called a face of σ.*

When a set of n-simplices is glued, a simplicial complex is formed.

Definition 3 (simplicial complex) *A simplicial complex K in $\mathbb{R}^d$ is a collection of simplices in $\mathbb{R}^d$ such that:*

1. *Every face of a simplex of K is in K;*
2. *the intersection of any two simplexes of K is a face of each of them.*

Any $L \subset K$ is called a subcomplex of K if L is a simplicial complex.

Next, the definition of n-chains and their boundaries is recalled. It is a key idea for formalizing the idea of *hole* in a multidimensional space.

Definition 4 (chain complexes) *Let K be a simplicial complex and n a dimension. A n-chain is a formal sum of n-simplices, $c = \sum_{i=1}^{m} a_i \sigma_i$, in K, where σ_i are n-simplices and $a_i \in \mathbb{Z}_2$ are coefficients. The sum between n-chains is defined componentwise, i.e., let $c' = \sum_{i=1}^{m} b_i \sigma_i$ be another n-chain, then $c + c' = \sum_{i=1}^{m} (a_i + b_i)\sigma_i$. The n-chains together with the addition form an abelian group denoted by C_n. To relate these groups with different dimension, the boundary of a n-simplex, $\sigma = \{v_1, \dots, v_n\}$, is defined as the sum of its $(n-1)$-dimensional faces, that is $\partial_n \sigma = \sum_{j=0}^{n} \{v_0, \dots, \hat{v}_j, \dots, v_n\}$, where the hat indicates that v_j is omitted. The boundary of a n-chain is the sum of the boundaries of its simplices. Hence, the boundary is a homomorphism that maps a n-chain to a $(n-1)$-chain, and we write $\partial_n : C_n \to C_{n-1}$. Then, a chain complex is the sequence of chain groups connected by boundary homomorphisms,*

$$\dots \xrightarrow{\partial_{n+2}} C_{n+1} \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} \dots$$

A crucial property of the boundary homomorphism is that the boundary of the boundary is null. Next, the chains with empty boundary are considered. From an algebraic point of view, they have a group structure.

Definition 5 (n-cycles and n-boundaries) *The group of n-cycles is the subgroup of the group of n-chains denoted by Z_n composed by those chains with empty boundary, $\partial c = 0$. The group of n-boundaries is the subgroup of the group of n-chains denoted by B_n composed by those chains that are in the image of the $(n+1)$-st boundary homomorphism, $B_n = im\partial_{n+1}$.*

Let us observe that since $\partial_{n+1}\partial_n = 0$ then B_n is a subset of Z_n. Therefore, we can already recall the definition of homology groups.

Definition 6 (homology groups) *The n-th homology group is the quotient of the n-boundaries over the n-cycles, that is, $H_n = Z_n / B_n$. The elements of H_n are called n-dimensional homology classes. The n-th Betti number is the rank of H_n.*

Next, the idea is to build a nested sequence of simplicial complexes in order to track the evolution of the homology groups throughout the sequence. The homology classes can merge among themselves following the "elder rule", that is, when merging two homology classes, to consider that the homology class that appeared first in the sequence persists while the other dies off. More formally, given a simplicial complex K and a monotonic continuous function $f : K \to \mathbb{R}$ which is the filtration function, we can define the sublevel set $K(a) = f^{-1}(-\infty, a]$ such that $K(a) \subseteq K(b)$ when $a \leq b$.

Definition 7 (filtration) *Let K be a simplicial complex and let $f : K \to \mathbb{R}$ be a non-decreasing function. A filtration $\mathcal{K}$ of K is a nested sequence of subcomplexes,*

$$\emptyset = K_0 \subset K_1 \subset \dots \subset K_m = K.$$

Such that if $a_1 < a_2 < \dots < a_n$ are the function values of the simplices in K and $a_0 = -\infty$ then $K_i = K(a_i)$ for each i.

The filtration that we will use in this paper is the so called Vietoris-Rips filtration. This filtration is usually applied to point clouds. The filtration function enlarges n-balls from each point. Then, when two of these n-balls intersect, a 1-simplex is built from these two points, establishing a relationship. The process is extrapolated for higher dimensions, i.e., if three balls intersect, a 2-simplex is built, and so on.

As previously mentioned, in general, for every $i \leq j$ we have an inclusion map from $K(a_i)$ to $K(a_j)$. Therefore, we have an induced homomorphism between $H_n(K_i)$ and $H_n(K_j)$.

Definition 8 (Persistent homology) *The corresponding sequence of homology groups connected by homomorphisms obtained from a filtration $\mathcal{K}$:*

$$0 = H_n(K_0) \to H_n(K_1) \to \cdots \to H_n(K_m) = H_n(K)$$

is called the n-th persistent homology of $\mathcal{K}$.

As a next step, the persistent Betti numbers are stocked as 2-dimensional points.

Definition 9 (persistence diagrams) *A persistence diagram is a multiset of 2-dimensional points in the extended real plane. Let $\mu_n^{i,j}$ be the number of n-dimensional homology classes born at $K(a_i)$ and dying entering $K(a_j)$, we have*

$$\mu_n^{i,j} = (\beta_n^{i,j-1} - \beta_n^{i,j}) - (\beta_n^{i-1,j-1} - \beta_n^{i-1,j}),$$

for all $i < j$ and all n. Then, the n-th persistence diagram of a filtration $\mathcal{K}$, denoted as $Dgm_n(\mathcal{K})$, is the multiset of points (a_i, a_j) with multiplicity $\mu_n^{i,j}$ (together with the points of the diagonal with infinity multiplicity by convention).

Finally, two persistence diagrams can be compared using a distance. The following can be considered the most common one, and the one that we will use in the next sections.

Definition 10 (bottleneck distance) *The bottleneck distance between two persistence diagrams A and B is:*

$$d_b(A, B) = \inf_{\phi: A \to B} \sup_{\alpha \in A} ||\alpha - \phi(\alpha)||_\infty$$

where ϕ is any possible bijection between A and B.

Let us describe now an undemanding example as an illustration of these concepts. It is composed by three different datasets showed in Figure 2. The first one samples a circumference (see Figure 2a), the second one is a noisy version of the previous circumference (see Figure 2b), and the last one is composed by two circumferences (see Figure 2c). Then, Vietoris-Rips filtration using the Euclidean metric was computed to obtain the persistence diagrams shown in Figure 3. The 2-dimensional blue and orange points of the persistence diagrams correspond, respectively, to the 0-th and 1-st persistent homology with birth and death time values as its coordinates. In the case of Figure 3a, the 1-st persistent homology H_1 presents just one point that corresponds to the *hole* of the circumference. However, in Figure 3b, some points appear close to the diagonal which can be considered as noise because they are components that live for short. Finally, the two orange points pictured in Figure 3c belong to the 1-st persistent homology H_1, and correspond to the two *holes*, one for each circumference of Figure 2c. A graphical description of the bottleneck distance is shown in Figure 4.

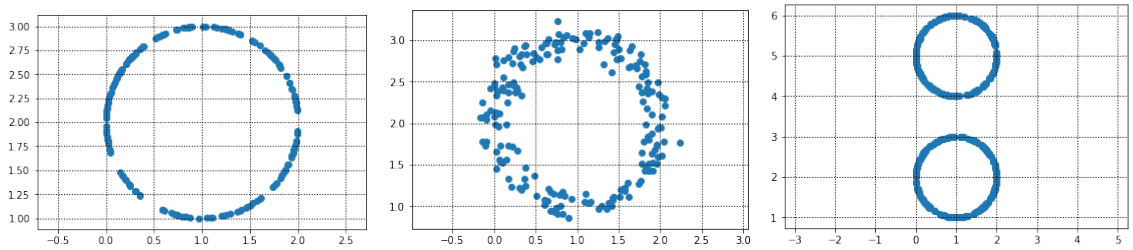
3 Description of the methodology

Next, we describe the methodology based on TDA techniques designed to automatically compare different literary styles. Broadly speaking, given a corpus composed by writings belonging to different categories (e.g., authors, styles, trends,...) a stemming process (which we call `stem`) is applied to each writing where the non-informative words (also called stop-words) are deleted. Then, the `skipgram` word embedding E (described in Section 2.1) is applied to the vocabulary of the corpus, obtaining a high-dimensional representation of the words as a point cloud. Finally, the Vietoris-Rips filtrations of the point clouds corresponding to the writings of the different categories are compared using the bottleneck distance. The pseudocode of this methodology applied to the experiment on Spanish Golden Age poets shown in Section 4 is described in Algorithm 1.

4 Experiment

In this section, we will justify the methodology presented above and describe thoroughly the experimentation process accomplished [2]. In the following subsections we proceed to describe each of the steps of the experiment in detail.

[2]The code is available in https://github.com/Cimagroup/Towards-a-Philological-Metric-Through-a-TDA-Approach

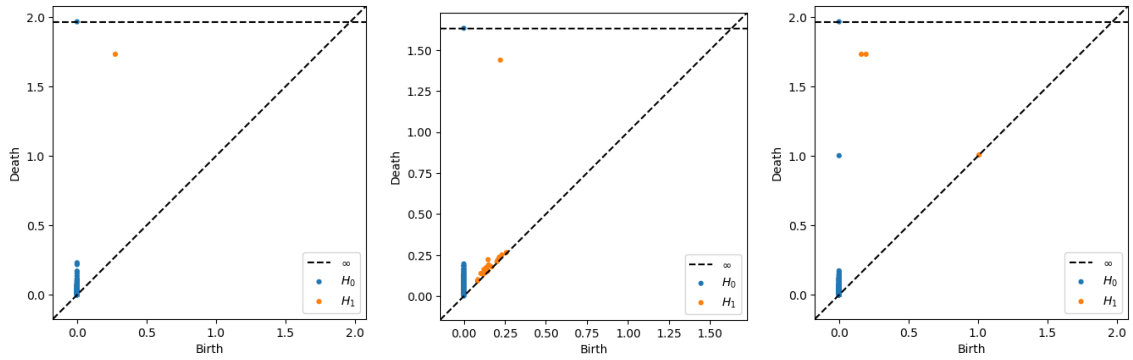


(a) A two-dimensional point cloud sampling a circumference.

(b) A two-dimensional point cloud sampling a noisy circumference.

(c) A two-dimensional point cloud sampling two circumferences.

Figure 2: Three datasets: a circumference, a noisy circumference, and two circumferences.



(a) Persistence diagram of Figure 2a.

(b) Persistence diagram of Figure 2b.

(c) Persistence diagram of Figure 2c.

Figure 3: Two persistence diagrams of the Vietoris-Rips filtration applied to a dataset of a random selection of points from a circumference and from two circumferences, respectively, with the H_0 and H_1 homology classes. We want to point out in Figure 3c there are two points corresponding to the two "holes" in H_1. To appreciate the colors in the images, please visit the online version of the paper.

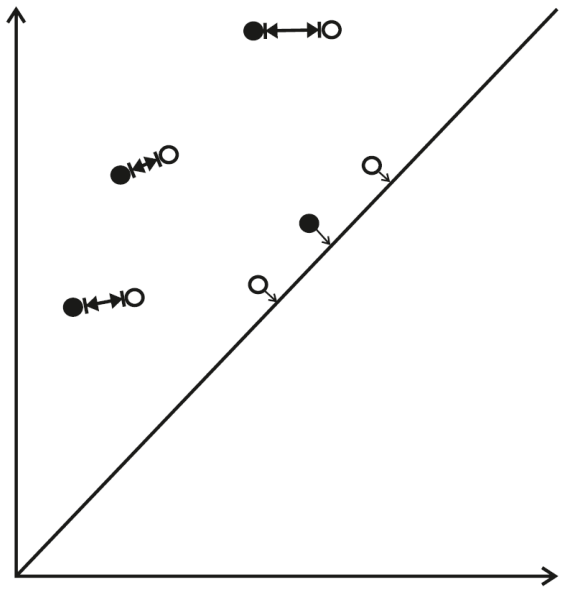


Figure 4: The set of arrows represents the optimum bijection between the black and white points that belong, respectively, to two different persistence diagrams, which are shown overlaid here.

Algorithm 1: Autonomous comparison of literary styles.

Input: A set of sonnets of P_i where P_i for $i \in \{1, \ldots, n\}$ is a given poet.
Output: The bottleneck distance $d_{i,j}$ between each pair of poets P_i and P_j, for $1 \leq i < j \leq n$.
for $i \in \{1, \ldots, n\}$ **do**
  $C_i = \{\text{verse} \mid \text{verse} \in \text{a sonnet of } P_i\}$;
end
$C = \bigcup_{i=1}^{n} C_i$;
$\hat{C} = \emptyset$;
for $v \in C$ **do**
  $\hat{v} = \{\texttt{stem}(w) \mid w \notin \text{stop-words}, w \in v\}$;
  $\hat{C} = \hat{C} \cup \{\hat{v}\}$;
end
$E = E(\hat{C}) \in \mathbb{R}^d$;
for $i \in \{1, \ldots, n\}$ **do**
  $W_i = \{w \in E \mid w = E(\text{word}),\ \text{word} \in \text{a sonnet of } P_i\}$;
  $\mathcal{K}_i = VietorisRipsFiltration(W_i)$;
  $Dgm_0(\mathcal{K}_i) = 0thPersistenceDiagram(\mathcal{K}_i)$;
  for $j \in \{i+1, \ldots, n\}$ **do**
    $d_{i,j} = d_b(Dgm(\mathcal{K}_i), Dgm(\mathcal{K}_j))$;
  end
end

4.1 The context: Spanish Golden Age literature

The Spanish Golden Age literature is a complex framework still alive in the sense that it remains an appealing subject for the literary experts. In this section, we will provide a justification from the literary experts that supports the following experimentation, and give the preliminary literary notions needed to understand it.

The main two concepts in the traditional philological techniques to study literature styles are the *signifier* and the *signified*, terms that come from Saussure's terminology [21]. Signifier and signified compose the so called linguistic *sign* which relates a concept with an abstract image in our mind.

$$\textit{Sign} = \textit{signifier} + \textit{signified}.$$

The signifier is both, the sound and its "acoustic image", and the signified is not just the concept, but a complex content that depends, in most cases, on the context. Both have importance in the desired effect of the poet. To establish comparisons between literary styles of poets of the Spanish Golden Age, we follow a basic philological reference [22], written by the 20-century Spanish poet Dámaso Alonso. According to Dámaso Alonso, some signifiers can evoke something specific. An example that Dámaso Alonso provides is the following verse from Quevedo:

> infame turba de nocturnas aves

both syllables *tur* from *turba* and *nocturnas* evoke obscurity. Besides, we can consider the last example as partial signifiers; any accent, syllable... can be considered as a signifier with its own signified. However, we are interested in studies related to what we consider the inner "stylistic configurations" of the sentences in order to capture them with the `word2vec` embedding. Following the study developed by Dámaso Alonso, poets draw on different stylistic configurations for their verses. The first one we would like to comment can be exemplified by the following sonnet:

> Afuera el fuego, el lazo, el hielo y la flecha
> de amor que abrasa, aprieta, enfría y hiere...

We can see that the main concepts of the first verse correspond member by member to the ones of the second verse, summarizing the following four sentences in the two verses: *Afuera el fuego de amor que abraza; afuera el lazo de amor que aprieta, afuera el hielo de amor que enfría, afuera la flecha de amor que hiere*. It can be described as the following formula:

$$\alpha(A_1 \ \ldots \ A_n)$$
$$\beta(B_1 \ \ldots \ B_n)$$

that summarizes the sentences $\alpha(A_i)\beta(B_i)$ for $i = 1 \dots n$. In the example we had before, α is *afuera* and β is *de amor*. Other kind of resource is the reiterative correlation plurality described in depth in [23]. Let us give an example with a sonnet of Lope de Vega (see Poem 4.1) where Lope de Vega applies the Dámaso Alonso's notion of correlation by dissemination and recollection. In this case, we have again correlation but it is not so rigid as in the first example we

El *humo* que formó cuerpo fingido, ←
que cuando está más denso para en *nada*; ←
el *viento* que pasó con fuerza airada ←
y que no pudo ser en red cogido;

el *polvo* en la región desvanecido ←
de la primera nube dilatada;
la *sombra* que, la forma al cuerpo hurtada, ←
dejó de ser, habiéndose partido,

son las palabras de mujer. Si viene
cualquiera novedad, tanto le asombra,
que ni lealtad ni amor ni fe mantiene.

Mudanza ya, que no mujer, se nombra,
pues cuando más segura, quien la tiene,
tiene *polvo*, *humo*, *nada*, *viento* y *sombra*. ←

Poem 4.1: Sonnet by Lope de Vega.

had, and it is harder to distinguish. The first correlation is disseminated in the quartets (the arrows pointing the verses in Poem 4.1), and the second is recollected in the last verse of the sonnet. By providing these techniques, we want to induce the following idea in the reader; the poets used different methods that concern the configurations of the verses, one example is the correlation we recalled here. Hence, our aim with the `word2vec` algorithm is to encapsulate this kind of configurations. We are concern that it is impossible, for now, to determine which exact literature methods an embedding algorithm catches, or even if it catches any. However, it is true that it can find similarities between words and their use taking into consideration the context of the words. Therefore, it seems natural, in a first approach, to see if the `word2vec` with its `skipgram` variation can imitate or be used instead of the traditional methods in order to distinguish different literature styles. Besides, looking at the mathematical formulation to study the architecture of the sonnets introduced by Dámaso Alonso and his comment [3] *"it would be a labour of a truly team of workers"* to apply such deep studies, in this paper, we take the chance to do that heavy work that Dámaso Alonso mentioned, with recent mathematical tools in a efficient and effective way.

Luis de Góngora and Lope de Vega are, both of them, important poets from the so called Spanish Golden Age. Traditionally, it is said that Luis de Góngora started the Culteranismo literature trend and that Lope de Vega is related to an opposite trend called Conceptismo which had its major representative in Francisco de Quevedo [24, 12]. See also [25] where it is claimed that both trends are related but with elements that distinguish them. However, there exists discrepancies between the literary experts. For example, in [22], Dámaso Alonso did a thorough study of Lope de Vega, and he even developed a study of the comparison of this author with Góngora. He stated that there existed a discontinuous influence by the Góngora's work on the Lope de Vega's work. So, it might not be possible (and it is natural not to be so) to establish rigid difference between such literary trends. In fact, poets present an evolution through their entire productive life, and the different literature trends can be inspired or fed by other trends. We also recommend [26] as an study of the context of these three poets.

4.2 The corpus and the preprocessing step

The corpus we used is a huge dataset composed by the sonnets of different Spanish Golden Age poets [27][4]. Besides, it provides some metrical annotations according to stressed syllables, type of rhyme... In our case we used the sonnets of the three poets we are interested in: Lope de vega, Quevedo, and Góngora. The latter produced less sonnets than the other two so, in order to avoid an unbalanced dataset, we kept 115 sonnets of each poet. These sonnets were chosen without taking into consideration the epoch, or any possible classification of the sonnets that the literary experts could consider, just the first 115 sonnets of the cited dataset.

[3] Free translation. Original comment in Spanish.

[4] The dataset can be found in `https://github.com/bncolorado/CorpusSonetosSigloDeOro`

Method	ϵ
Greenhouse-Geisser	0.563
Huynh-Feldt	0.565

Table 1: Sphericity is an assumption in repeated measures ANOVA designs. When ϵ does not reach 1, the F-score can be inflated and different corrections can be applied. In this case, we tried both Greenhouse-Geisser and Huynh-Feldt. Then,in Table 2, both corrections were applied as well with the sphericity assumption.

Then, each sonnet was pruned as a result of a stemming process. There exists some words that have no value in terms of meaning or that do not provide structure to the sentence such as prepositions: *de, el, la...* As they can be considered noise to the aim we follow, we erased them from the sonnets. Besides, some words are shortened to its root in order to avoid the `word2vec` algorithm to think that different verb tenses or words with different genre are different words. The procedure we applied to delete this non-informative words (also called stop-words) is implemented in the NLTK library [28].

4.3 Application of the `word2vec` algorithm

This step consists in the application of the `skipgram` variation of the `word2vec` algorithm. Specifically, we applied the implementation of this algorithm provided by the Python library `nltk`[5] which is specific for natural language processing tasks. Applying it, we obtained a high-dimensional embedding of the words of the 345 sonnets (115 of each poet). Specifically, the sonnets were embedded in a 150-dimensional space after a 250 iteration training using a window of 10 words. We used a window of 10 words because we wanted to catch patterns using the verses in their full extension, and 10 words is a possible upper bound to the number of words of a verse in a sonnet.

4.4 The filtration and the Bottleneck distance

Having the high-dimensional representation of the words that compose the different sonnets of the dataset, we compute the Vietoris-Rips filtration. The metric used to compute the Vietoris-Rips filtration is the cosine distance because it measures similarity between words by the angle of their vectors, and it is the common distance applied in the `word2vec` algorithm (see [16]). As a result, we have three different 0-th persistence diagrams, one for each poet.

4.5 Results

The methodology shown in Algorithm 1 with the specific procedures and parameters described in Subsection 4.2, Subsection 4.3, and Subsection4.4, was applied and repeated 100 times. The bottleneck distances obtained in these 100 repetitions are shown in Figure 5 using a box-plot representation[6]. There, we can see that the experimentation we applied can infer a significant difference between the bottleneck distances, being closer the persistence diagrams associated to the cloud points representing, respectively, Lope de Vega and Quevedo sonnets. In fact, the third quartile (i.e. the 75% of the dataset) is lower in the case of the bottleneck distance between the persistence diagrams associated to the cloud points representing, respectively, Lope and Quevedo sonnets. Finally, to decide if the differences between these bottleneck distances is significant, a repeated measures ANOVA was applied[7]. In Table 1, the Greenhouse-Geisser and Huynh-Feldt corrections, in case the sphericity assumption is violated, are shown. Then, in Table 2 the different values obtained by the application of the repeated measures ANOVA are displayed. There, we can infer that there exists a significant difference between the three groups of bottleneck distances as we expected by visualizing Figure 5. Finally, to specifically determine which of the groups is the different one, a pairwise comparison was computed in Table 3, concluding that the sample of the bottleneck distances of Quevedo and Lope de Vega are significantly different from the other two.

5 Conclusion

Extracting knowledge from more and more complex datasets is a hard work which requires the help of techniques coming from other fields of science. In this way, representing the data as points of a metric space open a bridge between

[5] https://www.nltk.org/

[6] In a box-plot, the higher horizontal line correspond to the maximum value and the lower horizontal line to the minimum value. The horizontal line in the middle of the box corresponds to the median, the top of the box is the third quartile, and the bottom of the box is the first quartile. Finally, the circumferences correspond to outliers.

[7] MedCalc software (https://www.medcalc.org/index.php) was used to do the statistical validation.

Source of variation		Sum of Squares	DF	Mean Square	F	p-value
Factor	Sphericity assumed	0.00834	2	0.00417	51.42	< 0.001
	Greenhouse-Geisser	0.00834	1.126	0.00741	51.42	< 0.001
	Huynh-Feldt	0.00834	1.130	0.00738	51.42	< 0.001
Residual	Sphericity assumed	0.0161	198	0.0000811		
	Greenhouse-Geisser	0.0161	111.452	0.000144		
	Huynh-Feldt	0.0161	111.850	0.000144		

Table 2: The repeated measures ANOVA was applied to infer if there exists a significant difference between the bottleneck distances. In the table, DF are the degrees of freedom. A p-value lower than 0.001 and a F-value of 51.42 were reached. So, we can say that there exists a significant difference and in Table 3, we did a pairwise comparison to determine which of the bottleneck distances is the different one.

Factors			Mean difference	Standard Error	p-value	95% CI
A	-	B	-0.000386	0.000442	1.0000	-0.00146 to 0.000690
		C	0.0110	0.00155	<0.0001	0.00721 to 0.0148
B	-	A	0.000386	0.000442	1.0000	-0.000690 to 0.00146
		C	0.0114	0.00150	<0.0001	0.00771 to 0.0150
C	-	A	-0.0110	0.00155	<0.0001	-0.0148 to -0.00721
		B	-0.0114	0.00150	<0.0001	-0.0150 to -0.00771

Table 3: A pairwise comparison was done. Here, A corresponds the sample of the bottleneck distances between Lope de Vega and Góngora, B corresponds to Quevedo and Góngora, and C corresponds to Quevedo and Lope de Vega. As it is shown, the p-value is lower than 0.001 when we compare with C. Therefore, the sample of the bottleneck distances between Quevedo and Lope de Vega is significantly different from the other two. The p-value and the confidence intervals were Bonferroni corrected.

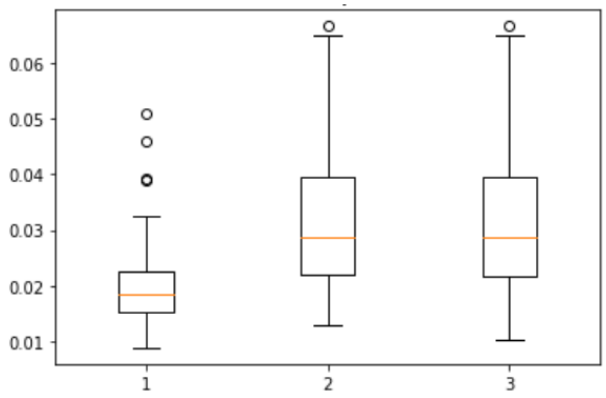


Figure 5: Box-plot showing the bottleneck distance results obtained from the sonnets of the three poets. (1) is the box-plot of the bottleneck distance obtained from the comparison between the sonnets of Quevedo and Lope, (2) is the box-plot of the bottleneck distance obtained from the comparison between the sonnets of Quevedo and Góngora, and (3) is the box-plot of the bottleneck distance obtained from the comparison between the sonnets of Lope de Vega and Góngora. From this box-plot, we can expect that the literary styles of Quevedo and Lope are significantly closer.

research fields which are apparently far away. The use of TDA techniques is a new research area which provides tools for comparing properties of point clouds in high-dimensional spaces, and therefore, for comparing the datasets represented by such point clouds.

In this paper, we propose the use of such TDA techniques in order to compare different stylistic trends in the literature. In this approach, bottleneck distance between the persistence diagrams of the Vietoris-Rips filtration obtained form the cloud points representing the sonnets of two different writers encode the differences among their literary styles and quantifies the nearness among them.

This novel approach opens a door for the interaction of TDA and philological research. TDA techniques can be applied in order to give a topological description of a work, a writer or an age and go deeper into their belonging to a greater trend. In addition, philology can suggest new ways to measure the nearness among styles which can be useful for applying TDA techniques in other application areas.

References

[1] Shusen Liu, Di Wang, Dan Maljovec, Rushil Anirudh, Jayaraman J. Thiagarajan, Sam Ade Jacobs, Brian C. Van Essen, David Hysom, Jae-Seung Yeom, Jim Gaffney, Luc Peterson, Peter B. Robinson, Harsh Bhatia, Valerio Pascucci, Brian K. Spears, and Peer-Timo Bremer. Scalable topological data analysis and visualization for evaluating data-driven models in scientific applications. *CoRR*, abs/1907.08325, 2019.

[2] Henri Riihimäki, Wojciech Chacholski, Jakob Theorell, Jan Hillert, and Ryan Ramanujam. A topological data analysis based classification method for multiple measurements. *CoRR*, abs/1904.02971, 2019.

[3] Karthikeyan Natesan Ramamurthy, Kush R. Varshney, and Krishnan Mody. Topological data analysis of decision boundaries with application to model selection. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5351–5360. PMLR, 2019.

[4] K. Johnson. *Quantitative methods in linguistics*. Blackwell Pub., 2008.

[5] Md Shidur Rahman. The advantages and disadvantages of using qualitative and quantitative approaches and methods in language "testing and assessment" research: A literature review. *Journal of Education and Learning*, 6(2):102–112, 2017.

[6] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168, 2013.

[7] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL, 2014.

[8] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *TACL*, 5:135–146, 2017.

[9] M.M. Deza and E. Deza. *Encyclopedia of Distances*. Encyclopedia of Distances. Springer Berlin Heidelberg, 2009.

[10] Shafie Gholizadeh, Armin Seyeditabari, and Wlodek Zadrozny. Topological signature of 19th century novelists: Persistent homology in text mining. *Big Data and Cognitive Computing*, 2(4), 2018.

[11] Tadas Temčinas. Local homology of word embeddings, 2018.

[12] J. Rutherford. *The Spanish Golden Age Sonnet*. Iberian and Latin American Studies. University of Wales Press, 2016.

[13] Zi Yin and Yuanyuan Shen. On the dimensionality of word embedding. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 895–906, 2018.

[14] Felipe Almeida and Geraldo Xexéo. Word embeddings: A survey. *CoRR*, abs/1901.09069, 2019.

[15] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA, 2013. Curran Associates Inc.

[17] David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. A closer look at skip-gram modelling. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias, editors, *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006*, pages 1222–1225. European Language Resources Association (ELRA), 2006.

[18] Jie Hu, Shaobo Li, Yong Yao, Liya Yu, Yang Guanci, and Jianjun Hu. Patent keyword extraction algorithm based on distributed representation for patent classification. *Entropy*, 20:104, 02 2018.

[19] Herbert Edelsbrunner and John L. Harer. *Computational Topology, An Introduction*. American Mathematical Society, 2010.

[20] James R. Munkres. *Elements of Algebraic Topology*. Addison Wesley Publishing Company, 1984.

[21] F. de Saussure, C. Bally, A. Sechehaye, A. Riedlinger, and A. Alonso. *Curso de lingüística general*. Filosofía y teoría del lenguaje. Editorial Losada, 1965.

[22] D. Alonso. *Poesía española: ensayo de métodos y límites estilísticos : Garcilaso, Fray Luis de León, San Juan de la Cruz, Góngora, Lope de Vega, Quevedo*. Biblioteca románica hispánica: Estudios y ensayos. Editorial Gredos, 1966.

[23] 1898-1990. Alonso, Dámaso. *Versos plurimembres y poemas correlativos : capítulo para la estilística del Siglo de Oro*. Sección de Cultura e Información Artes Gráficas Municipales, Madrid, 1944. Separata de: Revista de la Biblioteca, Archivo y Museo. Año XIII, núm. 49 (1944).

[24] Dámaso Chicharro Chamorro. Sobre los orígenes del conceptismo andaluz: Alonso de bonilla. *Boletín del Instituto de Estudios Giennenses*, (130):59–84, 1987.

[25] Santiago Molfulleda. Sobre la oposición entre culteranismo y conceptismo. *Universitas Tarraconensis. Revista de Filologia*, (6):55–62, 2018.

[26] Juan Manuel Rozas. Góngora, lope, quevedo. poesía de la edad de oro, ii, 2002.

[27] Borja Navarro, María Ribes Lafoz, and Noelia Sánchez. Metrical annotation of a large corpus of Spanish sonnets: Representation, scansion and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4360–4364, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[28] Edward Loper and Steven Bird. NLTK: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*, 2002.