frontiers
in Digital Humanities

ORIGINAL RESEARCH
published: 20 June 2018
doi: 10.3389/fdigh.2018.00015

On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry

*Borja Navarro-Colorado**

Software and Computing Systems, University of Alicante, Alicante, Spain

This paper analyzes the application of LDA topic modeling to a corpus of poetry. First, it explains how the most coherent LDA-topics have been established by running several tests and automatically evaluating the coherence of the resulting LDA-topics. Results show, on one hand, that when dealing with a corpus of poetry, lemmatization is not advisable because several poetic features are lost in the process; and, on the other hand, that a standard LDA algorithm is better than a specific version of LDA for short texts (LF-LDA). The resulting LDA-topics have then been manually analyzed in order to define the relation between word topics and poems. The analysis shows that there are mainly two kinds of semantic relations: an LDA-topic could represent the subject or theme of the poem, but it could also represent a poetic motif. All these analyses have been undertaken on a large corpus of Golden Age Spanish sonnets. Finally, the paper shows the most relevant themes and motifs in this corpus such as "love," "religion," "heroics," "moral," or "mockery" on one hand, and "rhyme," "marine," "music," or "painting" on the other hand.

Keywords: poetry, topic modeling, LDA, Spanish, sonnet, golden age, distant reading

OPEN ACCESS

Edited by:
Stan Szpakowicz,
University of Ottawa, Canada

Reviewed by:
Graeme Hirst,
University of Toronto, Canada
David Mimno,
Cornell University, United States
John Nerbonne,
Albert Ludwigs Universität Freiburg, Germany

****Correspondence:***
Borja Navarro-Colorado
borja@dlsi.ua.es

Specialty section:
This article was submitted to Digital Literary Studies, a section of the journal Frontiers in Digital Humanities

Received: *30 August 2017*
Accepted: *01 June 2018*
Published: *20 June 2018*

Citation:
Navarro-Colorado B (2018) On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry. Front. Digit. Humanit. 5:15. doi: 10.3389/fdigh.2018.00015

1. INTRODUCTION

The purpose of this paper is two-fold: first, to apply Latent Dirichlet Allocation (LDA) topic modeling (Blei et al., 2003) to a corpus of Spanish poetry (sonnets) and to extract a set of coherent and representative "topics" (meaning LDA topics), and then to analyze what kinds of LDA-topics[1] can be extracted from a corpus of poetry like this.

Although LDA has been applied with some reliability to literary prose (Jockers and Mimno, 2013) or drama (Schöch, 2017) before, its application to poetic text (Rhody, 2012; Herbelot, 2015) is still a subject of analysis. As well as other distributional semantic models, LDA topic modeling is based upon the contextual use of words: it assumes that if certain words tend to appear together in different texts, it is because they refer to the same topic. Although it is a model appropriate for scientific and many other types of texts, it is not clear if it is suitable for poetry, a genre in which a word related to a specific topic is frequently used to mean something else or to refer to another topic or topics. The main question of this paper is, then, if it is possible to apply Topic Modeling to poetry and what kind of formal topics will be obtained.

[1] "LDA-topic" will hereafter be referring to the output of an LDA algorithm (that is, a given set of words; see Blei et al., 2003; Blei, 2012 for a formal definition), and "topic" will hereafter be referring to the literary concept of topic: the subject or theme of a poem.

This analysis has been done on a Corpus of Spanish Golden Age Sonnets[2], composed of 5,078 sonnets, more than 71,092 lines and 52 poets (Navarro-Colorado et al., 2016). All these sonnets were written during the sixteenth and seventeenth centuries in Spanish (Castilian). There are several reasons why a corpus like this is appropriate for such an analysis. First of all, it is representative of a large literary period: the Spanish Golden Age, that cover the sixteenth and seventeenth centuries. Even though we can find only a single poetic form within it (the Spanish sonnet, composed by 14 hendecasyllables with consonant rhyme), it was used (and is still used) by a great amount of poets to talk about a large number of themes. Therefore, the corpus is structurally specific but semantically rich and diverse.

The paper is structured as follows. In the next section I will explain, first, the distributional-semantic foundations of topic modeling in order to clarify the kind of topics we can expect to find. Then I will show several applications of LDA to literary texts including poetry. The next two sections (4 and 5) are devoted to establishing a set of coherent LDA-topics from this corpus. Two LDA topic modeling algorithms will be compared: a standard LDA algorithm and a specific version of LDA for short texts based on word embeddings (Nguyen et al., 2015). Then the coherence of the extracted LDA-topics will be automatically evaluated applying common techniques such as nPMI measure or intruder detection. Section 6 presents the manual analysis of the most coherent LDA-topics extracted in the previous section. The analysis is focused on the relation between word topics and the main poems in which each LDA-Topic appears. I will show how the application of LDA to poetry extracts not only thematic topics but also poetic motifs. The paper will finish with a summary of the main conclusions.

2. TOPIC MODELING AND LITERARY STUDIES

LDA topic modeling is based on distributional semantic models (Turney and Pantel, 2010). These models define meaning as use (Wittgenstein, 1953): word meaning is determined by the contexts in which the word appears. It is usually represented as a vector in a vector space formed by the contextual words with which a word could appear in real texts. Through different vector similarity measures, distributional models are able to establish semantic relations between words according to the similarity of their contextual vectors. In this way, words with similar contextual vectors could be clustered together as semantically related words. When applied to literary texts, these models will show us how words are used for artistic purposes.

Based on this semantic model, LDA algorithms try to automatically discover topics in a set of documents according to how words share contexts. Formally, a "topic" (LDA-topic) is a distribution over a fixed vocabulary (Blei et al., 2003; Blei, 2012), that is, a high-probability set of words assigned to a topic (word topics or keywords). Like other distributional models, deep down it considers that if two words tend to appear in the same context, they are probably related to the same topic. LDA randomly assigns a topic to each word in the corpus, and then refines the assignment according to the most frequent topic in the document and the most frequent topic for the word in the whole corpus. It assumes that each document has more than one topic. Therefore, according to the words of each document, LDA specifies the probabilistic weights of each topic in each document[3].

Blei (2012) exemplifies how LDA works applying it to scientific texts from *Science* magazine. The paper shows how some LDA-topic keywords are clearly related with general themes such as "genetics," "evolution," or "computers." From the point of view of distributional semantics, topic modeling is appropriate for these types of texts because science texts have very specific terms for specific topics. I'm referring to terminology. It appears in specific contexts and it is related to specific topics. Given that topic modeling, as well as other distributional semantic models, is based upon the contextual use of words, the presence of specific words for specific topics used in specific contexts helps the algorithm extract clear and coherent topics.

In literary texts, and especially in poetry, however, words could share contexts in a non-conventional way: they are often used in contexts which are different from those that are generally used in non-literary texts. In other words, it is very common in poetry to use words related to a topic to talk about other topics (for instance metaphors, similes, figurative uses, etc.). Contextual use of words is different in scientific texts and in poetic texts. The question is, then, if it is possible to apply Topic Modeling to literary texts and what kind of topics will be obtained, keeping in mind that many words are being used out of its usual context.

During the last years several papers suggested the use of LDA topic modeling for literary text analysis, like Rhody (2012), Jockers and Mimno (2013), Tangherlini and Leonard (2013), Lou et al. (2015), Roe et al. (2016), and Schöch (2017).

Jockers and Mimno (2013), for example, use topic modeling to extract relevant themes from a corpus of nineteenth-century novels. They present a classification of topics according to genre, showing that in nineteenth-century English novels males and females tended to write about the same things but to very different degrees. For example, males preferred to write about guns and battles, while females preferred to write about education and children. From a computational point of view, this paper concludes that topic modeling must be applied with care to literary texts and it proves the needs for statistical tests that can measure confidence in results.

Roe et al. (2016) applied LDA to French *Encyclopédie* by Diderot and d'Alembert (1751–1772). LDA-topics are used here as as a form of exploratory data analysis to investigate the complex discursive makeup of the texts in the *Encyclopédie*. They compare the extracted LDA-topics with the original classification scheme of the articles of the *Encyclopédie*, showing several discrepancies between LDA-topics and the original classification of texts.

Schöch (2017) applies topic modeling to French Drama of the Classical Age and the Enlightenment, in an attempt to

[2] Freely available at https://github.com/bncolorado/CorpusSonetosSigloDeOro

[3] For a clear introduction to LDA Topics Modeling, see Blei (2012).

discover semantic types of topics and analyze if different dramatic subgenres have distinctive dominant topics. Eventually he finds subgenres with clearly distinctive topics.

Although LDA has been applied mainly to prose, some papers analyze its application to poetry. In this way, Rhody (2012) suggests that the outcome will be different from the application of topic modeling to non-figurative texts. When it is applied to figurative texts, some "opaque" topics (topics formed by words with no apparent semantic relation between them) really show symbolic and metaphoric relations. Rather than "topics," these topics represent symbolic meanings. She concludes that, in order to understand them, a close reading of the poems is necessary.

In this paper LDA topic modeling will be used for the analysis of poetry. Along with Rhody (2012), I will go further on these "opaque topics" in order to shed (some more) light on this.

3. A CORPUS OF SPANISH GOLDEN AGE SONNETS

As I said in section 1, the corpus used to develop this analysis is the Corpus of Spanish Golden Age Sonnets (Navarro-Colorado et al., 2016). It is composed of 5,078 sonnets of 52 poets (around 476,165 tokens, 28,599 types and more than 71,092 lines) written during the sixteenth and seventeenth centuries in Spanish.

This period of the Spanish Literature has also been the object of analysis of several literary studies (García Berrio, 1978; Rivers, 1993) whose goal is to find a common thematic typology for Golden Age Spanish sonnets, in order to show the present-day reader the complex thematic relations that exist between said poems. It is assumed that all these poems were composed according to a set of thematic principles and rules that the present-day reader finds difficult to understand (García Berrio, 1978). As a matter of fact, what the Golden Age poets do is to assume and rekindle a set of more or less specific topics established by the literary classics (Dante, Petrarch, etc.) and make them their own. These topics set the context for the Golden Age sonnets, and it is in this context that the poems are created and must be interpreted. These literary studies are based on a representative but short sample of poems: up to 1,500 sonnets were analyzed in García Berrio (1978). Although it is a great number of poems for a close reading, it is far from close to the thousands of sonnets composed during this period.

Following Moretti's Distant Reading framework (Moretti, 2007), the objective of this corpus is to be representative of the variety of sonnets that were written during that period. The corpus attempts to represent the kinds of sonnets written during Golden Age, and thus establishing the literary context for any Golden Age poem. Therefore, it includes all the authors of this period who wrote a significant number of sonnets. Authors who wrote but few sonnets (<10) have been rejected. **Table 1** shows the complete list of poets and the number of sonnets written by each one[4].

TABLE 1 | List of poets (chronological order).

Author	Number of Sonnets
Juan Boscán (c.1492–1542)	100
Garcilaso de la Vega (c.1499–1536)	38
Diego Hurtado de Mendoza (c.1503–1575)	60
Diego Hernando de Acuña (1520–1580)	84
Gutierre de Cetina (1519–1554)	247
Juan de Timoneda (c.1520–1583)	31
Diego Ramírez Pagán (1524–1562)	11
Fray Luis de León (1527–1591)	10
Juan de Almeida (d. 1572)	42
Baltasar de Alcázar (1530–1606)	18
Fernando de Herrera (1534–1597)	320
Francisco de Figueroa (c.1530–c.1588)	20
Diego Ximénez de Ayllón (1530–1590)	55
Francisco de la Torre (1534–1594)	78
Francisco de Aldana (1537–1578)	43
Miguel de Cervantes (1547–1616)	77
Andrés Rey de Artieda (1549–1613)	11
Pedro de Padilla (c.1550–1599)	43
Cristóbal de Virués (1550–614)	10
Luis de Góngora (1561–1627)	115
Lupercio Leonardo de Argensola (1559–1613)	60
Bartolomé Leonardo de Argensola (1562–1631)	158
Juan de Salinas (1559–1643)	18
Lope de Vega (1562–1635)	1346
Juan de Arguijo (1567–1623)	70
Francisco de Medrano (1570–1607)	51
Antonio Mira de Amescua (1577–1644)	32
Luis Martín de la Plaza (1577–1625)	21
Pedro de Espinosa (1578–1650)	20
Tirso de Molina (1579–1648)	56
Francisco de Borja y Aragón, príncipe de Esquilache (1581–1658)	142
Francisco de Quevedo (1580-1645)	517
Francisco López de Zárate (1580–1658)	58
Juan de Tassis y Peralta, conde de Villamediana (1582–1622)	203
Juan de Jáuregui y Aguilar (1583–1641)	23
Luis de Ulloa Pereira (1584–1674)	106
Pedro Soto de Rojas (1584–1658)	125
Luis Carrillo y Sotomayor (1585–1610)	50
Antonio Hurtado de Mendoza (1586–1644)	13
Esteban Manuel de Villegas (1589–1669)	12
Bernardino de Rebolledo (1597–1676)	54
Jerónimo de Cáncer y Velasco (c. 1599–1655)	17
Anastasio Pantaleón de Ribera (1600–1629)	18
Antonio Enróquez Gómez (1602–1660)	40
Gabriel Bocángel y Unzueta (1603–1658)	77
Jacinto Polo de Medina (1603–1676)	21
Francisco de Trillo y Figueroa (1618–1680)	49
Agustin de Salazar y Torres (1642–1673)	30
Sor Juana Inés de la Cruz (1651–1695)	72
José de Litala y Castelví (1672–1701)	152

[4] The corpus is unbalanced. However, for the objectives of this paper, the corpus will be processed and analyzed as a whole, without taking in mind the author of each poem.

The first and most influential poet is Garcilaso de la Vega. His 38 sonnets were widely imitated and commented. Together with Juan Boscán, he introduced the Italian sonnet and the Petrarquian topics into Spanish Literature. Among the other poets, Luis de Góngora stands out with him the Baroque period begins and he introduced a new style based on complex syntax and complex poetry images; there is also Lope de Vega, who was the most famous poet in that period; and Francisco de Quevedo, who developed several traditional and baroque topics. At the end of the period sor Juana Inés de la Cruz stands out, because she introduced a new feminine point of view about traditional topics (Terry, 1993).

Most sonnets have been obtained from the Miguel de Cervantes Virtual Library[5]. Texts have been modernized following current spelling rules of Spanish. Unfortunately, several sonnets had some typos that have been fixed. In order to ensure the re-usability of the corpus, each sonnet was marked with the standard TEI-XML[6]. The annotation includes the main metadata (title, source, encoding, etc.), the sonnet structure and the metrical pattern of each line.

4. CORPUS PRE-PROCESSING, NUMBER OF TOPICS AND ALGORITHMS

As I said before, the first objective of this paper is to extract coherent and representative LDA-topics from Spanish Golden Age poetry. Assuming that this corpus is representative of the period, the quality of the LDA-topics will depend on three factors:

- how the corpus is pre-processed,
- the number of LDA-topics extracted, and
- the specific LDA algorithm applied.

In order to extract these coherent topics, several experiments have been run modifying these parameters in each one of them. This section explains how they have been settled out and the next one will show the evaluation of the LDA-topic extracted in each experiment.

4.1. Corpus Pre-processing

A decisive factor when extracting LDA-topics from a corpus is how the corpus is pre-processed. It depends on what are you looking for. In some cases, as in Jockers and Mimno (2013) for example, LDA Topics are extracted taking into consideration only nouns. In other cases, as in Schöch (2017), the corpus is lemmatized; and many others just use raw texts.

For our purposes, two corpus pre-processing methods have been tested. In the first one, the stop words have been filtered. In this way, all words from closed Part of Speech (like articles, conjunctions, prepositions, etc.) have been deleted from the corpus. In the second one, besides using the stop-words filter, the whole corpus has been lemmatized, applying LDA to the lemma of each word. The corpus has been lemmatized with FreeLing[7] (Padró and Stanilovsky, 2012).

What I intend to do with these two pre-processing methods is to find out to what extent the coherence and representativity of LDA-topics extracted from poetry depend on word morphology, that is, whether it is preferable to consider that inflection creates different forms of the same words, or to disregard inflection and view inflected words as just a single one. In the first case, the LDA-topics will be affected by morphological structures (and, therefore, by poet style), but not in the second case.

4.2. Number of Topics

The second factor to consider when we want to obtain a coherent set of LDA-topics is the number of LDA-topics that must be extracted. LDA algorithms cannot find out on their own what is the appropriate number of LDA-topics that a given corpus must have. It must be stated previously. A small number of LDA-topics will result in topics that are too general and vague; a great number of LDA-topics will extract noisy topics. It dependes also on the granularity of the analysis, whether the LDA-topics need to be very specific or rather broad and general.

As I said in the introduction, in traditional literary studies there have been several attempts to establish and define the thematic typology for the Golden Age Spanish sonnets. However, there is no agreement about the number of topics that can be found: the proposals go from 20 topics (Rivers, 1993) to more than 90 (García Berrio, 1978).

Therefore, it is not clear what the best number of LDA-topics for the corpus of Golden Age sonnets is. Different tests have been run using different amounts of topics: 10, 25, 50, 100, and 250 LDA-topics. In this way, I will try to specify the appropriate level of granularity for this corpus of poetry.

4.3. Algorithms

The quality of the LDA-topics we obtain depends mainly on the LDA algorithm used to extract them. Topic modeling is not a specific algorithm, but a family of algorithms devoted to the extraction of topics (as a set of semantically related words) from texts. LDA (Blei et al., 2003) was the first topic modeling algorithm and today it is considered as the standard algorithm. Since then, many different topic modeling algorithms have been developed[8].

Far from testing a great number of topic modeling algorithms, in this paper I will focus mainly on two: the standard LDA implemented in MALLET[9] (McCallum, 2002) and a specific version of LDA developed for short documents and based on words embeddings called LF-LDA (Latent Feature LDA Nguyen et al., 2015).

The reason why I compare these two algorithms is the size of the poems of the corpus. A sonnet is a relatively short text: 14 lines of 11 syllables each, around 85–95 words per text. Since vector space models of semantics need large contexts and standard LDA was developed for more or less large documents,

[5] http://www.cervantesvirtual.com/

[6] www.tei-c.org/

[7] http://nlp.lsi.upc.edu/freeling/

[8] See http://www.cs.columbia.edu/~blei/topicmodeling_software.html for some of them

[9] http://mallet.cs.umass.edu/

I think that a specific LDA algorithm for short texts will extract better LDA-topics than the standard LDA.

LF-LDA tries to solve LDA's dependency on large contexts/documents by means of word embeddings. The LDA component that assigns words to topics is broadened with two matrices of latent features vectors trained on a large external corpus. One of these matrices associates latent-feature vectors with topics and the other one with words (Nguyen et al., 2015).

For our purposes, the algorithms have been run with the following parameters; in order to ensure a stable level of topic-to-word assignment, standard LDA was run over the corpus (both filtered and lemmatized) with more than 1,000 iterations and an interval optimization every 10 iterations. LF-LDA was run over the corpus with a similar configuration. The latent features vectors for LF-LDA were extracted with the Word2Vec model (Mikolov et al., 2013) implemented in Gensim[10] (Řehůřek and Sojka, 2010) from the same corpus of poetry and with a context window of 5 words. From these word embeddings, LF-LDA extracts the latent features needed to assign topics to words.

Latent features have been extracted from the same corpus of poetry in order to maintain the distributional relationships between words in this specific poetic context. Although, according to Herbelot (2015), it is appropriate to use a corpus of common word usages for a distributional analysis of poetry, in this paper I prefer to keep the corpus in isolation in order to maintain and represent only the poetic semantic relationships, regardless of the semantic word relationships in the common use of language. This way I will analyze how LF-LDA works with poetic texts only. The difference with the standard LDA is that each word is represented trough a latent-feature vector (based on Word2Vec) in a context of 5 words throughout the whole corpus. This 5-windows context is roughly the size of a line. Therefore, LF-LDA will extract LDA-topics looking for word relations not only inside every individual poem or all along throughout the whole corpus, but also within every single line of text.

Up to 20 experiments have been run in order to find the most coherent and representative LDA-topics of the Golden Age Spanish sonnets. Ten experiments were performed with the corpus filtered and ten with the corpus lemmatized. In each one, five experiments (one for each amount of LDA-topics: 10, 25, 50, 100, and 250) were run with the standard LDA algorithm, and the other five with the LF-LDA algorithm. Results are shown and discussed in the next sections (**Tables 2**, **3**).

5. AUTOMATIC EVALUATION OF TOPIC COHERENCE

LDA-topics have been both evaluated automatically and analyzed manually. Automatic evaluation is focused on specifying the coherence of LDA-topics, while manually analysis is designed to check their representativity. Since manual analysis of LDA-topics is a hard and time-consuming task, I have first performed the automatic evaluation in order to know which LDA configuration produces the most coherent LDA-topics. Once this set of coherent LDA-topics has been stablished, they have to be manually analyzed in order to see their representativity. This section is devoted to the automatic evaluation (coherence) and the next one to the manual analysis.

The evaluation of LDA-topics is not an easy task. It has been the subject of many research papers (Chang et al., 2009; Newman et al., 2010; Lau et al., 2014; Bhatia et al., 2017). All of them try to measure the coherence of the set of word-topics. Chang et al. (2009) first put forward the intruder method as a manual evaluation technique. It consists in introducing an "intruder" word (a word that has nothing to do with the topic) among the first five words of a topic, and then asking human evaluators if they are able to locate the intruder word. The coherence of an LDA-topic will depend on how easily a human annotator can detect the intruder word. The easier it is for him/her to find it, the more coherent the LDA-topic will be. Lau et al. (2014) developed the automatic version of this intruder-word detection technique. Newman et al. (2010), on their own, proposed to measure the coherence of LDA-topics through pointwise mutual information (PMI) for different pairings of topic words. It calculates word co-ocurrences on a sliding window over Wikipedia. This measure was improved by Lau et al. (2014) introducing a normalized PMI (nPMI) (Bouma, 2009), a measure that normalizes the co-ocurrence values between -1 (no co-ocurrence) to 1 (the highest co-ocurrence). Bhatia et al. (2017), finally, put forward a new approach to topic modeling evaluation: instead of evaluating the coherence of the LDA-topics, they suggest evaluating the coherence of the main documents related to each LDA-topic.

Following Lau et al. (2014), the LDA-topics extracted from the corpus of sonnets have been evaluated using the two most common measures: nPMI and the automatic version of intruder-word technique[11]. **Tables 2**, **3** show the results obtained in each experiment according to the configurations explained in the previous section. The first table shows the exact nPMI values obtained from the same corpus. I'm not so interested in absolute values as in the relative values that emerge from the comparison among the different experiments. The values of **Table 3** correspond to the number of coherent topics (that is, those LDA-topics in which the system has detected the intruder word) and the coherence percentage. Intruder words were randomly extracted from the same corpus.

Although these results are lower than the values obtained in standard texts (see Lau et al., 2014), these data show that the standard LDA algorithm achieves more coherent LDA-topics than the LF-LDA. Only using 10 topics from the lemmatized corpus LF-LDA outscores the standard LDA (in one point). The introduction of latent features with the context of a 5-words window (more or less the size of a verse) has not improved the coherence of the LDA-topics. A standard approach, using the poem as a contextual unit, has been better. In this respect, we can conclude that, from a distributional point of view, the poetic

[10] https://radimrehurek.com/gensim/

[11] In both cases I have used the code available at J. H. Lau GitHub page: https://github.com/jhlau/topic_interpretability

TABLE 2 | Results for nPMI evaluation technique.

	10 topics	25 topics	50 topics	100 topics	250 topics
LDA filtered	0.076	**0.103**	**0.129**	**0.149**	**0.15**
LDA lemmatized	**0.081**	0.089	0.112	0.124	0.135
LF-LDA filtered	0.046	0.072	0.088	0.105	0.108
LF-LDA lemmatized	0.058	0.061	0.069	0.079	0.07

Highest results in bold.

TABLE 3 | Results for "word intruder" evaluation technique.

	10 topics	25 topics	50 topics	100 topics	250 topics
LDA filtered	7 70%	**19 76%**	**41 82%**	**85 85%**	**211 84.4%**
LDA lemmatized	7 70%	18 72%	**41 82%**	74 74%	179 71.6%
LF-LDA filtered	7 70%	18 72%	40 80%	70 70%	150 60%
LF-LDA lemmatized	**8 80%**	16 64%	34 68%	57 57%	101 40,4%

Highest results in bold.

meaning of a word is obtained using the whole poem as a context, and not smaller poetic units such as the verse[12].

According to corpus pre-processing, results show that a simple stop-words filter is better than a complete corpus lemmatization. During lemmatization some morphological information is lost. It seems that this information is relevant for the coherence of poetic LDA-topics. Mainly the information related to verb inflection is important here. Spanish, as other romance languages, has a rich verbal inflection, and there is a great number of temporal, modal or aspectual features that are lost during lemmatization. It seems that such a morphological information is relevant in some way for the topic modeling of poetry[13].

So, when the corpus has not been lemmatized, LDA-topics are made up mainly of nouns. Whenever we come across a verb, it always shows some very specific topic feature related to time, tense, aspect, etc. For example, if we just take 10 topics from the non-lemmatized corpus, one time-related topic will appear, as it happens in Topic 7. This topic represents a common theme during the seventeenth century in Spain: the decline of the Spanish Empire. In these LDA-topics the verb "to be" appears in its form "fue" (was/were) next to other topic words such as "gloria" (glory), "valor" (courage), "fama" (fame), "mundo" (world), or "España" (Spain). The word "fue" (was/were) has a great relevance in the topic because these poems describe what the Spanish Empire was like. The same topic appears when the corpus is lemmatized, but this important word has disappeared.

Similarly, when 100 LDA-Topics have been extracted, two of them are related to verb tense: topic 61 shows topic words related to the simple past tense such as "fue" (he/she was), "vio" (saw), "pudo" (could), "dio" (gave), "tuvo" (got), or "quiso" (wanted); and topic 98 shows words related to the simple present tense, both in indicative and subjunctive, such as "sea" (he/she will be), "vea" (will see), "desea"(wishes), "emplea" (employs), "rodea" (surrounds), or "pelea" (fights).

Therefore, the information lost during the lemmatization process is indeed relevant for the extraction of coherent LDA-topics in a poetry corpus. This kind of information (such as verb inflection) is in some way related to the LDA-topics.

Finally, these data clearly show that topic coherence increases just as the number of topic increases up to 100 topics. The appropriate number of topics that should be assigned to this corpus must range from 50 to 100.

To conclude, the most coherent LDA-topics are obtained from the corpus of Golden Age Spanish sonnets by means of a standard LDA algorithm, a simple stop-words filter and aiming for a total of 50–100 topics. This set of topics will be manually analyzed in the next section.

6. MANUAL ANALYSIS OF LDA-TOPICS REPRESENTATIVITY

Once the most coherent set of LDA-topics has been established, in this section I will manually analyze its representativity[14]. Taking in mind the semantic model behind these algorithms (see section 2), the objective now is to see what topic modeling is really doing when it is applied to poetry: that is, whether it really classifies poems by topics or by any other phenomena. Considering the results about the analysis of LDA-topics coherence presented in

[12] Although it is not the subject of this paper, this conclusion is important in order to connect metrical patterns with meaning, as in Navarro-Colorado (2015). Meter is a poetic feature that depends on the line. In a poem there is a metrical pattern for each verse. If the main unit where words achieve their distributional meaning is the whole poem, then it is not possible to directly relate metrical patterns with distributional word meanings. This relation must be extracted only at poem/discourse level.

[13] About the effects of lemmatization in topic modeling, see Schofield and Mimno (2016).

[14] It is not my objective here to evaluate LDA-Topic Modeling in terms of precision and recall, but to analyze what Topic Modeling is showing us from the point of view of Literary Studies.

the previous section, the object of the analysis has been the set of 100 LDA-topics obtained with a standard LDA algorithm run over the non-lemmatized corpus.

The manual analysis has been developed in three steps:

- First, LDA-Topics have been compared to a close-reading thematic classification. Our aim is to make sure whether LDA topics are the same as or different from the concept of "topic" as reflected in literary studies.
- Second, the keywords of each LDA-topic have been examined in order to find common linguistic or literary features among them: a noun or a small description that justifies the LDA-topic. As result, a set of classes has been established to represent the common feature of each LDA-topic. It is based upon the manual analysis of each LDA-Topic and on previous studies of Golden Age Spanish poetry. A first manual classification of LDA-topics using these classes has been developed, and then all LDA-topics have been classified again by two annotators in order to evaluate the consistency of these classes and the analysis performed.
- Finally, the main poems of each LDA-topic (the sonnets with the highest Dirichlet parameter) have been analyzed in order to determine the relation or connection between the topic and the poems: if these poems share the same theme or if, on the contrary, they have different themes.

6.1. Comparison to Close-Reading Classifications

In order to show if topic modeling is really extracting topics or themes, similar to the kind of topics extracted when it is applied to scientific texts (as it is shown in Blei, 2012), the 100 LDA-topics extracted have been compared to the thematic classification of Spanish Golden Age sonnets developed by Rivers (1993). It is nowadays considered a standard thematic classification. Obviously, this classification was made from a small number of sonnets (150)[15], but I think it is suitable enough to allow a comparison so that we can find out if LDA-Topics and literary themes are similar or not. The thematic classes established by Rivers are the following: Love, "Beatus Ille," "Carpe Diem," Funeral, Metapoetic, Mythology, Recantation, Religion, Pastoral, Ruins, "Tempus Fugit," and Triumph. **Table 4** shows the relation of LDA-topic with each theme and the number of sonnets in each one (between brackets).

TABLE 4 | Comparison of Topic Models with Rivers (1993) classification.

LDA-topic	Total	Theme (Number of sonnets)
1	49	Ruins (6)
2	13	Recantation (1)
5	162	Religion (1)
9	83	Funeral (1), Ruins (1)
10	128	Mythology (1)
15	62	Love (1)
17	23	Funeral (1)
19	42	Love (1)
20	208	Funeral (3), Triumph (1), Ruins (1)
25	106	Love (1), Mythology (1), Religion (1)
28	128	Love (2), Recantation (1), Tempus fugit (1)
31	13	Love (2)
43	181	Carpe Diem (1), Love (1), Mythology (1)
44	139	Love (1)
45	15	Mythology (1)
47	412	Love (5), Religion (3), Mythology (2), Recantation (2)
48	16	Metapoetic (3), Triumph (1)
51	14	Triumph (1)
52	54	Love (2), Religion (1)
60	14	Love (1)
61	45	Love (1)
63	26	Love (1)
65	35	Love (2)
66	178	Love (1), Recantation (1)
68	8	Ruins (1)
69	17	Religion (1), Pastoral (1)
71	724	Love (5), Recantation (4), Religion (1), Triumph (1), Carpe Diem (1)
72	69	Ruins (1)
75	29	Love (1)
77	39	Love (1)
78	14	Beatus Ille (1), Mythology (1)
79	6	Love (1)
80	36	Mythology (1)
81	10	Beatus Ille (1)
84	46	Triumph (1), Religion (1)
85	254	Love (4), Beatus Ille (2), Mythology (2), Pastoral (1)
87	32	Love (1)
91	18	Pastoral (3)
92	58	Tempus fugit (1)
94	48	Carpe Diem (1), Love (1)
95	117	Carpe Diem (6), Love (6)

We cannot find a clear or constant relationship between a LDA-topic and a theme out of this comparison that is, the data do not indicate nor prove that a given LDA-topic will be clearly related with a specific theme. Maybe Topic 1, in some way, is related with the ruins' topic, but the same theme appears in topics 9, 20, 68, and 72. Love's theme, as well as mythologic and religious themes, appear distributed in several LDA-topics. The Horatian topic of "Carpe Diem" appears mainly related with topic 95 (6 sonnets), but it is also related with topics 43, 71, and 94. The three sonnets whose theme is poetry itself (metapoetics) are included in topic 48, but related with this topic there is a sonnet about triumph (war, empire, etc.).

The main conclusion of this comparison is that LDA topic modeling applied to poetry does not extract clear and defined themes as it does when applied to scientific texts. Topic modeling extracts LDA-topics according to the contextual use of words. In scientific texts there are words that are used exclusively in very specific contexts, related to specific topics (mainly terminology).

[15] For the porpoises of this paper I have used only 113 sonnets. The remaining sonnets were classified with syntactic or rhetorical features.

Thanks to these topic-specific words, LDA Topic Modeling can extract the topics. The linguistic uses in poetry work in the exact opposite way. In poetry a word related with a specific topic can be frequently used to refer to other topics (for instance metaphors, similes, figurative uses, etc.). For example, it is common to use war-related words (to shoot, to win, etc.) to talk about love. Therefore, the question now is to find out what kinds of relationships is LDA topic modeling extracting when it is applied to poetry.

6.2. LDA-Topics Classification

In order to clarify what topic modeling is extracting when it is applied to poetry, I have first developed a close analysis of the keywords of each topic. The objective is to find the common feature that justifies the LDA-Topic: a noun or a small description that explains the relationship between the keywords of each LDA-Topic. With this, I'm trying to figure out what the poems related with each LDA-Topic have in common.

A list of features has been obtained after the analysis of the 100 LDA-Topics and taking in mind previous semantic and literary analysis of Spanish Golden Age poetry such as (García Berrio, 1978; Rivers, 1993). The list is the following:

- Words related with love;
- words related with the Bible and with Catholic religion in general;
- words related with Greek and Roman Mythology;
- words related with satirical and mockery aspects;
- words related with eulogistic, dirges, funeral, or elegiac aspects;
- words related with moral and, specifically, with the passing of time ("tempus fugit");
- words related with water, the sea, or rivers;
- words related with music,
- words related with the poetry itself (metapoetics) or with literature in general;
- words related with nature,
- words related with the night,
- words related with painting,
- words related by similar phonetics (words that rhyme).

All these features, except the last one, are semantic classes. **Table 5** shows some words extracted from several LDA-Topics related with love, religion, eulogy, nature, or sea[16].

The last feature is not semantic but phonetic. It includes words that tend to appear together in the same documents (poems) because they rhyme with each other, that is, these words end with the same sound. Therefore, several of the LDA-Topics keywords are related because they rhyme with each other. Due to this rhyme, they tend to appear together in the same sonnets and LDA topic modeling groups them into the same sets of topics. **Table 6** shows two examples of these kinds of keyword topics.

The 100 LDA-Topics have been therefore classified into one of these classes, which cover the main topics of this corpus of poetry. In order to validate these classes and the classification made with them, the 100 topics have been classified again by two annotators following a double-blind process. The task consisted in labeling each LDA-Topic with at least one of the possible classes, but considering only the keywords of each LDA-Topic. If the annotators cannot find an appropriate class for a specific LDA-Topic, they can introduce the tag "others."

TABLE 5 | Examples of LDA keywords related with semantic classes.

Love	fuego "fire," amor "love," ciego "blind," llama "flame," ardiente "burning," pecho "chest," corazón "heart"...
Religion	dios "God," cielo "sky," santo "saint," padre "father," madre "mother," hijo "son," virgen "virgin," divino "divine"...
Eulogistic	vos "thou," valor "courage," fuerte "strong," fue "was/were," fama "fame," virtud "virtue," esfuerzo "effort,"...
Nature	flores "flowers," verde "green," prado "field," monte "hill," campo "countryside," hojas "leaves," fruto "fruit," sol "sun," árbol "tree," ...
Marine	mar "sea," viento "win," puerto "harbor," ondas "waves," nave "boat," cielo "sky," tormenta "storm," tierra "land," golfo "gulf,"...

TABLE 6 | Example of keywords with the same sound ("Rhyme Topics").

Topic 18	frío "cold," desvarío "nonsense," río "river," brío "spirit," desvío "drift," albedrío "free will," porfío "I strive," envío "I send," estío "summer," confío "I trust," ...
Topic 36	llama "flame," fama "fame," ama "housewife," nombre "name," derrama "spill," rama "branch," dama "lady (dame)," cama "bed," estima "admiration," voz "voice," inflama "ignite," ...

TABLE 7 | Inter-annotators agreement.

Annotators	IAA (%)
A1–A2	91
A1–A3	91
A2–A3	85

Before the manual annotation, all the LDA-Topics with low representativity were filtered. These topics were considered as "noise," defining noise empirically as those LDA-Topics with a low Dirichlet parameter (equal or <0.02) and few associated poems (equal or less than 10 sonnets). A total amount of 19 LDA-Topics were automatically classified as "noise" and filtered.

For the remaining 81 LDA-Topics we have three annotations: the original one (A1), and two more blind annotations (A2 and A3). **Table 7** shows the inter-annotators agreement two by two: the number of LDA-Topics that have been classified with the same class by two annotators.

IAA shows in general a high agreement. Only between annotators 2 and 3 the IIA go down to 85%.

The main cause of disagreement is semantic ambiguity: there are some LDA-Topics whose keywords have high ambiguity and, therefore, could be classified into two or more classes. For example, Topic 70 has been classified as "love," "nature," and "eulogy" by each one of the annotators. According to the keywords of this topic (see **Table 8**), it can express the three semantic features. In the same way, some other topics can be interpreted literally or metaphorically. For example, the keywords of LDA-Topic 95 are words related with the description

[16] The complete list of words related to each class is available at https://github.com/bncolorado/OnPoeticTopicModeling_Data

TABLE 8 | Some ambiguous topics.

Topic 70	belleza “beauty,” alma “soul,” naturaleza “nature,” grandeza “nobility,” cuerpo “body,” valor “courage,” dureza “strength,” virtud “virtue,” firmeza “firmness,”...
Topic 95	hermosa “lovely,” rosa “rose,” sol “sun,” nieve “snow,” flor “flower,” perlas “pearl,” blanco “white,” frente “forehead,” labios “lips,” rostro “face,”...

TABLE 9 | Number of topics classified into each class.

Theme	LDA-topics
Bible and religion	5
Eulogistic	9
Greek and Roman Mythology	4
Love	16
Marine	2
Metapoetic	1
Moral	7
Music	1
Nature and Pastoral	7
Night	1
Painting	1
Satirical and mockery	3
Rhyme	24
Noise	19

TABLE 10 | Examples of topics related to love.

Topic 4	fuego “fire,” amor “love,” ciego “blind,” llama “flame,” ardiente “burning,” pecho “chest,” corazón “heart,” ...
Topic 7	celos “jealousy,” amor “love,” cielos “sky,” desvelos “sleeplessness,” recelos “mistrust,” laura, ...
Topic 47	mal “evil,” dolor “pain,” bien “good,” triste “sad,” amor “love,” llanto “tears,” pena “shame,” dulce “sweet,” ...

of a woman (see **Table 8**). However, it could be classified as “love” (the description of the beloved woman) or as “moral” (following the Horatian topic of “carpe diem,” in which usually appears the description of a young woman).

In three topics (84, 87, and 89) there is a disagreement between the semantic class and the rhyme class (see **Table 8**). Actually, these topics can be classified into both classes because they are hybrid topics: they represent a recurrent rhyme and, at the same time, a semantic class. These hybrid topics will be commented in the next subsection.

Finally, although no annotator has used the tag “other,” the disagreement between annotators is showing the lack of a new semantic class. It occurs in Topic 83. It has been classified as both “moral” and “eulogy.” However, these keywords are related with the rural world and farm work. As well as there is a semantic class related with the pastoral world, this disagreement shows that farm work is also an important topic in this corpus.

The classes without disagreement are the Bible and Catholic religion; Greek and Roman mythology; satirical and mockery; sea; music; metapoetics; and painting. Classes with some disagreement are, on one hand, love, eulogistic, and moral; and on the other nature and the night. **Table 9** shows the number of topics classified into each class.[17].

In any case, assuming that these disagreements are due mainly to the proper ambiguity of semantics, the IAA is suitable, at least to support the final analysis of the next subsection. In conclusion, it seems that topic modeling is really extracting topics and grouping poems according to their topic. At least, the analysis of merely each LDA-Topic keywords allows us to relate the LDA-Topic with a literary topic or theme. To confirm or reject this idea it is necessary to analyze the theme of each sonnet.

6.3. Comparing LDA-Topics Classification and Sonnet’s Topic

The last manual analysis that I will present in this paper has to do with the semantic relationship among the sonnets grouped into each LDA-Topic, so that we can make sure whether all of them share the same topic or not—in other words, if the sonnets of a LDA-Topic express only one theme or, on the contrary, we can find two or more themes in the same LDA-Topic.

I have not analyzed all the sonnets of each LDA-Topic due to the great quantity of sonnets existing for some of them. I have however manually analyzed the main sonnets of each LDA-Topic, those with the highest Dirichlet parameter (up to 10 sonnets for each LDA-Topic). In this way I analyze the main semantic tendency of the LDA-Topics, assuming that in all cases there will be some exceptions. In some cases I have analyzed also sonnets with a minor parameter. For this analysis, the same semantic classes presented in section 6.2 have been used. Topics classified as “noise” and as “rhyme” have been excluded because clearly there is no semantic relation among their poems.

The results show that the sonnets of 25 LDA-Topics share the same theme or subject. In these cases, topic modeling is clearly arranging the sonnets according to their topic (similar, in some way, to the arrangement of prose texts presented in Blei, 2012). The six main themes classified in the corpus are the following:

- love (11 LDA-topics),
- heroic (mainly about the Spanish empire) (5 LDA-Topics),
- mockery (3 LDA-Topic)
- religion (2 LDA-Topics),
- nature (2 LDA-Topics),
- moral (1 LDA-Topics), and
- mythology (1 LDA-Topic).

The main topic of all Golden Age Spanish sonnets is love, and specifically the Petrarchan unrequited love. The 11 LDA-topics that refer to love show how love was treated during Golden Age in Spain. For example, love is seen as a “burning flame” that opposes the “frozen soul” of the beloved woman (topics 4 in **Table 10**), love and jealousy (topic 7 in **Table 10**), or love as “sweet suffering” (topic 47 in **Table 10**).

[17] The complete list of words related to each class is available at https://github.com/bncolorado/OnPoeticTopicModeling_Data.

TABLE 11 | Examples of topics.

Topic 5 (Religion)	Dios "God," cielo "sky," santo "saint," padre "father," madre "mother," hijo "son," virgen "virgin," divino "divine,"...
Topic 20 (Heroics, Spanish Empire)	espada "sword," valor "courage," España "Spain," rey "king," armas "weapons," marte "mars,"...
Topic 54 (Heroic, Eulogistic)	vos "thou," valor "courage," fuerte "strong," fue "was/were," fama "fame," virtud "virtue," esfuerzo "effort," ...
Topic 23 (Mockery, unfaithfulness)	casta "lineage," cuernos "cuckold," puta "prostitute," cornudo "cuckolded,"...
Topic 43 (Moral)	vida "life," muerte "death," suerte "fortune," fuerte "strong," fue "(he/she/it) was," fin "end," tiempo "time," advierte "advise,"...

TABLE 12 | Examples of motifs.

Topics 25 (Marine)	mar "sea," viento "win," puerto "harbor," ondas "waves," nave "boat," cielo "sky," tormenta "storm," tierra "land," golfo "gulf,"...
Topic 15 (Music)	voz "voice," dulce "sweet," canto "song," acento "accent," llanto "weeping," armonía "harmony,"...
Topic 17 Painting	pincel "paint brush," arte "art," colores "colors," pintor "painter," pintura "painting," retrato "portrait,"...

The remaining topics are common poetic themes as seen in heroic sonnets, religious, moral, etc. sonnets. See **Table 11** for some examples.

Therefore, in these cases topic modeling has grouped together texts with the same theme or topic. Assuming some exceptions, in general terms these LDA-topics are mainly used as themes.

In 20 LDA-Topics, however, the main sonnets of each LDA-Topic have different and diverse themes (two or more). In these cases, topic modeling is clearly not arranging texts by its theme or subject. More than a theme, the common feature in these LDA-Topics is a literary motif[18]. Actually, these sonnets use the keywords of the LDA-Topics as a poetic device to talk about several topics. Topic modeling is grouping poems that share recurrent words and poetic images that appear throughout the whole period and are used by different poets to talk about different subjects. For example, the use of words related to sea to talk about love. In this case, the LDA-topic is formed by sea-related words ("sea, win, harbor, waves...") but the topic of the sonnet is actually love.

These LDA-topics could include metaphors, metonymies, poetic tropes, allegories, symbols, etc., but not always. Since word topics express mainly recurrent poetic images that poets use to talk about different topics, I prefer to call this kind of relation between LDA-Topics and sonnets "motif," rather than "trope" or "metaphor."

This case includes, then, those LDA-topics whose main sonnets use the LDA-topic as a poetic motif to talk about diverse themes. The main motifs found in the corpus of Spanish Golden Age sonnets are the following:

- words related to heroism (4 LDA-Topics)
- words related to nature (4 LDA-Topics)
- words related to moral as the pass of time, ruins and similar (4 LDA-Topic)
- words related to the Bible and religion (2 LDA-Topics)
- words related to Mythology (2 LDA-Topics)
- words related to sea and water (marine) (1 LDA-Topic)
- words related to poetry itself and literature (1 LDA-Topic)
- words related to music (1 LDA-Topic)
- words related to painting (1 LDA-Topic)

It is interesting to note that some of the semantic classes defined in section 6.2 are used only as motifs (sea, poetry, music, painting) while others are used both as theme and motif (heroism, nature, moral, religion and Mythology). Only love is used just as theme. **Table 12** shows some examples of motifs.

Some themes expressed with these motifs are the following:

- Marine motif: love, moral, religious, and heroic topics, used sometimes as allegories.
- Musical motif: mainly love.
- Painting motif: love (about the beauty of the beloved), eulogistic, and satirical.

As I have said before, this classification is showing only the main tendency of these LDA-topics. I have analyzed only the main sonnets for each LDA-topic. Therefore, they may include some sonnets that are really depicting these themes—in these cases, sonnets whose main themes are actually the sea, music, or paintings. As well as themes, it is possible to conclude that these LDA-topics are used mainly as motifs.

In the LDA-Topic 25, for example, which is devoted to the marine motif, it is possible to find sonnets about love or morality, but in the main sonnets there are also sonnets whose theme is the sea itself. In this regard, this analysis shows the main tendency in the use of these terms in poetry: marine terms are generally used as symbols to talk about love or moral aspects, but also to talk about the sea as a topic.

Finally, some topics are hybrid topics in the sense that they include sonnets related to a theme or motif, but they are also rhyme topics. **Table 10**, for example, shows a love-related topic (jealousy, "celos") but, at the same time, it shows a lot of keywords that end in "-elos": "celos, cielos, desvelos, recelos, velos, hielos, etc." (jealousy, sky, sleeplessness, mistrusts, veils, ices, etc.).

In a nutshell, the analysis of the main sonnets grouped inside each LDA-Topis shows that topic modeling, applied to poetry, does not really extract topics (as in themes or subjects as in Blei, 2012) and does not group texts together according to a specific topic. More than this, topic modeling applied to poetry extracts poetic motifs: recurrent words and poetic images that usually appear together throughout the whole period and refer to one

[18] See, for example, the definition of "motif" in Poetry Foundation: https://www.poetryfoundation.org/learn/glossary-terms/motif.

or more themes. In this corpus of poetry, 56 of 100 LDA-Topics are motifs: 33 phonetic motifs (rhyme) and 20 semantic motifs. Only 25 LDA-Topics can be said to really purport a clear theme.

The analysis of LDA-topics representativity shows that they depend clearly on the kind of text to which LDA is applied. Just like Schöch (2017) shows that the application of LDA topic modeling to drama displays clear thematic topics together with others dramatic-specific topics related to character inventory or recurring dramatic actions; applying LDA topic modeling to poetry it is possible to find poetic-specific LDA-topics as rhyme topics or topics of motifs.

7. CONCLUSIONS

This paper explores the application of LDA topic modeling to a corpus of poetry. I have first developed several tests in order to find the most coherent set of LDA-topics. Three parameters have been tested: first, whether it is better to lemmatize the corpus of poetry or to simply use a stop-words filter; second, whether a standard LDA algorithm would be preferable over a specific LDA algorithm adapted to short texts; and third the most desirable number of topics (10, 25, 50, 100, or 250). To measure the coherence of the topics I have used two common techniques: the first one specifies the coherence of topics calculating the nPMI of word topics, the second one does so based on the discovery of an intruder word among the word topics.

Results showed that the most coherent set of topic words was obtained with a standard LDA algorithm applied over the non-lemmatized corpus and extracting between 50 and 100 LDA-topics. I have drawn two main conclusions from these data::

- During the lemmatization process, many poetic features disappear (at least in Spanish). This is the case with verbal inflection. As I showed in section 6, many LDA-topics show rhyme relations and many of them are based on verbal inflection or, in general, word morphology. This poetry-specific feature and its related topics are lost if the corpus is lemmatized. This is the reason why a simple stop-words filter is better in this case for the extraction of poetic LDA-topics.
- The LDA algorithm specifically developed for short texts tested here is based on word embeddings. These word embeddings have been extracted following the Word2Vec model with a 5-word window context, that is approximately the size of a verse. Therefore, one of the main differences between both algorithms is the contextual unit: in a standard LDA the contextual unit is the document (the sonnet in this case), in the LF-LDA algorithm it is also this 5-words window context (the line size). According to the results obtained, the distributional (poetic) meaning of each word is better stablished in the overall poem, without taking the line as semantic unit.

Once the 100 LDA-topics extracted with this configuration have been manually analyzed in order to examine its representativity. I have then first analyzed the word topics in each LDA-topic, trying to find a word that describes it, and then I have analyzed the relationships between the word topics of each LDA-topic and their main representative sonnets (the sonnets with a higher weight).

These analyses have allowed me to find two main relations between LDA-topics and sonnets: LDA-topics as themes and LDA-topics as motifs. In the first case, the LDA-topic clearly allude to the topic or theme of the poem. In the second case the LDA-topic is used as a poetic motif, considering motif here as a set of words that usually appear together. Two kinds of motifs have been defined: phonetic motifs and semantic motifs. Phonetic or "rhyme" motifs include Topics whose keywords have similar soundS, that is, they rhyme. Semantic motifs include Topics whose keywords express images that poets used to talk about different themes recurrently.

As Future Work, I plan first to test other topic modeling algorithms in order to find the most appropriate one for corpus of poetry. Second, I will go in depth into the manual analysis of the relationships between LDA-topics and their sonnets, in order to sketch a complete representation of the main topics and motifs of Spanish Golden Age sonnets. Finally, I plan to use these LDA-topics as keywords to improve the retrieval of sonnets from the corpus.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

FUNDING

This work was supported by the BBVA Foundation: grants for research groups 2016, project Distant Reading Approach to Golden Age Spanish Sonnets (Ayudas fundación BBVA a equipos de investigación científica, proyecto Análisis distante de base computacional del soneto castellano del Siglo de Oro): http://adso.gplsi.es. It was also partially conducted in the context of the COST Action Distant Reading for European Literary History (CA16204 - Distant-Reading): www.distant-reading.net.

ACKNOWLEDGMENTS

We would like to thank the reviewers for their helpful suggestions and comments.

REFERENCES

Bhatia, S., Lau, J. H., and Baldwin, T. (2017). "An automatic approach for document-level topic model evaluation," in *21st Conference on Computational Natural Language Learning (CoNLL)* (Vancouver, BC), 206–215.

Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM* 55, 77–84. doi: 10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022. Available online at: http://jmlr.csail.mit.edu/papers/v3/blei03a.html

Bouma, G. (2009). "Normalized (pointwise) mutual information in collocation extraction," in *Proceedings of German Society for Computational Linguistics (GSCL 2009)* (Postdam), 31–40.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., and Blei, D. M. (2009). "Reading tea leaves: how humans interpret topic models," in *NIPS Neural Information Processing Systems* (Vancouver, BC).

García Berrio, A. (1978). Lingüística del texto y tipología lírica (La tradición textual como contexto). *Rev. Española de Lingüíst.* 8, 19–75.

Herbelot, A. (2015). The semantics of poetry: a distributional reading. *Digit. Scholarsh. Human.* 30, 516–531. doi: 10.1093/llc/fqu035

Jockers, M. L., and Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics* 41, 750–769. doi: 10.1016/j.poetic.2013.08.005

Lau, J. H., Newman, D., and Baldwin, T. (2014). "Machine reading tea leaves: automatically evaluating topic coherence and topic model quality," in *14th Conference of the European Chapter of the Association for Computational Linguistics* (Gothenburg), 530–539.

Lou, A., Inkpen, D., and Tnsescu, C. (2015). "Multilabel subject-based classification of poetry," in *Proceedings of theTwenty-Eighth International Florida Artificial Intelligence Research Society Conference* (Hollywood, FL).

McCallum, A. K. (2002). *Mallet: A Machine Learning for Language Toolkit*. Available online at: http://mallet.cs.umass.edu

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality," in *Annual Conference on Neural Information Processing Systems (NIPS)* (Lake Tahoe).

Moretti, F. (2007). *Graphs, Maps, Trees: Abstract Models for a Literary History*. London, UK: Verso.

Navarro-Colorado, B. (2015). "A computational linguistic approach to Spanish Golden Age Sonnets : metrical and semantic aspects," in *4th Workshop on Computational Linguistics for Literature (CLfL 2015)* (Denver, CO), 105–113.

Navarro-Colorado, B., Ribes-Lafoz, M., and Sánchez, N. (2016). "Metrical annotation of a large corpus of Spanish sonnets: representation, scansion and evaluation," in *LREC 2016, Tenth International Conference on Language Resources and Evaluation* (Portorož), 4360–4364.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). "Automatic evaluation of topic coherence," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (Los Ángeles, CA).

Nguyen, D. Q., Billingsley, R., Du, L., and Johnson, M. (2015). Improving topic models with latent feature word representations. *Trans. Assoc. Comput. Linguist.* 3, 299–313. Available online at: https://transacl.org/ojs/index.php/tacl/article/view/582

Padró, L., and Stanilovsky, E. (2012). "FreeLing 3.0: towards wider multilinguality," in *Language Resources and Evaluation Conference (LREC 2012)* (Istanbul).

Řehůřek, R. and Sojka, P. (2010). "Software framework for topic modelling with large corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (Valletta: ELRA), 45–50.

Rhody, L. M. (2012). Topic modeling and figurative language. *J. Digit. Human.* 2. Available online at: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/

Rivers, E. L. (1993). *El Soneto Español en el Siglo de Oro*. (Madrid: Akal).

Roe, G., Gladstone, C., and Morrissey, R. (2016). Discourses and disciplines in the enlightenment: topic modeling the French encyclopédie. *Front. Digit. Human.* 2:8. doi: 10.3389/fdigh.2015.00008

Schöch, C. (2017). Topic modeling genre: an exploration of french classical and enlightenment drama. *Digit. Human. Q.* doi: 10.5281/zenodo.166356. [Epub ahead of print].

Schofield, A., and Mimno, D. (2016). Comparing apples to apple: the effects of stemmers on topic models. *Trans. Assoc. Comput. Linguist.* 4, 287–300. Available online at: https://transacl.org/ojs/index.php/tacl/article/view/868

Tangherlini, T. R., and Leonard, P. (2013). Trawling in the sea of the great unread: sub-corpus topic modeling and humanities research. *Poetics* 41, 725–749. doi: 10.1016/j.poetic.2013.08.002

Terry, A. (1993). *Seventeenth-Century Spanish Poetry*. Cambridge, UK: Cambridge University Press.

Turney, P., and Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *J. Artif. Intell. Res.* 37, 141–188. doi: 10.1613/jair.2934

Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Wiley-Blackwell (2010 reprint).

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.