A computational linguistic approach to Spanish Golden Age Sonnets: metrical and semantic aspects

Borja Navarro-Colorado
Natural Language Processing Group
University of Alicante
Alicante, Spain
borja@dlsi.ua.es

Abstract

Several computational linguistics techniques are applied to analyze a large corpus of Spanish sonnets from the 16th and 17th centuries. The analysis is focused on metrical and semantic aspects. First, we are developing a hybrid scansion system in order to extract and analyze rhythmical or metrical patterns. The possible metrical patterns of each verse are extracted with language-based rules. Then statistical rules are used to resolve ambiguities. Second, we are applying distributional semantic models in order to, on one hand, extract semantic regularities from sonnets, and on the other hand to group together sonnets and poets according to these semantic regularities. Besides these techniques, in this position paper we will show the objectives of the project and partial results.

1 Introduction

16th- and 17th-Centuries Spanish poetry is judge as one of the best period of the History of Spanish Literature (Rico, 1980 2000; Terry, 1993; Mainer, 2010). It was the time of great, famous and "canonical" Spanish poets such as Miguel de Cervantes, Lope de Vega, Garcilaso de la Vega or Calderón de la Barca, among others. Due to the importance given to this period, it has been deeply studied by scholars from the 19th century to the present. We are persuaded that new approaches based on a "distant reading" (Moretti, 2007; Moretti, 2013) or "macroanalysis" (Jockers, 2013) framework could shed new light on this period.

We have two general objectives: first, we will try to extract regular patterns from the overall period; and second, in order to analyze each author inside the broad literary context in which they wrote (García Berrio, 2000), we will look for chains of relationships between them.

Nowadays both objectives are focused on metrical and semantic aspects of Spanish Golden Age Sonnets. In this position paper we will present the computational linguistic techniques used to achieve these objectives.

Next section shows how a large corpus of Spanish Golden-Age sonnets has been compiled and annotated; Section 3 describes a hybrid scansion system developed to extract metrical patterns; Section 4 presents how we use distributional semantic models to extract semantic patterns from the corpus; finally, Section 5 shows some preliminar conclusions.

2 Corpus compilation and XML annotation

A corpus of 5078 sonnets has been compiled. It includes all the main poets of the Spanish Golden Age: Juan Boscán, Garcilaso de la Vega, Fray Luis de León, Lope de Vega, Francisco de Quevedo, Calderón de la Barca, Sor Juana Inés de la Cruz, etc. Our objetive is to include all the authors of this period who wrote a significant amount of sonnets. Authors who wrote but few sonnets (less than ten) have been rejected. Most sonnets have been obtained from the Miguel de Cervantes Virtual Library[1].

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

I must point out that sonnet quality is not taken into account. Following Moretti's Distant Reading framework (Moretti, 2007), we want a representative corpus of the kind of sonnets written in that period, not only the canonical sonnets. In other words, the corpus must represent the literary context of any Golden Age poet.

Each sonnet has been marked with the standard TEI-XML[2]. We have followed the standard TEI in order to ensure the re-usability of the corpus in further research. The main metadata annotated at the TEI-Header are:

- Project title and project manager (Title and Publication Statement),
- Title and author of each sonnet,
- Source publication and editor (Source Description).

Regarding the sonnet structure, we have annotated:

- Quatrains,
- Tercets,
- Verse line and number of line,
- Some extra lines included in the poem (called "estrambote")

The markup includes a representation of the metrical structure of each verse line. It will be explained in the next section.

Nowadays we have a first version of the corpus. We plan to publish it on the internet during 2016.

3 Metrical annotation and analysis

In order to extract the metrical pattern of each verse line, we have created a scansion system for Spanish based on Computational Linguistics techniques. In this section we will show how the metrical information is represented and what the main aspects of the scansion system are.

[2] www.tei-c.org/

3.1 Metrical representation

Spanish poetry measures poetic lines by syllables. The Spanish sonnet is an adaptation of the Italian sonnet, where each line has eleven syllables (hendecasyllable). The metrical pattern of each verse line is formed by a combination of stressed and unstressed syllables. There must be a stressed syllable in the tenth position. The rest of the stressed syllables can be combined in different ways, generating different rhythms or metrical patterns: Heroic (stressed syllables at position 2, 6 and 10), Melodic (at 3, 6, 10), sapphic (at 4, 8, 10), etc. (Quilis, 1984; Varelo-Merino et al., 2005)

For each verse line, we represent its metrical pattern by a combination of symbols "+" (stressed syllable) and "-" (unstressed syllable). For example:

```
<lg type="cuarteto">
<l n="1" met="---+---+-+-">
Cuando me paro a contemplar mi estado
</l>
```

"lg" tag represents the stanza (quatrain in this case), and "l" tag the line. Each line has the "$met =$" tag with the metrical pattern of the verse.

This verse from Garcilaso de la Vega has thirteen linguistic syllables, but it has only eleven metrical syllables. As we will show in the next section, "-ro a" (in "paro a") and "mi es-" (in "mi estado") conform a single syllable each due to the synaloepha phenomenon. Therefore this line is an hendecasyllable with stressed syllables in position 4, 8 and 10 (sapphic).

3.2 Scansion system

Metrical patterns extraction does not consist of a simple detection of syllables and accents. Due to the fact that there is not a direct relationship between linguistic syllables and metrical syllables, some ambiguity problems appear that must be solved by computational linguistics techniques. The main scansion problems are the following:

- The total amount of syllables could change according to the position of the last stressed syllable. If the last stressed syllable is the last one (oxytone), the line should have ten syllables and an extra syllable must be added. On contrary, if the last stressed syllable is the antepenultimate (proparoxytone), the line should

have twelve syllables and the last syllable must be removed. This is a fixed phenomenon and can be solved with rules.

- Not every word with linguistic accent has a metrical accent. It depends on the Part of Speech. Words like nouns, verbs, adjectives or adverbs have always a metrical accent; but prepositions, conjunctions and some pronouns have no metrical accent.

- A vocalic sound at the end of a syllable and at the beginning of the next one tends to be blended in one single syllable (syneresis if the syllables belong to the same word and synaloepha if they belong to different words). This phenomenon is not always carried out: it depends on several factors, mainly the intention during declamation.

- The opposite is possible too: a one single syllable with two vowels (normally semivowel like an "i" or "u") that can be pronounced as two separated syllables (dieresis).

These phenomena could change the metrical pattern extracted in two different ways: the amount of syllables and the type of each one of them (stressed or unstressed). The main problem are those verses in which it is possible to extract two or more different patterns, all of them correct.

For example, for a verse with twelve syllables and paroxitonal final stress it is necessary to blend at least two syllables in one through a phenomenon of synaloepha or syneresis. The problem appears when there are two possible synaloephas or syneresis: which of them must be carried out? The final metrical pattern will be completely different.

For example, the next verse line:

cuando el padre Hebrero nos enseña

It has 12 syllables. It is necessary to blend two syllables in one through synaloepha. However, there are two possibles synaloephas: "cuando+el" and "padre+Hebrero". Different metrical patterns are generated for each synaleopha:

- - + - - + - - - + -
- - - + - + - - - + -

A ranking of natural and artificial synaloephas has been defined by traditional metrical studies. For example, it is more natural to join two unstressed vowels than two stressed vowels (Quilis, 1984). From our point of view, this is a "deliberate" ambiguity (Hammond et al., 2013): both metrical patterns are correct, choosing one depends on how the verse line is pronounced.

An automatic metrical scansion system must resolve this ambiguity[3]. There are several computational approaches to metrical scansion for different languages (Greene et al., 2010; Agirrezabal et al., 2013; Hammond, 2014). For Spanish, P. Gervás (2000) proposes a rule-based approach. It applies Logic-programming to detect stressed and unstressed syllables. It has a specific module to detect and resolve synaloephas that is applied recursively up to the end of the verse. However, I think that this system is not able to detect ambiguities: if there are two possible synalephas, this system always chooses the first one. Therefore, it does not detect other possible metrical patterns.

We follow a hybrid approach to metrical scansion. First, rules are applied in order to separate words in syllables (hyphenation module), detect metrical syllables with a Part of Speech tagger[4], and finally blend or segment syllables according to synaloephas, dieresis or syneresis.

Before the application of synaloephas or syneresis rules, the system counts the number of syllables. If the line has eleven syllables, then these rules are not applied. If there are more than eleven syllables, then the system counts how many synaloephas or syneresis must be resolved. If resolving all synaloephas or syneresis the syllables amount to eleven, then the system applies them all. If resolving all synaloephas or syneresis the syllables amount to a number lower than eleven, the verse is ambiguous: the system must select which rules must be applied and which must not.

[3] Or al least the system must select the most appropriate one, even if it could detect and represent alternative patterns.

[4] We use Freeling as PoS tagger (`http://nlp.lsi.upc.edu/freeling/`) (Padró and Stanilovsky, 2012). For each word, the scansion system selects the most general PoS tag (noun, verb, etc.) Only for a few cases it is necessary a deeper analysis. For example, the system must distinguish between personal pronouns (stressed) and clitic pronouns (unstressed)

For these ambiguous verses (with two or more possible metrical patterns) we follow a statistical approach. First, the system calculates metrical patterns frequencies from non-ambiguous patterns. These patterns are extracted from lines in which it has not been necessary to apply the rules for synaloephas, or lines in which applying all possible rules for synalopehas, a unique patter of eleven syllables is obtained. Each time the system analyzes one of these lines, the frequency of its pattern is increased one.

From a total amount of 82593 verses[5], 6338 are ambiguous and 76255 non-ambiguos. Therefore, only 7,67% of lines are ambiguous. In these cases, from the possible pattern that can be applied to a specific line, the system selects the most frequent one: the pattern that has been used more frequently in non-ambiguous verses.

Our approach tends to select common pattern and reject unusual ones. It must be noted that we do not claim that the metrical pattern selected in ambiguous lines is the correct one. We claim that it is the most frequent one. As we said before, this is a "deliberate" ambiguity (Hammond et al., 2013) in which there are not correct or incorrect solutions.

Table 1 shows the most frequent patterns extracted from the corpus and its frequency.

Metrical Pattern	Name	Frequency
- + - - - + - - - + -	Heroic	6457
- + - + - - - + - + -	Sapphic	6161
- - + - - + - - - + -	Melodic	5982
- + - + - + - - - + -	Heroic	5015
- - - + - + - - - + -	Sapphic	3947
- + - - - + - + - + -	Heroic	3549
- + - + - + - + - + -	Heroic	3310
+ - - + - - - + - + -	Sapphic	3164
+ - - + - + - - - + -	Sapphic	3150
- - - + - - - + - + -	Sapphic	3105
- - + - - + - + - + -	Melodic	2940

Table 1: Most frequent metrical patterns.

Therefore, the previous example is annotated with the first metrical pattern (Melodic):

cuando el padre Hebrero nos enseña

```
<l n="1" met="--+--+---+-">
```

Nowadays we are manually reviewing the automatic annotation in order to correct errors, set up a Gold Standard and evaluate the system.

4 Semantic analysis

In order to develop a broad semantic analysis of Spanish Golden Age sonnets, we are applying Distributional Semantic Models (Turney and Pantel, 2010; Mitchell and Lapata, 2010). These models are based on the distributional hypothesis (Harris, 1951): words that occur in similar contexts have similar meanings. These models use vector space models to represent the context in which a word appears and, then, represent the meaning of the word.

Computational distributional models are able to establish the similarities between words according to the similarity of their contexts. Therefore, the application of these distributional models to corpora of sonnets can extract semantic similarities between words, texts and authors. A standard approach is based on a word-text matrix. Applying well-known distance metrics as Cosine Similarity or Euclidean Distance it is possible to find out the similarities between words or poems. In light of these similarities we can then establish the (distributional) semantic relations between authors.

We are applying two specific distributional semantic models: Latent Dirichlet Allocation (LDA) Topic Modeling (Blei et al., 2003) on one hand, and Distributional-Compositional Semantic Models (Mitchell and Lapata, 2010) on the other hand.

4.1 LDA Topic Modeling

During the last years several papers have proposed applying LDA Topic Modeling to literary corpora (Tangherlini and Leonard, 2013; Jockers and Mimno, 2013; Jockers, 2013; Kokkinakis and Malm, 2013) -among others-. Jokers and Mimno (2013), for example, use Topic Modeling to extract relevant themes from a corpus of 19th-Century novels. They present a classification of topics according to genre, showing that, in 19th-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 bat-

[5]This is the total amount of verses, including authors with less than ten sonnets that were rejected for the final version of the corpus.

tles, 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.

Rhody (2012) analyzes the application of Topic Modeling to poetry. The result 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 apparently no semantic relation between them) really shows symbolic and metaphoric relations. More than "topics", these topics represent symbolic meanings. She concludes that, in order to understand them, a closed reading of the poems is necessary.

We have run LDA Topic Modeling over our corpus of sonnets[6]. Using different configurations (10, 20, 50, 100 and 1000 topics), we are developing several analysis. In the next sections I will present these analysis together with some preliminary results and comments.

4.1.1 Common and regular topics

First, we have extracted the most common and regular topics from the overall corpus. We are analyzing them using as reference framework themes and topics established manually by scholars following a close reading approach (García Berrio, 1978; Rivers, 1993).

At this moment we have found four types of topics:

- Topics clearly related with classical themes. Table 2 shows some examples.

- Topics showing rhime relations: words that used to appear at the same sonnet because they rhyme between them. For example, "boca loca toca poca provoca" (Topic 14 of 100).

- Topics showing figurative and symbolic relations: words semantically related only in a symbolic framework. For example, topic 70 relates the words "río fuente agua" (river, fountain, water) with "cristal" (glass). This topic is showing the presence of Petrarchan tradition "rivers of glass"[7] in the Spanish poetry.

- Noise topics.

Topic Model	Traditional Theme
amor fuerza desdén arco niño cruel ciego flecha fuego ingrato sospecha	Unrequited Love
hoy yace sepulcro fénix mármol polvo ceniza ayer guarda muerta piedad cadáver	Funeral
españa rey sangre roma imperio grande baña valor extraña reino carlos hazaña engaña saña bárbaro	Decline of Spanish Empire

Table 2: Topic Models related to classical themes.

Once we detect an interesting topic, we analyze the sonnets in which this topic is relevant. For example, Topic 2 in table 2 represents clearly the funeral theme, sonnets composed upon the gravestone of a dead person. According to LDA, this topic is relevant in Francisco de Quevedo (10 poems), Góngora (6 poems), Lope de Vega (6 poems), Juan de Tassis y Peralta (6 poems), Trillo y Figueroa (3 poems), López de Zárate (3 poems), Bocángel y Unzueta (3 poems), Polo de Medina (2 poems), Pantaleón de Ribera (2 poems), etc. We can reach interesting conclusions from a close reading of these poems. For example,

- All these authors belong to 17th-century, the Baroque Period. This topic is related to the "brevity of life" theme, a typical Baroque topic. Topic Modeling is, then, confirming traditional studies.

- Most of these sonnets are really funeral sonnets, but not all of them. There are some love and satyrical sonnets too. However, these off-topic sonnets use words related to sepulcher, tomb, graves and death. In these cases, Topic Modeling is not showing topics but stylistic and figurative aspects of the poem. Francisco de Quevedo is an example of this aspect: he wrote quite a lot of funeral sonnets and, and the same

[6]We have used MALLET `http://mallet.cs.umass.edu/` (McCallum, 2002)

[7]"et giá son quasi di cristallo i fiumi" Petrarca *Canzoniere* LXVI.

time, he used words related to death in satyrical and mainly love sonnets. It is what Terry (1993) calls "ceniza amante" (loving ash), as a specific characteristic of Quevedos's sonnets.

Therefore, we benefit of the relations between sonnets and authors stablished by LDA Topic Modeling. Then we follow a close reading approach in order to (i) reject noise and random relations, (ii) confirm relations detected by manual analysis, and (iii) detect non-evident relations. This last situation is our main objective.

4.1.2 Cluster of sonnets and poets

Second, we are automatically clustering sonnets and authors that share the same topics. At this moment we have run a k-means cluster over an author-topic matrix [8]. Each author is represented by all the sonnets that they wrote. The matrix is formed by the weight that each topic has in the overall sonnets of each author[9]. Then a k-means cluster has been run using Euclidean distance and different amounts of clusters.

Some preliminary analysis shows that with 20 topics and clustering authors in only two groups, 16th-Century authors (Renaissance period) and 17th-Century authors (Baroque period) are grouped together. Only one poet (of 52) is misclassified. It shows that topic models are able to represent distinctive characteristics of each period. Therefore, we can assume some coherence in more fine clusters.

With 20 topics but clustering authors in ten groups, we have obtained coherent groups too. All poets grouped together wrote during the same period of time. The most relevant aspects of this automatic classification are the following:

- Íñigo López de Mendoza, Marqués de Santillana, was a pre-Renaissance poet. He was the first Spanish author who wrote sonnets. It appears isolated in a specific cluster. Topic Modeling has detected clearly that this is a special poet.

- The first generation of Renaissance poets are grouped together in the same cluster: Hernando de Acuña, Juan de Timoneda, Juan Boscán, Garcilaso de la Vega, Gutierre de Cetina and Diego Hurtado de Mendoza.

- There is another cluster that groups together poets of the second Renaissance generation: authors than wrote during the second half of the 16th Century as Miguel de Cervantes, Fray Luis de León, Francisco de Figueroa, Francisco de la Torre, Diego Ramírez Pagán, Francisco de Aldana and Juan de Almeida.

- One of the poets of this generation, Fernando de Herrera, appears in isolation in a specific cluster.

- Baroque poets (who wrote during 1580 to 1650) are grouped together in varios cluster. There are two main groups: the fist one includes poets born between 1560 to 1590[10], and the second one poets born from 1600 onwards[11].

This temporal coherence in the clusters, that appears in other clusters too, shows us that, on one hand, Topic Modeling could be a reliable approach to the analysis of corpora of poetry, and on the other hand, that there is some relation between topic models and the generations of poets during these centuries. Nowadays we are analyzing the relations between the poets grouped together in the same groups in order to know the reasons of this homogeneity. We plan to run other kinds of clusters in order to analyze other possibilities.

4.1.3 Topic timeline

Taking into account an author's timeline, we are analyzing how trendy topics change during the period. We want to know about the evolution of main

[8] We have used the cluster algorithm implemented in pycluster `https://pypi.python.org/pypi/Pycluster`

[9] Only a stop-list filter has been used to pre-process the corpus.

[10] Lope de Vega (b. 1562), Juan de Arguijo (b. 1567), Francisco de Medrano (b. 1570), Tirso de Molina (b. 1579), Francisco de Quevedo (b. 1580), Francisco de Borja y Aragón (b. 1581), Juan de Jáuregui (b. 1583), Pedro Soto de Rojas (b. 1584), Luis Carrillo y Sotomayor (b. 1585), Antonio Hurtado de Mendoza (b. 1586), etc.

[11] Jerónimo de Cáncer y Velasco (c. 1599), Pantaleón de Ribera (b.1600), Enríquez Gómez (b. 1602), Bocángel y Unzueta (b. 1603), Polo de Medina (b. 1603), Agustin de Salazar y Torres (1642), Sor Juana Inés de la Cruz (b. 1651), José de Litala y Castelví (b. 1672). Only Francisco López de Zárate (b. 1580) is misclassified.

topics during the period, which authors introduce new topics, to what extent these topics are followed by other poets, etc. Nowadays we have not preliminary results to illustrate this aspect.

4.1.4 Relations between metrical patterns and semantic topics

We are analyzing possible relations between metrical patterns and topics. Our hypothesis is that for specific topics, poets use specific metrical patterns and rhythms. At this moment this is an open question.

As a preliminary analysis, we have run a cluster of sonnets based on their metrical patterns. First, we have set up the most relevant metrical patterns of each author applying LDA to the metrical patterns. Instead of using the words, each sonnet is represented only with metrical patterns. Then we have run k-means cluster algorithm with Euclidean Distance and 10 clusters.

From these clusters we have some preliminary considerations:

- Íñigo López de Mendoza, Marqués de Santillana, appears again in isolation. As we said before, hos sonnets were written in a pre-Renaissance period: their meters and rhythm are very different from the others. The cluster is correctly detecting this special case.

- Other cluster is conformed mainly by Renaissance poets: from Garcilaso de la Vega to Fray Luis de León. Even though there are two Baroque poets in this cluster, it seems that Renaissance meters are quite stable and uniform.

- The other two clusters assemble Baroque poets together. At this moment we have not detected if there are any literary criteria that justify these clusters. It is noteworthy that one cluster includes Miguel de Cervantes and Lope de Vega, who tend to use more classical rhythms, and the other Góngora and Quevedo, that tend to use more Baroque rhythms.

These clusters based on metrical patterns are similar to the previous clusters based on distribution of words. Many poets appear together in both experiments: it seems that they are sharing the same distributional topics and metrical patterns. This suggest, albeit in a very speculative way, that there must be some kind of regularity between topics and meters.

In a nutshell, as we have shown in this section, applying LDA Topics Modeling and, in general, distributional model to our corpus of sonnets it is possible to extract not evident (latent) but reliable relations between words (specially figurative language), sonnets and poets. In any case, a final close reading is necessary in order to validate or reject the relations extracted automatically and justify them according to previous studies. These computational methods attract attention to possible latent relations, but it must always be manually validated.

4.2 Compositional-distributional semantic models

Recently a new model of computational semantics has been proposed: the compositional-distributional model (Baroni, 2013). The main idea of this model is to introduce the principle of compositionality in a distributional framework.

Distributional models are based on single words. Standard Vector Space Models of semantics are based in a term-document or word-context matrix (Turney and Pantel, 2010). Therefore, as we have shown in the previous section, they are useful models to calculate similarity between single words, but they cannot represent the meaning of complex expressions as phrases or sentences.

Following Frege's principle of compositionality (Montague, 1974), the meaning of these complex expressions is formed by the meaning of their single units and the relations between them. To represent compositional meaning in a distributional framework, it is necessary to combine word vectors. How semantic vectors must be combined to represent the compositional meaning is an open question in Computational Linguistics. Some proposals are vector addiction, tensor product, convolution, etc. (Mitchell and Lapata, 2010; Clarke, 2011; Blacoe and Lapata, 2012; Socher et al., 2012; Baroni, 2013; Baroni et al., 2014; Hermann and Blunsom, 2014).

From our point of view, compositional-distributional models are useful to detect semantic relations between sonnets based on stylistic features. These models are able to detect semantic similarity according to, not only the words used in a poem, but how the author combines these words.

The combination of words in a poem is the base of its literary style.

We plan to calculate semantic similarity according to specific phrases. For example, it is very specific of an author how they use the adjectives. Compositional-distributional models allow us to extract adjective-noun patterns from sonnets and to calculate the similarities between these patterns. If two poets tend to use similar adjective-noun patterns, then it is possible to establish an influential chain between them. We are working with standard tools as DISSECT (Dinu et al., 2013). Unfortunately, at this moment we have not results to show.

5 Conclusions

In this paper we have presented the computational linguistics techniques applied to the study of a large corpus of Spanish sonnets. Our objective is to establish chains of relations between sonnets and authors and, then, analyze each author in a global literary context. Once a representative corpus has been compiled and annotated, we have focused on two aspects: metrical patterns and semantic patterns.

Metrical patterns are extracted with a scansion system developed in the project. It follows a hybrid approach than combines hand-mande and statistical rules. With all these metrical patterns we plan, on one hand, to analyze the most relevant metrical patterns of the period, as well as the most relevant patterns of each author. On the other hand, we plan to cluster sonnets and authors according to the relevant metrical pattern they use, and establish metrical relational chains.

Semantic patters are extracted following a distributional semantic framework. First, we are using LDA Topic Modeling to detect the most relevant topics of the period and the most relevant topics of each author. Then we plan to group together authors and sonnets according to the topics they share. Finally we will establish the influential chains based on these topics.

We plan to combine both approaches in order to analyze the hypothesis that poets tend to use similar metrical patterns with similar topics. At this moment it is only a hypothesis that will be evaluated during the development of the project.

Finally, we want to go one step beyond Topic Modeling and try to relate authors not by what words they use, but by how they combine the words in sonnets. We plan to apply compositional-distributional models to cluster sonnets and authors with similar stylistic features.

As a position paper, we have presented only partial results of our project. Our idea is to establish a global computational linguistic approach to literary analysis based on the combination of metrical and semantic aspects; a global approach that could be applied to other corpora of poetry.

References

Manex Agirrezabal, Bertol Arrieta, Aitzol Astigarraga, and Mans Hulden. 2013. ZeuScansion : a tool for scansion of English poetry. In *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing*, pages 18–24, St Andrews Sctotland. ACL.

Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. 2014. Frege in Space: A Program for Compositional Distributional Semantics. *Linguistics Issues in Language Technology*, 9(6):5–110.

Marco Baroni. 2013. Composition in Distributional Semantics. *Language and Linguistics Compass*, 7(10):511–522.

William Blacoe and Mirella Lapata. 2012. A Comparison of Vector-based Representations for Semantic Composition. In *Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, number July, pages 546–556.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Daoud Clarke. 2011. A context-theoretic framework for compositionality in distributional semantics. *Computational Linguistics*, 38(1):41–71.

G. Dinu, N. Pham, and M. Baroni. 2013. DISSECT: DIStributional SEmantics Composition Toolkit. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. System Demonstrations.*

Antonio García Berrio. 1978. Lingüística del texto y tipología lírica (La tradición textual como contexto). *Revista española de lingüística*, 8(1).

Antonio García Berrio. 2000. Retórica figural. Esquemas argumentativos en los sonetos de Garcilaso. *Edad de Oro*, (19).

Pablo Gervas. 2000. A Logic Programming Application for the Analysis of Spanish Verse. In *Computational Logic*, Berlin Heidelberg. Springer Berlin Heidelberg.

Erica Greene, Lancaster Ave, Kevin Knight, and Marina Rey. 2010. Automatic Analysis of Rhythmic Poetry with Applications to Generation and Translation. In *Empirical Methods in Natural Language Processing*, pages 524–533, Massachusetts. ACL.

Adam Hammond, Julian Brooke, and Graeme Hirst. 2013. A Tale of Two Cultures : Bringing Literary Analysis and Computational Linguistics Together. In *Workshop on Computational Linguistics for Literature*, Atlanta, Georgia.

Michael Hammond. 2014. Calculating syllable count automatically from fixed-meter poetry in English and Welsh. *Literary and Linguistic Computing*, 29(2).

Zellig Harris. 1951. *Structural Linguistics*. University of Chicago Press, Chicago.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Models for Compositional Distributed Semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 58–68.

Matthew L Jockers and David Mimno. 2013. Significant Themes in 19th-Century Literature. *Poetics*, 41.

Matthew L. Jockers. 2013. *Macroanalysis. Digital Media and Literary History.* University of Illinois Press, Illinois.

Dimitrios Kokkinakis and Mats Malm. 2013. A Macroanalytic View of Swedish Literature using Topic Modeling. In *Corpus Linguistics Conference*, Lancaster.

José Carlos Mainer. 2010. *Historia de la literatura española*. Crítica, Barcelona.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit.

Jeff Mitchell and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science*, 34:1388–1429.

R. Montague. 1974. English as a formal language. In R. Montague, editor, *Formal philosophy*, pages 188–221. Yale University Press.

Franco Moretti. 2007. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.

Franco Moretti. 2013. *Distant reading*. Verso.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Language Resources and Evaluation Conference (LREC 2012)*, Istanbul.

Antonio Quilis. 1984. *Métrica española*. Ariel, Barcelona.

Lisa M. Rhody. 2012. Topic Modeling and Figurative Language. *Journal of Digital Humanities*, 2(1).

Francisco Rico, editor. 1980-2000. *Historia y crítica de la literatura española*. Crítica, Barcelona.

Elias L. Rivers. 1993. *El soneto español en el siglo de oro*. Akal, Madrid.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic Compositionality through Recursive Matrix-Vector Spaces. In *Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211.

Timothy R. Tangherlini and Peter Leonard. 2013. Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research. *Poetics*, 41:725–749.

Arthur Terry. 1993. *Seventeenth-Century Spanish Poetry*. Cambridge University Press.

PD Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Elena Varelo-Merino, Pablo Moíno-Sánchez, and Pablo Jauralde-Pou. 2005. *Manual de métrica española*. Castalia, Madrid.