Metrical annotation of a large corpus of Spanish sonnets: representation, scansion and evaluation.

Borja Navarro-Colorado[1], María Ribes Lafoz[2], Noelia Sánchez
[1]Department of Software and Computing System, [2]UA Language Centre
University of Alicante, Alicante (Spain)
borja@dlsi.ua.es, chitty@csidiomas.ua.es, noeliasanchezlopez@gmail.com

Abstract

The latest developments in Digital Humanities demand the creation of new and specific language resources. Specifically, the application of text-mining algorithms in a macro-analysis approach to the study of literary texts needs not only large but also richly annotated corpora. In order to analyze metrical and semantics aspects of poetry in Spanish with computational techniques, we have developed a large corpus annotated with metrical information. In this paper we will present and discuss the development of this corpus: the formal representation of metrical patterns, the semi-automatic annotation process based on an automatic scansion system, and the main annotation problems. Finally we will present the evaluation, in which an inter-annotator agreement of 96% has been obtained. The corpus is open and available.

Keywords: metre, scansion, corpus annotation, poetry, Spanish

1. Introduction

Even though the application of Natural Language Processing techniques to literary texts is not a new task, nowadays it is still undergoing a new stage of development.[1] For instance, the macroanalysis approach (Jockers, 2013), based on the study of large periods of the history of literature through the computational analysis of large corpora, needs the development of large and richly annotated corpora of literary texts.

Our interest is focused on the computational analysis of Spanish Literature, and specifically on the metrical and semantic aspects of Spanish Golden-Age Sonnets (Navarro Colorado, 2015). 16th- and 17th-Centuries Spanish poetry is considered one of the best periods 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 since the 19th century. We are persuaded that new approaches based on the macroanalysis framework (Jockers, 2013) (or "distant reading" (Moretti, 2007; Moretti, 2013)) could shed new light on this period.

We have compiled a large and representative corpus of Spanish sonnets written during the 16th and 17th centuries, in which the metrical pattern of each verse has been annotated following a semi-automatic process. In this paper we will present and discuss the three main aspects of this corpus:

1. a formal representation of metrical information;
2. the semi-automatic annotation process based on a new automatic scansion system;
3. the evaluation of the annotation.

The paper is organized as follows: after a brief State of the Art, section 3. shows how the corpus of Spanish Golden-Age sonnets has been compiled and and how the metrical patterns have been formally represented; section 4. describes the scansion system that has been specifically developed to annotate these metrical patterns, the manual revision and the main problems detected; section 5. presents the evaluation of the annotation. The paper finishes with some conclusions.

2. State or the art

Nowadays there are several projects devoted to the compilation and annotation of corpora of literary texts. For Spanish, González-Blanco and Rodríguez (2013) are developing a computer-based metrical repertory of Medieval Castilian poetry (from the 12th to the 15th century). The metrical information they provide is only about the size of the verse (the number of syllables), but they do not differentiate between stressed and unstressed syllables.

3. Corpus compilation and XML annotation

Our corpus is composed of 5078 sonnets written during the 16th and 17th centuries in Spanish (Castilian). 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[2].

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 that were written on that period, not only the canonical sonnets. In other words, the corpus must represent the literary context of any Golden Age poet.

[1] See, for example, the Computational Linguistics for Literature Workshop series.

[2] `http://www.cervantesvirtual.com/`

Each sonnet has been automatically marked with the standard TEI-XML[3]. 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);
- encoding description: how the metrical pattern is encoded (see below);
- metrical annotation status: that is, whether the metrical pattern has been manually checked or not.

Regarding the sonnet structure, we have annotated:

- quatrains (lg);
- tercets (lg);
- verse line (l), with the line number ($@n$) and the metrical pattern ($@met$);
- some extra lines included in the poem (called "estrambote").

3.1. Metrical representation

For each verse, the markup includes a representation of its metrical structure: the metrical pattern.

Spanish is considered a syllable-timed language (Abercrombie, 1990): poetic lines are measured in syllables. The Spanish sonnet is an adaptation of the Italian sonnet, where each line has eleven syllables (hendecasyllable). The metre of each verse line is formed by a combination of stressed and unstressed syllables. It is compulsory to have a stressed syllable in the tenth position. The rest of the stressed syllables can be combined in different ways, generating different rhythms: heroic (stressed syllables at position 2, 6 and 10), melodic (at 3, 6, 10), sapphic (at 4, 8, 10), etc. (Quilis, 1984; Varela-Merino et al., 2005)

The formal representation of metrical information is rooted in two assumptions:

1. Objectivity: the metrical pattern represents only metrical information that could be derived from the text objectively. Therefore, on one hand, the pattern is formed only by two kinds of syllables: stressed and unstressed. It will not represent secondary stresses. On the other hand, it will not represent potential pauses or silences, because they depend on how the lines are performed (a subjective decision of the performer).
2. Theoretical neutrality: as we did in previous corpus annotation projects (Navarro Colorado et al., 2003), in order to develop a corpus useful for as many researchers as possible, the metrical pattern is not based on any specific metrical theory. On the contrary, we will try to represent only the metrical aspects assumed by (or that appear in) any metrical theory. In this regard, the metrical pattern will not represent metre as a bracketed grid or any other specific aspect of generative metrics (Fabb and Halle, 2008). We use more traditional (Spanish) metrical studies as a theoretical framework, such as Quilis (1984) or Navarro Tomás (1995).

In spite of these principles, we must re-define the traditional metrical unit. The poetry metre is formed by a sequence of stressed and unstressed syllables organized in "metre groups" or "feet". According to Varela-Merino et al (2005), a "metre group" in the Spanish language is a sequence of unstressed syllables (zero o more) around a stressed one and delimited by a potential silence.

The problem is that, as we said before, this potential silence is subjective. It depends on how the verse is performed. The same verse could have different "metre groups" according to the different potential pauses a performer makes. The sole potential silence that is, to some extent, objective is the silence at the end of the line. It is more objective because, first, it depends on a fixed number of syllables (eleven in this case) and, second, it is explicitly marked on the paper with the change of line.[4]

Therefore, for corpus annotation purposes, we take the whole verse line as a metrical unit. A metrical pattern is, then, a sequence of stressed and unstressed syllables limited by the end of the line (the potential pause marked in the text). Formally, following the TEI standard, the metrical pattern is a combination of symbols "+" (stressed syllable) and "-" (unstressed syllable). It represents an abstract pattern formed by the compulsory metrical elements of any declamation and assumed by any metrical theory. Example 1 shows a verse (hendecasyllable) and its metrical pattern (sapphic: stressed syllables in position 4, 8 and 10):

(1) Cuando me paro a contemplar mi estado
$< met = \text{“} - - - + - - - + - + - \text{”} >$

4. Annotation process

In order to extract and annotate the metrical pattern of each verse line, we have followed a semi-automatic approach. First, all the verse lines have been analyzed with an automatic scansion system, then a set of metrical patterns has been manually checked and fixed. In this section we will present first the main scansion problems, then the automatic scansion system and finally the manual revision process.

4.1. Scansion problems

Scansion is the act of analyzing a line of verse in order to determine its metrical pattern. There are two main steps: first, splitting the verse line in syllables and, second, detecting the stressed syllables. However, the scansion of real

[3] www.tei-c.org/

[4] There would be cases in which the performer would not make a potential pause at the end of the line, for instance when there is an enjambment, but in any case, the separation of lines is a explicit mark which produces a closing cadence.

verse lines is more complex than it seems. Due to the fact that there is not a direct relationship between linguistic syllables and metrical syllables, some ambiguity problems appear. 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 will have ten syllables and an extra syllable must be added. On contrary, if the last stressed syllable is the antepenultimate (proparoxytone), the line will have twelve syllables and the last syllable must be removed.
- Not every word with a linguistic accent has a metrical accent. It depends on the Part of Speech of the word, and the context in which it appears. 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 (synaloepha or syneresis). This is a natural phonetic phenomenon, but it is not always carried out: it depends on several factors as, for example, the intention during declamation.
- The opposite is possible too: a single syllable with two vowels (normally with a semivowel like an “i” or “u”) that can be pronounced as two separate syllables (dieresis).

These phenomena could change the extracted metrical pattern in two different ways: the amount of syllables and the type of each one of them (stressed or unstressed). The main ambiguity problem arises when two or more patterns can be extracted from the same verse, all of them correct.

For example, for a verse with twelve syllables and a paroxytonal 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. Depending on which one we choose, the final metrical pattern will be completely different.

See, for example, the following verse line (2):

(2) *cuando el padre Hebrero nos enseña*[5]

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 (3):

(3) $< met =$ “$- - + - - + - - - + -$” $>$
$< met =$ “$- - - + - + - - - + -$” $>$

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 or interpreted.

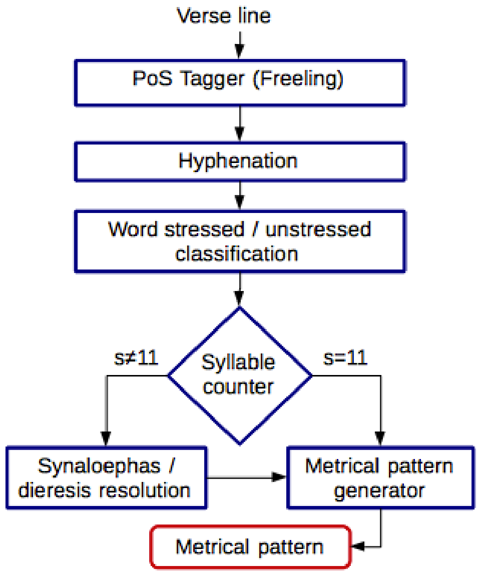
Verse line
PoS Tagger (Freeling)
Hyphenation
Word stressed / unstressed classification
s≠11
Syllable counter
s=11
Synaloephas / dieresis resolution
Metrical pattern generator
Metrical pattern

Figure 1: Automatic scansion system

4.2. Automatic scansion system

We have created a scansion system for Spanish based on Computational Linguistics techniques. An automatic metrical scansion system must resolve this ambiguity.[6] 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.

We have developed a rule-based system approach. The input of the system is a line and the output is a metrical pattern. The system consists of four main modules (see Figure 1):

1. Part of Speech tagger: we use the PoS tagger Freeling (Padró and Stanilovsky, 2012).[7] For each word, the scansion system selects the most general PoS tag (noun, verb, etc.) Only in a few cases a deeper analysis is needed. For example, the system must distinguish between personal pronouns (stressed) and clitic pronouns (unstressed).
2. Hyphenation module: it consists of a set of rules to split words into syllables and to detect the stressed syllable in each word.
3. Stressed / unstressed word classification module: as we said before, in the context of a line not all words are stressed. This module classifies the words of the line into stressed or unstressed according to their Part of Speech (Quilis, 1984).

[5] (ENG) “When father Hebrero teaches us”

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

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

4. Synaloepha resolution: if the line has more than eleven syllables, this module applies the rules to blend possible synaloephas. They are applied according to traditional metrical studies, which have defined a ranking of natural and artificial synaloephas. For example, it is more natural to join two unstressed vowels than two stressed vowels (Quilis, 1984). On the contrary, if the line has less than eleven syllables, then dieresis rules are applied in order to split diphthongs in two syllables.

All the corpus has been automatically annotated with this scansion system. At this moment we have around 71 100 metrical patterns that belongs to around 5078 sonnets.

4.3. Manual revision

Nowadays we are manually reviewing the automatic annotation in order to correct errors and set up a Gold Standard. At this moment, 1% of the metrical patterns of the corpus has been manually checked.

The manual revision is carried out in three steps:

1. Annotators training. In this stage, annotators learn how to annotate metrical patterns in the corpus. 38 sonnets have been annotated in parallel and all the problems have been discussed among annotators. From this stage an annotation guide has been created.[8]

2. Evaluation. Inter-annotator agreement has been calculated in order to evaluate the manual annotation. See below.

3. Checking and validation of potentially ambiguous patterns. In order to reduce the time-consuming manual annotation, from now to the end of the project we will manually review only those patterns potentially ambiguous. Once we have detected the main automatic annotation problem, we are extracting all those metrical patterns with some kind of ambiguity. For example, only 7,67% of lines have some kind of ambiguity due to the synaloepha phenomenon.

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

Table 1: Most frequent metrical patterns.

5. Evaluation

Both the automatic scansion system and the manual annotation have been evaluated. 100 sonnets (1400 metrical patterns) from all the corpus have been first automatically annotated, and then manually checked by two annotators in parallel.

In order to evaluate the automatic annotation system, we have compared the patterns extracted by the system with the patterns manually annotated. The system has achieve a 92% precision: from 1400 verses, 108 metrical patterns were extracted incorrectly. The main problems are the ones reported previously.

In order to evaluate the manual annotation, we have calculated the inter-annotator agreement (IAA) comparing the metrical patterns annotated by each annotator. From 1400 verses, in 1347 cases the same metrical pattern has been assigned to the same verse, and in 54 cases the metrical pattern assigned is different. It is a 96% of IAA.

The disagreement among annotators has been produced by three main reasons:

1. Problems with the guidelines. Some aspects of the annotation were not clearly explained in the annotation guideline. For example, how to annotate the Spanish adverbs finished in "-mente". This kind of adverbs has always two stressed syllables. The guidelines have been rewritten attending to these issues (0,7% of the patterns).

2. Problems with semantic ambiguity. Some terms in the verses have been interpreted differently by each annotator. Given that this is a corpus of poetry, where wordplay is a must, this problem has no easy solution (1,6% of the patterns).

3. Problems resulting from the application of the annotation guidelines. In some cases, it was just a mistake made by an annotator. (1,4% of the patterns).

With these data we can conclude than the annotation is being properly developed and the annotation guidelines have been correctly settled and applied.

6. Conclusions and Future Work

In this paper we have presented a corpus of Golden-Age Spanish sonnets and the metrical annotation of each verse. The metrical information has been formalized through a metrical pattern that represents the abstract metre associated with a specific verse. The corpus has been annotated following a semi-automatic approach, profiting from of a new automatic scansion system. We have reported the main annotation problems. Finally, the annotation has been evaluated, obtaining a 96% of the Inter-Annotator Agreement. The corpus is available at the project web page:

`http://www.dlsi.ua.es/~borja/mpb/`

As a Future Work, we plan to develop a machine learning module to improve both the automatic scansion system and the manual annotation through an iterative process.

[8] The annotation guide is available at the project web site: `http://www.dlsi.ua.es/~borja/GuiaAnotacionMetrica.pdf` (in Spanish)

7. References

Abercrombie, D. (1990). *Elements of General Phonetics*. Edinburgh University Press.

Agirrezabal, M., Arrieta, B., Astigarraga, A., and Hulden, M. (2013). ZeuScansion : a tool for scansion of English poetry. In *11th International Conference on Finite State Methods and Natural Language Processing*, St Andrews, Scotland.

Fabb, N. and Halle, M. (2008). *Meter in Poetry. A new theory.* Cambridge University Press, Cambridge (UK).

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

González-Blanco García, E.; Rodríguez, J. L. (2013). ReMetCa: a TEI based digital repertory on Medieval Spanish poetry. In *The Linked TEI: Text Encoding in the Web*, Rome.

Greene, E., Ave, L., Knight, K., and Rey, M. (2010). Automatic Analysis of Rhythmic Poetry with Applications to Generation and Translation. In *Empirical Methods in Natural Language Processing*, Massachusetts.

Hammond, A., Brooke, J., and Hirst, G. (2013). A Tale of Two Cultures : Bringing Literary Analysis and Computational Linguistics Together. In *Workshop on Computational Linguistics for Literature*, Atlanta, Georgia.

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

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

Mainer, J. C. (2010). *Historia de la literatura española*. Crítica, Barcelona.

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

Moretti, F. (2013). *Distant reading*. Verso.

Navarro Colorado, B., Civit, M., Martí, A., Marcos, R., and Fernández, B. (2003). Syntactic, semantic and pragmatic annotation in Cast3LB. In *Proceedings of The Shallow Processing of Large Corpora Workshop (SProLaC 2003)*, Lancaster.

Navarro Colorado, B. (2015). A computational linguistic approach to Spanish Golden Age Sonnets: metrical and semantic aspects. In *Workshop on Computational Linguistics for Literature.*, Denver.

Navarro Tomás, T. (1995). *Métrica española*. Labor.

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

Quilis, A. (1984). *Métrica española*. Ariel, Barcelona.

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

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

Varela-Merino, E., Moíno-Sánchez, P., and Jauralde-Pou, P. (2005). *Manual de métrica española*. Castalia, Madrid.