Enjambment Detection in a Large Diachronic Corpus of Spanish Sonnets

Pablo Ruiz, Clara Martínez Cantón, Thierry Poibeau, Elena Gonzalez-Blanco

To cite this version:

Pablo Ruiz, Clara Martínez Cantón, Thierry Poibeau, Elena Gonzalez-Blanco. Enjambment Detection in a Large Diachronic Corpus of Spanish Sonnets. LaTeCH-CLFL 2017, Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature., Association for Computational Linguistics, Aug 2017, Vancouver, Canada. pp.27 - 32. hal-01575168

HAL Id: hal-01575168
https://hal.archives-ouvertes.fr/hal-01575168
Submitted on 18 Aug 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Enjambment Detection in a Large Diachronic Corpus of Spanish Sonnets

Pablo Ruiz Fabo\textsuperscript{1,3}, Clara I. Martínez Cantón\textsuperscript{2,3}, Thierry Poibeau\textsuperscript{1} and Elena González-Blanco\textsuperscript{2,3}

\textsuperscript{1}Laboratoire LATTICE. CNRS, ENS, U Paris 3, PSL Research U, USPC 92120 Montrouge, France
\{pablo.ruiz.fabo,thierry.poibeau@ens.fr\}

\textsuperscript{2}Department of Spanish Literature and Literary Theory. UNED 28040 Madrid, Spain

\textsuperscript{3}LINHD: Digital Humanities Innovation Lab. UNED 28040 Madrid, Spain
\{cimartinez,egonzalezblanco@flog.uned.es\}

Abstract

Enjambment takes place when a syntactic unit is broken up across two lines of poetry, giving rise to different stylistic effects. In Spanish literary studies, there are unclear points about the types of stylistic effects that can arise, and under which linguistic conditions. To systematically gather evidence about this, we developed a system to automatically identify enjambment (and its type) in Spanish. For evaluation, we manually annotated a reference corpus covering different periods. As a scholarly corpus to apply the tool, from public HTML sources we created a diachronic corpus covering four centuries of sonnets (3750 poems), and we analyzed the occurrence of enjambment across stanzaic boundaries in different periods. Besides, we found examples that highlight limitations in current definitions of enjambment.

1 Introduction

Enjambment takes place when a syntactic unit is broken up across two lines of poetry (Domínguez Caparrós, 1988, 103), giving rise to different stylistic effects (e.g. increased emphasis on elements of the broken-up phrase, or contrast between those elements), or creating double interpretations for the enjambed lines (García-Page Sánchez, 1991).

The literature shows a debate on the stylistic effects emerging from a mismatch between syntactic and metrical units (Martínez Cantón, 2011). The types of effects possible and the syntactic units where the effects can be said to be attested are a matter of current research. Quilis (1964) characterized enjambment as occurring in a series of very specific syntactic contexts. The definition is still considered current, however, some aspects in it have been questioned: Are these the only syntactic configurations where such effects are observed? Are syntactic criteria enough to predict when these effects arise?

Given these unclear points, it is relevant to systematically collect large amounts of enjambment examples, according to current definitions of the phenomenon. This can provide helpful evidence to assess scholars’ claims. To this end, we developed a system to automatically detect enjambment in Spanish, applying it to a corpus of ca. 3750 sonnets by 1000 authors (15th to 19th century).

We are not aware of a systematic large-sample study of enjambment across periods, literary movements, or versification types in Spanish, or other languages. Automatic detection can help answer interesting questions in verse theory, which would benefit from a quantitative approach, complementing small-sample analyses, e.g.: “To what extent is enjambment used differently in free verse vs. traditional versification?” or “Does the use of enjambment increase in movements that seek distance from traditional forms?”

Finally, our study complements automatic metrical analyses of Spanish Golden Age sonnets by
Navarro-Colorado (2016; 2017), by focusing on enjambment and covering a wider period.

The paper is structured thus: First we provide the definition of enjambment adopted. Then, our corpus and system are described, followed by an evaluation of the system’s outputs. Finally, findings on enjambment in our diachronic sonnet corpus are discussed. Our project website provides details omitted here for space reasons.1

2 Enjambment in Spanish

Syntactic and metrical units often match in poetry. However, this trend has been broken since antiquity for various reasons (Parry (1929) on Homer, or Flores Gómez (1988) on early classical poetry).

Enjambment is considered to take place when a pause suggested by poetic form (e.g. at the end of a line or across hemistichs) occurs between strongly connected lexical or syntactic units, triggering an unnatural cut between those units.

Quilis (1964) carried out reading experiments, proposing that several strongly connected elements give rise to enjambment, should a poetic-form pause break them up:

1. Lexical enjambment: Breaking up a word.
2. Phrase-bounded enjambment: Within a phrase, breaking up sequences like noun + adjective, noun + prepositional phrase complementing it, verb + adverb, auxiliary verb + main verb, among others. For instance, the italicized words in the following lines by Matthew Arnold would be an enjambment, as a line-boundary intervenes between the noun roar and the prepositional phrase complementing it (Of pebbles): “Listen! you hear the grating roar // Of pebbles which the waves draw back, and fling, // At their return, up the high strand”.
3. Cross-clause enjambment: Between a noun antecedent and the pronoun heading a defining relative clause that complements the antecedent (e.g. “people // who persevere may succeed”).

Besides the enjambment types above, Spang (1983) noted that if a subject or direct object and their related verbs occur in two different lines of poetry, this can also feel unusual for a reader, even if the effect is less remarkable than in the environments identified by Quilis. To differentiate these cases from enjambment proper, Spang calls these cases enlace, translated here as expansion.

The procedure in Quilis (1964, 55ff.) for assessing the strength of the cohesion within syntactic elements was as follows: Around 50 participants were asked to read literary prose excerpts. Syntactic units within which it was rare for participants to produce a pause were considered to be strongly cohesive (see the list above). The unnaturalness of producing a pause within these units was seen as contributing to an effect of mismatch between meter and syntax, should the units be interrupted by a metrical pause.

Quilis (1964) was the only author so far to gather reading-based experimental evidence on Spanish enjambment. His typology is still considered current, and was adopted by later authors, although complementary enjambment typologies have been proposed, as Martínez Cantón (2011) reviews. Our system identifies Quilis’ types, in addition to Spang’s expansion cases.

Above we listed Quilis’ three broad types, but there are subtypes for each, equally annotated by our system; a detailed description and examples for each type and subtype is on our site.2

3 Diachronic Sonnet Corpus

The corpus is based on two public online collections (García González, 2006a,b). The first one covers 1088 sonnets by 477 authors from the 15th–17th centuries. The second one contains 2673 sonnets by 685 authors from the 19th century. We created scripts to download the poems, remove HTML and extract dates of birth and death for the authors. The corpus covers canonical as well as minor authors, inspired in distant reading approaches (Moretti, 2005, 2013). The distribution of sonnets and authors over periods is given on the project’s site.3

3.1 System Description

The system has three components: a preprocessing module to format input poems uniformly, an NLP pipeline, and the enjambment-detection module itself.

1https://sites.google.com/site/spanishenjambment
2https://sites.google.com/site/spanishenjambment/enjambment-types
3https://sites.google.com/site/spanishenjambment/our-large-sonnet-corpus
We used the IXA Pipes library as the NLP pipeline (Agerri et al., 2014), obtaining part-of-speech tags, syntactic constituents and syntactic dependencies with it.

In the absence of data annotated for enjambment, that may allow applying a machine learning approach, we created a rule and dictionary-based system that exploits the information provided by the NLP pipeline. A total of ca. 30 rules identify enjambed lines, assigning them a type among a list of 11 types, based on the typology in section 2. Some of the rules are very shallow, only taking the part-of-speech sequences around a line boundary into account. Some other rules additionally exploit constituency information. Dependency parsing results are used to detect among other cases subject/object/verb relations, relevant for the expansion cases defined by Spang (see section 2). For any type of rule, custom dictionaries can restrict rule application to a set of terms. E.g. certain verbs govern arguments introduced by one specific preposition; we itemized these verbs and their prepositions in a dictionary, to complement information provided by the NLP pipeline or to correct parsing errors. The lists of verbs and prepositions were obtained from online resources on the descriptive grammar of Spanish.⁴

An example of a rule would be the following: If line \( n \) contains a verb \( v \), and line \( n+1 \) has a prepositional argument \( pa \) governed by \( v \), and \( v \) is listed in the custom dictionary as accepting arguments introduced by \( pa \)'s preposition, assign enjambment type \( verb_{,}\ prep \) to line-pair \( \langle n, n+1 \rangle \).

It is possible, but rare in or corpus, for more than one enjambment type to be applicable to a line-pair. At the moment, the system annotates only one type per line, following a fixed rule order. In the future, criteria to output and evaluate multiple types per line could be developed.

The rules are currently implemented as Python functions. Future work that could benefit non-programmer users would be to make the rules configurable rather than written directly in code.

Enjambment annotations are output in a standoff format; the project’s site provides details.⁵

4 Evaluation and Result Discussion

We describe the evaluation method (the reference sets, the task and metrics), and present the results along with a brief discussion of error sources. Comments about the relevance of the results for literary studies are provided in section 5.

4.1 Test Corpora

To evaluate the system, we created two reference-sets (SonnetEvol and Cantos20th), which were manually annotated for enjambment by a metrics professor and a linguist.

1. SonnetEvol: 100 sonnets (1400 lines) from our diachronic sonnet corpus of ca. 3750 sonnets. This test-set contains 260 pairs of enjambed lines.
2. Cantos20th: 1000 lines of 20th century poetry (Colinas, 1983), showing natural contemporary syntax. We identified 277 pairs of enjambed lines.

The SonnetEvol diachronic test-set covers all centuries, with ca. 70% of sonnets from the 15th–17th centuries and 30% from the 19th. The test-sets cover all enjambment types, but some types are infrequent in them, as in Spanish poetry overall.

We annotated the Cantos20th corpus in order to assess the system’s performance on contemporary Spanish with natural diction, compared to its behaviour with the SonnetEvol corpus, which includes some archaic constructions and often shows an elevated register.

The distribution of enjambment types in both test-corpora is shown on Table 1. The enjambment types are described in detail, with examples, on our site.² The type labels generally stand for the constituents that take part in an enjambment, e.g. noun_prep and adj_prep mean, respectively, a noun or an adjective and the prepositional phrase complementing them.

To have an indication of the reliability of the annotation scheme, 50 sonnets of the SonnetEvol corpus were each tagged by two annotators. The ratio of matching labels across both annotators was 91.7%. Besides, a set of 120 sonnets (not from the test-sets) annotated by our students were later corrected by the professor; the ratio of matching labels was 89.7%. Getting several annotators’ input on more sonnets, and obtaining inter-annotator agreement metrics (e.g. Artstein and Poesio (2008)) is part of our planned future work.

---

4 http://www.wikilengua.org/index.php/Lista_de_complementos_de_regimen_A
5 https://sites.google.com/site/spanishenjambment/annotation-and-result-format
### 4.2 Enjambment-detection Tasks Evaluated

We defined two enjambment-detection tasks: **untyped match** and **typed match**. In **untyped match**, the positions of enjambed lines proposed by the system must match the positions in the reference corpus for a correct result to be counted. In **typed match**, for a correct result, both the positions and the enjambment type assigned by the system to those positions must match the reference.

The untyped match task can be seen as an enjambment **recognition** task, and typed match corresponds to an enjambment **classification** task.

### 4.3 System Results and Discussion

Precision, recall and F1 were obtained. Table 2 provides overall results for both corpora. Table 3 provides the per-type results on the diachronic test-corpus (SonnetEvol). The project’s site shows more detailed results. Lexical enjambment is not listed on the tables above, as no occurrences were found in the test corpora.

For untyped match, F1 reaches 80 points in the SonnetEvol corpus, whereas F1 for typed match is 66.31. For the contemporary Spanish corpus (Cantos20th), F1 is higher: 80.63 typed match, and 86.51 untyped match. This reflects additional difficulties posed by archaic language and historical varieties for the NLP system whose outputs our enjambment detection relies on.

A common source of error was hyperbaton: the displacement of phrases triggers constituency and dependency parsing errors. Prepositional phrase (PP) attachment also posed challenges: Verbal adjuncts get mistaken for PPs complementing nouns or adjectives. Creating a reparsing module to manage hyperbaton and improve PP attachment results may be fruitful future work.

Further interesting future work would be a detailed analysis of error sources. This would help determine the extent to which errors are due to the enjambment detection rules in themselves or to the NLP pipeline. In the second case, it would be useful to know the extent to which POS-tagging

---

6 [https://sites.google.com/site/spanishenjambment/evaluation](https://sites.google.com/site/spanishenjambment/evaluation)

7 PP attachment is a difficulty even in current languages (e.g. Agirre et al. (2008) for English). For historical varieties, Stein’s (2016) results for verbal adjuncts and prepositional complements in Old French also suggest this difficulty.
Figure 1: Percentage of enjambments per position in the 15th–17th centuries vs. the 19th. The y-axis represents line-positions; the x-axis is the percentage of enjambed line-pairs for a position over all enjambed line-pairs in the period. Enjambment across quatrains and across the octave-sestet divide is very rare, with a small increase in the 19th century. The division between the tercets blurs in the 19th century, in the sense that enjambment across them is clearly higher than in the previous period.

or parsing errors are due to archaic features and complex diction in some of the earlier sonnets in the corpus. The earlier varieties of Spanish covered in the corpus have a large lexical and syntactic overlap with contemporary Spanish, which justified applying NLP models for current Spanish to the entire corpus (besides the fact that we are not aware of NLP tools for 15th–17th century Spanish). However, it would be relevant to quantify error sources per period.

5 Relevance for Literary Studies

The system’s goal is detecting enjambment to help literary research on the phenomenon, via providing systematic evidence for its analysis. For instance, in our result validation, we find that the system annotates line-pairs that formally fit the description of an enjambment context (see section 2), but that we’d actually consider unlikely to yield a stylistic effect. Conversely, our annotators are sometimes surprised that line-pairs where they perceive an unnatural mismatch between syntactic and line-boundaries are not captured by our typology and left unannotated by the system.

Regarding the system’s potential for quantitative analyses, we consider our untyped detection results helpful, given an F1 of ca. 80 points on the diachronic test-set. As an example application, we examined the distribution of enjambment according to position in the poem, particularly in positions across a verse-boundary (lines 4–5, 8–9 and 11–12). Comparing the results for the 15th-to-17th centuries vs. the 19th century (Figure 1), we see that enjambment across the tercets increases clearly in the 19th century, with a small increase of enjambment across the quatrains (lines 4–5) and across the octave-sestet divide (lines 8–9). Performing such analyses on a large corpus opens the door for scholars to assess the literary relevance of the findings, and search for the best interpretation.

6 Outlook

With automatic enjambment detection, our goal is to help gather systematic large scale evidence to study the complex phenomenon of enjambment, which poses challenges for metrical and stylistic theory to characterize, and for critical practice to apply. Our metrics students have so far manually annotated enjambment for 400 sonnets; their work will permit computing inter-annotator agreement, and performing new tests of the automatic system. As our manually annotated corpus grows, we will examine the possibility of using supervised machine learning to train a sequence labeling and classification model to complement our current rules. A specific goal is improving enjambment type detection for the typed match task.

Acknowledgments

Pablo Ruiz Fabo was supported by a PhD scholarship from Région Île-de-France. The research was also supported by Starting Grant ERC-2015-STG-679528 POSTDATA. We are grateful to the anonymous reviewers for their valuable comments.
References


María Esperanza Flores Gómez. 1988. Coincidiencia y distorsión (encabalgamiento) de la unidad rítmica verso y las unidades sintácticas. [Coincidence and distortion (enjambment) between the line as a rhythmic unit and syntactic units]. Estudios clásicos 30(94):23–42.


