Distant Rhythm: Automatic Enjambment Detection on Four Centuries of Spanish Sonnets

Pablo Ruiz Fabo[1], Clara Martínez Cantón[2], Thierry Poibeau[1]

[1]Laboratoire LATTICE, Paris (ENS, CNRS, U Paris 3, PSL Research U, USPC)
[2]Department of Spanish Literature and Literary Theory, UNED, Madrid

[Accepted at Digital Humanities 2017 in Montréal]

1. Introduction

Enjambment takes place when a syntactic unit is broken up across two lines of poetry (Domínguez Caparrós, 2000: 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-Paje, 1991).

In Spanish poetry, the syntactic configurations under which enjambment takes place have been described extensively, and detailed studies on the use of enjambment by individual authors exist (see Martínez Cantón, 2011 for an overview)[1]. However, a larger-scale study to identify enjambment across hundreds of authors spanning several centuries, enabling distant reading (Moretti, 2013), was not previously available.

Given that need, we have developed software, based on Natural Language Processing, that automatically identifies enjambment in Spanish, and applied it to a corpus of approx. 3750 sonnets by ca. 1000 authors, from the 15th to the 19th century. What is the interest of such large-scale automatic analyses of enjambment? First, the literature shows a debate about which specific syntactic units can be considered to trigger enjambment, if split across two lines, and whether lexical and syntactic criteria are sufficient to identify enjambment. Second, the stylistic effects that enjambment permits are also an object of current research (Martínez Fernández, 2010). Systematically collecting large amounts of enjambment examples provides helpful evidence to assess scholars' current claims, and may stimulate novel analyses. Finally, our study complements Navarro's (2016) automatic metrical analyses of Spanish Golden Age sonnets, by covering a wider period and focusing on enjambment.

The abstract 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. Finally, findings on enjambment in our diachronic sonnet corpus are discussed. The project's website provides details omitted here for space reasons.[2]

[1] Among others, Quilis (1964), Domínguez Caparrós, (2000), Paraíso, (2000), Spang (1983) for a description of enjambment, and Alarcos (1966), Senabre (1982), Luján (2006), Martínez Fernández (2010) for case-studies on a single author.
[2] http://sites.google.com/site/spanishenjambment contains samples for the corpus and results, and other details.

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).

In Spanish tradition, enjambment[3] 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) performed reading experiments, proposing that the following strongly connected elements give rise to enjambment, should a poetic-form pause break them up:

1. **Lexical enjambment**:[4] Breaking up a word.
2. **Phrase-bounded enjambment**: Within a phrase, breaking up sequences like "noun + adjective", "verb + adverb", "auxiliary verb + main verb", among others.[5]
3. **Cross-clause enjambment**: Between a noun antecedent and the pronoun heading the relative clause that complements the antecedent.

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 pronounced than in the environments identified by Quilis. To differentiate these cases from enjambment proper, Spang calls these cases "enlace", translated here as **"expansion"**.

Quilis (1964) was the only author so far to gather recitation-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, besides Spang's expansion cases.

[3] The term for enjambment in Spanish studies is "encabalgamiento".
[4] We translate Quilis' terms thus: "lexical enjambment" stands for "encabalgamiento léxico" or "tmesis". "Phrase-bounded enjambment" stands for "encabalgamiento sirremático", and "cross-clause enjambment" stands for "encabalgamiento oracional".
[5] For Quilis' complete list of syntactic environments that can trigger enjambment, and the types identified by our system, see https://sites.google.com/site/spanishenjambment/enjambment-types

3. Corpus

The corpus is based on two public online collections[6] (García González, R. (ed.), 2006a, 2006b). 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.[7] Table 1 shows the distribution of authors and poems by century. The corpus covers canonical as well as minor authors, inspired in distant reading approaches (Moretti, 2007, 2013).

Period *	Sonnet Count	Sonnet %	Author Count	Author %
14.5	43	1.14	2	0.17
15	2	0.05	2	0.17
15.5	8	0.21	5	0.43
16	141	3.75	58	4.99
16.5	411	10.93	108	9.29
17	478	12.71	300	25.82
17.5	5	0.13	2	0.17
18.5	13	0.35	6	0.52
19	1150	30.58	361	31.07
19.5	1510	40.15	318	27.37
Total	*3761*	*100*	*1162*	*100*

TABLE 1: DISTRIBUTION OF SONNETS AND AUTHORS PER PERIOD.

*** Exact dates of birth and death are available for a minority of authors; often only the century was provided in the corpus sources. Periods ending in ".5" cover authors who lived in two centuries. E.g. period "15.5" covers authors born in the 15th and deceased in the 16th century**

[6] From Biblioteca Virtual Cervantes, http://www.cervantesvirtual.com/

[7] About 30% of the 15th to 17th century authors had exact dates of birth and death, for the rest only the centuries were available. Among the 19th century authors, ca. 45% had exact dates of birth and death.

4. System description

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

The NLP pipeline is IXA Pipes (Agerri et al., 2014). Its results for contemporary Spanish are competitive. Our system uses it to obtain part-of-speech tags, syntactic constituency (e.g. verb-phrase, noun-phrase) and syntactic dependencies (e.g. direct object).

The enjambment detection module is rule and dictionary-based, and exploits the information provided by the NLP pipeline. Rules (30 in total) of different characteristics identify enjambed lines, assigning them a type among a list of 12 types, based on the typology in Section 2.[8]

- Some rules are very shallow and only take **parts of speech** into account.
- Some rules additionally exploit **constituency** info.
- Some rules use **dependency** information, e.g. to detect "subject / object / verb" relations.
- 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 correct parsing errors.

Enjambment annotations are output in standoff format. The project's site provides details.[9]

[8] The full list of types identified by the system is at the project's site: https://sites.google.com/site/spanishenjambment/enjambment-types
[9] https://sites.google.com/site/spanishenjambment/annotation-and-result-format

5. System evaluation and discussion

5.1. Test-corpus

To evaluate the system, we created two reference-sets (SonnetEvol and Cantos20th), manually annotating enjambment in them.

1. **SonnetEvol**: 100 sonnets (1400 lines) from our diachronic sonnet corpus of ca. 3750 sonnets (Table 1). This test-set contains 260 pairs of enjambed lines.[10]
2. **Cantos20th**: 1000 lines of 20th century poetry (Colinas, 1983), showing natural contemporary syntax. We identified 277 pairs of enjambed lines.

The distribution of enjambment types in the test-corpora is balanced (Table 2). The SonnetEvol diachronic test-corpus is balanced across periods (Table 3).[11]

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.

For the evaluation reported here, each sonnet was annotated by a single annotator. Obtaining multiple annotators' input on the same sonnet to assess inter-annotator agreement (Artstein and Poesio, 2008) is part of our ongoing work.

	Test-Corpus			
	SonnetEvol		*Cantos20th*	
Enjambment Types *	**Count**	**%**	**Count**	**%**
Total Phrase-Bounded	*104*	*40.00*	*175*	*63.18*
adj_adv	2	0.77	1	0.36
adj_noun	29	11.15	54	19.49
adj_prep	14	5.38	11	3.97
adv_prep	0	0	3	1.08
noun_prep	39	15.00	85	30.69
relword	1	0.38	2	0.72
verb_adv	5	1.92	7	2.53
verb_cprep	9	3.46	2	0.72
verb_chain	5	1.92	10	3.61
Total Cross-Clause[12]	*23*	*8.85*	*31*	*11.19*
Total Expansions	*133*	*51.15*	*71*	*25.63*
dobj_verb	65	25.00	39	14.08
subj_verb	68	26.15	32	11.55
Total All Types	*260*	*100*	*277*	*100*

TABLE 2: DISTRIBUTION OF ENJAMBMENT TYPES IN THE MANUALLY ANNOTATED REFERENCE CORPORA, providing counts and each type's percentage of the total enjambments per corpus. Counts refer to pairs of enjambed lines

SonnetEvol Test-corpus	
Period **	**Sonnet Count**
Total 15th-17th	*72*
14.5	3
15	2
15.5	2
16	14
16.5	21
17	27
17.5	3
Total 19th	*28*
18.5	3
19	17
19.5	8
Total All Periods	*100*

TABLE 3: DISTRIBUTION OF SONNETS BY PERIOD IN THE MANUALLY ANNOTATED SONNETEVOL CORPUS. The 16th, 17th and 19th centuries cover ca. 30% of the corpus each, and the 15th century covers ca. 10% of the sonnets

*** The project site describes each enjambment type: http://sites.google.com/site/spanishenjambment/enjambment-types**

**** Exact dates of birth and death are available for a minority of authors; often only the century was provided in the corpus sources. Periods ending in ".5" cover sonnets for authors who lived in two centuries. E.g. period "15.5" covers sonnets for authors born in the 15th and deceased in the 16th century**

[10] In other words, if there is an enjambment between lines 1 and 2, we consider that as "pair of enjambed lines" in the reference corpus.
[11] Balancing across periods does not apply to the Cantos20th test-corpus: it covers the 20th century only.
[12] We did not define subtypes for cross-clause enjambment

5.2. Enjambment-detection tasks evaluated

We defined two enjambment-detection tasks:

1. **Span-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.
2. **Typed span-match:** for a correct result, both the positions and the enjambment type assigned by the system to those positions must match the reference.

5.3. System results and discussion

Precision, recall and F1 were obtained.[13] Table 4 provides overall results for both corpora. Table 5 provides the per-type results on the diachronic test-corpus (SonnetEvol). The project's site contains more detailed results.[14]

Corpus	Task	N	P	R	F1
SonnetEvol	*span-match*	260	74.18	87.64	80.35
	typed span-match		61.24	72.31	66.31
Cantos20th	*span-match*	277	84.01	89.17	86.51
	typed span-match		78.04	83.39	80.63

TABLE 4: OVERALL ENJAMBMENT DETECTION RESULTS. Number of test-items, Precision (P), Recall (R) and F1 in our two test-corpora, for the span-match and typed span-match enjambment detection tasks

Enjambment or Expansion Type *	N	P	R	F1
Phrase-Bounded (all types)	*104*	*66.19*	*88.46*	*75.72*
adj_adv	2	100	50.00	66.67
adj_noun	29	54.55	82.76	65.75
adj_prep	14	58.82	71.43	64.52
noun_prep	39	55.36	79.49	65.26
relword	1	100	100	100
verb_adv	5	50.00	100	66.67
verb_cprep	9	83.33	55.56	66.67
verb_chain	5	100	80.00	88.89
Cross-Clause[12]	*23*	*76.00*	*82.61*	*79.17*
Expansions (all types)	*133*	*61.54*	*66.17*	*63.77*
dobj_verb	65	60.00	69.23	64.29
subj_verb	68	63.24	63.24	63.24

TABLE 5: ENJAMBMENT DETECTION RESULTS PER TYPE. On the SonnetEvol corpus. Number of items per type, Precision (P), Recall (R) and F1 on the typed span-match enjambment detection task.
*** The types are described on our site: http://sites.google.com/site/spanishenjambment/enjambment-types**

[13] The definitions for Precision (P), Recall (R) and F1 were the usual:
$F1 = 2\frac{P \cdot R}{P + R}$; $P = \frac{\text{nbr. of correct outputs}}{\text{nbr. of system outputs}}$; $R = \frac{\text{nbr. of correct outputs}}{\text{nbr. of reference outputs}}$

[14] E.g. per-type results for the Cantos20th corpus, or breakdowns for SonnetEvol per period (overall and per type), see https://sites.google.com/site/spanishenjambment/evaluation

For untyped detection (span-match), F1 reaches 80% in the SonnetEvol corpus, whereas F1 for typed detection is 66.31%. For the contemporary Spanish corpus (Cantos20th), F1 is higher: 80.63% typed detection, 86.51% span-match. This reflects additional difficulties posed by archaic language and historical varieties for the NLP system whose outputs our enjambment detection relies on.[15]

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.[16]

Creating a reparsing module to manage hyperbaton and improve PP attachment results may be fruitful future work.

6. Scholarly results and discussion

The system's goal is detecting enjambment to help literary research on the phenomenon, via providing systematic evidence for its analysis.

We consider our untyped enjambed-line detection results helpful, given an F1 of ca. 80% 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 (Table 6 and 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).[17]

The value of the tool is helping perform such analyses on a large corpus. This opens the door for scholars to assess the literary relevance of the findings, and search for the best interpretation.

Enjambed line positions	Scholarly relevance	15th-17th cent.		19th cent.	
		Count	%	Count	%
4-5	across quatrains	2	0.07	19	0.26
8-9	across octave-sestet divide	2	0.07	12	0.16
11-12	across tercets	20	0.72	147	2.01

TABLE 6: PAIRS OF ENJAMBED LINES ACROSS VERSE BOUNDARIES IN THE 15TH-17TH VS. THE 19TH CENTURIES: Counts of enjambed line-pairs and percentages over the total number of enjambed line-pairs for each period. An example of the types of analyses stimulated by automatic enjambment detection

[15] Expansions get lower F1 than phrase-bounded types overall. But we do not think that the F1 difference between SonnetEvol and Cantos20th is due to the higher proportion of expansions in SonnetEvol (Table 2): Results per-type (see project's site[14]) show that phrase-bounded enjambment detection is 10 points of F1 lower in SonnetEvol than in Cantos20th. Also, phrase-bounded enjambment results for the 15th-17th period (with more archaic language) are 10 points of F1 lower than in the 19th century.

[16] This is a common problem in syntactic parsing, even for contemporary languages (see Agirre et al, 2008, for English). For historical varieties, Stein's (2016) results for verbal adjuncts and prepositional complements in Old French also suggest the difficulties posed by prepositional phrases.

[17] Given the manageable data volume, we validated the counts for enjambment across a verse boundary (Table 6) manually (but not the more voluminous data for all other positions).

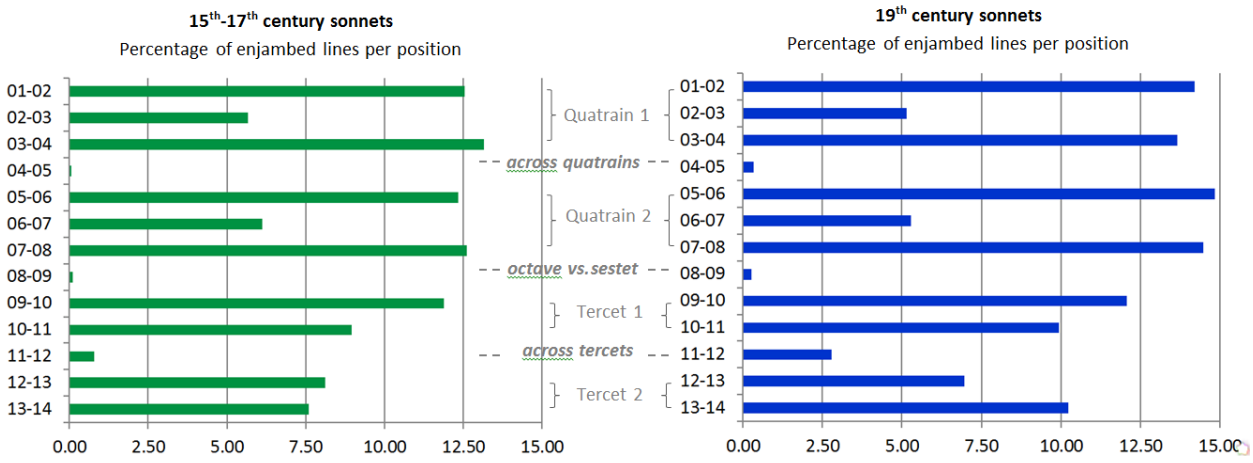

FIGURE 1: PERCENTAGE OF ENJAMBMENTS PER POSITION IN THE 15TH-17TH CENTURIES VS. THE 19TH.
The y-axis represents line-positions; the x-axis is the percentage of enjambed line-pairs for a position over all enjambed line-pairs in the period. Enjambment across quatrains and across the octave-sestet divide 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

7. Outlook

The characterization of enjambment in Spanish literary theory has unclear points. Systematically obtaining enjambment examples is helping us find additional evidence to analyze these unclear points. Moreover, 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 an extent is enjambment used differently in free verse vs. traditional versification?

Students in our metrics classes are currently annotating enjambment for 450 sonnets. These annotations will permit inter-annotator agreement computation. We will also examine the possibility of using supervised machine learning to train a sequence labeling and classification model to complement our current detection rules.

Acknowledgements
Pablo Ruiz Fabo was supported by a PhD scholarship from Région Île-de-France. We also thank Clara Martínez Cantón's and Borja Navarro's metrics students for their ongoing sonnet annotation work, and Borja Navarro for introducing his students to the annotations required.

Bibliography

Agerri, R., Bermudez, J. and Rigau, G. (2014). IXA Pipeline: Efficient and ready to use multilingual NLP tools. In *Proceedings of LREC 2014, the 9th International Language Resources and Evaluation Conference*. Reykjavik, Iceland.

Agirre, E., Baldwin, T. and Martinez, D. (2008). Improving Parsing and PP Attachment Performance with Sense Information. In *Proceedings of ACL 2008, Conference of the Association for Computational Linguistics*, 317-325. Columbus, Ohio, US.

Alarcos Llorach, E. (1966). *La Poesía de Blas de Otero [por] E. Alarcos Llorach*. Madrid, Anaya.

Artstein, R., and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics 34.4*: 555-596.

Colinas, A. (1983). *Noche más allá de la noche*. Madrid, Visor Libros.

Domínguez Caparrós, J. (2000). *Métrica española.* UNED, Spain.

García González, R. (ed.) (2006a). *Sonetos del siglo XV al XVII*. Alicante, Biblioteca Virtual Miguel de Cervantes. Retrieved from http://www.cervantesvirtual.com/nd/ark:/59851/bmc2r439

García González, R. (ed.) (2006b). *Sonetos del siglo XIX*. Alicante, Biblioteca Virtual Miguel de Cervantes. Retrieved from http://www.cervantesvirtual.com/nd/ark:/59851/bmc4q861

García-Page, M. (1991) En torno al encabalgamiento. Pausa virtual y duplicidad de lecturas. *Revista de literatura* 53.105: 595-618.

Flores Gómez, M. E. (1988). Coincidencia y distorsión (encabalgamiento) de la unidad rítmica verso y las unidades sintácticas. *Estudios clásicos*, *30*(94): 23-42.

Luján Atienza, Á. L. (2006). *Desde las márgenes de un río: la poesía coral de Diego Jesús Jiménez*. Córdoba, Litopress.

Martínez Cantón, C. (2011). *Métrica y poética de Antonio Colinas* (PhD Dissertation from UNED, Spain). Sevilla, Padilla Libros.

Martínez Fernández, J. E. (2010): *La voz entrecortada de los versos.* Barcelona, Davinci Continental.

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

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

Navarro-Colorado, Borja, Lafoz, M. R. and Sánchez, N. (2016). Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation. *Proceedings of LREC, Tenth International Conference on Language Resources and Evaluation:* 4630-4634. Portorož, Slovenia.

Paraíso, I. (2000). *La métrica española en su contexto románico.* Madrid, Arco Libros.

Parry, M. (1929). The distinctive character of enjambement in Homeric verse. In *Transactions and Proceedings of the American Philological Association* (60: 200-220). Johns Hopkins University Press, American Philological Association.

Quilis, A. (1964). *La estructura del encabalgamiento en la métrica española: Contribución a su estudio experimental.* Consejo Superior de Investigaciones Científicas.

Senabre, R. (1992). El encabalgamiento en la poesía de Fray Luis de León. *Revista de Filología Española, 62(1)*. Consejo Superior de Investigaciones Científicas.

Spang, K. (1983). *Ritmo y versificación. Teoría y práctica del análisis métrico.* Universidad de Murcia, Spain.

Stein, A. (2016). Old French dependency parsing: Results of two parsers analyzed from a linguistic point of view. In *Proceedings of LREC the 11th International Language Resources and Evaluation Conference*: 707-713. Portorož, Slovenia.