The Diachronic Spanish Sonnet Corpus: TEI and linked open data encoding, data distribution, and metrical findings

Pablo Ruiz Fabo
Université de Strasbourg, France

Helena Bermúdez Sabel
Université de Lausanne, Switzerland

Clara Martínez Cantón and Elena González-Blanco
UNED (Universidad Nacional de Educación a Distancia), Spain

Abstract

How has the sonnet form in Spanish evolved over the centuries? What is the distribution of metrical patterns and combinations thereof, considering diachronic, geographical, and social factors? What rhyme schemes are favoured in different periods and regions? How is enjambment distributed within the sonnet? Providing quantitative answers to such questions requires a corpus spanning several centuries, annotated for the relevant literary features and containing author metadata. The absence of appropriate digital resources to undertake a macroanalytic study of the evolution of the sonnet in Spanish led us to create the Diachronic Spanish Sonnet Corpus. This article presents how the corpus was designed for providing quantitative evidence on the evolution of sonnets in Spanish, and our findings regarding metrics and enjambment. The corpus contains 4,085 sonnets by 1,204 Spanish and Latin American authors (15th to 19th centuries), encoded in TEI, with RDFa attributes. The corpus aims at breadth, including many peripheral authors besides some major ones. Author metadata were encoded (dates, origin, gender). Scansion and enjambment were annotated automatically, with the ADSO and ANJA tools. The range of authors and periods, the use of TEI and RDFa for interoperability, and the combination of metrical and enjambment annotations goes beyond previously available digital resources. The corpus allowed us to examine the evolution of metrical patterns and their combinations after the Golden Age, complementing earlier studies. We also observed an increase in enjambment across the tercets in the 19th century, which may indicate increased variety in the discourse organization of sonnets in the period.

Correspondence:
Pablo Ruiz Fabo,
Université de Strasbourg,
LiLPa UR 1339, 61 av. des
Vosges, 67000 Strasbourg,
France.
E-mail:
ruizfabo@unistra.fr

doi:10.1093/llc/fqaa035 Advance Access published on 6 September 2020

1. Introduction

Digital resources in poetry in S1panish, freely available and open to the community, are scarce, particularly for certain periods. There are important projects and digital libraries such as Biblioteca Virtual Cervantes[1] or the corpora developed by Real Academia Española,[2] besides comprehensive Medieval archives like ADMYTE (Marcos Marn and Faulhaber, 1992) or BiDTEA (Gago Jover, 2015). These projects *include* texts in verse, but are not exclusively devoted to poetry. Moreover, while these resources provide efficient search engines to explore the texts, the datasets are not always exportable or encoded in formats that would ease their computational analysis.

Some digital editions of Spanish medieval poetry exist like the verse subcorpus 'Textos poéticos espaoles' in BidTEA, besides metrical repertoires containing partial editions like ReMetCa (González-Blanco and Rodríguez, 2014). For the Golden Age, Navarro-Colorado *et al.* (2016) presented the *Corpus of Spanish Golden-Age Sonnets*.[3] For later periods, we are not aware of Spanish poetry collections meant as a resource for literary scholarship.

This lack of open resources poses difficulties for Digital Humanities studies in Spanish (Agenjo, 2015). It hinders collecting quantitative evidence relevant for questions on the evolution of poetic forms, like the sonnet. In order to compensate for such deficiencies and help gain new knowledge on the sonnet in Spanish, we have developed Diachronic Spanish Sonnet Corpus (DISCO), a corpus of sonnets featuring content and data formats not available in existing resources.[4] Section 2 describes the corpus and how it was designed to address research questions on the evolution of the sonnet form. Section 3 discusses how the corpus contributes evidence towards such questions: in Spanish, like in other traditions, variation in rhyme schemes occurs in both quatrains and tercets. To what extent are the different rhyme schemes favoured in each period? What is the distribution of metrical patterns and combinations thereof, considering diachronic, diatopic, and social parameters? How is enjambment distributed within the sonnet? The DISCO corpus spans several centuries and countries, comprising 4,085 sonnets with prosodic annotations. It can provide quantitative data on the occurrence of specific formal features diachronically, towards a general description of the sonnet in Spanish, along the lines of Jockers' (2013) macroanalysis approach. Earlier literature on the evolution of the Spanish sonnet was based on smaller samples containing eminent examples from different periods (Quilis, 1975). A macroanalytic approach was pursued by Navarro-Colorado (2015, 2016), however, it was restricted to the Golden Age. Data from DISCO can thus reinforce or refine conclusions in earlier literature.

2. Corpus description

DISCO intends to increase the representativeness of electronic corpora for sonnets in Spanish, besides providing both descriptive and analytical annotations. To promote accessibility, it is available as a GitHub repository[5] and through Zenodo,[6] following good practices for data use, reuse, and conservation.

DISCO contains 4,085 sonnets in Spanish by 1,204 authors (15th to 19th century),[7] extracted from HTML sources at Biblioteca Virtual Cervantes (anthologies by García González, 2005, 2006a, 2006b) and Wikisource.[8]

The corpus includes both canonical and non-canonical poets, from different Spanish-speaking countries. Without intending to diminish the complexity of canonicity, we exploited sources that aim at breadth, favouring the inclusion of lesser known and peripheral authors. The majority of authors in our source for the Golden Age (García González, 2006a) are not covered in the corpus by Navarro-Colorado *et al.* (2016), which targets major Golden Age poets (Navarro-Colorado *et al.*, 2016, p. 4360). Only 17 of our 477 Golden Age authors are part of that corpus, with only 333 sonnets in common. In the 18th century, where less sonnets were produced, ca. 42% of our Spanish authors appear in a standard 18th-century anthology of Spanish poetry (Polt, 1979). Concerning the small number of 18th-century Latin American authors in DISCO, four out of those six authors are found in Latin American poetry anthologies (Caillet Bois, 1965; De Onís, 1956; Montes de Oca, 1968). Regarding the 19th century, our source anthology covers many minor authors. Around 18%

Table 1 Sonnet and author distribution per period, including the number of female and male authors, and the continent where they developed their literary activity

Period	No. of sonnets	Nbr of authors					Sources
		By gender		By continent		Total	
Golden Age (15th–17th)	1,088	Female	31	America	13	477	García (2006a)
		Male	446	Europe	464		
18th century	321	Female	1	America	6	42	García (2005) Wikisource
		Male	41	Europe	36		
19th century	2,676	Female	48	America	334	685	García (2006b)
		Male	637	Europe	351		

of DISCO's 19th century authors (120 authors) appear in anthologies for Spanish or Latin American poetry.[9]

A synopsis of the dataset is shown in Table 1. Regarding our data distribution, our low coverage for Latin American authors is due to a skew towards European authors in our source, which we aim at correcting in the future. In the 18th century, the small number of sonnets results from lack of interest in this form by the period's authors (Quilis, 1975, p. 135). Our 19th century source (García González, 2006b) is balanced geographically. However, in spite of our efforts to create a broad corpus, the representation of female authors could be improved (see Section 4 for intended means to counter this).

2.1 Encoding paradigms: TEI and linked open data

Two versions of the texts are provided: one collecting every sonnet per author, the other with a single sonnet per file.[10] The poems are encoded in XML-TEI P5. A plain-text version is also provided. For corpus preparation, we closely followed the TEI guidelines and RIDE's criteria for Digital Text Collections (Henny-Krahmer and Neuber, 2017).

Besides the TEI-semantics, this corpus provides a layer of Linked Open Data (LOD) expressed in RDFa (Herman *et al.*, 2015).[11] To our knowledge, no out-of-the-box tools exist for publishing literary TEI corpora as LOD.[12] Enriching TEI with RDFa attributes is thus a solid approach in order to translate TEI semantics to the web—see precedents like Jewell (2010) and Tittel *et al.* (2018)—and benefit from the possibilities offered by the Semantic Web: First, we enrich our dataset by linking to third-party ones (as DBpedia), providing additional resources to complement the corpus. Secondly, we enhance the publication of our data openly using standard schemas, supplying semantic interoperability that allows third-party applications to automatically use our data. We consider creating corpora of these characteristics as a first step to overcome the problem of fragmentary access to literary resources, and thus a contribution to literary scholarship.

2.2 Author metadata

DISCO has collected metadata about the texts and authors and made it explicit and accessible, using TEI and LOD. The HTML sources contained (sometimes incomplete) information relevant for identifying author metadata, within a non-standardized text-string, e.g. birthplace, date of birth and death, profession. The project collected these data and labeled them explicitly. Missing information was inferred programmatically or actively researched manually.

As a result of this process of metadata curation, the author's lifetime is expressed through the years of birth and death (or centuries of activity in the case of early lesser known authors, where such data are unavailable). Places of birth and death are included whenever possible, besides author gender.

Author origin is relevant to determine linguistic usage or literary trends according to national traditions. We tried to be detailed when defining origins, providing the settlement and country (if possible) and continent (obligatorily). We tried to establish the place where the author actually developed their literary activity. Thus, in some cases, we added the element **<floruit>** to better define the geographic identification of a given author.[13]

We augmented the TEI annotation with URIs and both class and property information, in RDFa.[14] The relation between authors and their works is defined using the DCMI Metadata Terms (DCMI Usage Board, 2012). Most features regarding poets' biographical data were formalized with the FOAF vocabulary (Brickley and Miller, 2014). Besides implementing different RDF vocabularies to transmit the semantics of the TEI annotation, we supplied links to other resources, enriching the corpus and increasing its (inter)operability. Authors were assigned VIAF identifiers, by querying VIAF's API, supplemented with manual validation. Furthermore, besides linking to VIAF and, in a small number of cases, to the Wikidata datasets[15] for author identification, we added links to the esDBpedia dataset,[16] in order to refer to some metrical concepts available in this resource.

2.3 Metrical encoding and enjambment

Using the @**met** attribute, each line was annotated for scansion with the ADSO tool[17] (Navarro-Colorado, 2017), which specializes in Spanish fixed-meter forms, attaining a per-line accuracy of 95%; the error rate is similar to human performance (Navarro-Colorado, 2017, p. 123).

We also performed a simple analysis of the rhyme scheme in the quatrains. This is generally ABBA (enclosed) in Spanish sonnets. However, alternate rhyme (ABAB) sometimes occurs. The rhyme scheme was detected using a heuristic, and encoded in the @**type** attribute of **<lg>** elements. The Spanish terminology was used (i.e. *cuarteto* for ABBA quatrains and *serventesio* for ABAB quatrains).

Enjambment, an important device relevant for metrics and style, was equally encoded. Using the @**enjamb** attribute, lines were annotated for enjambment with the ANJA tool.[18] The tool was evaluated in Ruiz Fabo *et al.* (2017), using manually annotated samples from DISCO besides non-sonnet materials. The tool's performance at detecting enjambment is above 0.8 F1,[19] and its efficacy at classifying enjambment types varies across periods and types. Based on this evaluation, a @**cert** attribute specifies the expected certitude for each enjambment type annotated in the corpus.[20]

3. Contributions to Literary Scholarship

Thanks to this work, we provide both primary sources and analytical annotations for digital scholarship on Spanish poetry. Besides, the corpus allows us to quantify the occurrence of metrical features and enjambment, complementing what is known about the sonnet form.

3.1 New annotated data

The metadata mentioned in Section 2.2 were unavailable in structured, machine-readable format in the corpus sources, or in other sonnet collections, like *Sonnett-Archiv* (Elf Edition). Regarding coverage, the corpus complements the *Corpus of Spanish Golden-Age Sonnets* by Navarro-Colorado *et al.* (2016), by including minor Golden Age authors. Only 8.15% of DISCO's sonnets are shared across both corpora (333 sonnets). For later periods, we cover more poems and authors than existing digital corpora, up to the 19th century. Finally, combining the annotation of metrical patterns, stanza types, and enjambment is not offered by prior corpora. The metadata collected for DISCO allow us to create subcorpora based on author period, origin, and gender, enabling comparative analyses across author groups.

3.2 Metrical findings

Descriptive statistics on the corpus reinforce generalizations drawn in earlier studies for Golden Age sonnets.[21] Additionally, thanks to the diachronic nature of our corpus, we can determine which generalizations still apply to later periods or whether new trends emerge.

Corpus data on stress patterns (Table 2 and Fig. 1) agree with existing descriptions of the Spanish hendecasyllable based on both large-sample (Navarro-Colorado, 2016, for the Golden Age) and small-sample analyses (Henríquez Ureña, 1919; Domínguez Caparrós, 2014): *a maiori* patterns (with 6th-syllable stress) predominate, and *a minori* patterns (with stresses on the 4th and 8th syllables) follow. However, our data show an increase of *a minori* patterns in the 19th century, which might suggest an interest in metrical variety in that period.[22]

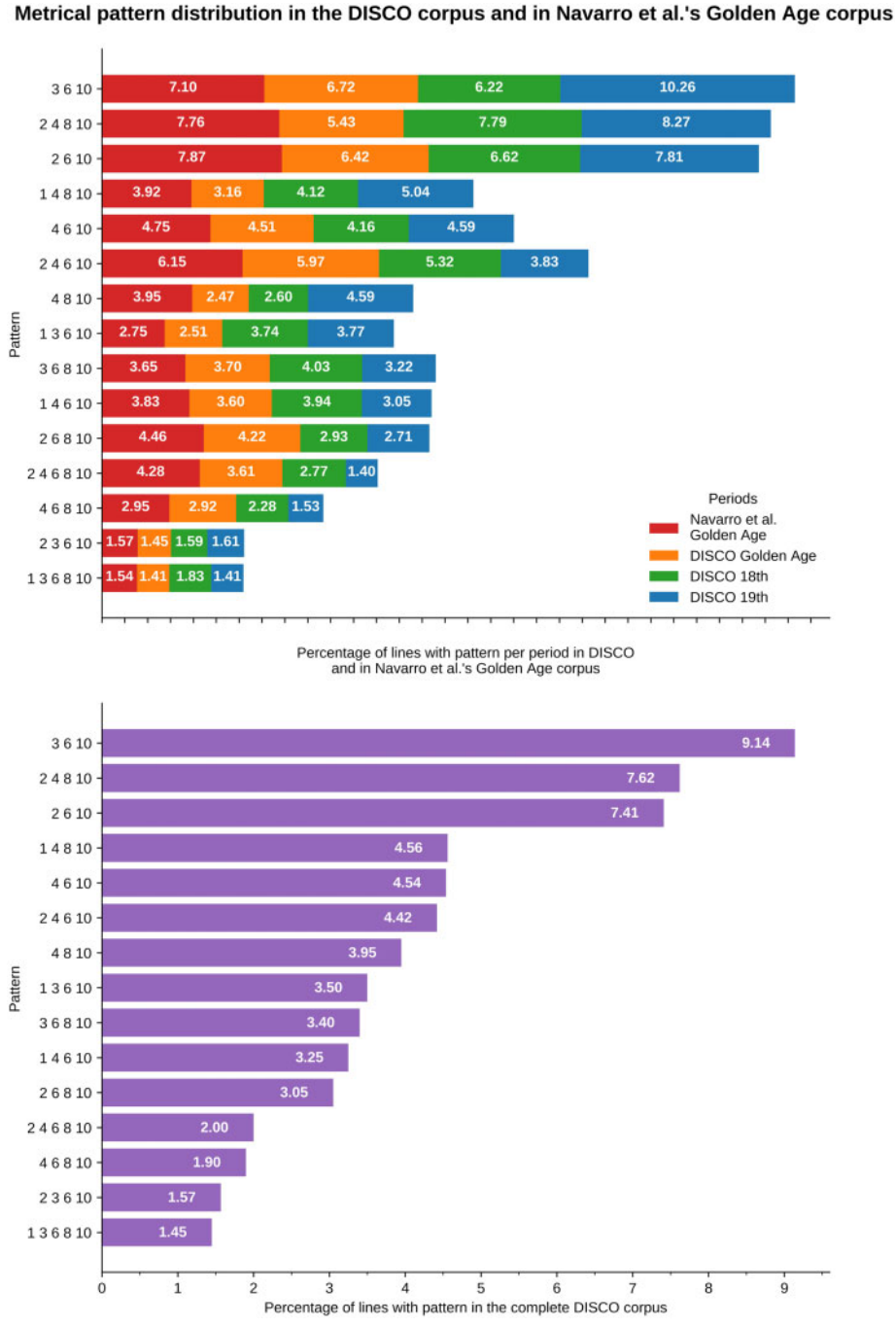

Fig. 1 Metrical pattern distribution in hendecasyllable sonnets. On the left, percentage of lines for the 15 most frequent patterns in each period of DISCO and in Navarro *et al.*'s (2016) corpus. On the right, percentage of lines with those patterns in the complete DISCO corpus (all periods). Data are sorted by frequency in the complete DISCO corpus, in decreasing order

Table 2 Distribution of metrical patterns per period in hendecasyllable sonnets (percentage of lines for each pattern) for the 15 most frequent patterns in the DISCO corpus, besides in Navarro *et al.*'s (2016) corpus, sorted by decreasing percentage of occurrence in the complete DISCO corpus. Pattern classes are also provided (*mai*: *a maiori*, i.e. stress on 6th syllable, *min*: *a minori*, i.e. stress on 4th and 8th syllable). Rows for *a minori* patterns are in italics. *Stress count* refers to the number of stresses in the pattern. Patterns with three stresses are widely used in any period. Most *a maiori* patterns with four stresses decrease in the 19th century, whereas *a minori* patterns increase in that century

Pattern	Pattern Class	Stress Count	DISCO				Navarro *et al.*
			Percentage of lines				
			All periods	19th	18th	Golden Age	Golden Age
3 6 10	mai	3	9.14	10.26	6.22	6.72	7.10
2 4 8 10	*min*	*4*	*7.62*	*8.27*	*7.79*	*5.43*	*7.76*
2 6 10	mai	3	7.41	7.81	6.62	6.42	7.87
1 4 8 10	*min*	*4*	*4.56*	*5.04*	*4.12*	*3.16*	*3.92*
4 6 10	mai	3	4.54	4.59	4.16	4.51	4.75
2 4 6 10	mai	4	4.42	3.83	5.32	5.97	6.15
4 8 10	*min*	*3*	*3.95*	*4.59*	*2.6*	*2.47*	*3.95*
1 3 6 10	mai	4	3.50	3.77	3.74	2.51	2.75
3 6 8 10	mai	4	3.40	3.22	4.03	3.70	3.65
1 4 6 10	mai	4	3.25	3.05	3.94	3.60	3.83
2 6 8 10	mai	4	3.05	2.71	2.93	4.22	4.46
2 4 6 8 10	mai	5	2.00	1.40	2.77	3.61	4.28
4 6 8 10	mai	4	1.90	1.53	2.28	2.92	2.95
2 3 6 10	mai	4	1.57	1.61	1.59	1.45	1.57
1 3 6 8 10	mai	5	1.45	1.41	1.83	1.41	1.54

Table 3 Count of hendecasyllable versus alexandrine sonnets according to the authors' continent of production, in the 19th century (alexandrine sonnets are very rare before). The type of rhyme scheme in the quatrains (enclosed or alternate) is also specified. The alexandrine sonnet is preferentially used by American authors, and a large proportion of alexandrine sonnets show alternate rhyme

Meter	Quatrain Rhyme	Sonnet Count		
		Total	American	European
Hendecasyllable	Enclosed	2,287	1,223	1,064
	Alternate	122	96	26
	Total	**2,409**	**1,319**	**1,090**
Alexandrine	Enclosed	145	121	24
	Alternate	122	98	24
	Total	**267**	**219**	**48**

Regarding diachronic data on the number of stressed positions (Table 2), patterns with three stresses are highly used across periods. However, most *a maiori* patterns with four stresses decrease in the 19th century. This might indicate a 19th-century preference for 'lighter' patterns, with stresses further apart from each other.

Whereas the predominant meter for sonnets is naturally the hendecasyllable, alexandrines are attested, mostly in the 19th century, preferentially used by American authors.[23] A large proportion of alexandrine sonnets use an alternate rhyme scheme (ABAB) instead of the usual enclosed scheme (ABBA); see Table 3. These results support findings by earlier scholars (Quilis, 1975, pp. 135–7), showing that they generalize to a larger corpus and providing quantitative data on the trends.

DISCO sonnets after the Golden Age allow us to complement findings from Navarro-Colorado (2016) regarding metrical pattern combinations. His study (p. 101) notes a striking proportion of ternary-rhythm lines (stressed on the 3rd and 6th syllables) followed by binary-rhythm lines (stressed on the 2nd and 6th or 4th and 8th syllables). The question is posed whether this is a distinctive trait of Golden Age sonnets, but no answer can be provided given lack of a comparison group with later sonnets.

Table 4 Ternary (*ter*) and binary (*bin*) rhythm combinations in two-line sequences where the first line has a ternary rhythm. *N* represents the count for each sequence type and % its percentage. *Q* is the ratio between ternary–binary and ternary–ternary line-counts. Lines of type *other* fall into neither category. Data restricted to hendecasyllable sonnets

	Navarro *et al.* Golden Age			DISCO Golden Age			DISCO 18th			DISCO 19th		
Sequence	***N***	**%**	***Q***	***N***	**%**	***Q***	***N***	**%**	***Q***	***N***	**%**	***Q***
ter–bin	7,262	11.07	*2.42*	908	9.98	*1.93*	487	11.76	*1.89*	3,426	11.84	*1.46*
ter–ter	2,996	4.57		471	5.18		242	6.23		2,349	8.11	
ter–other	3,009	4.59		533	5.86		217	5.58		1,746	6.03	

DISCO allows us to compare. Table 4 displays per-period counts for two-line sequences where the first line has a ternary rhythm, showing whether the second line follows a ternary, binary, or other pattern. Type *other* refers to lines that do not conform to the requirements cited above from Navarro-Colorado (2016) for a line to count as binary or ternary.[24] Column *N* shows counts, % is the percentage for each category and *Q* is the ratio of ternary–binary sequences over ternary–ternary. The counts show that the ratio of ternary-rhythm lines followed by binary lines has a higher value in Navarro-Colorado *et al.*'s (2016) Golden Age corpus (covering major authors) than in DISCO's Golden Age subcorpus (covering lesser-known authors), where the ratio is close to the value in DISCO 18th century. The ratio does not decrease clearly until the 19th century. Our corpus suggests that the feature may distinguish pre-19th century sonnets from earlier ones.

3.3 Enjambment

The discourse structure or rhetorical organization of sonnets tends to be tied with stanzas. For instance, the sestet can represent a contrast or a way to provide closure to points laid out in the octave. Different discourse relations are possible between both quatrains within the octave and between both tercets within the sestet. Spiller (1992, pp. 1–63) discusses several patterns in the light of individual poems from the Italian and English tradition. Enjambment at a stanza boundary would allow discourse units, such as a statement, a contrast, a concession or a conclusion, to spread over more than one stanza. Sonnets where this occurs are also discussed by Spiller (1992, pp. 49, 190), who shows examples from Petrarch.

We focused on cases of enjambment across stanza boundaries in the different periods of our diachronic corpus, besides Navarro's Golden Age corpus, given the interest of such cases for analyzing discourse relations in the sonnet.[25] Only 14-line sonnets were considered; around 70 sonnets with extra lines and were excluded, e.g. sonnets with an extra final stanza called *estrambote*, sometimes found in the Golden Age. The trend that emerges in all periods is that enjambment is virtually absent across stanza boundaries, i.e. between lines 04–05, 08–09, and 11–12 (cf. Fig. 2 and Table 5). However, in our 19th century corpus, enjambment across the tercets clearly increases. An examination of the poems involved shows that these enjambments generally result in the entire sestet being used to develop the point that brings closure to the sonnet, as opposed to other sonnets where each tercet more independently develops a single point within the sonnet's discourse structure. Additionally, in about 10% of cases, an enjambment between lines 11 and 12 actually precedes a final couplet rather than representing an enjambment across tercets. The coexistence of these different patterns in 19th-century sestets, aided by enjambment, may indicate more variety in discourse organization of the sonnet form in this period than in earlier periods, where enjambment across stanza boundaries was virtually absent.

If we look at trends within stanza boundaries rather than across them, a general trend visible in the corpus is that enjambment starting at an odd position within the stanza (first or third line in quatrains, first line in tercets) is more frequent than enjambment starting at an even position (second line). This trend can however be reversed in the final tercet, where an enjambment between both closing lines (second and third in the tercet) can be more frequent than between the tercet's first and second line.

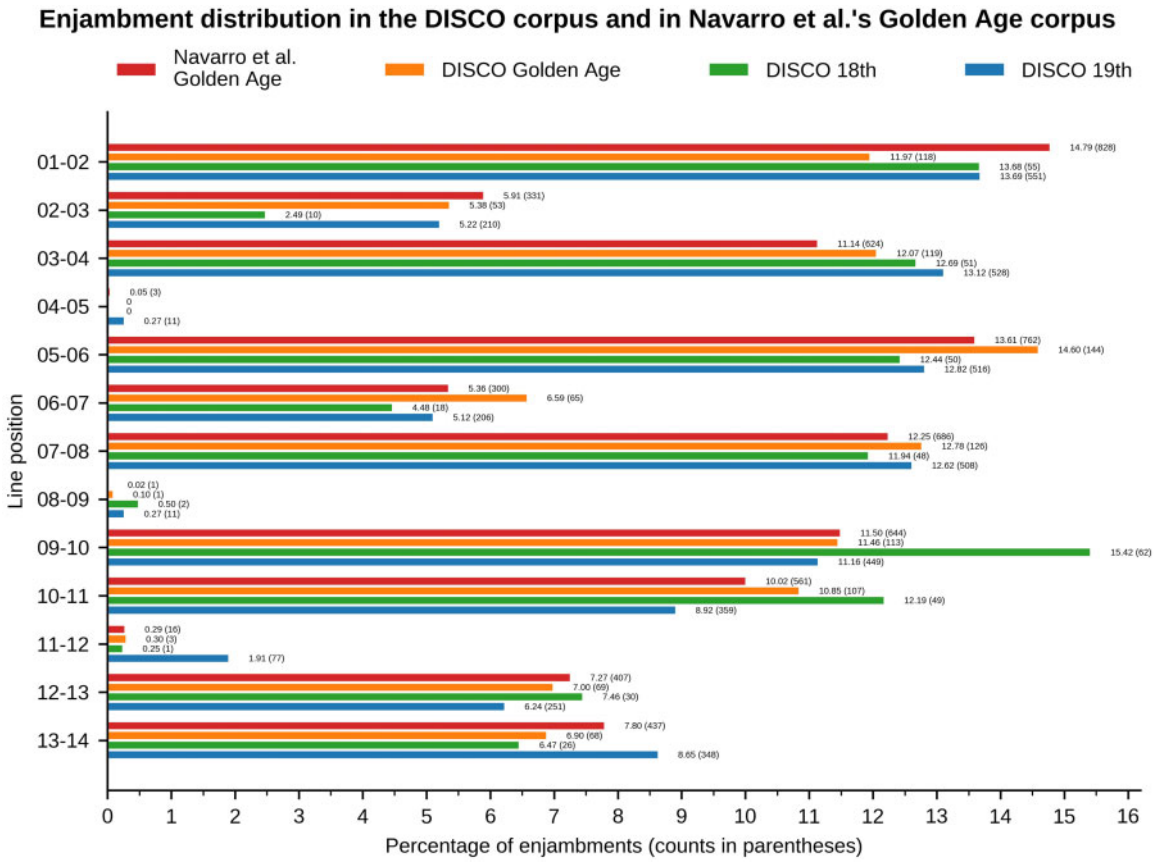

Fig. 2 Percentage of enjambments occurring at each line-pair in each period of DISCO and in Navarro *et al.*'s (2016) Golden Age corpus. Bars are labeled with the percentage and, in parentheses, the count. Enjambment across stanza boundaries is virtually absent, with the exception of 19th century tercets (between lines 11 and 12). Enjambment starting at an odd position (first or third line in a stanza) is more frequent than enjambment starting at even positions (second or fourth lines); this trend can however be reversed in the final tercet

Table 5 Enjambment across stanza boundaries (counts and percentages) in DISCO's different periods besides in Navarro *et al.*'s (2016) Golden Age corpus. The annotations were manually checked. Enjambment across stanzas is very rare in most periods, attested in <0.3% of sonnets (<20 sonnets). However, in the 19th century, there is an increase in enjambment across the tercets, observed in 1.91% of sonnets (77 sonnets). This may indicate more variety in discourse organization in the sonnet-form in the 19th century compared to earlier periods. Only 14-line sonnets were analyzed (e.g. without *estrambote*)

	Navarro *et al.* Golden Age		**DISCO Golden Age**		**DISCO 18th**		**DISCO 19th**	
Line Pair	**Count**	%	**Count**	%	**Count**	%	**Count**	%
04–05	3	0.05	0	0	0	0	11	0.27
08–09	1	0.02	1	0.1	2	0.5	11	0.27
11–12	16	0.29	3	0.3	1	0.25	77	1.91

4. Conclusions and Future Work

Corpora providing a wide coverage of periods and authors, annotated with the relevant metadata, are crucial to answer questions in literary history. The DISCO corpus is helping address research questions that would otherwise not be approachable given lack of appropriate data, such as large-sample quantitative analyses of the metrical evolution of the sonnet form. We reported findings regarding the evolution of metrics and enjambment. The distribution of metrical patterns observed in DISCO confirms, based on a large corpus, generalizations from earlier literature based on smaller samples. In earlier studies, large-sample metrical analyses were available for the Golden Age sonnet. Our research complements these, providing large-sample analyses until the 19th century, for both metrics and enjambment. Regarding enjambment, we focused on its distribution at stanza boundaries, finding that enjambment across the tercets increases in the 19th century. This may reflect a variety in the way discourse is organized within the sestet in the period: in a larger proportion of sonnets than in earlier periods, the entire sestet, rather than each isolated tercet, is used to develop the same point.

Our future work will include curating author metadata. At present, attributes encode the level of certainty regarding the identification of lesser-known poets. Nevertheless, an individual study of uncertain cases will increase metadata accuracy. Further analysis of non-canonical authors and sources may point us to authors not yet included, towards enlarging the corpus. We will pay special attention to including more female poets and increasing the representativity of Latin American writers.

We are also working on a web-based interface, to enhance the corpus usability and dissemination. We are especially interested in the use of digital tools in an educational environment. A user-friendly interface that allows the exploration of certain prosodic elements can become a great educational resource, not only to provide students with multiple examples of sonnets, but also to let them explore formal characteristics of literary devices (e.g. enjambment) and practice their identification aided by an annotated reference set.[26]

Funding

This work was supported by the European Research Council [ERC-2015-STG-679528], POSTDATA project, Poetry Standardization and Linked Open Data.

References

Agenjo, X. (2015). Las bibliotecas virtuales españolas y el tratamiento textual de los recursos bibliográficos. *Ínsula: Revista de Letras y ciencias Humanas*, **822**: 12–15.

Aguirre, R. G. (1979). *Antología de la Poesía Argentina. Tomo I, 1604 - 1918*. Buenos Aires: Fausto.

Bermúdez-Sabel, H., Curado Malta, M., and González-Blanco, E. (2017). Towards interoperability in the European poetry community: The standardization of philological concepts. In Gracia Jorge et al. (eds), *Proceedings of Language, Data, and Knowledge: First International Conference (LDK 2017)*. Cham: Springer, pp. 156–65.

Biblioteca Virtual Miguel de Cervantes. (2007). *Biblioteca del Soneto [Sonnet Library]*. http://www.cervantesvirtual.com/bib/portal/bibliotecasoneto/(accessed 22 September 2018).

Brickley, D. and Miller, L. (2014). FOAF Vocabulary Specification. http://xmlns.com/foaf/spec/(accessed 17 September 2018).

Caillet Bois, J. (1965). *Antología de la Poesía Hispanoamericana*. Madrid: Aguilar.

Chiarcos, C. and Erjavec, T. (2011). *Owl/dl Formalization of the Multext-east Morphosyntactic Specifications. Proceedings of the 5th Linguistic Annotation Workshop, LAW V* '11. Stroudsburg: PA, USA, pp. 11–20.

Cimiano P., Chiarcos C., McCrae J. P. and Gracia J. (2020). *Linguistic Linked Data in Digital Humanities. Linguistic Linked Data*. Cham: Springer.

Ciotti, F., Peroni, S., Tomasi, F., and Vitali, F. (2016). *An OWL 2 Formal Ontology for the Text Encoding Initiative. Digital Humanities 2016: Conference Abstracts*, pp. 151–3.

Dauster, F. (1970). *Antología de la Poesía Mejicana*. Zaragoza - Madrid: Ebro.

DCMI Usage Board (2012). DCMI Metadata Terms. http://dublincore.org/documents/dcmi-terms (accessed 22 September 2018).

Delente, É. (2019). Le traitement automatique du rythme régulier. Un cas particulier: L'enjambement/To what an extent is automated processing of enjambment in French

versified discourse possible? Presented à *Plotting Poetry (and Poetics) 3 - Machiner la poésie (et la poétique) 3*. 26 September, 2019, Nancy, France. https://plottingpoetry.wordpress.com/slides

De Onís, F. (1956). *Anthologie de la Poésie ibéro-américaine. Collection Unesco D'œuvres Représentatives*. Série Ibéro-Américaine, 9. Paris: Nagel.

Dés, M. (1993). *Noche insular. Antología de la Poesía Cubana*. Barcelona: Lumen.

Domínguez Caparrós, J. (2009). *El Moderno Endecasílabo Dactílico, Anapéstico o de Gaita Gallega*. Sevilla: Padilla Libros.

Domínguez Caparrós, J. (2014). *Métrica Española*. Madrid: UNED.

Elf Edition: *Sonett-Archiv*. http://sonett-archiv.com (accessed 22 September 2018).

Fernández, C. and Becco, J. (1968). *Antología lineal de la poesía argentina*. Biblioteca Románica Hispánica. Antología Hispánica 26. Madrid: Gredos.

Ferro, H. (1965). *Antología Comentada de la Poesía Hispanoamericana*. New York: Las Américas.

Gago Jover, F. (2015). La biblioteca digital de textos del español antiguo (BiDTEA). *Scriptum Digital* **4**: 5–36.

García González, R. (2005). *Sonetos del siglo XVIII*. Biblioteca Virtual Miguel de Cervantes. http://www.cervantesvirtual.com/obra-visor/sonetos-del-siglo-xviii–0/html/ (accessed 22 September 2018).

García González, R. (2006a). *Sonetos del siglo XV al XVII*. Biblioteca Virtual Miguel de Cervantes. http://www.cervantesvirtual.com/obra-visor/sonetos-del-siglo-xv-al-xvii–0/html/ (accessed 22 September 2018).

García González, R. (2006b). *Sonetos del siglo XIX*. Biblioteca Virtual Miguel de Cervantes. http://www.cervantesvirtual.com/obra-visor/sonetos-del-siglo-xix–0/html/ (accessed 22 September 2018).

González-Blanco, E. and Rodríguez, J. L. (2014). ReMetCa: A Proposal for Integrating RDBMS and TEI-Verse. *Journal of the Text Encoding Initiative*, **8**. https://jtei.revues.org/1274 (accessed 22 September 2017).

Henny-Krahmer, U. and Neuber, F. (2017). Criteria for reviewing digital text collections, version 1.0. *A Review Journal for Digital Editions and Resources*, **6**. https://www.i-d-e.de/publikationen/weitereschriften/criteria-text-collections-version-1-0/ (accessed 22 September 2018).

Henríquez Ureña, P. (1919). El endecasílabo castellano. *Revista de Filología Española*, **6**: 132–57.

Herman, I., Asida, B., McCarron, S., and Birbeck, M. (2015). RDFa Core 1.1 - Third Edition, https://www.w3.org/TR/rdfa-core (accessed 22 September 2018).

Jewell, M. O. (2010). Semantic screenplays: Preparing TEI for Linked Data. In *Digital Humanities 2010*. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-878.html (accessed 22 September 2018).

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

Marcos Marín, F. and Faulhaber, C. B. coordinators. (1992). *ADMYTE. Archivo Digital de Manuscritos y Textos Españoles*. http://www.admyte.com/admyteonline/contenido.htm (accessed *22* September 2018).

Martínez Cantón, C. and Ruiz Fabo, P. (2019). La evaluación de herramientas como modo de aprendizaje e introducción de las Humanidades Digitales en el aula universitaria. La experiencia docente "Poesía distante". *Didáctica (Lengua y Literatura)*, **31**: 171–90. https://hal.archives-ouvertes.fr/hal-02421983v2 (accessed 24 May 2020)

Martínez Cantón, C., Ruiz Fabo, P., and Navarro-Colorado, B. (2017). A caballo entre el verso y las humanidades digitales. La evaluación de herramientas como modo de aprendizaje en el aula. In *III Congreso Internacional de Humanidades Digitales Hispánicas*. Málaga, Spain. October 2017. http://e-spacio.uned.es/fez/eserv/bibliuned:352-Cmartinez3/Martinez_Clara_Acaballo.pdf (accessed 22 September 2018).

McCrae, J., Spohr, D., and Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. *Proceedings of the 8th Extended Semantic Web Conference on The Semantic Web: Research and Applications, V (Part I)*: pp. 245–59, Berlin: Springer.

Montes de Oca, F. (1968). *Poesía Mexicana*. Mexico City: Porrua.

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

Navarro-Colorado, B. (2016). Hacia un análisis distante del endecasílabo áureo : Patrones métricos, frecuencias y evolución histórica. *Rhythmica: Revista Española de métrica Comparada*, **14**: 89–118.

Navarro-Colorado, B. (2017). A metrical scansion system for fixed-metre Spanish poetry. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqx009 (accessed 22 September 2018).

Navarro-Colorado, B., Ribes Lafoz, M. and Sánchez, N. (2016). *Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation. Proceedings of the Language Resources and Evaluation Conference*. http://www.lrec-conf.org/proceedings/

lrec2016/pdf/453_Paper.pdf (accessed 22 September 2018).

Navas Ruiz, R. (2010). *El Siglo XIX*. Colección Poesía Española, Antología crítica dirigida por Francisco Rico. Madrid: Visor.

Polt, J. H. R. (1979). *Poesía del Siglo XVIII*. Madrid: Castalia.

Quilis, A. (1964). *Estructura del encabalgamiento en la métrica española*. Consejo Superior de Investigaciones Científicas, Patronato Menéndez y Pelayo, Instituto Miguel de Cervantes.

Quilis, A. (1975). *Métrica Española*. Madrid: Ediciones Alcalá.

Rodríguez, M. (2011). *Otra Cuba Secreta. Antología de poetas cubanas del XIX y del XX*. Madrid: Verbum.

Ruiz Fabo, P., Martínez Cantón, C., Poibeau, T., and González-Blanco, E. (2017). *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*. Vancouver, Canada.

Schulman, I. A. and Picon Garfield, E. (1986). *Poesía Modernista Hispanoamericana y Española*. Madrid: Taurus.

Serna, M. and Castany, B. (2008). *Antología Crítica de Poesía Modernista Hispanoamericana*. Madrid: Alianza.

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

Spiller, M. R. (1992). *The Development of the Sonnet: An Introduction*. London: Routledge.

Tittel, S., Bermúdez-Sabel, H., and Chiarcos, C. (2018). Using RDFa to Link Text and Dictionary Data for Medieval French. In McCrae, J. et al. (eds), *Proceedings of the 6th Workshop on Linked Data in Linguistics (LDL-2016): Towards Linguistic Data Science. European Language Resources Association* (ELRA), Paris, France, Miyazaki, Japan. ISBN 979-10-95546-19-1 http://lrec-conf.org/workshops/lrec2018/W23/pdf/10_W23.pdf (accessed 22 September 2018).

Wikisource. *Categoría: Sonetos*. https://es.wikisource.org/w/index.php?title=Categor%C3%ADa:Sonetos (accessed 15 October 2018).

Yáñez, M. (1988). *Antología del Soneto Hispanoamericano*. Havana: Editorial Arte y Literatura.

Notes

1 Biblioteca Virtual Miguel de Cervantes, http://www.cervantesvirtual.com/ (accessed 22 September 2017).

2 Corpus Diacrónico del Español (CORDE) (http://corpus.rae.es/cordenet.html), Corpus de Referencia del Español Actual (CREA) (http://corpus.rae.es/creanet.html), and Corpus del Español del Siglo XXI (CORPES XXI) (http://web.frl.es/CORPES) (accessed 10 October 2018).

3 Corpus at https://github.com/bncolorado/CorpusSonetosSigloDeOro.

4 The authors would like to credit José Calvo Tello as one of the main contributors of the first version of the corpus presented in this article.

5 https://github.com/pruizf/disco/tree/v2.1 (accessed 24 May 2020).

6 https://doi.org/10.5281/zenodo.1012567 (accessed 10 October 2018).

7 About 125 sonnets by approx. 20 authors whose production took place in the early 20th century (with date of death prior to 1936) are also included in the corpus; the documentation on the GitHub repository provides more details: https://github.com/pruizf/disco/tree/v2.1#data-distribution (accessed 24 May 2020).

8 Sonnet category in Spanish Wikisource: https://es.wikisource.org/wiki/Categor%C3%ADa:Sonetos

9 We have consulted diachronic anthologies (De Onís, 1956; Ferro, 1965; Caillet Bois, 1965), period-specific ones (Schulman and Picon Garfield, 1986; Serna and Castany, 2008; Navas Ruiz, 2010), specific to the sonnet (Yáñez, 1988) and country-specific ones. Authors from Spain are already covered in the works just cited. We also reviewed country-specific anthologies for the three Latin American countries with more than 25 authors in DISCO: Cuba (Dés, 1993; Rodríguez, 2011), Mexico (Montes de Oca, 1968; Dauster, 1970), and Argentina (Fernández and Becco, 1968; Aguirre, 1979).

10 The corpus sources group the sonnets by author. Author order in the sources is alphabetical and can be recovered from author names. The order of sonnets by each author in the corpus sources can be recovered from their order in our per-author files or following our numeric poem IDs sequentially.

11 RDFa is a way of encoding RDF-style relationships (that is, a subject–predicate–object model) using simple attributes in existing markup languages, in this case, XML-TEI.

12 Whereas the publication of literary corpora in Linked Open Data formats is not widespread, inspiration could be drawn from the linguistics community, which has been especially successful in building the means to convert resources with linguistic annotations to the

Resource Description Framework model (see Chiarcos and Erjavec, 2011; McCrae *et al.*, 2011). For an overview of LOD resources in the Digital Humanities framework community and a description of strategies for their integration with the TEI see Cimiano *et al.* (2020). In addition, more general projects, not limited to linguistic analysis, are being developed as well: see work on building a TEI ontology in Ciotti *et al.* (2016).

13 Although 'floruit' is a convention most commonly used to describe the peak of activity of a person (or group of people), we have broadened its use here. On the one hand, we use it as it was intended for to specify periods of activity. On the other hand, we reinterpret the original meaning of 's/he flourished' to geographically ascribe the production of authors for which birth place does not correspond to the place where they 'flourished'. For instance, we find authors who were born in Spain but emigrated to South America in their childhood, entering, years later, the literary circles of their host country. In addition, it is useful to annotate authors for which we do not have any information regarding their origin (or death place), but for whom we know in which city they were publishing.

14 For more details about the encoding of the corpus, please see the documentation available on https://github.com/pruizf/disco/tree/v2.1. Details on the RDFa attributes are at https://github.com/pruizf/disco/tree/v2.1#rdfa-attributes (accessed 24 May 2020).

15 https://www.wikidata.org (accessed 10 October 2018).

16 http://es.dbpedia.org/ (accessed 10 October 2018). In a future version, our corpus will be fully compliant with the POSTDATA ontology, which will provide means to publish European poetry materials as Linked Open Data. See Bermúdez-Sabel et al. (2017) for a description of methodology and goals of this project. A draft version of the POSTDATA network of ontologies is available at http://postdata-prototype.linhd.uned.es/ontology.php (accessed 11 November 2019). *POSTDATA* stands for *Poetry Standardization and Linked Open Data.*

17 https://github.com/bncolorado/adsoScansionSystem (accessed 10 October 2018).

18 The tool is available at https://bitbucket.org/pruizf/anja_public/src, with a demo at http://prf1.org/anja/index/. It detects different types of enjambment (i.e. a mismatch between syntactic and metrical structure) as characterized by Quilis (1964). The tool also detects the concept of *enlace* (Spang, 1983), which takes place when a subject or direct object occur in a line adjacent to their governing verb's line, and which triggers a less noticeable effect than the enjambment types defined by Quilis.

19 F1 is the harmonic mean of precision (P) and recall (R), defined as follows: $F1 = 2PR/(P+R)$. *Precision* is the number of correct annotations produced by the tool divided by all annotations produced by the tool. *Recall* is the number of correct annotations produced by the tool over all annotations in the test corpus.

20 The evaluations carried out in the publications just cited provide an indication of the tools' annotation quality. Annotations in the corpus were not revised manually, with the exception of cases of enjambment across a stanza boundary, used for Table 5 in this article, which were manually checked by one scholar. For details on the certitude values encoded in the corpus for each enjambment type, see https://sites.google.com/site/spanishenjambment/ for details (accessed 10 October 2018).

21 Note that the 333 sonnets common to both Navarro-Colorado et al.'s Golden Age corpus and DISCO were only counted within Navarro *et al.*'s corpus when compiling data for all tables and figures in this article, in order to better highlight the differences between the corpora. More generally, the code to reproduce our analyses and create the tables and figures for the article was made public at https://github.com/pruizf/disco-ms.

22 See Domínguez Caparrós (2014, p. 143) or Henríquez Ureña (1919, p. 132ff.) for details on *a maiori* and *a minori* patterns. The main *a maiori* variants as described in previous studies are *2 6 10* and *3 6 10*; this is confirmed in our data. Patterns are formalized as a series of numbers indicating stressed syllables, e.g. *2 6 10* for the second, sixth, and tenth syllables. Note that 10th-syllable stress is mandatory in all patterns.

23 In Spanish, the alexandrine has 14 metrical syllables. In sonnets, the hendecasyllable predominates almost exclusively. However, particularly since the 19th century, alexandrine sonnets have been written.

24 The criteria in Navarro-Colorado (2016a, p. 101) require stresses on syllables 3 and 6 for the rhythm of a line to be considered ternary, and stresses on either syllables 2 and 6 or 4 and 8 for the rhythm to be considered binary. To avoid potentially ambiguous cases, our implementation excludes from the results on Table 4 lines that fulfill conditions for both binary and ternary rhythm. Alexandrine sonnets are also excluded; the question applies only to hendecasyllable meter.

25 Enjambment was annotated in the corpus as per Quilis' (1964) typology, which covers cases where different

types of very cohesive phrases spread over two lines. Additionally, the corpus has annotations for a related phenomenon called *enlace* by Spang (1983), which we call *expansion* in our work (Ruiz *et al.*, 2017). This involves cases where a subject or direct object occur in the adjacent line to their governing verb, and the stylistic effect is less perceptible than with enjambment. Whereas we annotated expansion cases for completeness, there is no consensus on whether expansion is a relevant stylistic phenomenon (cf. Delente, 2019). Accordingly, for our quantitative analyses of enjambment, we only considered cases of enjambment "proper" and left out expansion from our counts in Table 5 and Fig. 2. Note that we did compile and manually validate data on expansion cases across stanza boundaries. If we consider enjambment and expansion together instead of enjambment only, an equivalent increase at tercet boundaries in the 19th century is still observed. As elsewhere in the article, the 333 sonnets shared by Navarro *et al.*'s (2016) Golden Age corpus and DISCO were only counted within Navarro *et al.*'s corpus, to better isolate corpus characteristics.

26 See Martínez Cantón *et al.* (2017), Martínez Cantón and Ruiz Fabo (2019) in which a tool for enjambment analysis was explored in an educational environment.