The Diachronic Spanish Sonnet Corpus (DISCO): TEI and Linked Open Data Encoding, Data Distribution and Metrical Findings

Pablo Ruiz Fabo (pablo.ruiz@linhd.uned.es), UNED (Universidad Nacional de Educación a Distancia), Spain
Helena Bermúdez Sabel (helena.bermudez@linhd.uned.es), UNED (Universidad Nacional de Educación a Distancia), Spain
Clara Martínez Cantón (cimartinez@flog.uned.es), UNED (Universidad Nacional de Educación a Distancia), Spain
Elena González-Blanco (elenagbg@gmail.com), UNED (Universidad Nacional de Educación a Distancia), Spain
Borja Navarro Colorado (borja@dlsi.ua.es), Universidad de Alicante, Spain

[Short paper presented at DH2018 in Mexico City]

1. Introduction

Digital resources in poetry in Spanish are scarce, particularly for certain periods. This poses difficulties for Digital Humanities studies in Spanish.

Some digital editions of medieval poetry exist, e.g. BiDTEA (Gago Jover et al, 2015), ADMYTE (Marcos Marin and Faulhaber, 1992), PoeMetCa (Escribano et al, 2016), besides resources containing partial editions like ReMetCa (González-Blanco and Rodríguez, 2014). For the Golden Age, Navarro-Colorado et al. (2015) presented the *Corpus of Spanish Golden-Age Sonnets*. For later periods, we are not aware of poetry collections, although other genres are covered in Textbox (Schöch et al, 2017), BETTE (Santa María Fernández et al, 2017), Aracne (Álvarez and Martín, 2015) or Revistas Culturales 2.0 (Ehrlicher and Rißler-Pipka, 2015).

This paper describes the DISCO corpus and how it complements available digital materials for poetry in Spanish in several respects: First, the author and period range. Second, metadata concerning the authors and their works expressed in TEI-RDFa, given the importance of interoperability between literary datasets and the advantages of Linked Open Data as a paradigm. Finally, example findings that can be obtained with our corpus are provided, regarding metrical patterns diachronically.

The corpus is available on GitHub and Zenodo.[1]

[1] https://doi.org/10.5281/zenodo.1012567

2. Corpus description

The corpus contains 4087 sonnets in Spanish by 1204 authors (15th to 19th century),[2] extracted from HTML sources at Biblioteca Virtual Cervantes (García, 2005, 2006a, 2006b) and Wikisource. Sonnets were chosen given the form's importance in European poetry, where it is even considered as its own genre. The form's clear restrictions make it easily amenable to computational treatment, facilitating meaningful comparison across poems. Several computational linguistics studies on the sonnet exist (Navarro-Colorado et al., 2015, 2016, 2017a, 2017b; Agirrezabal, 2017). A new sonnet corpus complements earlier work on both traditional and computational poetry analyses.

We focused on canonical and non-canonical authors, from different Spanish-speaking countries (Figure 1).

Period	Nbr of Sonnets	Nbr of Authors			Sources
Golden Age (15th-17th)	1088	477	Female	31	García (2006a)
			Male	446	
			America	12	
			Europe	458 (+7)	
18th century	323	42	Female	1	García (2005) Wikisource
			Male	41	
			America	6	
			Europe	36	
19th century	2676	685	Female	48	García (2006b)
			Male	637	
			America	334	
			Europe	348 (+3)	

Figure 1: Sonnet and author distribution per period, including the number of female and male authors, and the continent where they developed their literary activity. Numbers in parentheses indicate authors which were probably active in Europe.

[2] 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 (footnote 1 above) provides more details.

2.1. Encoding Paradigms: TEI and Linked Open Data

The poems are encoded in XML-TEI P5. A plain-text version is also provided. Together with the TEI-semantics, this corpus provides a layer of Linked Open Data (LOD) expressed in RDFa (Herman et al., 2015). To our knowledge, no out-of-the-box tools exist for publishing literary TEI corpora as LOD.[3] In this context, the enrichment of TEI with RDFa attributes is a solid approach to translate TEI semantics to the web (see precedents like Jewell, 2010) and benefit from the wide range of possibilities of the Semantic Web: First, we enrich our dataset by linking to third-party ones (as DBpedia), providing additional resources to complement the corpus. Second, we publish our data openly using standard schemas, thus supplying semantic interoperability that allows third-party applications to automatically use our data.

2.2. Author metadata

Author metadata were extracted or inferred from unstructured source content, and specified in the teiHeader: Year, place of birth and death, and gender. Two versions of the texts are available: one collecting every sonnet per author, the other with a single sonnet per file.

For the current corpus release we augmented the TEI annotation with URIs and class/property information, expressing them in RDFa. The most straightforward information concerns authors and their works, and the DCMI Metadata Terms (DCMI Usage Board, 2012) provides an appropriate scheme. Most features regarding authors' biographical data were formalised with the FOAF vocabulary (Brickley and Miller, 2014). Links to other resources were supplied. For instance, authors were assigned *Virtual International Authority File* (VIAF) identifiers, by querying VIAF's API supplemented with manual validation. Since the corpus includes non-canonical authors, LOD is an important asset to share their work thanks to the enhanced display of this type of data implemented by search engines.

Our documentation[1] provides further details.

2.3. Metrical encoding and enjambment

Using the **met** attribute, each line was annotated for scansion (strong and weak syllables) with the ADSO tool[4] (Navarro-Colorado, 2017), which specializes in Spanish fixed-meter forms, attaining a performance of 0.95 F1. A heuristic was used to automatically annotate the quatrains' rhyme-scheme, i.e. enclosed (ABBA) or alternate (ABAB).

[3] 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 McCrae et al., 2011; Chiarcos and Ejavec, 2011). 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).

[4] https://github.com/bncolorado/adsoScansionSystem

Using an **enjamb** attribute, lines were annotated for enjambment[5] with the ANJA tool[6] (Ruiz-Fabo et al., 2017). The tool's performance at detecting enjambment is above 0.8 F1, and its efficacy at classifying enjambment types varies across periods and types. A *cert* attribute specifies the expected certitude for each enjambment type annotated.

The corpus documentation[1] provides more details.

2.4. How's this corpus different?

The metadata mentioned in 2.2. were unavailable in structured, machine-readable format in the corpus sources, or in other sonnet collections, like Sonnet-Archiv (Elf Edition). Regarding coverage, the corpus complements Navarro-Colorado et al's (2015) Golden Age Sonnet corpus, by including minor Golden Age authors. For later periods, we cover more poems and authors than existing digital corpora, up to the 19th century. Our corpus integrates RDFa annotations, which in a second version will be fully compliant with the POSTDATA model.[7] This is a pioneering model that will provide means to publish European poetry materials as Linked Open Data. Finally, combining the annotation of metrical patterns, stanza types and enjambment is not offered by prior corpora.

3. Some metrical findings

Corpus data on stress patterns (Figure 2) agree with existing descriptions[8] of the Spanish hendecasyllable based on small-sample analyses: A maiori patterns (with 6th-syllable stress) predominate, and a minori patterns (with 4th-syllable stress) 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.

Regarding diachronic data on the number of stressed positions (Figure 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.

---

[5] The tool detects different types of enjambment (i.e. a mismatch between syntactic and metrical structure) as characterized by Quilis (1964). The tool also detect's Spang's (1983) concept of *enlace*, 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

[6] See https://sites.google.com/site/spanishenjambment/ for details

[7] See Bermudez-Sabel et al. (2017). Version 0.2 of the POSTDATA model is available at https://doi.org/10.5281/zenodo.832906

[8] See Domínguez Caparrós (2014: 143) or Henríquez Ureña (1919: 132) for details on *a maiori* and *a minori* patterns. The main *a maiori* variants as described in previous literature 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.

Whereas the predominant meter for sonnets is naturally the hendecasyllable, alexandrines[9] are attested, mostly in the 19th century, preferentially used by American authors. The alexandrine sonnet uses an alternate rhyme scheme (ABAB) more often than the usual enclosed scheme (ABBA). See Figure 4.

Pattern	Pattern Class	Stress Count	Percentage of lines			
			All periods	19th	18th	Golden Age
3 6 10	mai	3	8.76	10.23	6.16	6.21
2 6 10	mai	3	8.55	7.82	6.65	5.75
2 4 8 10	*min*	*4*	*8.41*	*8.26*	*7.77*	*6.32*
2 4 6 10	mai	4	7.14	3.83	5.30	5.90
2 6 8 10	mai	4	6.37	2.71	3.10	4.07
4 6 10	mai	3	6.23	4.61	4.16	4.30
1 4 6 10	mai	4	6.17	3.03	3.87	3.49
3 6 8 10	mai	4	6.03	3.19	3.98	3.55
1 4 8 10	*min*	*4*	*5.88*	*5.02*	*4.2*	*3.38*
4 8 10	*min*	*3*	*5.76*	*4.56*	*2.62*	*2.73*
1 3 6 10	mai	4	5.73	3.76	3.79	2.40

Figure 2: Distribution of stress patterns per period (percentage of lines for each pattern) for the 10 most frequent patterns in the corpus, sorted by decreasing percentage of occurrence in the complete 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 4 stresses decrease in the 19th century, whereas *a minori* patterns increase in that century.

[9] 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.

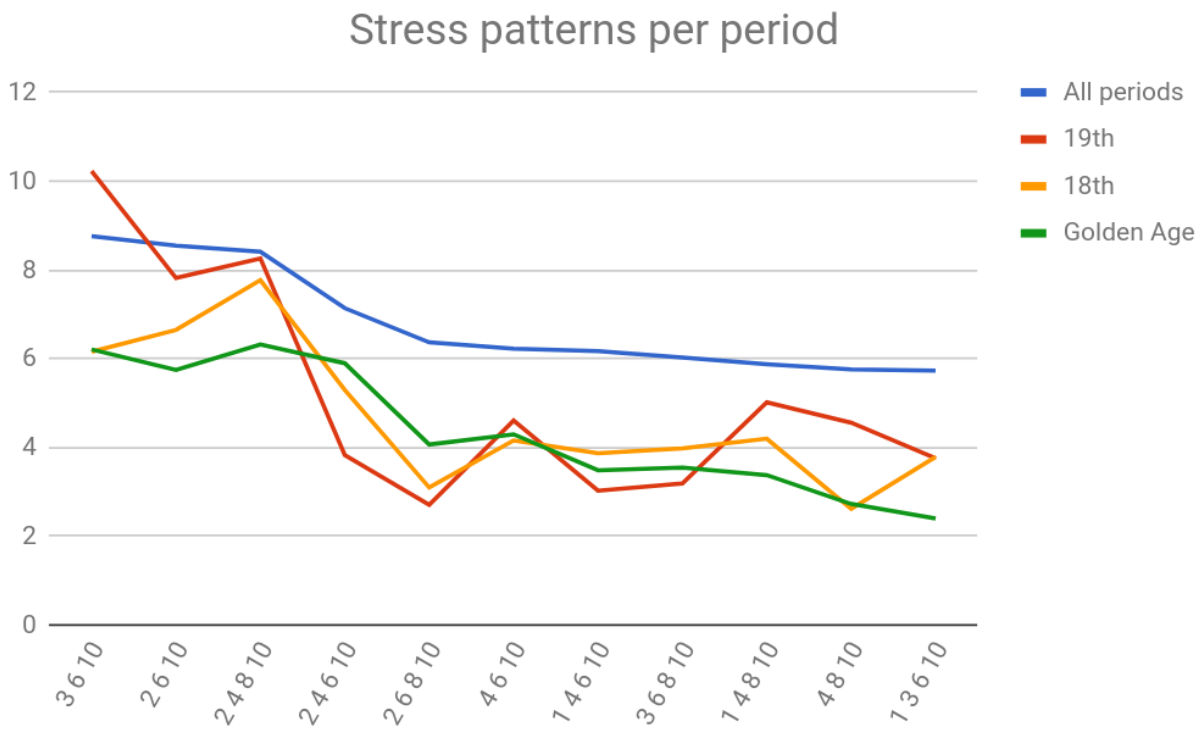

Figure 3: Distribution of stress patterns per period (percentage of lines for each pattern) for the 11 most frequent patterns in the corpus.

Meter Length	Quatrain Rhyme	Sonnet Count		
		total	American	European
hendecasyllable	Enclosed	2269	1218	1051
	Alternate	122	96	26
	Total	**2391**	**1314**	**1077**
Alexandrine	Enclosed	122	98	24
	Alternate	145	121	24
	Total	**267**	**219**	**48**

Figure 4: Count of hendecasyllable vs. 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 there's a preference for alternate rhyme for this meter length.

Acknowledgements

Supported by the project 'Poetry Standardization and Linked Open Data: POSTDATA' (ERC-2015-STG-679528), funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, and led as a Principal Investigator by Dr. Elena González-Blanco, LINHD-UNED (http://postdata.linhd.es/).

References

Agenjo, X. (2015): Las bibliotecas virtuales españolas y el tratamiento textual de los recursos bibliográficos. *Ínsula: revista de letras y ciencias humanas* 822: pp. 12–15.

Agirrezabal, M. (2017): *Automatic Scansion of Poetry. PhD Thesis*. University of the Basque Country.

Álvarez Mellado, E. and Martín-Fuertes, L. (2015): *Aracne Project,* http://www.fundeu.es/aracne/ (Accessed 22 Sep. 2017).

Bermúdez-Sabel, H., Curado Malta, M. and González-Blanco, E. (2017): Towards Interoperability in the European Poetry Community: The Standardization of Philological Concepts, in Jorge Gracia et al. (ed.) *Proceedings of Language, Data, and Knowledge: First International Conference (LDK 2017)*: pp. 156–65. Springer International Publishing doi:10.1007/978-3-319-59888-8_14.

Biblioteca Virtual Miguel de Cervantes (1999): *Biblioteca Virtual Miguel de Cervantes,* http://www.cervantesvirtual.com/ (Accessed 22 Sep. 2017).

Biblioteca Virtual Miguel de Cervantes (2007): *Biblioteca del Soneto [Sonnet Library]*, http://www.cervantesvirtual.com/bib/portal/bibliotecasoneto/ (Accessed 22 Sep. 2017).

Brickley, D. and Miller, L. (2014): FOAF Vocabulary Specification, http://xmlns.com/foaf/spec/ (Accessed 22 Nov. 2017).

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, pp. 11–20 Stroudsburg: PA, USA,

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. 151–153

DCMI Usage Board (2012): DCMI Metadata Terms, http://dublincore.org/documents/dcmi-terms (Accessed 22 Nov. 2017).

Domínguez Caparrós, J. (2009). *El moderno endecasílabo dactílico, anapéstico o de gaita gallega*. Sevilla: Padilla Libros.

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

Ehrlicher, H., and Rißler-Pipka, N. (2015): *Revistas Culturales 2.0*, https://www.revistas-culturales.de/es. (Accessed 22 Sep. 2017).

Elf Edition: *Sonett-Archiv,* http://sonett-archiv.com. (Accessed 22 Sep. 2017).

Escribano, J., González-Blanco, E. and Río Riande, G. del (2016). *PoeMetCa—Recursos digitales para el estudio de la Poesía Medieval Castellana,* http://poemteca.linhd.es (Accessed 22 Sep. 2017).

Gago Jover, F. (2015): La biblioteca digital de textos del español antiguo (BiDTEA). *Scriptum Digital* 4: pp. 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 26 Nov. 2017).

García González, Ramón (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 26 Nov. 2017).

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 26 Nov. 2017).

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 Sep. 2017), doi:10.4000/jtei.1274.

Henríquez Ureña, P. (1919). El endecasílabo castellano. *Revista de Filología Española*, 6: pp. 132–157.

Herman, Ivan, Asida, B., McCarron, S., Birbeck, M. (2015): RDFa Core 1.1 - Third Edition, https://www.w3.org/TR/rdfa-core (Accessed 22 Nov. 2017).

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 Nov. 2017).

Marcos Marín, F. and Faulhaber, C. B. (coord.) (1992): *ADMYTE*. *Archivo Digital de Manuscritos y Textos Españoles*, http://www.admyte.com/admyteonline/contenido.htm (Accessed 22 Sep. 2017).

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. 245–259, Berlin: Springer-Verlag.

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. (2017): *ADSO project – Análisis distante del soneto castellano de los Siglos de Oro [Distant analysis of the Spanish Golden Age sonnet]*, http://adso.gplsi.es/index.php/es/proyecto-adso (Accessed 22 Sep. 2017).

Navarro-Colorado, B., Ribes Lafoz, M. and Sánchez, N. (2015): *Corpus of Spanish Golden-Age Sonnets*. Alicante: University of Alicante, https://github.com/bncolorado/CorpusSonetosSigloDeOro (Accessed 22 Sep. 2017).

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 Sep. 2017)

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 Sep. 2017)

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

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.

Santa María Fernández, M. T., Jiménez Fernández, C. M. (2017): *Biblioteca Electrónica Textual Del Teatro Español, 1868-1936*. Universidad Internacional de la Rioja, Spain.

Schöch, C. Henny, U., Calvo Tello, J. Popp, S. (2015): *The CLiGS Textbox*, https://github.com/cligs/textbox (Accessed 22 Sep. 2017)

Wikisource: *Categoría: Sonetos.*, https://es.wikisource.org/w/index.php?title=Categor%C3%ADa:Sonetos (Accessed 26 Nov. 2017)