DISCO: Diachronic Spanish Sonnet Corpus

This poster presents a corpus of 19th-century sonnets in Spanish in XML-TEI (685 authors, 2677 sonnets). It includes well-known authors, like Bécquer, Delmira Agustini or "Clarín", but also less canonized authors. Texts and authors are enriched with identifiers and metadata. See https://github.com/postdataproject/disco

1. Introduction

A fundamental difficulty for Digital Humanities studies on Spanish literature is a scarcity of digital resources (Agenjo, 2015).

Some resources do however exist. BiDTEA (Gago Jover et al, 2015), ADMYTE (Marcos Marin and Faulhaber, 1992), ReMetCa (González-Blanco and Rodríguez, 2014) and PoeMetCa (Escribano et al, 2016) have digitized Spanish Medieval texts. Navarro-Colorado et al. (2015) presented the *Corpus of Spanish Golden-Age Sonnets*.

Regarding 19th-century Spanish literature, available collections covering different genres are Textbox (Schöch et al., 2015), BETTE (Santa María Fernández, 2017), Aracne (Álvarez-Mellado and Martín-Fuertes, 2015), and Revistas Culturales 2.0 (Ehrlicher and Rißler-Pipka, 2015). Nevertheless, none of these projects are working on poetry.

DISCO complements this growing ecosystem by adding a meaningful representation of 19th-century sonnets, with more periods under validation, to be published shortly.

2. Corpus description

2.1. What is DISCO

Our corpus collects 2677 sonnets in Spanish from the 19th century, by 685 authors (Spain or Latin America). It intends to provide a wide sample, inspired by distant reading approaches (Moretti, 2005). The raw texts were extracted from Biblioteca Virtual Miguel de Cervantes (1999).

The texts have been encoded in XML-TEI P5, given this standard's benefits in terms of reuse, storage and retrieval. Author metadata have been extracted or inferred from unstructured content in the sources, and placed 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 encoding a single sonnet per file. For corpus preparation, we closely followed the TEI guidelines and RIDE's criteria for Digital Text Collections (Henny and Neuber, 2017).

Additionally, authors have been assigned VIAF identifiers. This gives the corpus an entry-point the to the linked open data cloud, enhancing its findability. The corpus is available on GitHub and saved in Zenodo, adopting good practices for data use, reuse, and conservation.

We have also obtained sonnets from other centuries, since the 15th century to the present. These are under validation and will be published shortly within the DISCO corpus, which intends to give a wide perspective on the sonnet in Spanish diachronically.

2.2. Why sonnets

The sonnet has had great importance in European poetry; the relevance of the corpus for literary scholarship is guaranteed. It is a "manageable" form to treat computationally, obeying clear restrictions. Variability stays within bounds, making meaningful comparison across poems easier, regarding scansion or rhyme types. Besides, some digital collections of sonnets already exist (with different features to ours, as discussed below) as well as automatic analyses of this form. The sonnet has received attention from the computational linguistics community (Navarro-Colorado et al, 2015, 2016, 2017; Agirrezabal, 2017) including the ADSO project (Navarro-Colorado 2017). The DISCO corpus will also be useful for that audience. For these reasons, a new sonnet corpus allows us to engage in a dialogue with earlier work in traditional literary studies, in digital corpus development, and in computational poetry analyses.

Concerning digitally available sonnet corpora, Sonnet-Archiv (Elf Edition) is organized as a forum, and its coverage is less wide than ours. The “Sonnet Library” (Biblioteca Virtual Miguel de Cervantes, 2007) is organized alphabetically, rather than using meaningful criteria for literary scholarship, like periods. Both are traditional websites. Finally, the Corpus of Spanish Golden Age Sonnets (Navarro-Colorado et al., 2015) covers major authors from the 15th to the 17th century, with an automatic metrical annotation. Author metadata in these corpora are very limited and unavailable in a machine-readable format (see Calvo Tello, 2017, for discussion of related issues). With the DISCO corpus, we are offering a wider period and author range, from major to minor authors, encoded in XML-TEI, available as repository, with richer structured and standard metadata.

3. Summary

With the DISCO corpus, while focusing on sonnets, we intend to increase available digital resources in Spanish poetry, by addressing additional periods, covering minor as well as canonical authors, and including materials from several Spanish-speaking countries. Choosing the sonnet complements existing work on this form, in traditional and computational literary studies. TEI was adopted in order to serve the large community using this format. The corpus can be made available as linked open data as it includes VIAF IDs. It is published at https://github.com/postdataproject/disco.

Acknowledgements

The work was supported by Starting Grant ERC-2015-STG-679528 POSTDATA, PI Elena González-Blanco. We also thank Helena Bermúdez from the LINHD-UNED lab for later contributions to the corpus

References

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

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

Álvarez Mellado, Elena / Martín-Fuertes, Leticia (2015): *Aracne Project* [online]. Available at: http://www.fundeu.es/aracne/ [Accessed 22 Sep. 2017].

Biblioteca Virtual Miguel de Cervantes (1999): *Biblioteca Virtual Miguel de Cervantes* [online]. Available at: http://www.cervantesvirtual.com/ [Accessed 22 Sep. 2017].

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

Calvo Tello, José. (2017). Review of Corpus of Spanish Golden Age Sonnets by Borja Navarro Colorado, María Ribes Lafoz and Noelia Sánchez (ed.), in *RIDE, 6*. Institut für Dokumentologie und Editorik, Köln. [Online]. Available at: http://ride.i-d-e.de/issue-6/corpus-of-spanish-golden-age-sonnets/ [Accessed 22 Sep. 2017].

Ehrlicher, Hanno / Rißler-Pipka, Nanette (2015). *Revistas Culturales 2.0*. Augsburg: Universität Augsburg. [Online]. Available at: https://www.revistas-culturales.de/es [Accessed 22 Sep. 2017].

Elf Edition: *Sonett-Archiv* [online]. Available at: http://sonett-archiv.com [Accessed 22 Sep. 2017].

Escribano, Juanjo / González-Blanco, Elena / Río Riande, Gimena del (2016). *PoeMetCa—Recursos digitales para el estudio de la Poesía Medieval Castellana*. [Online]. Available at: http://poemteca.linhd.es [Accessed 22 Sep. 2017].

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

González-Blanco, Elena / Rodríguez, José Luis (2014): "ReMetCa: A Proposal for Integrating RDBMS and TEI-Verse", in *Journal of the Text Encoding Initiative* 8 [online]. Available at: https://jtei.revues.org/1274 [Accessed 22 Sep. 2017], doi:10.4000/jtei.1274.

Henny, Ulrike / Neuber, Frederike (2017): “Criteria for Reviewing Digital Text Collections, version 1.0”. IDE, Institut for Dokumentologie und Editorik, [online]. Available at: https://www.i-d-e.de/publikationen/weitereschriften/criteria-text-collections-version-1-0/ [Accessed 22 Sep. 2017].

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

Moretti, Franco (2005): *Graphs, Maps, Trees: Abstract Models for a Literary History*. London and New York: Verso.

Navarro-Colorado, Borja (2015): A computational linguistic approach to Spanish Golden Age Sonnets: metrical and semantic aspects. In *ACL Workshop on Computational Linguistics for Literature* 105.

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

Navarro-Colorado, Borja / Ribes Lafoz, María / Sánchez, Noelia (2015): *Corpus of Spanish Golden-Age Sonnets*. Alicante: University of Alicante [online]. Available at: https://github.com/bncolorado/CorpusSonetosSigloDeOro [Accessed 22 Sep. 2017].

Navarro-Colorado, Borja / Ribes Lafoz, María, / Sánchez, Noelia (2016): “Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation”, in *Proceedings of the Language Resources and Evaluation Conference* [online]. Available at: http://www.lrec-conf.org/proceedings/lrec2016/pdf/453_Paper.pdf [Accessed 22 Sep. 2017]

Navarro-Colorado, Borja (2017): “A metrical scansion system for fixed-metre Spanish poetry”, in *Digital Scholarship in the Humanities.* https://doi.org/10.1093/llc/fqx009 [Accessed 22 Sep. 2017]

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

Schöch, Christof / Henny, Ulrike / Calvo Tello, José / Popp, Stefanie (2015): *The CLiGS Textbox*. Würzburg: University of Würzburg. [Online]. Available at: https://github.com/cligs/textbox [Accessed 22 Sep. 2017]