Quantitative automatic analysis of stanzas and metrical structures in Spanish Poetry: a first approach

Laura Hernández Lorenzo, Javier de la Rosa, Álvaro Pérez Pozo, Aitor Díaz Medina, Salvador Ros Muñoz, Elena González-Blanco

Automatic analysis of poetic texts has received notable attention from the scholarly community in recent times. For Spanish Poetry, approaches for scansion and syllabification, as well as rhyme analysis, have been implemented. However, automatic extraction and quantification of stanzas and metrical structures –a complex task given the abundance, variety, and frequent irregularity of structures– remains unexplored.

Within the context of the ERC POSTDATA Project, that aims to provide a standard model for European Poetry, our approach to stanza and structure identification builds in top of Rantanplan, an open-source Python package and tool for syllabification and scansion. Levering the information on metrical syllables and the stressed endings of each line of verse, our algorithm assigns rhyme per line with a customizable window. Both metrical lengths and rhyme patterns are then fed to a pattern matching algorithm using regular expressions that search for matches within a growing knowledge base of existing named structures and stanzas. The most complex match is then returned in both machine and human-readable formats, enabling scholars to pursue further investigations.

We tested our system on a corpus of over 25,000 unlabelled poems, finding that at least 10% had a recognizable structure in the Spanish poetry tradition. After randomly sampling 150 poems for manual verification, we obtained an accuracy of 75%. Most errors were due to regular expression subsumption and complex structures with intertwined rhymes between the composite stanzas.

González-Blanco, E., Ros, S., Diez Platas, Ma. L., Ruiz Fabo, P., Bermúdez Sabel, H., Ayciriex, L., & Martinez Canton, C. (2018, octubre 9). *Poetry Lab* [Poster]. CLARIN Annual Conference 2018, Pisa, Italy. https://zenodo.org/record/1465500

Navarro-Colorado, B. (2017). A metrical scansion system for fixed-metre Spanish poetry. *Digital Scholarship in the Humanities*, *33*(1), 112-127.

Plecháč, P. (2018). A collocation-driven method of Discovering Rhymes (in Czech, English, and French Poetry). En M. Fidler & V. Cvrček (Eds.), *Taming the Corpus. From Inflection and Lexis to Interpretation* (pp. 79–95). Springer.

Ruiz Fabo, P., Martínez Cantón, C. I., Poibeau, T., & González-Blanco, E. (2017). Enjambment Detection in a Large Diachronic Corpus of Spanish Sonnets. *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 27-32.