Towards a system for automatic scansion for Catalan poetry

The scansion of poetry is a task that involves acting on a string of text, the verse, in order to detect its rhythm and establish a metrical pattern. In the case of Catalan verse, the person in charge of the scansion must be able to identify the tonic syllables of each word, resolve the cases of accentual clashes and valleys and detect and apply, if necessary, the phenomena of syntactic phonetics or stylistic resources. A set of decisions that condition the result of the scansion: number of syllables and accentual positions, in short, rhythmic, and metrical distribution. This paper presents the first automatic scansion proposal for Catalan. Although several proposals for its scansion are known for different languages (to cite some cases close to Catalan: Gervás (2000), Navarro-Colorado et al. (2016) and Marco & Gonzalo (2021) for Spanish, or Mittmann et al. (2018) and Carvalho et al. (2020) for Portuguese). The program we present written in Python offers a rule-based system for scanning verses. The program has been tested for the analysis of decasyllabic verse, one of the most frequent in the Catalan poetic tradition, through a corpus built especially for the research. The corpus was made possible thanks to the collaboration of the Càtedra Màrius Torres of the Universitat de Lleida, which provided part of the texts for the corpus. Another part has been compiled from digital samples of poetry books from various internet repositories. Preliminary results show a high accuracy and efficiency of the programme, as well as pointing out the main difficulties to be faced in improving it.

Keywords: scansion, metre, Catalan, poetry

Carvalho, R., Loula, A., & Queiroz, J. (2020). Identificação computacional de estruturas métricas de versificação na prosa de Euclides da Cunha / Computational identification of versification metric structures in Euclides da Cunha's prose. *REVISTA DE ESTUDOS DA LINGUAGEM*, *28*(1), 41–68. https://doi.org/10.17851/2237-2083.28.1.41-68

Gervás, P. (2000). A Logic Programming Application for the Analysis of Spanish Verse. In J. Lloyd, V. Dahl, U. Furbach, M. Kerber, K.-K. Lau, C. Palamidessi, L. M. Pereira, Y. Sagiv, & P. J. Stuckey (Eds.), *Computational Logic — CL 2000. CL 2000. Lecture Notes in Computer Science* (pp. 1330–1344). Springer. https://doi.org/10.1007/3-540-44957-4_89

Marco, G., & Gonzalo, J. (2021). Escansión automática de poesía española sin silabación. *Procesamiento Del Lenguaje Natural*, *66*, 77–87. https://doi.org/10.26342/2021-66-6

Mittmann, A., von Wangenheim, A., & dos Santos, A. L. (2018). Aoidos: A system for the automatic scansion of poetry written in Portuguese. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9624 LNCS*, 611–628. https://doi.org/10.1007/978-3-319-75487-1_46/COVER

Navarro-Colorado, B., Ribes Lafoz, M., & Sánchez, N. (2016). *Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation* (pp. 4360–4364). https://aclanthology.org/L16-1691