PoetryLab as Infrastructure for the Analysis of Spanish Poetry

Javier de la Rosa
LINHD
UNED
Madrid, Spain
versae@linhd.uned.es

Álvaro Pérez
LINHD
UNED
Madrid, Spain
alvaro.perez@linhd.uned.es

Laura Hernández
LINHD
UNED
Madrid, Spain
laura.hernandez@linhd.uned.es

Aitor Díaz
Control and Communication Systems
UNED
Madrid, Spain
adiazm@scc.uned.es

Salvador Ros
Control and Communication Systems
UNED
Madrid, Spain
sros@scc.uned.es

Elena González-Blanco
School of Human Sciences and Technology
IE University
Madrid, Spain
egonzalezblanco@faculty.ie.edu

Abstract

The development of the network of ontologies of the ERC POSTDATA Project brought to light some deficiencies in terms of completeness in the currently available European poetry corpora. To tackle the issue in the realm of the Spanish poetic tradition, our approach consisted in designing a set of tools that any scholar could use to automatically enrich the analysis of Spanish poetry. The effort crystallized in the PoetryLab, an extensible open source toolkit for syllabification, scansion, enjambment detection, rhyme detection, stanza identification, and historical named entity recognition for Spanish poetry. We designed the system to be interoperable, compliant with the project ontologies, easy to use by tech-savvy and non-expert researchers, and requiring minimal maintenance and setup. Furthermore, we propose the integration of the PoetryLab as a core functionality in the tool catalog of CLARIN for Spanish poetry.

1 Introduction

The main goal of the ERC-funded POSTDATA Project (Curado Malta and González-Blanco, 2016)[1] was to formalize a network of ontologies capable of expressing any poetic expression and its analysis at the European level, thus enabling scholars all over Europe to interchange their data using Linked Open Data. The POSTDATA Project bridges the digital gap between traditional cultural assets and the growing world of data. It is focused on poetry analysis, classification, and publication, applying Digital Humanities methods of academic analysis –such as XML-TEI encoding (Dombrowski and Denbo, 2013; Flanders and Hamlin, 2013)– in order to look for standardization, as well as innovation by using semantic

[1]Starting Grant research project Poetry Standardization and Linked Open Data: POSTDATA (ERC-2015-STG-679528) funded by the European Research Council (ERC, https://erc.europa.eu) under the research and innovation program Horizon2020 of the European Union: http://postdata.linhd.uned.es/

Javier De La Rosa, Salvador Ros, Álvaro Pérez, Aitor Díaz, Laura Hernández, Mirella De Sisto and Elena González-Blanco 2021. PoetryLab as Infrastructure for the Analysis of Spanish Poetry. *Selected papers from the CLARIN Annual Conference 2020*. Linköping Electronic Conference Proceedings 180: 180 75–82.

web technologies (Cigarrán-Recuero et al., 2014) to link and publish literary datasets in a structured way in the linked data cloud. The advantages of making poetry available online as machine-readable linked data are threefold: first, the academic community would have an accessible digital platform to work with poetic corpora and to contribute to its enrichment with their own texts; second, this way of encoding and standardizing poetic information will be a guarantee of preservation for poems published only in old books or even transmitted orally, as texts will be digitized and stored; third: datasets and corpora will be available and open access to be used by the community for other purposes, such as education, cultural diffusion or entertainment.

However, varied research interests result in corpora that might not share the same facets of an analysis. To alleviate this concern and foster the completeness of the interchanged corpora, our team set to build a software toolkit to assist in the analysis of poetry. This paper details the first iteration of the PoetryLab, an extensible open source toolkit for syllabification, scansion, enjambment detection, rhyme detection, stanza identification, and historical named entity recognition for Spanish poetry, that achieves state of the art performance in the tasks for which reproducible alternatives exist.

2 PoetryLab

Despite a long and rich tradition (Bello, 1859; Navarro Tomás, 1991; Caparrós, 2014), not many computational tools have been created to assist scholars in the annotation and analysis of Spanish poetry. With ever increasing corpora sizes and the popularization of distant reading techniques, the possibility to automate part of the analysis became very attractive. Although solutions exist, they are either incomplete, e.g., scansion of fixed-metre poetry (Agirrezabal et al., 2016; Navarro-Colorado, 2017; Gervas, 2000; Agirrezabal et al., 2017), not applicable to Spanish (Agirrezabal et al., 2017; Hartman, 2005), or not open or reproducible (Gervas, 2000). Moreover, disparate input and output formats, operating system requirements and dependencies, and the lack of interoperability between software packages, further complicated the limited ecosystem of tools to analyze Spanish poetry. These limitations guided the design of the PoetryLab as a two layer system: a REST API that operates as middleware connecting the different tools, and a consumer web-based UI that exposes the functionality to non-experts users. All tools are released as independent Python packages with their own command line interface applications (where appropriate), and are ready to produce RDF triples compliant with the POSTDATA Project network of ontologies. Figure 1 shows a diagram of the general architecture of the system.

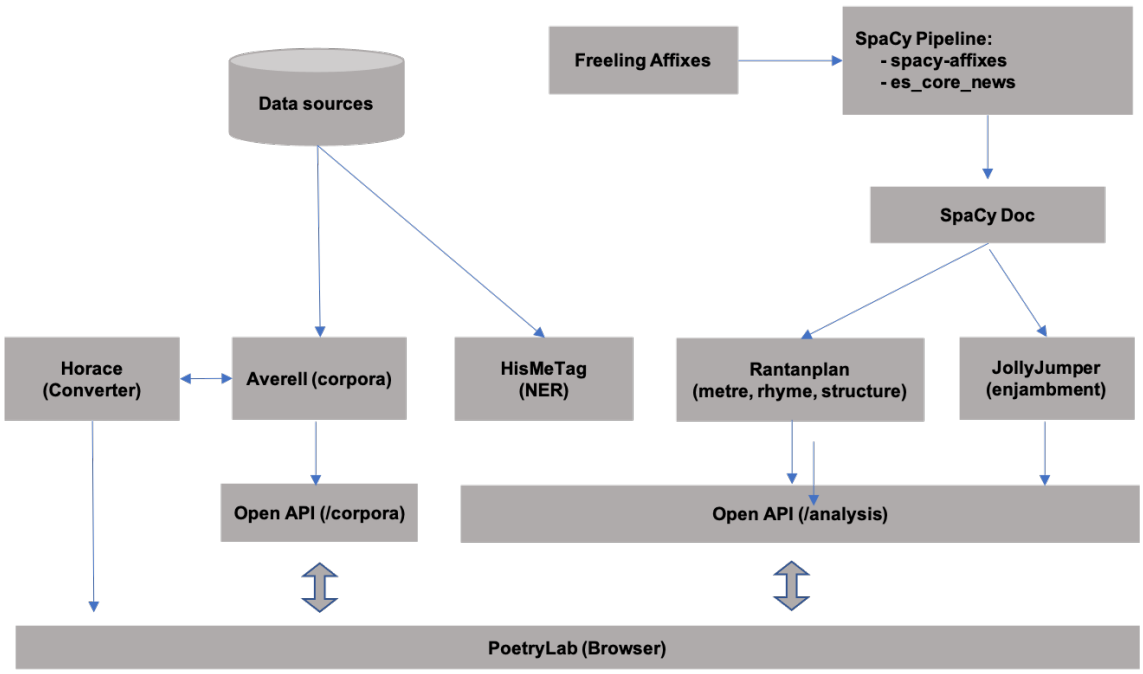

Figure 1: General architecture of the PoetryLab.

This granular design allows for each component of the PoetryLab to be used and deployed as a set of Docker images, which makes managing the different tools lifecycle and versioning a less problematic issue. We tested this approach by using ouroboros[2], a service to automatically update running docker containers with the newest available image, and the demo site of the PoetryLab has been running without major incidents over a year now[3]. We feel hosting the PoetryLab as one of the tools in the catalog of software tools available in CLARIN would be a good addition to its ecosystem, since it requires little effort to setup and the maintenance of the different tools is deferred to their own maintainers, as it usually happens in the Open Source ecosystem, making it easy for hot-replacement when new versions become available. Moreover, the use of Docker containers as deployment strategy and the fact that the tools are stateless, allow the use of lambda architectures to minimize the running costs.

2.1 PoetryLab API

At its core (see Figure 1), the PoetryLab API provides a self-documented Open API (OpenAPI Initiative, 2017) that connects the independent packages together and exposes their outputs in different formats. Two main endpoints provide functionality to analyze texts uploaded by an user (`/analysis`), and to work with a catalog of existing corpora (`/corpora`)[4].

2.1.1 Endpoint `/analysis`

The first endpoint of the PoetryLab API, `/analysis`, leverages three tools to perform several aspects of the analysis of a poem: scansion and rhyme identification, enjambment detection, and named entity recognition (i.e. Rantanplan, Jollyjumper, and Hismetag). AnCora (Taulé et al., 2008), the corpus spaCy is trained on for Spanish, splits most affixes thus losing the multi-token word information and causing some failures in the part of speech tags it produces. To circumvent this limitation and to ensure clitics were handled properly, we integrated Freeling's affix rules via a custom built pipeline for spaCy. The resulting package, spacy-affixes[5], splits words with affixes so spaCy can handle their part of speech correctly (Padró and Stanilovsky, 2012). Getting this information right was crucial to identify the stress of some monosyllabic and disyllabic words, and to find a special kind of enjambment called sirrematic in which a grammatical unit is divided in two lines (see Table 1 for a summary of the performance of our scansion system). The outputs of these two tools are then transformed to accommodate to the definitions given in the network of ontologies developed within the POSTDATA Project.

Method	**Accuracy**
(Gervas, 2000)	88.73
(Navarro-Colorado, 2017)	94.44
(Agirrezabal et al., 2017)	90.84
Rantanplan (ours)	**96.23**

Table 1: Scores on Navarro-Colorado's fixed-metre 1,400 verses corpus. Best scores in bold.

Lastly, the PoetryLab API provides a pluggable architecture that allows for the integration of external packages developed in languages other than Python. This is the case for our named entity recognition system, HisMeTag (Platas et al., 2021), developed in Java and connected to the PoetryLab API through an internal REST API exposed via Docker. The only requirement is to consume raw plain text and to produce both a JSON output and RDF triples compliant with the POSTDATA Project network of ontologies.

2.1.2 Endpoint `/corpora`

The second available endpoint, `/corpora`, aims to facilitate working with existing repositories of annotated poetry. Averell, the tool that handles the corpora, is able to download an annotated corpus and

[2] `https://github.com/pyouroboros/ouroboros`
[3] `http://postdata.uned.es/poetrylab`
[4] A demo with the Open API user interface is available at `http://postdata.uned.es:5000/ui/`.
[5] `https://github.com/linhd-postdata/spacy-affixes/`

reconcile different TEI entities to provide a unified JSON output and RDF triples at the desired granularity. That is, for their investigations some researchers might need the entire poem, poems split line by line, or even word by word if that is available. Averell allows to specify the granularity of the final generated dataset, which is a combined JSON or RDF with all the entities in the selected corpora.

Name	Size	Docs	Words	Granularity	License
Disco V2	22M	4,088	381,539	stanza, line	CC-BY
Disco V3	28M	4,080	377,978	stanza, line	CC-BY
Sonetos Siglo de Oro	6.8M	5,078	466,012	stanza, line	CC-BY-NC 4.0
ADSO 100 poems corpus	128K	100	9,208	stanza, line	CC-BY-NC 4.0
Poesía Lírica Castellana del Siglo de Oro	3.8M	475	299,402	stanza, line, word, syllable	CC-BY-NC 4.0
Gongocorpus	9.2M	481	99,079	stanza, line, word, syllable	CC-BY-NC-ND 3.0 FR
Eighteenth Century Poetry Archive	2.4G	3,084	2,063,668	stanza, line, word	CC BY-SA 4.0
For Better For Verse	39.5M	103	41,749	stanza, line	Unknown
Métrique en Ligne	183M	5,081	1,850,222	stanza, line	Unknown
Biblioteca Italiana	242M	25,341	7,121,246	stanza, line, word	Unknown

Table 2: Available corpora in Averell

Each corpus in the catalog must specify the parser to produce the expected data format. At the moment, there are parsers for five corpora, all using the TEI tag set (see Table 2). For corpora not in our catalog, the researcher can define her own or reuse one of the existing ones to process a local or remote corpus.

Moreover, for plain text local corpora Averell allows to post-process the raw texts with Rantanplan to enrich poems with their metrical and structural information as detected by the tool. The result of this process can still be combined seamlessly with the existing corpora in the catalog.

2.2 PoetryLab UI

The PoetryLab API is then used to provide with functionality to a React-based web interface that non-technical scholars can use to interact with the packages in a graphical way (see Figure 2). The frontend gives the option to download the generated data in both JSON and POSTDATA Project RDF triples formats[6].

The web interface is run entirely in the browser as a stateless application. However, the collection of analyzed poems are saved to the browser local storage which persists between sessions and restarts. Unfortunately, it lacks a user management system that could provide with persistent storage in a backend.

2.3 PoetryLab tools

Most tools included in PoetryLab are generally available to the public as standalone libraries or applications. Rantanplan, JollyJumper, HisMeTag and Horace are all orchestrated to be used by tech-savvy and non-expert researchers, through the Poetrylab UI or as standalone applications. Moreover, Averell and spacy-affixes are auxiliary packages that the rest of the toolkit builds upon.

2.3.1 Rantanplan

Rantanplan[7] (de la Rosa et al., 2020) is a Python library for the automated scansion of Spanish poetry. Scansion is the measurement of the rhythm of verses of a poem. It is comprised of four modules that work together to perform scansion of both fixed-metre as well as mixed-metre poetry: part-of-speech (PoS) tagger, syllabification, stress assignment, and metrical adjustment (see Algorithm 1). Rantanplan is fast and accurate as it is built using SpaCy while leveraging Freeling rules to handle clitics and other

[6] http://postdata.uned.es/poetrylab
[7] https://github.com/linhd-postdata/rantanplan/

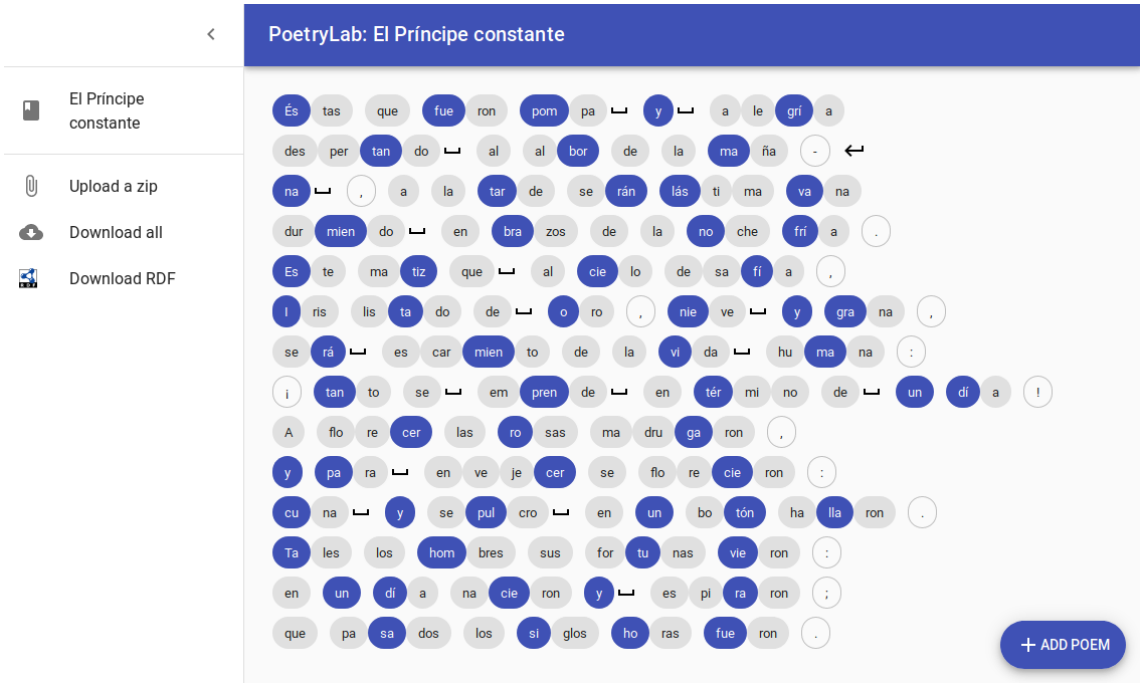

Figure 2: PoetryLab showing stressed syllables (blue), sinalefas (‿) and enjambments (↵).

nuances of the Spanish language through spacy-affixes. Rantanplan is the current state-of-the-art both in terms of speed and accuracy on the reference corpus used by Navarro-Colorado (Navarro-Colorado et al., 2016), yielding a 96.23% of accuracy on a fixed-metre corpus of hendecasyllabic verses.

Rantanplan is also able to identify up to 45 different types of the most significant Spanish stanzas (see Figure 3). Stanzas are structural units formed by lines of verses, and therefore they are related to the author style and even historical preferences that make identifying them a complicated task. Rantanplan first analyzes the verses that comprise a stanza to gather information about their lengths, rhyme pattern and rhyme type. With this information, it checks a set of rules crafted and sorted by domain experts in order to propose a stanza type. Rantanplan is able to correctly identify 78.63% of the stanza types in a corpus of over 5000 stanzas manually annotated.

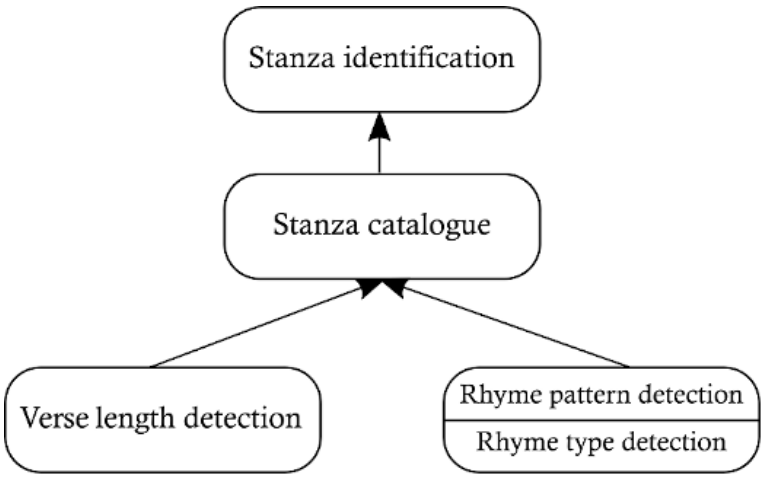

Figure 3: Stanza detection architecture

Algorithm 1: Rantanplan main algorithm

Input: Poem as a sequence $\mathcal{L}$ of lines $\langle l_1, l_2, \ldots, l_n \rangle$,
each with a sequence $\mathcal{W}$ of words $\langle w_1, w_2, \ldots, w_n \rangle$
Output: Data structure with metrical information
for $l_i \in \mathcal{L}$ **do**
  for $w_i \in \mathcal{W}$ **do**
    $tag_i \leftarrow$ pos(w_i)
    $syllables_i \leftarrow$ syllabify(w_i)
    $stresses_i \leftarrow$ stress($syllables_i$, tag_i)
  end
  $groups \leftarrow$ phonological($syllables$, $stresses$)
  $pattern \leftarrow$ transform($groups$)
  if $length$ **then**
    while $|pattern| < length$ **do**
      $g \leftarrow$ generate_phonological($\mathcal{W}$)
      $pattern \leftarrow$ transform(g)
    end
  end
  $patterns \leftarrow$ push($pattern$)
end
$rhyme \leftarrow$ extract($patterns$)
$stanza \leftarrow$ identify ($pattern$, $rhyme$)
return $patterns$, $rhyme$, $stanza$

2.3.2 JollyJumper

Jollyjumper[8] is our enjambment detection Python library for Spanish (see Figure 4). Enjambment is the metric phenomenon that occurs when there is a disagreement between the syntactic unit and metric unit, that is, when the syntactic unit exceeds the verse pause and overflows into the next verse, or when elements of the unity of sense which constitutes the next verse are anticipated at the end of a verse. Automatic detection allows for large scale quantitative analyses on the phenomenon. As an example, we ran the system on approx. 9,000 sonnets from the 15th to the early 20th century, examining patterns of evolution in the distribution of the use of enjambment according to line-position in the sonnet.

2.3.3 HisMeTag

HisMeTag[9] is a Java tool for the identification and tagging of place names in Medieval Spanish texts. It combines lexical, syntactic, and semantic analysis with NLP technologies. This task involves specific challenges: the complex morphosyntactic characteristics in proper-noun use in medieval texts, the lack of strict orthographic standards, and the diachronic and geographical variations in Spanish from the 12th to the 15th century. The system is also integrated in Poetrylab API as a Docker image which then exposes its functionality using the same common API.

2.3.4 Horace

Horace[10] is a translation tool between the natively produced PoetryLab JSON format consumed internally and the semantic formats of the POSTDATA Project. This tool is capable of reading the outputs of most of the tools in PoetryLab to build knowledge graphs in RDF format compliant with the POSTDATA set of ontologies. The step is crucial as it allows exposing the information produced by the toolkit through a SPARQL endpoint, thus enabling the interoperability and sharing of both the data included in the different public corpora and the automated annotations produced by our tools.

[8] https://github.com/linhd-postdata/jollyjumper
[9] https://github.com/linhd-postdata/hismetag
[10] https://github.com/linhd-postdata/horace

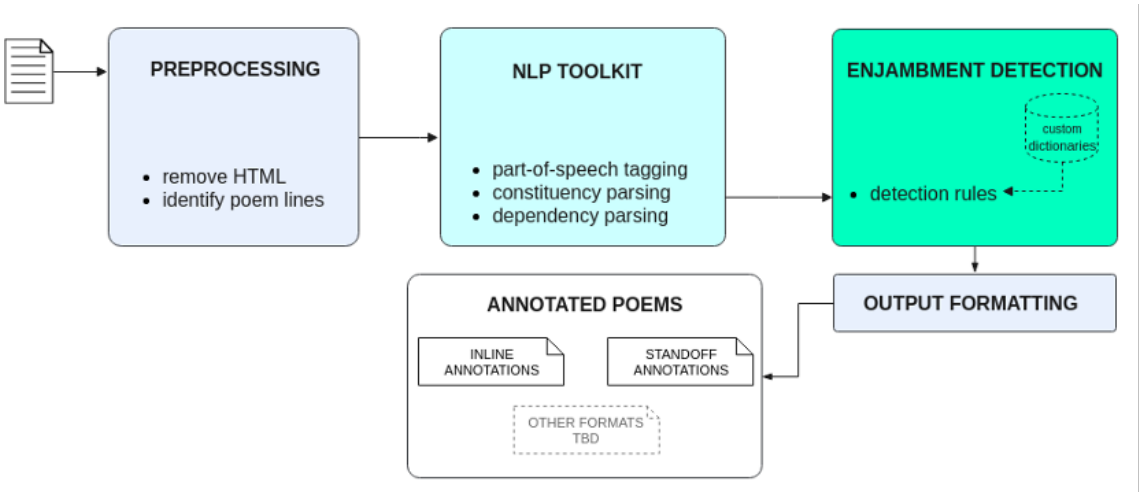

Figure 4: Jollyjumper General architecture

2.3.5 Averell

This tool is a one-stop command line interface application to gather existing corpora. Averell[11] is able to download annotated poetic corpora from different sources, parse them, and turn them into a single JSON or CSV file ready for analysis. This is especially useful when setting benchmarks, since it reduces the burden of cleaning and parsing from the researchers. Since Averell makes filed translation internally for each corpus in its catalog, it also allows researchers to create a corpus that is the combination of several other corpora, selecting sets of poems that meet a specific set of conditions (e.g., only corpora in Italian with manually verified annotations of metrical patterns). Moreover, the granularity at which this selection can be made is to the preference of the researcher, being able to choose between poem, stanza, verses, words, and even syllables (whenever available in the specific corpus, see Table 2).

3 Conclusion

The PoetryLab has proven useful in that it provides an integrated set of tools for Spanish poetry scholars. It might become useful at several stages of the research cycle. Averell helps build ad-hoc corpora, which may include metrical information generated by Rantanplan, rhetorical devices as detected by JollyJumper, and even historical named entities as recognized by HisMeTag. It also produces machine readable and interoperable data suitable to be ingested into a triple store compliant with the POSTDATA Project network of ontologies (i.e., Horace). In fact, this approach is already being tested as we export the analysis of poems and feed them into a Virtuoso Universal Server that integrates with the POSTDATA Project network of ontologies to produce repertoires knowledge graphs.

The PoetryLab will be eventually integrated into the larger POSTDATA Project public website, making working with European repositories of poetry a more pleasant task, and assisting whenever possible with the metrical and rhetorical side of the analysis. Moreover, more attention needs to be put into building a friendly web user interface useful for different user profiles.

Acknowledgements

Research for this paper has been achieved thanks to the Starting Grant research project Poetry Standardization and Linked Open Data: POSTDATA (ERC-2015-STG-679528) obtained by Elena González-Blanco. This project is funded by the European Research Council (https://erc.europa.eu) (ERC) under the research and innovation program Horizon2020 of the European Union.

[11] https://github.com/linhd-postdata/horace

References

Manex Agirrezabal, Aitzol Astigarraga, Bertol Arrieta, and Mans Hulden. 2016. Zeuscansion: a tool for scansion of english poetry. *Journal of Language Modelling*, 4.

Manex Agirrezabal, Iñaki Alegria, and Mans Hulden. 2017. A comparison of feature-based and neural scansion of poetry. *arXiv preprint arXiv:1711.00938*.

Andrés Bello. 1859. *Principios de la ortolojía i métrica de la lengua castellana.*. la Opinión.

José Domínguez Caparrós. 2014. Teoría métrica del verso esdrújulo. *Rhythmica: revista española de métrica comparada*, 12:55–96.

Juan Cigarrán-Recuero, Joaquín Gayoso-Cabada, Miguel Rodríguez-Artacho, María-Dolores Romero-López, Antonio Sarasa-Cabezuelo, and José-Luis Sierra. 2014. Assessing semantic annotation activities with formal concept analysis. *Expert Systems with Applications*, 41(11):5495–5508.

Mariana Curado Malta and Elena González-Blanco. 2016. Postdata. towards publishing european poetry as linked open data. In *International Conference on Dublin Core & Metadata Applications*. DCMI.

Javier de la Rosa, Álvaro Pérez, Laura Hernández, Salvador Ros, and Elena González-Blanco. 2020. Rantanplan, fast and accurate syllabification and scansion of spanish poetry. *Procesamiento del Lenguaje Natural*, 65:83–90.

Quinn Dombrowski and Seth Denbo. 2013. Tei and project bamboo. *Journal of the Text Encoding Initiative*, 5.

Julia Flanders and Scott Hamlin. 2013. Tapas: building a tei publishing and repository service. *Journal of the Text Encoding Initiative*, 5.

Pablo Gervas. 2000. A logic programming application for the analysis of spanish verse. In *International Conference on Computational Logic*, pages 1330–1344. Springer.

Charles O Hartman. 2005. The scandroid 1.1. *Software available at http://oak. conncoll. edu/cohar/Programs. htm*.

Borja Navarro-Colorado, María Ribes Lafoz, and Noelia Sánchez. 2016. Metrical annotation of a large corpus of spanish sonnets: representation, scansion and evaluation. In *International Conference on Language Resources and Evaluation*, pages 4360–4364.

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

Tomás Navarro Tomás. 1991. Métrica española. *Reseña histórica y descriptiva*, 50.

OpenAPI Initiative. 2017. Openapi specification. *Retrieved from GitHub: https://github. com/OAI/OpenAPI-Specification/blob/master/versions/3.0*, 1.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *International Conference on Language Resources and Evaluation*.

Mª Luisa Díez Platas, Salvador Ros Muñoz, Elena González-Blanco, Pablo Ruiz Fabo, and Elena Álvarez Mellado. 2021. Medieval Spanish (12th–15th centuries) named entity recognition and attribute annotation system based on contextual information. *Journal of the Association for Information Science and Technology*, 72(2):224–238. _eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24399.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *International Conference on Language Resources and Evaluation*.