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 existing 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, 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 (Ros and González-Blanco, 2018; Curado Malta and González-Blanco, 2016) 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. 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, 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 (Moretti, 2013; Jockers, 2013), the possibility of automating part of the analysis became very appealing. 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 connects the different tools, and a consumer web-based UI that exposes the functionality to non-experts users. Independently, all tools are released as 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 (Ros and González-Blanco, 2018; Bermúdez-Sabel et al., 2017). Figure 1 shows a diagram of the general architecture of the system.

This granular design allows for each component of the PoetryLab to be used and deployed as a Docker image, 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 great addition, 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 a deployment strategy and the fact that the tools are stateless, allows the

\(^2\)See https://github.com/pyouroboros/ouroboros
\(^3\)See http://postdata.uned.es/poetrylab
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).

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. First, built on top of the industrial-strength natural language processing framework spaCy (Honnibal and Montani, 2017), two Python packages perform scansion and rhyme analysis, and enjambment detection, namely, Rantaplan and JollyJumper. 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 properly handled, we integrated Freeling’s affixes rules via a custom built pipeline for spaCy. The resulting package, spacy-affixes, 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 bisyllabic 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)                | 70.88        |
| (Navarro-Colorado, 2017)      | 94.44        |
| (Agirrezabal et al., 2017)    | 90.84        |
| Rantanplan (ours)             | 96.23        |

Table 1: Percentages of accuracy 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., 2017), 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 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 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 specifying the granularity of the final generated dataset, which is a combined JSON or RDF file containing all the entities in the selected corpora.

---

4 A demo with the Open API user interface is available at [http://postdata.uned.es:5000/ui/](http://postdata.uned.es:5000/ui/).
5 See [https://github.com/linhd-postdata/rantanplan/](https://github.com/linhd-postdata/rantanplan/) and [https://github.com/linhd-postdata/jollyjumper](https://github.com/linhd-postdata/jollyjumper)
6 See [https://github.com/linhd-postdata/spacy-affixes/](https://github.com/linhd-postdata/spacy-affixes/)
7 Gervás’ method was only evaluated on 9,643 verses of the 10,000 verses corpus.
8 See [https://github.com/linhd-postdata/averell/](https://github.com/linhd-postdata/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 seamless 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 Projet RDF triples formats.\(^9\)

![PoetryLab: El Principe constante](image)

**Figure 2:** PoetryLab showing stressed syllables (blue), sinalefas (edian) and enjambments (edian).

---

\(^9\)http://postdata.uned.es/poetrylab
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 persist between sessions and restart. Unfortunately, there is no a user management system implemented which would provide with persistent storage in the backend.

3 Conclusion

Although at an early stage, the PoetryLab has proven useful in that it provides an integrated set of tools for Spanish poetry scholars. It also produces machine readable and interoperable data suitable to be ingested into a triple store compliant with the POSTDATA Project network of ontologies. In fact, this approach is already being tested as we export the analysis of poems and feed them into an Omeka that integrates with the POSTDATA Project network of ontologies.

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.

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.

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 ortoñoja i métrica de la lengua castellana.. la Opinión.

Helena Bermúdez-Sabel, Mariana Curado Malta, and Elena Gonzalez-Blanco. 2017. Towards interoperability in the european poetry community: the standardization of philological concepts. In International Conference on Language, Data and Knowledge, pages 156–165. Springer.

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

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.

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.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings. Convolutional Neural Networks and Incremental Parsing, 7.

Alexander Huber. 2018. Eighteenth-century poetry archive.

Matthew L Jockers. 2013. Macroanalysis: Digital methods and literary history. University of Illinois Press.

Franco Moretti. 2013. Distant reading. Verso Books.

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*.

M Luisa Díez Platas, Mª de los Llanos Tobarra Abad, Salvador Ros Muñoz, Elena González-Blanco García, Antonio Robles Gómez, Agustín Caminero Herráez, and Gimena del Rio Riande. 2017. Hispanic medieval tagger (hismetag): una aplicación web para el etiquetado de entidades en textos medievales. Zenodo.

Salvador Ros and Elena González-Blanco. 2018. Poetry and digital humanities making interoperability possible in a divided world of digital poetry: Postdata project.(presentation).

Pablo Ruiz Fabo, Helena Bermúdez Sabel, Clara Isabel Martínez Cantón, Elena González-Blanco García, and Borja Navarro-Colorado. 2018. The diachronic spanish sonnet corpus (disco): Tei and linked open data encoding, data distribution and metrical findings.

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*. 