



Building infrastructure for annotating medieval, classical and pre-orthographic languages: the Pyrrha ecosystem

Thibault Clérice, Vincent Jolivet, Julien Pilla

► To cite this version:

Thibault Clérice, Vincent Jolivet, Julien Pilla. Building infrastructure for annotating medieval, classical and pre-orthographic languages: the Pyrrha ecosystem. Digital Humanities 2022 (DH2022), Jul 2022, Tokyo, Japan. hal-03606756

HAL Id: hal-03606756

<https://hal.science/hal-03606756>

Submitted on 24 Mar 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Building infrastructure for annotating medieval, classical and pre-orthographic languages: the Pyrrha ecosystem

Thibault Clérice thibault.clerice@chartes.psl.eu

Vincent Jolivet vincent.jolivet@chartes.psl.eu

Julien Pilla julien.pilla@chartes.psl.eu

2021-11-28

1 Introduction

For the past five years, we have been working on the development of infrastructure to build corpora and machine learning models for lemmatisation and morphosyntactic tagging. Ancient and medieval languages with rich morphology and high spelling variation represent a hanging fruit in the domain of these corpora. However, producing “gold” corpora is a tedious and costly task: even when the automatically produced annotations gain in quality thanks to ever more efficient models and vice versa, a significant amount of manual correction and validation work remains.

To reduce the cost and guarantee the interoperability of our corpora, we have built an ecosystem: (1) Pyrrha, a post-correction webapp for lemmatisation and morphosyntactic tags, (2) PyrrhaCI, a continuous integration tool for validating corpora, (3) Protogenie for merging and standardizing sometimes heterogeneous corpora, (4) Pie-Extended, a tagger taking into account the difference between real-world data training corpora and (5) Deucalion, a web service for annotation.

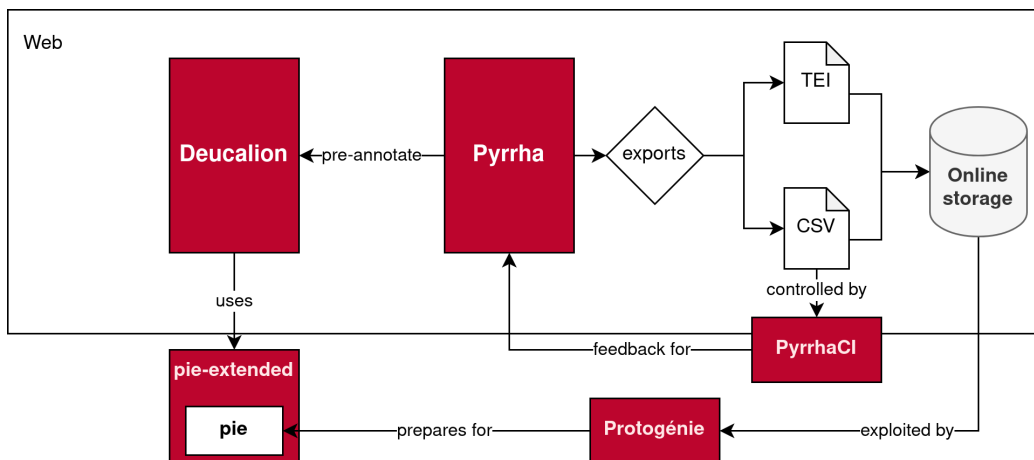


Figure 1: Infrastructure developed at the École nationale des Chartes.

2 Producing data

Pyrrha (Clérice & Pilla, 2021) is designed to accelerate the correction of lemmatisation and morphosyntactic annotation. When we started our work, our team members were using spreadsheets, which have the ability to display all tags and context at the same time. The Pyrrha web application takes up this principle of a tabular interface but adds powerful validation functionalities thanks to checklists (lexicons of lemmas and morphosyntactic tags) guaranteeing the interoperability of the newly produced corpora, as well as batch correction functionalities, inspired by PoCoto (Vobl et al., 2014), a correction interface for OCR. Both of these functionalities are at the core of Pyrrha and have proven to be useful in speeding up the correction of out-of-domain corpora (cf. 2). The application also allows collaboration, both for corpora and checklists, logging of corrections and export to multiple standards such as TEI and TSV.

3 Curating Corpora

While Pyrrha produces data that should be compliant with standard reference sets, mistake happens. *PyrrhaCI* (Clérice et al., 2021) is meant for testing the following attributes of datasets

1. respect of reference sets;

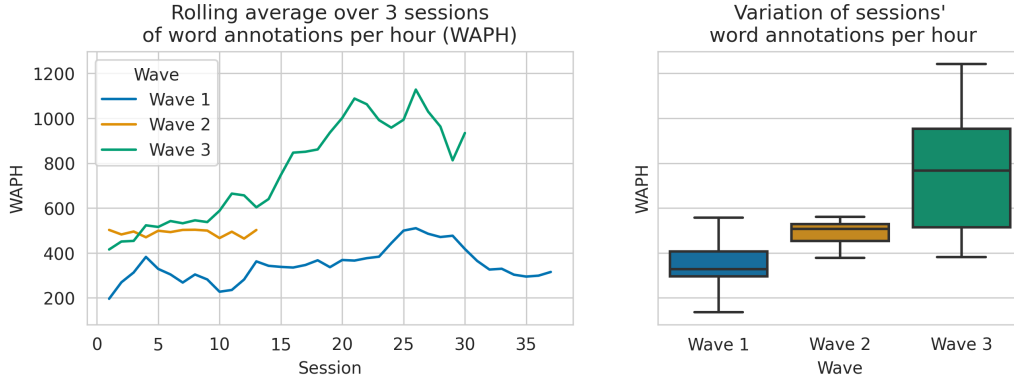


Figure 2: Rolling average of the number of checked tokens per hour. Checking a token includes verifying its correctness and correcting it if necessary. Three waves of correction are visible, the first corpus was completely out-of-domain compared to the lemmatizer (Classical Latin), with the end of the corpus being very different from the rest of the corpus in terms of spelling (letters K and W appeared), themes and syntax. The second and third wave benefits of a new model, retrained on the data produced in wave 1: as a result, there is less corrections and a faster checking rate on waves 2 and 3. Wave 3 is composed of two blocs: one is Thomas More’s *Utopia* (beginning of W3) and the other the *Legenda Aurea* which was nearly 100% correct, hence the effectiveness. Each wave / corpus was corrected and proofread on all categories that Pyrrha allows: Lemma, POS and morphological tags.

2. cross-categorical annotations (*e.g.* POS(dog)!=Verb);
3. n-gram tagging (*e.g.* ADJ should not be followed by VERcon).

Each test failure can be manually ignored for further tests, allowing for a more agile interpretation of grammar. PyrrhaCI is meant to be used as a continuous integration tool, through Github Actions or TravisCI, to validate datasets in open repositories and track the issues raised by editions.

Protogenie (Clérice, 2020b) is focused on preparing datasets for training. It is meant for the following:

1. keeping track of and using the same original train/dev/test splits while adding new data in order to have “uniform” evaluation,
2. allowing for normalization of datasets that come from different projects in different formats,
3. adding transformation to the original dataset (while respecting (1)), such as removing the distinction between U and V in Latin, replacing labels, splitting multi-categorical tags, etc.

While (1) is easily taken care of, it is, in our experience, common to find datasets with different formatting choices or data-based variations such as punctuation, capitalization, morphological tags. Protogenie enables normalization of the “behavior” of different corpora, without having to work with pre-modified files, facilitating easy update of the latter and ensuring the stability of training and performances evaluation.

4 Producing new data: our lemmatization pipeline

Our models are trained with Pie (Manjavacas et al., 2019)¹, a lemmatizer with state-of-the-art results on pre-orthographic and morphologically rich languages and a relatively flexible and stable python API. Once trained, our models are served through Pie-Extended. Its first function is to bridge the gap between the real-world data and the curated training data by normalizing

¹We moved to a small fork of Pie, <https://github.com/lascivaroma/PaPie>, which includes tailored functionalities for our training sets.

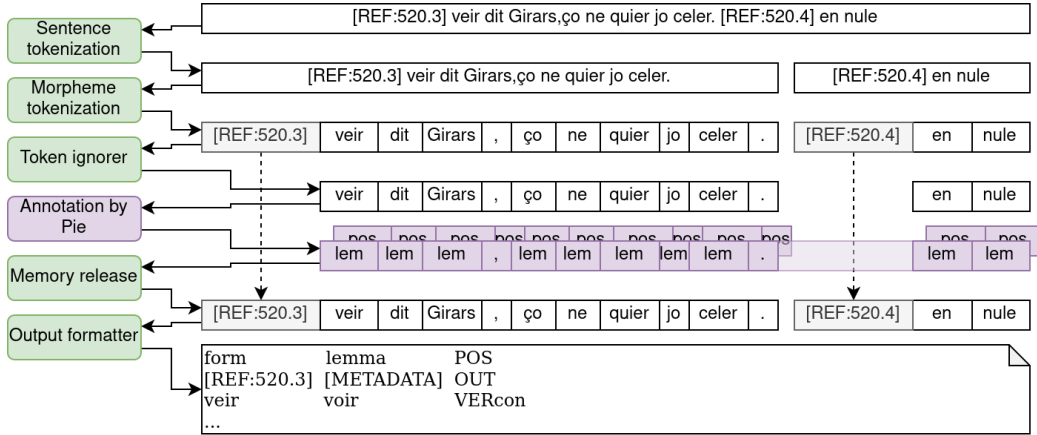


Figure 3: Steps for token pass-through in Pie-Extended.

the first according to the latter². It also provides features such as token³ passthrough (*cf.* 3).

Finally, as not everyone knows how to install and run a python program in a shell, we produced the Deucalion interface (Clérice, 2020a), meant both for documenting (with complete bibliography for each model and software) and tagging. It is a software layer allowing to use Pie-Extended on the web. This Deucalion interface can be used as a stand-alone web application or an API.

5 Conclusion

Over the last five years, this infrastructure has allowed us to build a 1+ million token corpus of Old French (Camps et al., 2021), a couple of datasets in both classical and late Latin (Clérice, 2021; Glaise & Clérice, 2021), one for pre-orthographic Early Modern French (Gabay et al., 2020), and others. There are improvements we would still like to make (*e.g.* the user-friendliness and capacities of PyrrhaCI) and we now have our eyes on *data valorization*,

²*E.g.* in our Latin dataset shared with us by the LASLA, there was no punctuation. Unknown character can trigger weird behaviors in neural networks system, from our experience, creating issues for both the context of other lemma and its own lemmatization.

³*e.g.* metadata token with text identifiers.

through the reuse of tools such as Blacklab (de Does et al., 2017)⁴. Pyrrha has made lemmatization easier for our collaborators and made it a simpler task to produce data and share them across projects. This paper will be an opportunity to present a proven ecosystem, and also to assess its benefits, its costs and shortcomings.

⁴A demo for Latin is available at <https://blacklab.alpheios.net/latin-texts/search> thanks to Alpheios.

References

- Camps, J.-B., Clérice, T., Duval, F., Ing, L., Kanaoka, N., & Pinche, A. (2021).
Corpus and models for lemmatisation and pos-tagging of old french.
- Clérice, T. (2020a).
flask_pie, a pie-extended wrapper for flask (Version 0.1.0).
https://github.com/hipster-philology/flask_pie
- Clérice, T. (2020b). *Protogenie, post-processing for nlp dataset*. Zenodo.
<https://doi.org/10.5281/zenodo.3883585>
- Clérice, T. (2021).
Lemmatisation et analyse morpho-syntaxique des Priapées.
<https://github.com/lascivaroma/priapea-lemmatization>
- Clérice, T., Blotière, É., & Schmied, M.-C. (2021). *pyrrhaCI* (Version 0.1).
<https://github.com/hipster-philology/pyrrhaCI>
- Clérice, T., & Pilla, J. (2021). *Pyrrha* (Version 3.0.0).
<https://doi.org/10.5281/zenodo.2325427>
- de Does, J., Niestadt, J., & Depuydt, K. (2017).
Creating research environments with blacklab.
CLARIN in the Low Countries, 245–258.
- Gabay, S., Clérice, T., Camps, J.-B., Tanguy, J.-B., & Gille-Levenson, M. (2020). Standardizing linguistic data: Method and tools for annotating (pre-orthographic) french. *Proceedings of the 2nd International Conference on Digital Tools & Uses Congress*, 1–7.
- Glaire, A., & Clérice, T. (2021). Du IIème siècle à Thomas More, un corpus gold de latin lemmatisé et annoté en morpho-syntaxe.
<https://doi.org/10.5281/zenodo.1234>
- Manjavacas, E., Kádár, Á., & Kestemont, M. (2019). Improving lemmatization of non-standard languages with joint learning. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1493–1503.
<https://doi.org/10.18653/v1/N19-1153>
- Vobl, T., Gotscharek, A., Reffle, U., Ringlstetter, C., & Schulz, K. U. (2014). Pocoto-an open source system for efficient interactive postcorrection of ocred historical texts. *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 57–61.