AcTo: How to Build a Network of Integrated Projects for Medieval Occitan

Why Medieval Occitan?
While being the ancestor of Modern Occitan, a modern language spoken by minorities in France, Spain (Catalonia), and Italy, Medieval Occitan is also, and crucially, the language of a corpus of texts fundamental for the pre-modern cultural history of Europe. Indeed the corpus of Old Occitan literature, and especially the texts of the Troubadours have had a great influence in the development of Moden European literature and beyond.

AcTo https://acto.hypotheses.org/

AcTo stands for *Acolhir e Tornar*, ``to welcome and return", and is a network of data and resource centers for the study of Medieval Occitan headed by Université Paul-Valéry in Montpellier, France, which gathers together projects from different countries (France, Italy, Spain, Germany, UK).

The aim of the project is to federate existing resources (**digital editions, lexicons, but also tools**), harmonising the data and metadata encoding within the projects as well as with international standards

"Per solatz revelhar,
qu'es tròp endormitz,
et per Pretz, qu'es faiditz,
acolhir e **to**rnar,
me cudèi trabalhar"
Guiraud de Bornèlh

List of AcTo projects

AcTo objectives
- ❏ Alignment of metadata and documentation in collaboration with CIRDOC and Occitanica.eu
- ❏ Annotation and referencing of place and persons' names in digital edition
- ❏ Copyright and legal issues
- ❏ **Orthographic normalisation and lemmatisation across projects**

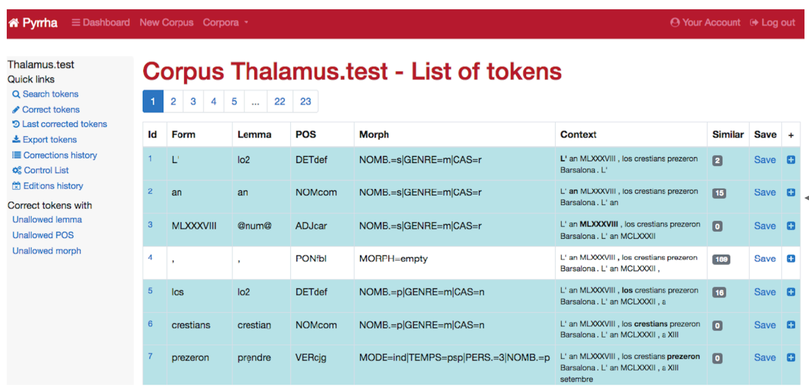
Current work on Orthographic Normalisation and Lemmatisation

Medieval Occitan orthography was not standardised. Digital editions, while preserving the verbatim transcription, should also allow for search by normalised forms. Harmonising the normalisation as well as the lemmatisation choices is a crucial pre-requisite for a federated search throughout all existing corpora.

Current experiment on automatically analysing the text based on
- ❏ Rule based methods
- ❏ Machine learning methods re-training models derived from Old French

Le Petit Thalamus de Montpellier

Link each token in the TEI digital edition to the reference lexicon, Dictinnaire d'Occitan Médieval (DOM).



OMÉLiE project
Outils et méthodes pour l'édition linguistique enrichie

- ❏ Use of a post-correction environment to create a training and test corpus

AcTo and CLARIN

- ❏ Better metadata
- ❏ Visibility of resources in the VLO
- ❏ Federated content search
- ❏ Use of best practices and standards (TEI, LMF, ...)
- ❏ Collaboration for the development of NLP pipelines

Gilda Caïti-Russo Laboratoire LLACS Univ Paul-Valéry Montpellier
Jean-Baptiste Camps Centre Jean-Mabillon École nationale des chartes Université PSL, Paris
Gilles Couffignal Université Paris-Sorbonne
Francesca Frontini Laboratoire PRAXILING Univ Paul-Valéry Montpellier
Hervé Lieutard Laboratoire LLACS Univ Paul-Valéry Montpellier
Elisabeth Reichle Ludwig Maximilian University of Munich
Maria Selig Universität Regensburg