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

Gilda Caïti-Russo, Jean-Baptiste Camps, Gilles Couffignal, Francesca Frontini, Hervé Lieutard, Elisabeth Reichle, Maria Selig

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

Gilda Caïti-Russo
Laboratoire LLACS
Univ Paul-Valéry Montpellier 3
gilda.russo@univ-montp3.fr

Jean-Baptiste Camps
Centre Jean-Mabillon
École nationale des chartes
Université PSL, Paris
jean-baptiste.camps@chartes.psl.eu

Gilles Couffignal
Université Paris-Sorbonne
gilles.couffignal@paris-sorbonne.fr

Francesca Frontini
Laboratoire PRAXILING
Univ Paul-Valéry Montpellier 3
francesca.frontini@univ-montp3.fr

Hervé Lieutard
Laboratoire LLACS
Univ Paul-Valéry Montpellier 3
herve.lieutard@univ-montp3.fr

Elisabeth Reichle
Ludwig Maximilian University of Munich
Elisabeth.Reichle@dom.badw.de

Maria Selig
Universität Regensburg
Maria.Selig@sprachlit.uni-regensburg.de

Abstract

We present AcTo, a network of integrated projects for the development of language resources and tools for Medieval Occitan. This abstract illustrates the resources in the network, as well as the first steps towards their integration, aiming towards the harmonisation and interoperability of NLP and lexical resources for the annotation of digital editions.

1 Introduction

Computational linguistics methods and digital language resources are becoming more and more important for philology. Computational philology approaches and infrastructures develop and adapt tools and methods specifically for the needs of scholars working with historical languages, and the development of computational corpora and lexicons is flourishing in this domain (Crane, 2012; Passarotti et al., 2019). Medieval philologists are not lagging behind digital classicists in the development of new approaches and solutions, with successful experiments in the application of OCR and textual analysis techniques to their manuscripts (Pinche et al., 2019). We concentrate here on computational philology approaches for Medieval or Old Occitan. While being the ancestor of 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 [1]. As proof of this, Medieval Occitan is taught and studied at academic level in many European and American universities, and beyond.

In this abstract we illustrate the first activities of AcTo[2], a network of data and resource centers for the study of Medieval Occitan headed by Université Paul-Valéry in Montpellier, France, which gathers

[1] It would be impossible to trace here the influence of Occitanism in literature. Troubadours have been dubbed the inventors of modern verse (Wilhelm, 1970) and their influence has extended to contemporary American authors such as Ezra Pound and more recently W.S. Merwin, poet winner of the Pulitzer price for poetry in 2008.

[2] AcTo stands for *Acolhir e Tornar*, "to collect and return", a line from troubadour Guiraud de Bornèlh.

together projects from different countries (France, Italy, Spain, Germany, UK)[3]. 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[4]. During a first project meeting, several working groups were constituted, dedicated to the alignment of metadata and documentation[5], to the annotation and referencing of place and persons' names in digital editions, and to legal issues. The project plans to draw experience and help from CLARIN, Huma-Num and other infrastructures in order to align itself to existing best practices.

Within this framework, one specific effort, which constitutes the main object of this abstract, concerns orthographic normalisation across projects. This is a particularly important issue, since 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. Here we shall illustrate ongoing activities which have the aim of automatically linking a lexicographic resource to an existing corpus by means of an ad hoc morphological analyser, which automatically reconducts non standard forms to the normalised lemma.

2 The resources

2.1 The Thalamus project and corpus

The Thalamus ANR project carried out the digitising and TEI encoding of the manuscript corpus of the government books of the medieval city of Montpellier. The critical digital edition is available online (Carrasco et al., 2014…), with the various manuscripts displayed in parallel, aligned by year, something which allows scholars to investigate how successive chronicles have re-written and edited past events of the city in the light of contemporary matters. This synoptic edition makes it possible for scholars to study the diachronic evolution of pre-diglossic Occitan from 1260 to 1426 as no other document can do. So far the normalisation and annotation of the text has been limited to place and person names, which are searchable from two dedicated indices, independently from their written form, which may vary. The current objective is to implement a search by forms and lemmas, in order to manage and study graphical variation. For this reason we are currently looking into making the TEI Thalamus corpus, the Medieval Occitan dictionary (*Dictionnaire de l'occitan médiéval*, see 2.2) and the OMÉLiE project (2.3) all interoperable.

2.2 The DOM, a reference lexicon for Medieval Occitan

The *Dictionnaire de l'occitan médiéval* (DOM), is a project coordinated by the Bayerische Akademie der Wissenschaften. It is a reference lexicographic resource for Medieval Occitan philologists. Based on PostgreSQL, the DOM is available online (Stempel et al., 1996…). The lexical entries, completed with bibliographic references, list the lemmas and all of their variants, the (polysemic) meanings, and a list of attestations. The dictionary provides a separated alphabetic list of lemmas and variants, so that search is also possible by all of the variants. The articles are connected by hyperlinks to the *Französisches Etymologisches Wörterbuch* (FEW) for further etymological research (Wartburg, 1922 1967). Linking editions to the DOM has been identified as an important task in the overall goal of federating the various digital editions projects within AcTo. The DOM will provide the necessary lexicographic information for normalisation and lemmatisation of Old Occitan texts when digitalised and prepared for annotation. DOM entries are provided with a unique URI, which could be used as a unique lemma reference by digital editions of texts. The DOM lexicon therefore offers the opportunity of creating a platform for access, analysis and interpretation of Old Occitan digitised texts. It facilitates cross-collection search inside the Old Occitan corpus and offers the possibility to make use of the existing network of lexicographic sources provided by

[3] See the project's web site for the complete list, as well as descriptions of individual projects and proceedings of past meetings (Caiti-Russo et al., 2019).

[4] The project focuses on the Medieval stage of the Occitan language, but in a diachronic perspective the relationship to modern Occitan is crucial; AcTo is supported by AIEO (Association Internationale d'études Occitanes) and by the CIRDÒC (Centre Interrégional de développement de l'occitan), which maintains a repository, `Occitanica.eu`, hosting a number of language resources for the Occitan language and will be eventually related to *Lo Congrès*, the most important repository of language resources for Modern and Contemporary Occitan (`https://www.locongres.org`).

[5] Discussions over these aspects will be headed by the CIRDÒC, a partner in AcTo; notice that the `Occitanica.eu` datacenter is already harvested by Europeana; harvesting to the VLO of relevant resources is currently being considered.

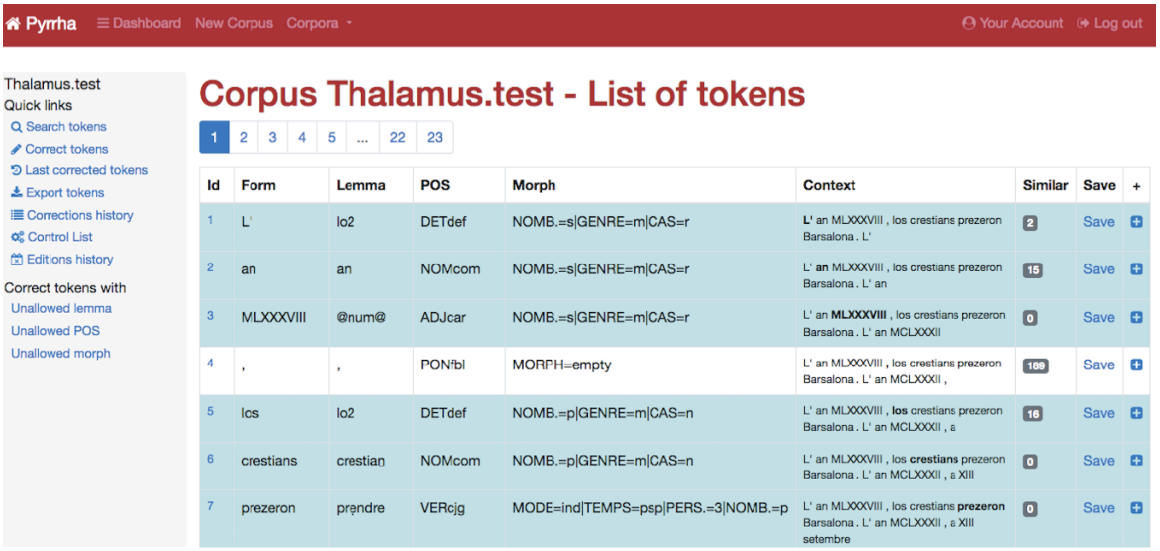

Figure 1: The Pyrrha post-correction interface.

the DOM. In the long run, the aim is to create a hybrid system of lexicographic devices and electronic corpora ("*Digitales lexikalisches System*") hosted by the Bayerische Akademie der Wissenschaften/Leibniz Rechenzentrum.

2.3 The OMÉLiE project

OMÉLiE (*Outils et méthodes pour l'édition linguistique enrichie*) is a project of the *École des chartes* in Paris, with support from SCRIPTA (Université PSL) and the DIM *Sciences du texte et connaissances nouvelles* (Région Île-de-France). Its objectives are to offer tools and methods for the linguistic enrichment and analysis of ancient and medieval texts. Currently the research concentrates on Old French and Occitan. The aim is to offer an environment in which TEI editions can be uploaded, automatically lemmatised and annotated with morphosyntactic tags (Part-of-Speech, morphological analysis) and later post-corrected by humans to produce better models. The annotation system is based on deep learning methods, and in particular uses the `Pie` tagger (Manjavacas et al., 2019). It is integrated in a post-correction environment, Pyrrha (Clérice et al., 2019), that allows for close inspection as well as batch corrections, and can handle reference lists of lemmas and tags (see Figure 1). Both are available as open source software. This environment, which had initially been tested on Medieval French corpora, has now been applied to Medieval Occitan texts such as the romance of *Flamenca* and, crucially, the *Thalamus*. Pyrrha could easily be extended to support more recent varieties of Occitan, by ensuring interoperability with the CORLIG project (*Corpus de la Renaissance Littéraire gasconne*) coordinated by the Sorbonne University in Paris.

3 The lemmatisation project

The current project aims at bringing together DOM and OMÉLiE for an improved lemmatisation of the Thalamus. A first lemmatisation strategy has been developed, which is based on *LemmaGen* (Juršic et al., 2010) learning lemmatisation rules from existing lemma-wordform pair examples extracted from the DOM articles. In order to improve on that, and to provide for the full morphological analysis and lemmatisation of word forms, an annotation campaign is currently being carried out by the Thalamus and the OMÉLiE teams, to create an annotated corpus using Pyrrha. The annotation is performed by correcting the output of a first basic model, and will serve as training and test set for the creation of a better one. Following the annotation guidelines, first the lemmatisation is corrected, strictly following the DOM orthography; missing lemmas are recorded and set aside for their integration in the DOM; then the morphosyntactic annotation of the token in context is carried out, using the Cattex tagset (Prévost et al., 2013;

Guillot et al., 2013). The poster presentation will show the first results of the lemmatisation model, and show how the link between the digital edition and the DOM can be encoded in the TEI edition.

4 Future work

Due to its influence that goes well beyond the borders of historical Occitania and modern day France, Medieval Occitan can be seen as part of a shared European heritage. For this reason we intend to integrate the AcTo community within the activities of CLARIN ERIC, as well as those of various national consortia, in order to ensure the visibility and interoperability of our digital resources as well as to exploit and adapt existing solutions and technologies. Future objectives of AcTo are:

- to make TEI editions of the whole Troubadour corpus in order to render lemmatisation and morphosyntactic annotation possible for a larger corpus until exhaustivity is reached,
- to develop a cartography of Medieval Montpellier (from the Thalamus), and more generally a cartography of the Troubadour space in Europe,
- to ensure the alignment between the Medieval Occitan lexical resources and their modern and contemporary counterparts.

References

[Caiti-Russo et al.2019] Gilda Caiti-Russo, Francesca Frontini, and Hervé Lieutard. 2019. Acolhir e Tornar – AcTo: Ressorsas numericas per l'occitan medieval [carnet de recherche]. `https://acto.hypotheses.org/`.

[Carrasco et al.2014…] Raphaël Carrasco, Vincent Challet, Gilda Caïti-Russo, Stéphane Durand, Marc Conesa, Yves Mausen, Daniel Le Blévec, Chantal Wionet, and Florence Clavaud, editors. 2014/…. *Le « Petit Thalamus » de Montpellier: édition critique numérique du manuscrit AA9 des Archives municipales de Montpellier dit Le Petit Thalamus*. Université Paul Valéry Montpellier-III, Montpellier. `http://thalamus.huma-num.fr/`.

[Clérice et al.2019] Thibault Clérice, Julien Pilla, and Jean-Baptiste-Camps. 2019. hipster-philology/pyrrha: 2.0.0. `https://doi.org/10.5281/zenodo.2541730`.

[Crane2012] Gregory Crane. 2012. The Perseus Project. In *Leadership in Science and Technology: A Reference-Handbook*, pages 644–652. SAGE Publications, Thousand Oaks.

[Guillot et al.2013] Céline Guillot, Sophie Prévost, and Alexei Lavrentiev. 2013. *Manuel de référence du jeu Cattex09*. École normale supérieure de Lyon, Lyon. Version 2.0. `http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_manuel_2.0.pdf`.

[Juršic et al.2010] Matjaz Juršic, Igor Mozetic, Tomaz Erjavec, and Nada Lavrac. 2010. Lemmagen: Multilingual lemmatisation with induced ripple-down rules. *Journal of Universal Computer Science*, 16(9):1190–1214.

[Manjavacas et al.2019] Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 2019. Improving lemmatization of non-standard languages with joint learning. *CoRR*, abs/1903.06939.

[Passarotti et al.2019] Marco Passarotti, Flavio Massimiliano Cecchini, Greta Franzini, Eleonora Litta, Francesco Mambrini, and Paolo Ruffolo. 2019. LiLa: Linking Latin. Building a Knowledge Base of Linguistic Resources for Latin.

[Pinche et al.2019] Ariane Pinche, Jean-Baptiste Camps, and Thibault Clérice. 2019. Stylometry for Noisy Medieval Data: Evaluating Paul Meyer's Hagiographic Hypothesis. In *Digital Humanities Conference 2019 - DH2019*, Utrecht, Netherlands. ADHO and Utrecht University.

[Prévost et al.2013] Sophie Prévost, Céline Guillot, Alexei Lavrentiev, and Serge Heiden. 2013. *Jeu d'étiquettes morphosyntaxiques CATTEX2009*. École normale supérieure de Lyon, Lyon. version 2.0. `http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_2.0.pdf`.

[Stempel et al.1996…] Wolf-Dieter Stempel, Maria Selig, Claudia Kraus, Renate Peter, and Monika Tausend. 1996/…. *Dictionnaire de l'occitan médiéval (DOM en ligne)*. Bayerische Akademie der Wissenschaften, Munich. `http://www.dom-en-ligne.de/`.

[Wartburg1922 1967] Walther von Wartburg. 1922–1967. *Französisches Etymologisches Wörterbuch: eine Darstellung des galloromanischen Sprachschatzes*. ATILF, Leipzig. `https://apps.atilf.fr/lecteurFEW/`, *eFEW: FEW informatisé*, ed. Pascale Renders.

[Wilhelm1970] James J. Wilhelm. 1970. *Seven Troubadours: Creators of Modern Verse*. Pennsylvania State University Press, University Park.