# Building an integrated workflow for reviving a minority culture with Wikimedia

Luís Trigo
CODA [1] - FLUP [2]

Carlos Silva
CLUP [3] - FLUP[2]

Ana Afonso
FLUP [2]

## Abstract

In this research, we aim to develop an integrated method to revitalize a minority language community through the use of several Wikimedia projects and tools. Mirandese is our case study. The main target audience will be language students and native, heritage and new speakers. We will promote sharing resources with Wikimedia projects from the same linguistic family in the Iberian Peninsula, with an emphasis on Asturian. We will work closely with Wikimedia Portugal and in cooperation with Wikimedia Spain. We also expect to engage with other minority languages from Portuguese-speaking countries.

## Introduction

The linguistic description of the Mirandese language and community engagement are long-standing problems. This project aims at promoting academic engagement, filling this gap, and benefit:

- Institutional minority language preservation projects of Mirandese;
- Other institutional minority language preservation projects;

More specifically, we seek to answer the questions:

- What impact does a Mirandese Wikimedia have on language conversation?
- How do Wikimedia projects contribute to language description, conservation, and prestige?

**Date**: June 1, 2024 - May 31, 2025.

## Related work

All team members have extensive experience in language documentation. Luís Trigo and Carlos Silva have created and maintained an open dictionary of Sri Lanka Portuguese [4]. They also study phoneme frequency for language description and automatic identification. Carlos Silva worked as a linguist at the Galician and Portuguese word bank [5]. Marta Afonso is a heritage speaker of Mirandese, a Mirandese language teacher at FLUP, and a translator from Portuguese and French to Mirandese.
This collaboration is ideal to tackle the main goals of this project:

- Create a scientific grammar of Mirandese;
- Free resources about this language through Wiki projects and tools.

These goals are in line with previous work on comparative Iberian and creole linguistics, and with partnerships that are already in place, e.g. Academia de la Llingua Asturiana [6].

## Methods

This project comprises the description and conservation of the Mirandese language and culture. Several researchers collected data but they had little or no impact in the community. This project differs from the previous ones by putting in place a cohesive workflow that evolves data collection in a structured way, tidying previous resources, and making everything available through Wikimedia projects and tools. Thus, we intend to develop the following tasks:

- Data collection: We start by assessing and extracting the data within Asturian and Portuguese Wiktionary, Wikidata, Lexemes (using Lingua Libre, Listeria, and Ordia), Commons, and Wikipedia. After tyding this data, we will identify the gaps to be filled through the surveys.
- Surveys: Words and expressions that are missing in the Wikimedia projects previously mentioned will be collected in the Mirandese community through elicitation.
- Classes/workshops: To test the quality of the data and to identify more gaps, we plan a series of workshops in the community, as well as classes on Mirandese and comparative Ibero-Romance linguistics.
- Translations: We take a comparative approach involving Asturian phonology and Portuguese orthography to enable a first model of automatic translation of Mirandese. We will also build parallel corpora with Asturian and Portuguese - from Wikipedia.

## Expected output

The expected outputs include:

- I: Insights to inform decision making (L2 didactic strategies from phoneme frequency, rule ordering and pairing, cognate assessment).
  - Audience: Language teachers
- II: Scientific publications (comparative linguistics, phonology, NLP)
  - Audience: Scientific community
- III: Datasets: lexicon (Wikidata Lexemes and Wikitionary) and corpora (Wikipedia articles)
  - Audience: general public
- IV: Events (Classes and workshops, Porto Meeting - networking with other minority languages, Editathons, Wiki Takes X)
  - Audience: students, researchers, people interested in Mirandese culture and Mirandese community.

## Risks

The fact that most contemporary materials about Mirandese are under closed licenses is a challenge because we cannot introduce them directly into WikiSource and Wiki Commons.

## Community impact plan

We are already actively working with scholars and students, not only from language studies but also from other social sciences. We are incorporating Wikimedia platforms as a reference in the research data planning at FLUP. Beginning in the next semester, we will promote regular intercultural editathons with the support of Wikimedia Portugal. This grant would enable us to focus and develop effective workflows, methods, and case studies that would greatly increase our impact and activity level regarding the use of Wikimedia platforms.

Beyond the academic community and Mirandese speakers, we expect to reach minority languages from Portuguese-speaking countries and Iberian-related languages.

## Evaluation

We will evaluate the success of the projects' implementation through the number of:

- Scientific publications;
- Participation in international conferences;
- Wikimedia edits (Lingua Libre, Wikipedia articles and lexemes);
- Participants in the proposed events.

## Budget

The budget would pay for a full academic year of an MSc student in the dissertation phase and four-month pay for a Ph.D. researcher who would advise and support this student, and further develop linguistic and pedagogical studies.

PhD advisor: 4 x 2414.90 euros = 9659.6 euros = 10625.56 dollars
MsC student: 12 x 1549.00 euros = 18588 euros = 20446.8 dollars
Travels: 1000 euros = 1100 dollars
Overheads (15%): 4387.14 euros = 4825.85 dollars

## Prior contributions

The research team for this project has already developed some relevant research work and literature about Mirandese and other minority languages, e.g.:

- Gramaticografia do Mirandês [7]
- Palatals frequency through Wiktionary [8][9]
- Wikcionário IPA fixing - tba
- CreoPhonPT [10]
- L Princepico [11] (Mirandese translation of "The Little Prince")

Carlos Silva and Luís Trigo have closely worked to integrate student and community contributions in Phonology and Romance Linguistics classes from the Bachelor's and Master degree curricula [12] [13]. These were the basis for the tutorial [14] that was accepted in PROPOR 2024 [15]. Their work was also distinguished by the prize "Innovative Pedagogical Practice" [16]. All team members have participated in, assisted, and hosted some Wikimedia Portugal events, including the Porto Meeting [17] and the Catalan Wikimarathon [18]. We are also supporting Wikimedia Portugal in the formation of local minority language Wikipedia editing groups, that, for now, include East Timor students (Tetum). Lingua Libre was also highlighted in the Phonology I from the Linguistics MSc course (Carlos Silva) and used in Mirandese classes (Ana Afonso) at FLUP.

## References

[1] CODA - Centre for Digital Culture and Innovation : https://coda.letras.up.pt/
[2] FLUP - Faculty of Arts and Humanities of the University of Porto: https://letras.up.pt/
[3] CLUP - Centre of Linguistics of the University of Porto: https://clup.pt/
[4] PtLanka: https://github.com/Portophon/PtLanka.
[5] Álvarez, R. (2023). Tesouro do léxico patrimonial galego e portugués. Santiago de Compostela: Instituto da Lingua Galega. http://ilg.usc.es/Tesouro
[6] Academia de la Llingua Asturiana https://alladixital.org/
[7] Silva, C. S. (2019). A gramatização do mirandês: estudo sobre a gramaticografia de uma língua minoritária. Linred: Lingüística en la Red, (17), 25.
[8] Trigo, L., & Silva, C. (2022, March). Comparing lexical and usage frequencies of

palatal segments in portuguese. In International Conference on Computational Processing of the Portuguese Language (pp. 353-362). Cham: Springer International Publishing.

[9]Silva, C., Trigo, L. & (2024, March). Frequency, overlap and origins of palatal sonorants in three Iberian languages. In International Conference on Computational Processing of the Portuguese Language. Cham: Springer International Publishing. - Accepted paper - to be published

[10]Silva, C. S., & Trigo, L. (2023). CreoPhonPt: a collaborative database saving Portuguese creoles from digital obliteration. Digital humanities 2023: collaboration as opportunity: Book of Abstracts.

[11] L Princepico (translated by Ana Afonso): https://www.cm-mdouro.pt/pages/246?poi_id=230

[12] Trigo, L., Silva, C. S., Almeida, V. M. D., & Marques, D. (2023). People first-testing integrated digital research/teaching concepts from the ground up (CT). Programming and data infrastructure in digital humanities: book of abstracts.

[13] Silva, C. S., Trigo, L., Pichel, J. R., & Almeida, V. M. D. (2023). A linguística comparativa ibérica na sala de aula com recurso a métodos de investigação digital. In 44th ACIS Conference: book of abstracts.

[14] Silva, C. S., Pichel, J. R., Granja, F. & Trigo, L. (2024) Automatic measurement of distances between languages using Swadesh lists and big text corpora. In International Conference on Computational Processing of the Portuguese Language. - Tutorial accepted

[15] Call for workshops, tutorials & shared tasks: https://propor2024.citius.gal/index.php/call-for-workshops-proposals/

[16] Prémio "Prática Pedagógica Inovadora" 2023 distingue 4 docentes e investigadores FLUP: https://sigarra.up.pt/flup/pt/noticias_geral.ver_noticia?p_nr=161866

[17] Porto Meeting - Wikimedia minority languages meeting: https://meta.wikimedia.org/wiki/Porto_Meeting_2023

[18] Catalan Wikimarathon: https://sigarra.up.pt/flup/en/noticias_geral.ver_noticia?p_nr=167586