CLingS: Cross-lingual information retrieval for scientific datasets in less-resourced languages

Valentina Fedchenko,
Perrine Quennehen
INALCO / Paris
name.surname
@inalco.fr

Ka-I Lim
NAER / Taipei
liz462
@mail.naer.edu.tw

Milan Rusko
SAS / Bratislava
milan.rusko
@savba.sk

Abstract

This document presents an initial overview of the CLingS project, currently in its early development stage. It outlines a collaborative effort to build a cross-lingual information retrieval platform for scientific literature in underrepresented languages. The project CLingS aims to develop datasets, tools, and methods to improve multilingual access to scientific knowledge.

1 Project Overview

Project Title: CLingS: Cross-lingual information retrieval for scientific datasets in less-resourced languages Languages.[1]

Project Reference: ANR-25-CHR4-0003

Project Duration: January 2026 – December 2028

Partners and Funding Agencies:

- **National Institute for Oriental Languages and Civilizations (INALCO)**. Agence National de la Recherche, Paris, France (Project Coordinator). PI: Valentina Fedchenko
- **Institute of Informatics, Slovak Academy of Sciences**. Slovak Academy of Sciences, Bratislava, Slovakia. PI: Milan Rusko
- **Constantine the Philosopher University in Nitra** (non-funded partner). Nitra, Slovakia. PI: Martin Diweg-Pukanec
- **National Academy for Educational Research**. National Science and Technology Council, Taipei, Taiwan. PI: Ka-I Lim

Built as a CHIST-ERA ERA-NET consortium (European coordinated research on long-term ICT and ICT-based scientific Challenges), the CLingS project brings together academic partners with complementary expertise in natural language processing, speech and language technologies, linguistics, and educational research.

2 Project Objectives

The CLingS project aims to improve access to scientific knowledge in languages that remain underrepresented in digital infrastructures and natural language processing tools (Ranathunga and de Silva, 2022), (Helm et al., 2023). These languages (to be described in Section 3) present diverse situations. Some are associated with long-standing scientific traditions and substantial bodies of scholarly texts, which, however, remain difficult to access in digital form, as Yiddish. Others, such as Taiwanese, have only a limited amount of scientific literature available today.

In this context, the project not only seeks to improve access to existing resources, but also, where necessary, to support the development of new ones. In the case of Taiwanese, one objective is to foster the creation of scientific content through a combination of AI-based methods and human expertise.

To address this, the project will explore several strategies, depending on the language: leveraging the best available multilingual pretrained language models (Hedderich et al., 2020); continuing pretraining on curated scientific corpora (Micallef et al., 2022); applying data augmentation techniques (Marivate et al., 2020); and potentially employ-

[1] https://www.inalco.fr/en/cross-lingual-information-retrieval-scientific-datasets-less-resourced-languages-clings

ing transfer learning from related, slightly better-resourced languages within the set (e.g., transfer from Chinese to Taiwanese for our project; (Tars et al., 2021), (Heffernan et al., 2022)).

The main objective of CLingS is to develop a cross-lingual information retrieval platform that enables users to search scientific literature written in several moderately resourced languages using natural-language queries. The system will return answers grounded in scientific sources and accompanied by precise bibliographic references.

To achieve this goal, the project pursues several complementary objectives. First, it will build and annotate specialized scientific corpora across multiple academic domains, including linguistics and philology, medicine, mathematics, geography, and law. These corpora will serve as the foundation for training and evaluating language technology models.

Second, the project will develop cross-lingual information retrieval methods capable of identifying relevant documents and passages across languages. These methods will rely on modern neural architectures and embedding-based representations of scientific texts.

Third, CLingS will design terminological alignment tools that can automatically identify correspondences between scientific terms across languages, helping to bridge lexical and conceptual gaps between different scientific traditions.

Finally, the project will implement a user-centered evaluation framework involving domain experts and linguists, who will assess the usefulness, accuracy, and usability of the system in realistic research scenarios.

3 Project Languages

The CLingS project focuses on seven languages: Belarusian, Estonian, Punjabi, Slovak, Tâigí (Taiwanese), Ukrainian, and Yiddish. These languages differ significantly from a typological perspective, including their writing systems, morphological structures, and the degree to which they are represented in current language technology models.

Despite their diversity, they share an important common characteristic: their use as languages of scientific communication remains relatively underdeveloped. For historical and sociolinguistic reasons, these languages have long existed in environments dominated by larger scientific languages. In the case of Punjabi, English has played this dominant role, while in the former Soviet space Russian historically dominated scientific communication for languages such as Belarusian, Ukrainian, and partly Estonian. In Taiwan, the development of scientific discourse in Taiwanese has been overshadowed by the dominance of Mandarin Chinese.

4 Expected results

The main outcome of the CLingS project will be a platform for cross-lingual exploration of scientific literature, enabling users to access and navigate scientific publications written in multiple languages that currently remain difficult to search.

Technologically, the project will produce a set of advanced language technology tools, including multilingual scientific embeddings, dense retrieval models adapted to scientific domains, and methods for cross-lingual terminology alignment. These tools will facilitate the discovery of relevant documents and passages even when queries and source texts are written in different languages.

Another important result will be the creation of curated and annotated scientific corpora in several moderately resourced languages.

The project will also generate terminological graphs linking scientific concepts across languages, enabling more precise mapping between scientific vocabularies and improving cross-lingual knowledge discovery.

From a practical perspective, CLingS will deliver interfaces for multilingual querying and APIs that can be integrated with existing scientific databases and research infrastructures. These interfaces will make it possible for researchers, students, and educators to access scientific knowledge beyond the boundaries of a single language.

All datasets, models, and tools developed within the project will be released under permissive open-source licenses, ensuring transparency, reproducibility, and long-term impact. The project will also contribute to the development of evaluation protocols for multilingual scientific information retrieval.

Overall, CLingS aims to demonstrate how modern artificial intelligence and language technologies can help diversify access to scientific knowledge, strengthen multilingual scholarship, and support research communities working in languages that are currently underrepresented in global digital infrastructures.

References

[Hedderich et al.2020] Hedderich, Michael A. and Adelani, David I. and Zhu, Dawei and Alabi, Jesujoba and Markus, Udia and Klakow, Dietrich. 2020. Transfer Learning and Distant Supervision for Multilingual Transformer Models: A Study on African Languages. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. 2580–2591.

[Heffernan et al.2022] Heffernan, Kevin and Çelebi, Onur and Schwenk, Holger. 2022. Bitext Mining Using Distilled Sentence Representations for Low-Resource Languages. *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics. 2101–2112.

[Helm et al.2023] Helm, Paula and Bella, Gábor and Koch, Gertraud and Giunchiglia, Fausto. 2023. Diversity and Language Technology: How Techno-Linguistic Bias Can Cause Epistemic Injustice. *arXiv preprint arXiv:2307.13714*.

[Marivate et al.2020] Marivate, Vukosi and Sefara, Tshephisho and Chabalala, Vongani and Makhaya, Keamogetswe and Mokgonyane, Tumisho and Mokoena, Rethabile and Modupe, Abiodun. 2020. Investigating an Approach for Low Resource Language Dataset Creation, Curation and Classification: Setswana and Sepedi. *Proceedings of the First Workshop on Resources for African Indigenous Languages*. European Language Resources Association. 15–20.

[Micallef et al.2022] Micallef, Kurt and Gatt, Albert and Tanti, Marc and van der Plas, Lonneke and Borg, Claudia. 2022. Pre-training Data Quality and Quantity for a Low-Resource Language: New Corpus and BERT Models for Maltese. *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*. Association for Computational Linguistics. 90–101.

[Ranathunga and de Silva2022] Ranathunga, Surangika and de Silva, Nisansa. 2022. Some Languages are More Equal than Others: Probing Deeper into the Linguistic Disparity in the NLP World. *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and IJCNLP*. Association for Computational Linguistics. 823–848.

[Tars et al.2021] Tars, Maali and Tättar, Andre and Fišel, Mark. 2021. Extremely Low-Resource Machine Translation for Closely Related Languages. *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. 41–52.