AI-TraLow: AI-Driven Translation for Low-Resource Languages and Cultures

Antoni Oliver[†], **Maite Melero**[‡], **Felipe Sánchez-Martínez**[◇], **Víctor M. Sánchez-Cartagena**[◇]
[†]Universitat Oberta de Catalunya (UOC),
[‡]Barcelona Supercomputing Center (BSC), [◇]Universitat d'Alacant (UA)
aoliverg@uoc.edu, maite.melero@bsc.es, {fsanchez,vm.sanchez}@ua.es

Abstract

In this paper, we present AI-TraLow, a project dedicated to advancing AI-driven translation for low-resource languages and cultures. The research is structured around three primary objectives: firstly, the development of advanced data curation techniques designed to refine parallel corpora and detect machine-generated content; secondly, the exploration of integrating linguistic resources—such as dictionaries and grammatical rules—directly into model prompts and fine-tuning techniques to enhance translation precision; and thirdly, the mitigation of hardware constraints through knowledge distillation to produce efficient models viable for standard desktop environments. By targeting specific linguistic groups, including Iberian varieties (Aranese, Aragonese, Asturian and Eonavian), Mayan languages, and languages of vulnerable migrant communities, AI-TraLow seeks to foster linguistic diversity and digital inclusion. Ultimately, this initiative delivers open-source tools and models that ensure cultural heritage is both preserved and accessible within the contemporary digital landscape.

Keywords: Large Language Models, Machine Translation, Low-Resource Languages, Data Curation, Linguistic Integration, Knowledge Distillation

1. Introduction

The rapid advancement of Large Language Models (LLMs) has fundamentally transformed the landscape of Machine Translation (MT). Despite these breakthroughs, languages with limited digital resources continue to lag behind due to the heavy data dependency of modern architectures. The AI-TraLow project (*AI-Driven Translation for Low-Resource Languages and Cultures*) was established to tackle these disparities by hypothesizing that the emergent few-shot capabilities and multimodal nature of LLMs can be leveraged to bridge the resource gap, even when traditional parallel data is scarce.

AI-TraLow is structured as a coordinated project with a total duration of three years, having officially launched on September 1, 2025. The initiative is powered by a consortium of three institutions, each leading a specialized subproject:

- **Universitat d'Alacant (UA):** is the coordinating institution and leads the subproject *Curation and Exploitation of Heterogeneous Resources for Translating Low-resource Languages* (CEHR-TraLowLa). This unit focuses on data acquisition, curation, and the integration of linguistic information into LLM.
- **Barcelona Supercomputing Center (BSC):** Responsible for the subproject *Multimodal Large Language Models for Translating Low-Resource Languages* (MLLM4TRA). Their research centers on building multimodal models and developing strategies for adapting LLMs to low-resource scenarios.
- **Universitat Oberta de Catalunya (UOC):** Manages the subproject *Large Language Models for Translating Low-Resource Romance Languages of the Iberian Peninsula*. This team focuses on the development of linguistic resources and the rigorous evaluation of translation systems.

Ultimately, AI-TraLow aims to empower marginalized linguistic communities—ranging from the Iberian Peninsula (Aranese, Aragonese, Asturian, Eonavian) to Mayan varieties and migrant languages such as Wolof or Amazigh—ensuring they remain viable and visible in the contemporary digital era.

2. Related Work

The AI-TraLow project is situated at the intersection of LLMs and MT for low-resource scenarios. The following areas represent the current scientific landscape and the specific gaps this initiative aims to address.

2.1. LLMs in Machine Translation

Recent advancements in LLMs, particularly those based on the Transformer decoder-only architecture, have demonstrated remarkable capabilities in translation tasks (Alves et al., 2024; Xu et al., 2024; Yang et al., 2023). However, as highlighted in the project's rationale, these models often suffer from the "curse of multilinguality," where model capacity

Proceedings of LANLP: Bridging Ibero and Latin American NLP Communities, pages 15–20

is diluted across numerous languages, leading to suboptimal performance for those with limited digital presence. AI-TraLow builds upon the potential of instruction tuning (Alves et al., 2024) to specialize these models for neglected linguistic varieties.

2.2. Data Curation and Synthetic Content

A primary bottleneck for low-resource MT is the prevalence of noisy and machine-translated content in web-crawled data. This project will create high-quality datasets-encompassing text and, where applicable, image data, refined through state-of-the-art filtering and curation tools. Because the increasing presence of machine-generated text on the web poses a risk of "model collapse," a central focus of our work is the development of robust classifiers to distinguish between human-authentic and synthetic content. By ensuring that datasets are representative of human-authored language, and by employing innovative LLM-based approaches to clean existing parallel corpora, we will establish a reliable foundation for advancing translation technologies in data-scarce scenarios.

2.3. Integration of Heterogeneous Linguistic Resources

While current LLMs rely heavily on massive datasets, they often underutilize high-quality symbolic resources. Research has shown that providing in-context linguistic descriptions can significantly aid the learning of endangered languages. AI-TraLow extends this by exploring a hybrid methodology: integrating linguistic resources—such as dictionaries and transfer rules from *Apertium* (Forcada et al., 2011)—directly into both model prompts and fine-tuning techniques to teach models "unseen" languages on the fly (Tanzer et al., 2024; Zhang et al., 2024a,b; Hus and Anastasopoulos, 2024; Elsner et al., 2024).

2.4. Multimodal and Efficient Modeling

The state of the art in multimodality suggests that moving away from traditional subword tokenization can benefit languages with non-standardized orthographies. By employing pixel-based models (Salesky et al., 2024; Caglayan et al., 2019) and byte-to-byte architectures (Choe et al., 2019; Xue et al., 2022; Clark et al., 2022), AI-TraLow addresses the limitations of fixed vocabularies. Finally, to ensure the usability of these models in standard desktop environments, the project draws on knowledge distillation and efficiency strategies to mitigate the hardware constraints typically associated with large-scale models.

3. Objectives, methodology, and work plan

3.1. Objectives

The main objective of this project is to advance the development of MT systems for low-resource languages using decoder-only LLMs, thereby enabling these languages to join the wave of adoption of LLMs, specifically for MT. This global objective can be broken down into several specific objectives:

O1 Obtain and curate high-quality resources for low-resource languages.

O2 Leverage linguistic information to improve the translation of low-resource languages with LLMs.

O3 Build large language models tailored for translating a subset of the Mayan languages.

O4 Develop advanced methods for image-to-text translation of low-resource languages with multimodal LLMs.

O5 Develop advanced strategies for adapting LLMs for translating low-resource languages

O6 Build multimodal LLMs for the translation of low-resource languages spoken by migrants in vulnerable situation.

O7 Develop advanced methods for the automatic enhancement of existing linguistic resources used in rule-based MT and the generation of synthetic data with them.

O8 Build LLMs for the translation of low-resource languages of the Iberian Peninsula.

O9 Evaluation and comparison of encoder-decoder translation models and decoder-only models for translating low-resource languages of the Iberian Peninsula.

3.2. Methodology

The project is structured around a three-phase iterative cycle, which will be executed twice throughout its duration. These phases comprise: (i) data compilation and curation, (ii) research into the training and fine-tuning of LLMs for MT, and (iii) the release of the high-performance models.

During the *data compilation* phase, the consortium will develop new corpora for the target languages while simultaneously refining existing datasets. The generation of new resources will involve diverse methodologies, including document scanning and optical character recognition, alongside targeted web crawling. To safeguard data integrity, the project will develop techniques to detect and filter machine-generated or automatically translated content. Additionally, this stage encompasses the acquisition of visual data to support

image-to-text translation tasks and the creation of novel evaluation benchmarks tailored to the languages under study.

The second phase is dedicated to the exploration of LLM training and fine-tuning strategies for translation. The research team will employ established methodologies while remaining adaptable to emerging techniques within the field. A key focus will be investigating the integration of linguistic resources (e.g. dictionaries and grammars) with monolingual and parallel corpora. Furthermore, we will assess the impact of synthetic data on model performance and the efficacy of multimodal inputs in mitigating translation ambiguity. To this end, a specialized task is devoted to the development of multimodal LLMs. Throughout this process, models will be continuously monitored using a suite of automatic evaluation metrics.

In the final phase, the most effective models from the preceding stage will undergo rigorous human evaluation to ensure translation quality. Once validated, all models and tools will be released to the public under open-source licenses.

3.3. Work plan

The project is structured into five interconnected work packages (WPs). **WP1** is dedicated to the acquisition and curation of linguistic resources, with a specific emphasis on developing automated mechanisms to detect machine-generated or synthetically translated content, thereby ensuring the integrity of the training data. **WP2** explores the integration of linguistic resources (dictionaries, translation rules, grammar books) to augment the translation capabilities of LLMs in low-resource scenarios. **WP3** investigates multimodal architectures to facilitate text translation directly from visual inputs, a strategy aimed at mitigating contextual disambiguation challenges in MT. **WP4** focuses on the design of robust strategies for training and adapting LLMs; this includes advanced fine-tuning methodologies, the generation of high-quality synthetic corpora, and the distillation of knowledge from massive models into compact, efficient encoder-decoder architectures. Finally, **WP5** will apply the methodologies and tools derived from previous WPs to the development and release of optimized LLMs for a targeted subset of languages, thereby maximizing the project's impact and fostering digital inclusion within the respective linguistic communities.

WP1. Data acquisition and curation

This work package focuses on supplying resources for training and fine-tuning LLMs for low-resource language pairs. We will pursue several lines of work: (i) *corpus generation:* scanning existing books in extremely low-resource languages and conducting optical character recognition; and gathering text and image data in low-resource languages; (ii) *data quality enhancement:* researching methods for the automatic detection of text generated or translated automatically, and utilizing LLMs to clean existing parallel corpora; and (iii) *induction of linguistic resources:* developing methods for automatically generating linguistic resources that can be used later for augmenting training data (WP4) or for in-context learning (WP2).

WP2. Leveraging linguistic information in LLMs for translation

Low-resource languages often lack the substantial amounts of monolingual and bilingual data required to train competitive MT systems. Leveraging explicit linguistic information presents a promising approach to enhancing translation quality for these languages. In this WP, we will focus on exploiting underutilized resources, such as components from existing rule-based MT systems and monolingual dictionaries. Additionally, we will develop advanced methods to integrate a wide range of linguistic resources, including grammar books, into LLMs, and perform an analysis of the semantic representations to better understand which resources are useful and why.

WP3. Image-to-text translation of low-resource languages with multimodal LLMs

Low-resource languages frequently face tokenization challenges and are underrepresented in state-of-the-art LLMs (Petrov et al., 2024), resulting in significant performance disparities when accurately representing all languages (Ali et al., 2023). Pixel-based models have shown potential in overcoming tokenization challenges and improving cross-lingual transfer. Resolving ambiguous situational translations is a practical application of pixel-based models that involves translating source sentences that are directly influenced by the image given as input to the model. This work package focuses on: (i) investigating the suitability of pixel-based translation models for low-resource languages, and (ii) improving the reasoning abilities of multimodal LLMs for dealing with ambiguous situational translations that require visual cues to resolve it.

WP4. Strategies for training and adapting LLMs for translating low-resource languages

This WP focuses on developing and evaluating advanced methodologies to optimize LLMs for effective machine translation in data-scarce scenarios. A key priority is the development of gender-inclusive MT systems. Given that low-resource datasets are often sourced from uncurated web content,

they frequently reflect and amplify historical gender biases. Following the project's ethical roadmap, this WP will investigate fine-tuning techniques and debiasing prompts to ensure that translations do not perpetuate harmful stereotypes. Additionally, this WP explores an array of innovative, scalable, and resource-efficient solutions—such as knowledge distillation into compact architectures—that will then be applied in WP5

WP5. Building models for translating low-resource languages

In this WP, we will develop LLMs for translating a subset of the languages of interest to the project. We will apply, when possible, the methods developed and the lessons learnt in the previous WPs.

The resulting models will undergo a comprehensive evaluation, where they will be compared against encoder-decoder models in terms of error typology and characteristics of the produced translations. This comparison will determine whether training/fine-tuning encoder-decoder systems or LLMs is more effective for the languages of interest, and which strategies work best. The findings will guide future efforts.

4. Resources released during the project

The project focuses on three primary low-resource language groups: Mayan languages,[1] Romance languages of the Iberian Peninsula (Aranese, Aragonese, Asturian, and Eonavian), and languages spoken by migrant communities in Spain (Wolof, Amazigh, and Pashto). The main objective is to develop and release LLM-based translation models for select languages within each of these clusters.

In addition to these models, the project will produce the following group-specific resources:

Mayan languages

- Digitized dictionaries and descriptive grammars.
- Curated corpora of literary and web-crawled text.
- Translations of the FLORES+ (Goyal et al., 2022) evaluation dataset into K'iche', Kaqchikel, Q'eqchi', and Mam.

[1]The following languages will be addressed: Achi, Awakateko, Ch'orti', Chuj, Kaqchikel, Itza', Ixil, Jakalteko, Q'eqchi', Q'anjob'al, Akateko, Mam, Mopan, Poqomam, Poqomchi', K'iche', Sipakapense, Sakapulteko, Tekiteko, Tzeltal, Tz'utujil, Uspanteko, and Yucatec Maya.

Romance languages of the Iberian Peninsula

- Expanded Apertium lexical resources.
- Enhanced versions of the FLORES+ evaluation dataset.
- New translations for the NTREX (Federmann et al., 2022) and WMT24++ evaluation benchmarks.

Migrant community languages

- Multimodal image-text datasets designed specifically for pixel-based translation models.

5. Current Status

Although AI-TraLow officially launched in September 2025, several key milestones have already been achieved across the project's work packages:

- WP1. Data Acquisition: We have compiled and started the preprocessing of a comprehensive set of monolingual and bilingual Apertium dictionaries for Asturian, Aragonese, and Aranese. Simultaneously, we have completed the scanning and OCR of 29 Mayan dictionaries. We have developed a new precise method for automatically detecting machine translated text (García-Romero et al., 2025) and studied the ability of bilingual speakers to distinguish between human and machine translated text (García-Romero et al., 2026). To ensure high-quality evaluation, the human translation of the following datasets is currently underway: FLORES+, NTREX, and WMT24++ into Asturian, Aragonese, Aranese, and Eonavian; and FLORES+ into K'iche', Kaqchikel, Q'eqchi' and Mam.
- WP2. Integration of Linguistic Information: the challenges of adding dictionaries to the prompt have been deeply analysed and methods based on *Group Relative Policy Optimization* to inject terminology have been explored (Garcia Gilabert et al., 2025).
- WP3. Multimodality: preliminary experiments have established initial baselines for emoji-based disambiguation in vision-language models, providing a foundation for pixel-based translation tasks.
- WP4. Training Strategies: we devised and extensively evaluated methods for knowledge distillation leveraging multiple translations from the initial (teacher) model (Galiano-Jiménez et al., 2025).

6. Conclusions

The AI-TraLow project represents a strategic effort to ensure that low-resource languages are not left behind in the current era of Large Language Models. Beyond the technical development of translation systems, this initiative establishes a scalable paradigm for linguistic preservation by treating linguistic knowledge and multimodal signals as essential anchors for neural architectures.

Our approach shifts the focus from purely data-driven methods to a hybrid model where AI-driven curation and symbolic expertise safeguard the integrity of minority languages. Ultimately, AI-TraLow is committed to the principles of Open Science. By releasing high-performance, efficient models and curated tools, we aim to empower local communities and researchers, ensuring that cultural heritage and linguistic diversity can actively flourish and remain visible within the contemporary digital landscape.

Acknowledgements

Coordinated project funded by the Spanish Ministry of Science, Innovation and Universities (MICIU), the Spanish Research Agency and the European Regional Development Fund (FEDER) through R+D+i project grants PID2024-158157OB-C31, PID2024-158157OB-C32 and PID2024-158157OB-C33.

Bibliographical References

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Schulze Buschhoff, Charvi Jain, Alexander Arno Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. 2023. Tokenizer choice for LLM training: Negligible or crucial?

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota.

Dokook Choe, Rami Al-Rfou, Mandy Guo, Heeyoung Lee, and Noah Constant. 2019. Bridging the gap for tokenizer-free language models.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Micha Elsner et al. 2024. Shortcomings of llms for low-resource translation: Retrieval and understanding are both the problem. In *Proc. of the Ninth Conference on Machine Translation*, pages 1332–1354.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proc. of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25:127–144.

Aarón Galiano-Jiménez, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, and Víctor M Sánchez-Cartagena. 2025. Multi-hypothesis distillation of multilingual neural translation models for low-resource languages. *arXiv preprint arXiv:2507.21568*.

Javier Garcia Gilabert, Carlos Escolano, Xixian Liao, and Maite Melero. 2025. Terminology-constrained translation from monolingual data using GRPO. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1335–1343, Suzhou, China. Association for Computational Linguistics.

Cristian García-Romero, Miquel Esplà-Gomis, and Felipe Sánchez-Martínez. 2025. Automatic machine translation detection using a surrogate multilingual translation model. *arXiv preprint arXiv:2511.02958*.

Cristian García-Romero, Miquel Esplà-Gomis, and Felipe Sánchez-Martínez. 2026. When translations surprise: Human awareness of predictability in translations. In *Proceedings of the 15th biennial Language Resources and Evaluation Conference (in press)*, Palma de Mallorca, Spain.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Jonathan Hus and Antonios Anastasopoulos. 2024. Back to school: Translation using grammar books. In *Proc. of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20207–20219, Miami, Florida, USA.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2024. Language model tokenizers introduce unfairness between languages. *Advances in Neural Information Processing Systems*, 36.

Elizabeth Salesky, Philipp Koehn, and Matt Post. 2024. Benchmarking visually-situated translation of text in natural images. In *Proc. of the Ninth Conference on Machine Translation*, pages 1167–1182, Miami, Florida, USA.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024a. Teaching large language models an unseen language on the fly.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024b. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand.