Teaching Machine Translation Technologies with MTUOC

Antoni Oliver
Universitat Oberta de Catalunya
aoliverg@uoc.edu

Sergi Álvarez-Vidal
Universitat Autònoma de Barcelona
sergi.alvarez@uab.cat

Abstract

This paper presents the pedagogical integration of MTUOC, an open-source project developed at Universitat Oberta de Catalunya (UOC)—a distance-learning institution—to facilitate the training, fine-tuning, and integration of Neural Machine Translation (NMT) and Large Language Models (LLMs). The project consists of a modular suite of tools designed to streamline complex technical workflows for translation purposes. These components are currently utilised across research, industry knowledge transfer, and formal education. Specifically, the tools have been successfully implemented in a Bachelor's degree in Translation and Interpreting and a Master's degree in Translation Technologies. Furthermore, a pilot open course based on this framework received significant interest, reaching over 100 participants. This paper outlines the core components of the project, discusses the teaching experiences gathered in asynchronous environments, and describes the organisation of a forthcoming open course scheduled for October 2026. The results suggest that providing students with accessible, high-level interfaces for AI-based translation technologies enhances their technical autonomy and professional readiness.

1 Introduction

The integration of Artificial Intelligence (AI) in translation workflows has moved from being a specialized skill to a core competency in the language industry (Olohan, 2007). This shift has led to new professional profiles that require a specific set of skills beyond traditional translation (Cid et al., 2020). This section outlines the institutional context and the motivation behind the development of a dedicated technological framework for translation pedagogy.

1.1 The Universitat Oberta de Catalunya

The Universitat Oberta de Catalunya (UOC) is an online institution based in Barcelona (Catalonia, Spain), offering a comprehensive range of undergraduate, postgraduate, and doctoral programmes. The university operates within a multilingual framework, using Catalan, Spanish and English as its primary working languages.

Within the Arts and Humanities Department, the academic offer includes a Bachelor's degree in Translation, Interpreting, and Applied Languages, as well as a Master's degree in Translation Technologies, both aligned with the latest European standards for translator training (EMT Board, 2022). For several years, these programmes have placed a strategic emphasis on teaching the latest advancements in Machine Translation (MT) through a hands-on, practical approach. This commitment to technical proficiency led to the creation of the MTUOC project (Oliver and Álvarez, 2023), an initiative designed to bridge the gap between theoretical AI concepts and professional translation practice, following the principle of empowering all users in the age of AI (Kenny, 2022). The project is briefly introduced in the following subsection, with its core technical components detailed in section 2.

1.2 The MTUOC project: Context and scope

The MTUOC project was established to address a recurring challenge in translation technology:

While most components required to train and deploy translation models are available under free and permissive licenses, they often lack comprehensive documentation or user-friendly interfaces. This project aims to facilitate the use of existing open-source tools while developing auxiliary software to streamline the training, evaluation, deployment, and integration of Neural Machine Translation (NMT) and Large Language Models (LLMs).

By providing a more accessible framework, the project bridges the gap between complex computational requirements and the practical needs of the translation community. Currently, it is employed in three main areas:

- Research: Providing a stable framework for linguistic and computational experimentation.
- Knowledge transfer: Facilitating collaboration with industry partners to implement customized translation solutions.
- Education: Serving as the technological backbone for specialized courses in translation technology.

By integrating these tools into the curriculum, the university ensures that students do not merely learn how to operate commercial interfaces, but rather understand the underlying processes of data curation, model evaluation, and system integration. This approach prepares future professionals to adapt to a rapidly changing technological landscape.

1.3 Subjects and courses

The MTUOC project is currently integrated into several levels of higher education, following a modular pedagogical design that adapts to different student profiles.

In the Bachelor's degree in Translation, Interpreting, and Applied Languages, the toolkit is used in the Machine Translation and Post-editing course. The mandatory curriculum focuses on foundational concepts and basic operations. However, the course is structured to allow students interested in technical specialization to complete a series of advanced optional activities, which involve more complex data manipulation and engine customization.

At the postgraduate level, the framework is a core component of the Translation and Technologies course within the Master's degree in Translation Technologies. In this context, all students are required to engage with more advanced tasks, such as model fine-tuning and technical evaluation. Similar to the undergraduate level, an additional track of specialized activities is available for those seeking to further their expertise in MT development.

Finally, these materials were adapted for a free online course (MOOC). This initiative met with high demand, attracting a diverse range of participants and confirming the interest of the broader translation community in accessible, high-level technical training. The success of this pilot has led to the planning of a permanent open course, further described in section 4.

2 Components of the MTUOC project

The MTUOC project is designed as a modular ecosystem that covers the entire machine translation lifecycle. Rather than a single monolithic application, it consists of several interconnected components and independent scripts that can be used separately or as part of a complex workflow. This modularity is a key pedagogical feature, as it allows students to isolate and understand each stage of the translation process.

The software is primarily developed in Python, featuring clean, well-documented code that facilitates customization and adaptation according to specific needs. Most components are cross-platform, ensuring compatibility across Linux, Windows, and macOS. Furthermore, specific processes are handled through simple, adaptable Bash scripts. This architecture ensures that all components are immediately accessible for standard use while remaining flexible enough for advanced pedagogical activities that require code modification or pipeline adjustment.

All the components can be freely downloaded from the project's GitHub.[1]

2.1 Server, translators, editors, and plugins

At the core of the project is a client-server architecture designed to facilitate the deployment of MT models. The MTUOC-server acts as a central hub capable of hosting multiple NMT or LLM-based engines simultaneously. Furthermore, the server can be configured to support various popular communication protocols, ensuring compatibility with

[1] https://mtuoc.github.io/

a wide range of third-party client applications. To interact with this server, the project provides several dedicated interfaces:

- **Translators and editors:** The project offers both desktop and web-based applications for translating text and documents. These include a dedicated post-editing interface that provides essential Computer-Assisted Translation (CAT) functionalities alongside specialized features, such as automated quality control and the integration of quality estimation metrics.

- **Plugins:** These integration modules allow students to deploy their custom-trained models within professional environments. At the time of writing, the project provides plugins for OmegaT, Trados Studio, and Okapi Tools, ensuring that the technology is compatible with widely used industry standards.

2.2 Tools for corpus creation, alignment, and preprocessing

Effective MT depends on high-quality data. This sub-module includes a suite of tools focused on the preparation of linguistic assets, providing functionalities for:

- **Data acquisition:** Tools for web crawling to facilitate corpus creation and for extracting comparable and parallel corpora from Wikipedia.

- **Data alignment:** Modules designed to align parallel corpora and to perform automated searches for parallel segments within comparable datasets.

- **Corpus preprocessing:** Scripts for cleaning, tokenising, and filtering data to ensure it meets the quality standards required for NMT training. These tools allow students to empirically observe the impact of data noise on the performance of the final output.

2.3 Tools and scripts for training and fine-tuning NMT and LLMs

These components provide the necessary interfaces to interact with state-of-the-art training and fine-tuning toolkits, such as Marian, OpenNMT, Eole, and Axolotl, among others. The project simplifies hyperparameter configuration and the management of training environments by providing numerous documented configuration templates. Key features include:

- **NMT training:** Streamlined scripts that guide users through the process of developing Neural Machine Translation models from scratch, reducing the technical barriers associated with command-line environments.

- **NMT and LLM fine-tuning:** Specialized tools designed to adapt Large Language Models for translation tasks through supervised fine-tuning. This allows students to engage directly with modern AI paradigms and observe how pre-trained models can be specialized for specific domains or language pairs.

2.4 Scripts and programs for MT evaluation

The final stage of the workflow involves assessing the quality of the generated translations. The project includes a comprehensive set of evaluation tools designed to provide both automated and human-centered feedback:

- **Automatic metrics:** The toolkit provides scripts and interfaces that facilitate the use of industry-standard metrics, such as sacreBLEU (Post, 2018) and COMET (Rei et al., 2020). These tools ensure that students can produce reproducible and comparable results when evaluating model performance.

- **Post-editing tool:** The project offers a dedicated environment for post-edition that allows for the collection of fine-grained data, including various post-editing effort metrics (Oliver et al., 2020). This allows a critical approach to technology assessment, moving beyond purely automated scores to include human-in-the-loop evaluation.

3 Pedagogical experience and evaluation

The MTUOC project has been deployed in diverse educational settings, ranging from formal university degrees to open-access online training. This multi-layered implementation has allowed for a comprehensive evaluation of the tools' versatility and their impact on student learning. By integrating these components into different curricula, we have been able to observe how students with varying levels of technical background

engage with AI-driven translation workflows in an asynchronous, distance-learning environment. The following subsections detail the specific experiences within formal and informal education, as well as the recurring challenges identified during these processes.

3.1 Implementation in Bachelor's and Master's degrees

The integration of the MTUOC toolkit in formal degrees followed a tiered pedagogical strategy. Given that translation curricula are often densely packed with linguistic and core translation modules, it is challenging to introduce extensive technical training without overwhelming the students. As previously noted in the literature, overcoming these technical obstacles is crucial for student empowerment (Kenny and Doherty, 2014). To this end, we adopt a problem-based learning approach (Mellinger, 2018) where students gain confidence through practical, hands-on tasks. Consequently, our approach focuses on a mandatory introductory level, complemented by optional advanced activities for students with a specific interest in technology. This elective track is particularly relevant today, as it prepares students for highly sought-after professional roles that require a deep understanding of AI-driven translation architectures.

The implementation revealed several key insights regarding the students' technical backgrounds:

- **The command-line barrier**: We observed a general lack of familiarity with Command-Line Interfaces (CLI). For many students, the initial interaction with a terminal can be discouraging, representing a significant technical hurdle.

- **Operating system diversity**: The student body primarily uses Windows, followed by macOS, while Linux users are almost non-existent. This diversity necessitates a "cross-platform first" approach.

- **Scaffolded learning**: To mitigate initial friction, we designed preliminary activities that are easily executable across all operating systems. Once students gain confidence and understand the underlying logic of the processes, they are encouraged to engage with Linux-based environments and terminal commands through more advanced, specialized tasks.

3.2 Implementation in a free online course

The MTUOC project was also the cornerstone of an open-access online course designed for the broader translation community. The initial iteration was structured as a comprehensive 12-week programme, with an estimated weekly workload of five hours. While the course generated significant interest—attracting over 100 enrollees—the completion rate was notably low, with only five participants reaching the end.

An analysis of this pilot experience revealed two primary factors behind the high attrition rate:

- **Misaligned expectations and commitment**: The initial enthusiasm led many to register without fully accounting for the long-term time commitment required.

- **Diverse user interests**: Participants exhibited varied professional needs. While some were focused on engine deployment and server management, others were primarily interested in corpus acquisition or NMT training. A monolithic course structure forced participants to navigate sections that were not always relevant to their specific goals.

The primary takeaway from this experience is that a modular approach is more effective for professional development in this field. Rather than a single, extensive course, it is more sustainable to offer a series of shorter, specialized modules. For instance, four modules of four weeks each provide a more manageable and focused learning path than a single 16-week programme. This flexibility allows participants to enroll only in the segments that align with their professional interests, thereby increasing engagement and completion rates.

3.3 Coordination problems

A significant challenge identified in previous iterations was the lack of temporal and structural synchronization between the different educational tracks. The Bachelor's degree, the Master's degree, and the open online course were conducted as independent sessions with different start dates and distinct student cohorts. This fragmentation resulted in several operational inefficiencies:

- **Instructional and pedagogical overload**: Managing three staggered timelines imposed

a substantial burden on the teaching staff. Beyond server maintenance, the primary drain on resources was the continuous duplication of pedagogical efforts, including the moderation of forums, and the provision of technical troubleshooting for three distinct student cohorts. This fragmentation prevented the faculty from streamlining feedback and consolidating teaching materials effectively.

- **Reduced peer interaction**: In the formal degree programmes, the number of students per group remained relatively small (ranging from 5 to 10). This limited size hindered the potential for peer-to-peer learning and the creation of a robust community of practice.

- **Resource fragmentation**: Technical issues or updates in the MTUOC toolkit had to be addressed independently across three different contexts, increasing the likelihood of versioning conflicts.

As discussed in the following sections, future iterations of these courses will be conducted simultaneously across all three levels. This strategic alignment aims to optimize teaching resources and foster a larger, more dynamic learning community.

3.4 Learning environments and infrastructure

The delivery of learning materials and the facilitation of communication have relied on a combination of open-access platforms and institutional systems. To date, the infrastructure has been organized as follows:

- **Learning materials and software**: A dedicated Wiki and a GitHub repository served as the primary hubs for all pedagogical content, including documentation, scripts, and source code. This ensured that students always had access to the latest versions of the MTUOC toolkit while fostering familiarity with standard version-control practices.

- **Institutional channels**: For students enrolled in the Bachelor's and Master's degrees, the Universitat Oberta de Catalunya Virtual Campus was used for formal assignments, grading, and official communications.

- **Open-access channels**: For the free online course, the same Wiki and GitHub resources were used, but communication and community support were managed via Google Groups.

While this setup provided a functional baseline, the use of Google Groups for the open course proved to be sub-optimal for fostering real-time interaction and technical troubleshooting. Consequently, for future iterations, the project will maintain GitHub as the central repository for materials but aims to migrate community communication to more efficient, modern open-source or free platforms that allow for better-organized discussion threads and more agile peer-to-peer support.

4 New course design

To address the challenges identified in previous iterations—namely student attrition in long-format courses and the instructional overload caused by fragmented schedules—a redesigned pedagogical framework will be implemented starting in October 2026. This new design moves away from a monolithic structure toward a highly modular and synchronized set or shorter courses.

The primary objective of this overhaul is to optimize teaching resources while creating a more resilient learning community. By decoupling the project's components into specialized units and aligning the academic calendars of formal and informal tracks, the project aims to foster a collaborative environment where students from different backgrounds can interact. The following subsections detail the architectural pillars of this new approach: the transition to independent modules, the implementation of a unified calendar, and the adoption of a more agile communication infrastructure.

4.1 Modules

The new curriculum is structured into four interconnected modules, each designed to function as an independent learning unit while maintaining a logical progression. This modularity allows students to customize their learning path based on their specific professional needs or technical background. Each module spans four weeks with an estimated workload of five hours per week (20 hours per module).

- **Module 1: Deployment, integration, and evaluation.** This introductory unit focuses

on the operational side of translation technology. Participants learn to host NMT and LLM engines using the MTUOC-server and integrate them into professional workflows via CAT tools such as OmegaT and Trados Studio. The module also covers basic quantitative evaluation.

- **Module 2: Corpus engineering.** This module addresses the data layer of machine translation. It covers the exploration of existing parallel datasets and the creation of custom comparable and parallel corpora through web crawling and Wikipedia extraction. Emphasis is placed on essential preprocessing tasks, such as data alignment and noise reduction.

- **Module 3: NMT training and fine-tuning.** Moving into model development, this unit guides students through the end-to-end process of training NMT engines from scratch using frameworks like Marian, Eole, and fairseq2. It also introduces techniques for fine-tuning existing NMT models to specific domains.

- **Module 4: LLMs for translation—in-context learning and specialization.** The final module explores cutting-edge AI paradigms. Students learn to implement In-Context Learning (ICL) by leveraging translation memories and terminological databases. Furthermore, it introduces the fundamentals of LLM specialization through both Full Fine-tuning and Parameter-Efficient Fine-Tuning (PEFT) methods, such as LoRA.

4.2 Unified calendar

To maximize instructional efficiency and foster a synchronized learning community, a unified calendar has been established for all three tracks (Bachelor's, Master's, and the open online course). The modules are offered twice per academic year, with start dates aligned to the postgraduate schedule, typically beginning in mid-October and mid-March.

For Bachelor's students, this synchronization entails a specific timeline: since their regular semester begins earlier, they will start these optional technical modules approximately one month into their coursework. Furthermore, as the Bachelor's semester also concludes earlier, the final module will take place during the break between semesters. Crucially, even during this period, these students will remain integrated into the shared communication platform, ensuring they have continuous access to peer support and instructor guidance alongside participants from the other tracks.

The following table outlines the synchronized schedule for both semesters:

Module	1st semester	2nd semester
Module 1	Mid-Oct -Mid-Nov	Mid-Mar - Mid-Apr
Module 2	Mid-Nov - Mid-Dec	Mid-Apr - Mid-May
Module 3	Mid-Jan - Mid-Feb	Mid-May - Mid-Jun
Module 4	Mid-Feb - Mid-Mar	Mid-Jun - Mid-Jul

Table 1: Unified calendar

4.3 Teaching and communication platform

To facilitate this cross-track interaction, the project will transition from Google Groups to Discord, leveraging its specific templates for educational environments. This platform offers a more dynamic and persistent communication space, which is essential for maintaining engagement across the staggered schedules of the different student cohorts. The adoption of Discord is driven by three primary pedagogical goals:

- **Real-time technical support:** The use of channel-based categories and organized threads for specific technical issues allows for faster troubleshooting and creates a searchable knowledge base for all students.

- **Peer-to-peer mentorship:** Discord's informal environment encourages interaction between tracks. More experienced participants, such as those from the Master's degree or the professional open course, can easily support Bachelor's students, fostering a collaborative community of practice.

- **Integration and agility:** The platform allows for direct integration with GitHub webhooks, providing real-time notifications for repository updates and Wiki changes.

By utilizing Discord's classroom templates, the project can establish a structured yet flexible virtual campus. This shared space will act as the "social glue" of the initiative, ensuring that no student works in isolation, regardless of their specific academic calendar.

5 Conclusions

The rapid evolution of Artificial Intelligence and its application to translation workflows necessitates a shift in pedagogical models. As demonstrated throughout the various iterations of the MTUOC project, teaching these technologies requires more than just updated software; it demands an agile and modular infrastructure capable of adapting to a shifting technological landscape.

The transition to a four-module design for October 2026, synchronized across different academic tracks and supported by a dynamic communication environment like Discord, addresses the primary challenges identified in our previous experiences: instructional overload, student attrition, and technical fragmentation. This new framework ensures that students—whether in formal degrees or open-access tracks—can engage with the latest advancements in NMT and LLMs within a cohesive community of practice.

A key factor in the sustainability of this initiative is the use of a GitHub-based Wiki for all teaching materials. This format allows the teaching staff to update content, scripts, and documentation in real-time, which is essential given that AI methodologies can become obsolete in a matter of months. Furthermore, we reiterate that all materials and tools developed within this project are released under open-source licenses, allowing for their free use, distribution, and modification without restrictions.

Finally, the authors extend an invitation to other universities and translation departments to adopt this curriculum or adapt it to their specific institutional needs. By sharing these open pedagogical resources, we aim to contribute to a more technically proficient and autonomous generation of translation professionals, capable of not only using AI but also understanding and shaping the tools of their trade.

Acknowledgements

The preparation and implementation of the courses described in this paper have been partially supported by the research projects TamTAS PCI2025-167063-2, funded by MICIU/AEI/10.13039/501100011033 and European Union in the Chist-era call 2025 Science in your own language; LLMTrad-IBE: Large Language Models for translating Low-Resource Languages of the Iberian Peninsula, funded by MCIU/AEI/10.13039/50110011033/FEDER,UE with reference PID2024-18157OB-C33. The authors would like to thank the institutional support provided during the development and testing phases of the technological framework by the Universitat Oberta de Catalunya and Arts and Humanities department.

References

Cid, Clara Ginovart, Carme Colominas, and Antoni Oliver. 2020. Language industry views on the profile of the post-editor. *Translation Spaces*, 9(2):283–313.

EMT Board. 2022. European graduate competences framework in translation. Technical report, European Commission, Directorate-General for Translation.

Kenny, Dorothy and Stephen Doherty. 2014. Statistical machine translation in the translation curriculum: Overcoming obstacles and empowering translators. *The Interpreter and Translator Trainer*, 8(2):276–294.

Kenny, Dorothy. 2022. *Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence*. Language Science Press, Berlin.

Mellinger, Christopher D. 2018. Problem-based learning in computer-assisted translation pedagogy. *HERMES-Journal of Language and Communication in Business*, (57):195–208.

Oliver, Antoni and Sergi Álvarez. 2023. Training and integration of neural machine translation with MTUOC. In *Proceedings of the 1st Workshop on Open Community-Driven Machine Translation*, pages 5–13.

Oliver, Antoni, Sergi Alvarez, and Toni Badia. 2020. PosEdiOn: Post-editing assessment in Python. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 403–410.

Olohan, Maeve. 2007. Economic trends and developments in the translation industry: What relevance for translator training? *The Interpreter and Translator Trainer*, 1(1):37–63.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.