EAMT 2026

1st International Workshop on Teaching AI-Based Translation and Technologies (TAITT 2026)

Proceedings of the Workshop

June 15, 2026

ISBN 9789403901398
DOI 10.26116/9789403901398

Message from the Organising Committee

This volume contains the proceedings of the 1st International Workshop on Teaching AI-Based Translation and Technologies (TAITT 2026)[1], hosted as part of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026)[2]. TAITT is the first dedicated forum focusing on pedagogy, curriculum design, and professional training in the rapidly evolving landscape of AI-driven translation technologies. The first edition of TAITT carries some historical significance, as 2026 marks the 25th anniversary of the first workshop on teaching machine translation (MT), held at MT Summit VII—an event which Kenny (2020) identifies as "the beginning of systematic thinking about the teaching of translation technology". Since then, translation education has undergone profound technological, social, and pedagogical transformations. Today, teachers and trainers must integrate a vast and growing range of content—from traditional translation technologies to machine learning fundamentals, MT evaluation, and large language models—while also fostering critical thinking and combating technological hype. The first edition of TAITT intends to bring together researchers, educators, industry practitioners, and translators to address these challenges collaboratively. The workshop provides a forum for contributions that explore how AI-based translation and technologies can be taught effectively across academic, professional, and industry contexts, and how curricula can be adapted to ensure that future translators, linguists, and language professionals develop the necessary competencies for working with modern translation technologies.

TAITT 2026 welcomed research papers and extended abstracts addressing pedagogical innovation, technological integration, empirical insights, curriculum design, and critical approaches to AI-based translation teaching. The workshop received 17 submissions (12 research papers and 5 extended abstracts). Of these, 14 submissions were evaluated positively in the peer-review process, amounting to an acceptance rate of 82%. To accommodate as many contributions as possible as part of the workshop, TAITT will host a combination of oral and poster presentations.

The accepted submissions cover a broad spectrum of topics, ranging from didactic resources and approaches for teaching the technical foundations of MT and language-oriented AI in a translation context via strategies for implementing these technologies in computer-assisted translation teaching to proposals focusing on developing relevant digital literacies such as data, MT, and AI literacy.

Complementing the regular workshop programme are two keynotes by our invited speakers, Lynne Bowker from Université Laval and Miquel Esplà-Gomis from Universitat d'Alacant. Lynne and Miquel will be joined by Pilar Sánchez-Gijón from Universitat Autònoma de Barcelona, Arda Tezcan from Ghent University and Erik Angelone from TH Köln in a panel discussion that concludes the workshop.

We would like to take this opportunity to express our heartfelt thanks to all the people and institutions involved in the first edition of TAITT: the authors, keynote speakers and panelists for their highly insightful and relevant contributions, the members of the Programme Committee for peer-reviewing the submissions, the conference team from Tilburg University for providing organisational and infrastructural support and the European Association for Machine Translation for supplying the overall thematic framework for the conference and our workshop.

We hope you enjoy reading the proceedings of TAITT 2026 and are looking forward to a day of insightful and engaging discussions on teaching AI-based translation and technologies.

June 2026,

Ralph Krüger, Dorothy Kenny, Sheila Castilho, Sergi Álvarez-Vidal, Nora Aranberri, María Isabel Rivas Ginel, and Janiça Hackenbuchner

[1] https://sites.google.com/view/taitt2026/home
[2] https://eamt2026.org/

Organising Committee

Organising Committee & Workshop Chairs

Ralph Krüger, Institute of Translation and Multilingual Communication, TH Köln – University of Applied Sciences Cologne, Germany
Dorothy Kenny, School of Applied Language & Intercultural Studies, Dublin City University, Ireland
Sheila Castilho, School of Applied Language & Intercultural Studies, Dublin City University, Ireland
Sergi Álvarez-Vidal, Department of Translation and Interpreting and East Asia Studies, Universitat Autònoma de Barcelona, Spain
Nora Aranberri, Department of English and German Philology and Translation and Interpreting Studies, Euskal Herriko Unibertsitatea, Spain
María Isabel Rivas Ginel, Faculty of Philosophy, Arts and Letters, UCLouvain, Belgium
Janiça Hackenbuchner, Language and Translation Technology Team, Department of Translation, Interpreting and Communication, Ghent University, Belgium

Programme Committee and Invited Guests

Programme Committee

Lynne Bowker, Université Laval
Vicent Briva-Iglesias, Dublin City University
Dragoș Ioan Ciobanu, University of Vienna
Loïc De Faria Pires, Université de Mons
Gokhan Dogru, Universitat Pompeu Fabra
Ana Guerberof Arenas, University of Groningen
Marie-Aude Lefer, UCLouvain
Lieve Macken, Ghent University
Joss Moorkens, Dublin City University
Jean Nitzke, University of Agder
Celia Rico Pérez, Universidad Complutense de Madrid
Miguel A. Rios Gaona, University of Vienna
Caroline Rossi, Université Grenoble Alpes
María del Mar Sánchez Ramos, Universidad de Alcalá
Pilar Sánchez-Gijón, Universitat Autònoma de Barcelona
Antonio Toral, Universitat d'Alacant
Ester Torres Simón, Universitat Autònoma de Barcelona
Vincent Vandeghinste, KU Leuven
Mireia Vargas Urpí, Universitat Autònoma de Barcelona

Keynote Speakers

Lynne Bowker, Université Laval in Québec
Miquel Esplà-Gomis, Universitat d'Alacant

Panelists

Lynne Bowker, Université Laval in Québec
Miquel Esplà-Gomis, Universitat d'Alacant
Arda Tezcan, Ghent University
Erik Angelone, TH Köln – University of Applied Sciences
Pilar Sánchez-Gijón, Universitat Autónoma de Barcelona

Keynote Talk

"In 2026, We Are Friction-Maxxing" - Scaffolding Friction as Part of Teaching AI-Based Translation and Technologies

Lynne Bowker
Université Laval in Québec
2026-06-15 09:10:00 – Room: **MindLabs, Room: MLZ 1.21**

Bio: Lynne Bowker is Full Professor and Canada Research Chair in Translation, Technologies and Society at Université Laval in Québec. Her teaching and research interests lie at the intersection of language and technologies, including the pedagogy of translation and translation tools. Some of her recent publications in this area include Teaching machine translation literacy to non-translation students" (in Translation, interpreting and technological change: Innovations in research, practice and training, M. Winters, S. Deane-Cox & U. Böser, eds, Bloomsbury, 2024) and "Teaching translation students about data in the age of generative AI" (in Teaching translation in the age of generative AI, JC Penet, J. Moorkens & M. Yamada, eds, Language Science Press, 2026). In addition, she has developed several textbooks and Open Educational Resources (OER) to support teaching and learning various aspects of translation and translation tools. Many of these resources are freely available on the website of the Machine Translation Literacy Project.

Keynote Talk

Owning the Infrastructure: AI Sovereignty as a Key Competence for Next-Generation Translators

Miquel Esplà-Gomis
Universitat d'Alacant
2026-06-15 13:30:00 – Room: **MindLabs, Room: MLZ 1.21**

Bio: Miquel Esplà-Gomis is an Associate Professor in the Department of Software and Computing Systems at the Universitat d'Alacant. After completing his PhD in Computer Science, he has focused his career on the intersection of technology and language, both as a researcher and an educator. For over a decade, he has taught translation technologies to translation students, while also introducing Master's students from both Computer Science and Linguistics to the fields of computational linguistics and multilingual technologies. His research interests center on parallel-corpora collection and curation, as well as translation quality estimation. Dr. Esplà-Gomis has contributed to over 60 publications and has participated in several significant European initiatives, such as the ParaCrawl series of projects and the GoURMET project, which focused on machine translation for under-resourced languages. Between 2021 and 2022, he also coordinated the international consortium for the MaCoCu project, aimed at expanding digital resources for European languages. Beyond his research and teaching, he contributes to the broader academic community as a member of the executive committee of the European Association for Machine Translation (EAMT).

Table of Contents

Programme

Monday, June 15, 2026

09:00 - 09:10 Opening of the Workshop

09:10 - 10:00 **Keynote 1** - **Lynne Bowker** - "In 2026, We Are Friction-Maxxing" - Scaffolding Friction as Part of Teaching AI-Based Translation and Technologies

Oral Presentations (AI-Based Translation)

10:00 - 10:15 **Gokhan Dogru** - Evaluative Judgement in Teaching AI-Based Translation: A Classroom Case Study of AI-Mediated Translation and Post-Editing

10:15 - 10:30 **Katrin Menzel** - Teaching Linguistic Prompt Control for LLM-Based Translation: A Classroom Approach to Developing Critical and Responsible AI Literacy

10:30 - 11:00 Coffee Break

Oral Presentations (Technology and Digital Literacies)

11:00 - 11:15 **Vincent Vandeghinste** - Mimicking Neural Machine Translation History for Pedagogic Reasons

11:15 - 11:30 **Antoni Oliver** and **Sergi Álvarez-Vidal** - Teaching Machine Translation Technologies with MTUOC

11:30 - 11:45 **Ralph Krüger** - A Technical Curriculum on Language-Oriented Artificial Intelligence in Translation and Specialised Communication

11:45 - 12:00 **Pilar Sánchez-Gijón** - Teaching Data Management to Translation Students: From Documentation Practices to Data Literacy

12:00 - 12:15 **Yu Hao**, **Elise Wu** and **Ester Leung** - Translator Competence in the Age of Agentic AI Orchestration: A "Backcasting" Perspective

12:15 - 12:30 Open Discussion

12:30 - 13:30 Lunch Break

13:30 - 14:20 **Keynote 2** - **Miquel Esplà-Gomis** - Owning the Infrastructure: AI Sovereignty as a Key Competence for Next-Generation Translators

Oral Presentations (Project-Based Pedagogy)

Monday, June 15, 2026 (continued)

14:20 - 14:35 **Koen Kerremans** - Integrating AI-Based Technologies into Translation Workflows through a Simulated Translation Bureau (STB)

14:35 - 14:50 **Alina Karakanta** - Beyond Post-Editing: A Project-Based Module on MT and LLM Integration for Trainee Translators

14:50 - 15:00 Open Discussion

15:00 - 15:30 Coffee Break

Poster Session

15:30 - 16:15 **Esmée Bennison** and **Lynne Bowker** - Teaching AI-based Translation Technologies: Drawing Inspiration from Engineering Design Education to Learn about Sustainability

15:30 - 16:15 **Jeevanthi Liyana Pathirana**, **Pierrette Bouillon**, **Jonathan Mutal**, **Sabrina Girletti** and **Lise Volkart** - COPECO-Speech: Multimodal Post-Editing with Speech and LLMs for Translation Teaching

15:30 - 16:15 **Erik Angelone** - Facilitating Interaction-Oriented AI Literacy in Translator Training: A Process-Oriented Approach

16:15 - 17:15 **Panel Discussion: Exploring the Jagged Frontier of AI in Translation Didactics** - Lynne Bowker, Miquel Esplà-Gomis, Arda Tezcan, Erik Angelone, Pilar Sánchez-Gijón

17:15 - 17:30 End of the Workshop

A Technical Curriculum on Language-Oriented Artificial Intelligence in Translation and Specialised Communication

Ralph Krüger
Institute of Translation and Multilingual Communication
TH Köln – University of Applied Sciences Cologne, Germany
ralph.krueger@th-koeln.de

Abstract

This paper presents a technical curriculum on language-oriented artificial intelligence (AI) in the language and translation (L&T) industry. The curriculum aims to foster domain-specific technical AI literacy among stakeholders in the fields of translation and specialised communication by exposing them to the conceptual and technical/algorithmic foundations of modern language-oriented AI in an accessible way. The core curriculum focuses on 1) vector embeddings, 2) the technical foundations of neural networks, 3) tokenization and 4) transformer neural networks. It is intended to help users develop computational thinking as well as algorithmic awareness and algorithmic agency, ultimately contributing to their digital resilience in AI-driven work environments. The didactic suitability of the curriculum was tested in an AI-focused MA course at the Institute of Translation and Multilingual Communication at TH Köln. Results suggest the didactic effectiveness of the curriculum, but participant feedback indicates that it should be embedded into higher-level didactic scaffolding – e.g., in the form of lecturer support – in order to enable optimal learning conditions.

1 Introduction

The recent emergence of general-purpose AI (GPAI) technologies in the form of large language models (LLMs) has widened the scope of automation in the L&T industry to a considerable extent, as discussed, e.g., in Lambson and Han (2025) and in Krüger and Angelone (forthcoming). For example, a single LLM can perform translation-technological tasks that were originally distributed over different expert technologies (e.g., intra- and interlingual machine translation, terminology extraction or quality estimation). Furthermore, LLMs can augment traditional translation technologies with new features that were previously beyond the scope of these technologies (e.g., augmenting traditional automatic translation quality control with meaning-based source- and target-segment comparisons) and they can (partially) automate tasks which largely resisted being automated to date (e.g., autonomous text production at high proficiency levels or intersemiotic machine translation/interpreting).

The current LLM-driven automation cycle in the L&T industry is shaping up to be both more extensive and more invasive than previous automation cycles, as evidenced, for example, by the *LangOps Manifesto*'s *AI-first* principle.[1] It could be claimed that – in order to retain high levels of agency and control in AI-first environments in the spirit of the human-centered AI approach[2] – L&T industry stakeholders require domain-specific AI literacy, as conceptualised, e.g., in the AI literacy framework proposed in Krüger (2024) and Krüger (forthcoming). In this framework, overarching AI literacy is further decomposed into the individual

[1] "AI systems are already good enough for many use cases and their performance will continue to improve. We always start by trying AI solutions to solve language problems." (https://langops.org/our-manifesto/)

[2] For a general overview of the concept of human-centered AI, see Shneiderman (2020), for a translation-specific perspective, see Jiménez-Crespo (2025).

Proceedings of the 1st International Workshop on Teaching AI-Based Translation and Technologies (TAITT 2026), pages 1–11
June 15, 2026 ©2026 Association for Computational Linguistics

subdimensions of technical, performance-focused, interaction-focused, implementation-focused and ethical/societal AI literacy, which are interrelated in various ways.

The curriculum proposed in this paper focuses on developing technical AI literacy, which involves knowledge of the basic operating principles of modern – currently to a large extent transformer-based – AI technologies. It is certainly true that, as noted by Bowker (2026, 73) with regard to translation didactics, "[i]t is not necessary for most translation educators or students to understand all the details of how an artificial neural network operates in order to use AI-based translation." This view is complemented by Rivas Ginel and Moorkens (2025, 295), who found that translators' (dis)trust in AI technologies such as LLMs is mostly calibrated by their use of these systems rather than by a deeper understanding of the systems' technical foundations.

However, there are also several arguments in favour of fostering technical AI literacy in L&T industry stakeholders. Most importantly, technical AI literacy has an element of empowerment in that it helps in demystifying the black-box nature of modern machine-learning based AI technologies such as as LLMs, a point which was previously made by Doherty and Kenny (2014, 296) in the context of statistical machine translation (SMT) and by Kenny (2019, 438) in the context of neural machine translation (NMT). AI-literate stakeholders familiar with the operating principles of these technologies may have a better-informed picture of the tasks these technologies can currently be used for and the level of quality at which these tasks can be completed (aspects related to performance- and implementation-focused AI literacy). This may also enable them to assume technology-oriented consulting roles in L&T industry contexts; see, e.g., the "consulting competence" in the post-editing competence model proposed by Nitzke et al. (2019, 248). At a higher level of abstraction, technical AI literacy may also help L&T stakeholders develop their computational thinking (Celik, 2023, 4) and, related to this, their algorithmic awareness (Gran et al., 2021) and algorithmic agency (Mills and Gutierrez, 2026, 42), ultimately contributing to their digital resilience (Kornacki and Pietrzak, 2024, 63, 65) in highly automated AI-first environments. These arguments in favour of technical AI literacy are supported, to some extent, by feedback on the didactic suitability of the technical curriculum as discussed in Section 3.3.

With its L&T stakeholder focus, the curriculum stands in a tradition of initiatives such as MultiTraiNMT (Kenny, 2022), DataLit^{MT} (Krüger and Hackenbuchner, 2024), adaptMLLM (Lankford et al., 2023) and LT-LiDER (Moorkens et al., 2024).

2 The curriculum

The technical curriculum on language-oriented AI in translation and specialised communication is provided as a GitHub repository[3], which consists of a series of Jupyter notebooks made available under a CC BY-SA 4.0 License (Attribution-ShareAlike 4.0 International) and hosted in Google's Colaboratory (Colab) online environment. The core curriculum presented below is complete. Further notebooks addressing additional aspects of language-oriented AI relevant to the L&T industry will be added to the curriculum in the future.

2.1 Structure of the curriculum

The core curriculum consists of four sections, each assigned to a separate Jupyter notebook. These notebooks cover 1) vector embeddings, 2) the technical foundations of neural networks, 3) tokenization (with a focus on subword tokenization) and 4) the inner workings of transformer neural networks.

Vector embeddings – Representing linguistic data as numbers

The first notebook is concerned with vector embeddings and illustrates to users in a detailed way how linguistic data can be represented as numbers to be processed by neural networks. The notebook covers static/decontextualized and dynamic/contextualized word embeddings as well as sentence embeddings and multilingual embeddings. Users can load a pre-trained static word embedding model and train a small word embedding model themselves, explore individual word and sentence embeddings and embedding relations through visualisations and by calculating the Euclidean Distance and the Cosine Similarity between individual embeddings. The notebook also illustrates how transformer language models – taking BERT (Devlin et al., 2019) as an example – start with static/decontextualized embeddings in their initial embedding layer and convert these into

[3] https://github.com/ITMK/AI_Literacy/

dynamic/contextualized embeddings in their output layer (this process is explored in more detail in the fourth notebook concerned with transformer language models). This notebook is intended to meet L&T stakeholders where they are in their capacity as (aspiring) experts in natural language communication and guides them towards developing a more computational perspective on natural language that can be broadened and deepened in the further sections of the curriculum.

To rephrase this approach in didactic terms, the first notebook leads users "from the zone of current development to the zone of proximal development – the space where one cannot quite master a content/task of their own, but they can with the help of an expert" (Angelone, 2026, 33). The content covered in the subsequent notebooks is then mostly situated in this zone of proximal development. The expert guidance required in this zone can be embedded into the notebooks, to some extent, through adequate didactic scaffolding as discussed in Section 2.2. If the notebooks are used as learning resources in an instructor-led course, the lecturer can provide further expert guidance (see also the discussion in Section 3.2).

Technical foundations of neural networks – Building blocks, forward and backward pass

The second notebook addresses the technical foundations of neural networks and introduces users to the basic building blocks of these networks (neurons, trainable parameters, activation functions), their structure (input layer, hidden layers, output layer) and the flow of information through these networks (forward and backward pass). In this notebook, which intentionally favours didactic simplicity and accessibility over functional completeness, users can create a small experimental neural network in Python by defining the individual layers and the mathematical operations performed within and between these layers. Users can then simulate a simplified forward pass through this neural network, which illustrates how such a network may perform a translation of a short example sentence. The notebook also illustrates through a simplified backward pass how, in neural network training, the network's parameters (weights and biases) would be updated via backpropagation in order to reduce network loss. Building upon the previous notebook, users learn how word vectors are combined into a matrix and how this matrix is processed by the individual layers of the neural network. The notebook is intended to further develop users' computational thinking and algorithmic awareness by illustrating how natural language is actually processed and how natural language semantics is represented within a neural network. At the same time, it introduces users to foundational machine learning concepts and operations such as matrices, tensors, dot-product calculation, non-linear and softmax activation functions etc., which they will encounter again in the context of transformer neural networks.

Tokenization – Reducing the vocabulary size of neural networks

The third notebook is concerned with tokenization, which is employed in language models in order to reduce the size of the vocabulary to be learned. The notebook first guides users through word- and character-based tokenization and discusses the advantages and disadvantages of each method. From there, the notebook guides users through the process of subword-based tokenization, which combines the advantages of word- and character-based tokenization and is the tokenization method actually used in state-of-the-art language models. In this context, users are introduced to three major subword tokenization algorithms, i.e., Byte-Pair Encoding (BPE) (Gage, 1994), WordPiece (Schuster and Nakajima, 2012) and Unigram (Kudo, 2018). Users can then load a BPE, a WordPiece and a Unigram tokenizer, have them each tokenize an example sentence and explore how the three algorithms differ in how they tokenize words and which special characters they employ to designate word boundaries or intra-word splits. Then, users are introduced to the concept of token IDs, which are unique identifiers associated with each token in a language model. Armed with this knowledge, users can access the vocabulary of the language model GPT-2 (Radford et al., 2019) and explore which tokens are stored in this vocabulary. Finally, the notebook illustrates the steps involved in going from subword tokenization to vector embeddings to be processed by a language model.

Tokenization is covered after word embeddings and the technical foundations of neural networks since particularly subword tokenization introduces an additional level of complexity which may be confusing to users at first, but which should be less daunting once they are familiar with word embed-

dings and how these are processed by neural networks and once their computational thinking has evolved accordingly. Also, having encountered the encoder-only transformer model BERT in the first notebook and the decoder-only transformer model GPT-2[4] in the current notebook, users are well-equipped to proceed to the next notebook, specifically concerned with transformer language models.

Transformer neural networks

The fourth notebook builds upon the foundations laid in the previous three notebooks and takes a deep dive into the inner workings of the transformer neural network architecture. Initial focus is placed on encoder-decoder vs. encoder-only vs. decoder-only models and on their respective areas of application and training regimes. In this context, the notebook illustrates how the Hugging Face transformers pipeline[5] defaults to encoder-decoder models for sequence2sequence tasks such as text summarization or translation, to encoder-only models for natural language understanding tasks such as named entity recognition or question answering and to decoder-only models for natural language generation tasks such as text generation/completion. Then, the notebook explores in detail the individual components of the encoder and the decoder side of an encoder-decoder transformer language model. Here, the focus is placed on the original transformer architecture proposed by Vaswani et al. (2017) in the context of NMT research because, a) current LLMs are still strikingly similar to this original transformer architecture (Raschka, 2025a) and b) early transformer models are a bit simpler to understand than current LLMs "but still complex enough to give you a solid grasp of how modern transformer models work" (Raschka, 2025b). When more profound differences exist between the original transformer architecture and current LLMs (e.g., in terms of how positional embeddings are created, which activation functions are used or how layer normalization is performed), the notebook provides an 'architecture update alert', where it briefly discusses the current techniques and points users to relevant literature. On the encoder side, the notebook first illustrates how inputs to the transformer are embedded (here, users can draw on their previous knowledge acquired in the notebook on vector embeddings) and how these embeddings are enriched with positional encoding vectors. Then, the notebook explores in detail the transformer self-attention process (splitting the original input embedding representation into query, key and value representations, dot-product calculation, scaling, padding/masking, softmax normalization etc.), which creates a dynamic/contextualized representation of the initially static/decontextualized input embedding representation. Armed with this conceptual and technical/algorithmic knowledge of transformer self-attention, users can then visualize and explore the self-attention process on their own by using the BertViz package (Vig, 2019). The discussion of the decoder side is more concise because the decoder is conceptually similar to the encoder, meaning that users will already be familiar with many of the components discussed here. Particular focus is placed on masked multi-head attention and on output generation in the final linear and softmax layers of the decoder. Here, users can have GPT-2 complete a text and visualize the top three token probabilities at each generation step. They can also access the user-friendly Hugging Face Decoding Visualizer[6] to further explore greedy decoding and the Hugging Face Beam Search Visualizer[7] to explore the more complex beam search decoding algorithm.

This core curriculum is currently being complemented by further notebooks covering training and adaptation strategies for language models as well as more recent developments in language-oriented AI, such as multimodal LLMs[8], large reasoning and action models, small language models or knowledge-enhanced LLMs (e.g., via RAG or Knowledge Graphs). The notebooks of the core curriculum are also updated on an ongoing basis according to user feedback.

[4]The curriculum relies on older language models such as BERT and GPT-2 because they are functionally less complex than current state-of-the-art models and, more importantly, because they are considerably smaller and can thus be loaded into a Google Colab environment and used in this environment at reasonable speed.

[5]`https://huggingface.co/docs/transformers/main_classes/pipelines`

[6]`https://huggingface.co/spaces/agents-course/decoding_visualizer`

[7]`https://huggingface.co/spaces/m-ric/beam_search_visualizer`

[8]Including models capable of processing spoken language, which is intended to extend the applicability of the curriculum to the field of interpreting as well.

2.2 Didactic makeup of the Jupyter notebooks

The curriculum explores the didactic potential of Jupyter notebooks in introducing audiences with little to no programming knowledge and correspondingly low levels of computational thinking and algorithmic awareness to complex concepts and algorithmic processes from fields such as natural language processing, as investigated in a translation context in Krüger (2022). Jupyter notebooks support *literate computing* (Millman and Pérez, 2018) through a combination of (multimodal) documentation sections providing conceptual information interleaved with executable code cells, which illustrate the algorithmic implementation of these concepts. This combination of documentation and executable code allows authors to use these notebooks to "tell an interactive, computational story" (Barba et al., 2019, 7) to their readers, with a didactic scaffolding tailored to specific target audiences.The target audience *L&T industry stakeholders* as referred to in this paper includes, e.g., students of MA programmes in translation, specialised communication (as central L&T industry fields) and related areas as well as professionals working in the L&T industry. It is assumed that these stakeholders have a professional interest in language-oriented AI but have not been exposed in detail to the technical foundations of these technologies. It is further assumed that these stakeholders have had moderate to no exposure to computer programming before, making them novices or advanced beginners in terms of computational thinking.

Figure 1 depicts a typical example of the didactic makeup of the Jupyter notebooks within the technical curriculum. The left screenshot is concerned with a forward pass of a simple example sentence through an experimental translation neural network and illustrates how the 'semantics' of a particular input word is processed and represented by a hidden layer in the network (excluding the activation function at this point). The right screenshot is concerned with self-attention in a transformer language model. The code cells implement BertViz to visualise the BERT model's self-attention process performed on a predefined example sentence (which users are encouraged to change in order to experiment with their own sentences). The documentation sections explain how to interpret the BertViz visualisation. Short exercises then invite users to engage more deeply with the topic, in the current example also highlighting points of contact with topics covered at a previous point in the curriculum.

3 Measuring the didactic suitability of the curriculum

The core curriculum presented above was tested in the MA course "Foundations of Artificial Intelligence for Specialised Communication" in the MA in Multilingual Specialised Communication and Specialised Translation programme (code: *MAFKÜ*) and the MA in Terminology and Language Technology programme (code: *MATS*) at TH Köln's Institute of Translation and Multilingual Communication. The course was intended for first-year students of the two MA programmes and was also attended by three external guest students who are experienced professional translators working in public-sector language services (code: *professionals*). Participants completed a pre-test questionnaire (October 2025) and a post-test questionnaire (January 2026). The study was conducted in accordance with TH Köln's Regulations for Safeguarding Good Scientific Practice[9] and Code of Conduct for Research and Transfer[10]. The questionnaires were derived from the draft version of the TransAI Literacy Scale (TrAILS) proposed in Krüger (forthcoming) and were administered in German.[11]

The pre-test questionnaire was completed by 24 participants. Of these, 12 (50%) were MAFKÜ students, 9 (37.5%) were MATS students and 3 (12.5%) were professionals (item 0.3). The latter group had a professional experience of more than 20 years, while the professional experience of the MATS and MAFKÜ students varied between <1 year (66.7%), 1 to <5 years (12.5%), 5 to <10 years (4.2%) and 10 to <20 years (4.2%) (item 0.2). The post-test questionnaire was completed by 15 participants, resulting in a survey attrition rate of 37.5%. The composition of the post-test group

[9] https://www.th-koeln.de/mam/bilder/forschung/2023-05-31_gwp_en.pdf

[10] https://www.th-koeln.de/mam/bilder/forschung/verhaltenskodex_maerz_2021_en-us.pdf

[11] Draft version of TrAILS: https://forms.gle/QdhPiP9D3F6rUca4A
Pre-test questionnaire: https://forms.gle/VNwL8LpNdsKh6HLd8
Post-test questionnaire: https://forms.gle/VBzvcFub3EtSLSAt6

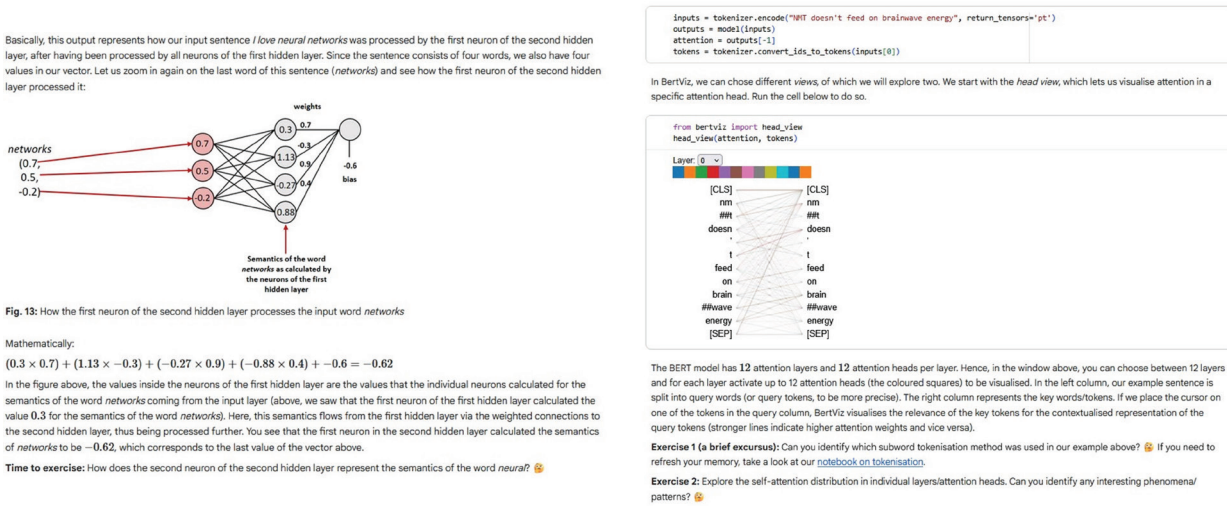

Figure 1: Example of the didactic makeup of the Jupyter notebooks – left: processing of embeddings in a hidden network layer (excluding activation); right: visualising and exploring self-attention in a transformer language model

shifted to 6 MATS students (40%), 5 MAFKÜ students (33.3%), 3 in-house professionals (20%, this group remained constant) and 1 freelance professional (6.7%).[12]

This shift is relevant for interpreting the results: The MATS curriculum includes a higher share of IT-related courses than the MAFKÜ curriculum (among other things, MATS students attended a course on markup languages, which ran parallel to the AI course and they will attend a separate course on Python programming in the second semester). Since attrition was considerably higher among MAFKÜ students compared to MATS students (58.33% vs. 33.33%), the post-test group likely exhibits a higher baseline of IT competence and technical disposition than the initial pre-test group. This is supported by the participants' self-assessment related to IT competence (item 0.6) and interest in the technical foundations of language-oriented AI (item 0.5). The share of participants reporting high or very high IT competence almost doubled from 20.9% in the pre-test to 40% in the post-test. Similarly, the share of participants reporting a high or very high interest in the technical foundations of language-oriented AI increased from 70.8% in the pre-test to 80% in the post-test.[13]

3.1 Overall didactic effectiveness of the curriculum

The participants' language-oriented AI knowledge was measured using three questionnaire items (items 1.1 to 1.3) with an 11-point Likert scale ranging from 0 = *Hardly/Not at all* to 10 = *Perfectly* without intermediate scale labelling.[14] The questions were asked in both the pre- and the post-test in order to measure participants' self-reported knowledge gain. Results are presented in table 1.

We focus here on the composite index, which represents the mean of the responses to items 1.1 to 1.3 for each participant (serving as an aggregate of participants' self-assessed knowledge of the technical foundations of language-oriented AI). The increase in participants' self-assessed technical knowledge from $M = 3.72$ ($SD = 2.13$, $N = 24$) in the pre-test to $M = 6.76$ ($SD = 1.43$, $N = 15$) in the post-test is highly statistically significant ($p<0.001$)[15], with a very large effect size

[12] Since the study did not use constant participant IDs for pre- and post-test, it is unclear whether the participant who identified as a freelance professional in the post-test questionnaire identified as a MAFKÜ or a MATS student in the pre-test questionnaire (both questionnaires asked participants for their 'primary role' [item 0.3]; no further external guest students joined the course between the pre-test and the post-test). The attrition rate may stem from the course being offered every two semesters; consequently, some enrolled students may have decided to defer their participation to a future term.

[13] There may be a combination of a selection effect (higher retention rate of students with a higher initial IT affinity; [very] high IT competence pre-test MAFKÜ group: 8.33% vs pre-test MATS group: 22.22% vs. pre-test professionals: 66.67%) and a treatment effect (increase in IT competence through contents covered in the technical curriculum) at work here. This will be discussed in more detail in Section 3.1.

[14] *I can explain the basics of how modern language-oriented AI technologies work* (item 1.1), *I can explain how modern language-oriented AI technologies are trained and fine-tuned* (item 1.2), *I know if language-oriented AI supports the translation technologies I use and, if so, how* (item 1.3).

[15] Welch's t-test for independent samples was used to account for the lack of consistent participant IDs across pre- and post-test and to address unequal group sizes and variances

Knowledge area	Pre-test ($M \pm SD$)	Post-test ($M \pm SD$)	p-value	Cohen's d
1.1 Basic operating principle	3.67 ± 2.24	6.73 ± 1.33	< 0.001	1.58
1.2 Training/Fine-tuning	3.04 ± 2.18	5.87 ± 1.77	< 0.001	1.39
1.3 AI-supported trans tech	4.46 ± 2.77	7.67 ± 2.09	< 0.001	1.27
Composite index	$\mathbf{3.72 \pm 2.13}$	$\mathbf{6.76 \pm 1.43}$	$\mathbf{< 0.001}$	**1.60**

$N_{Pre} = 24$, $N_{Post} = 15$, 11-point Likert scale from 0 (*Hardly/Not at all*) to 10 (*Perfectly*)

Table 1: Self-reported knowledge gain of participants with regard to the technical foundations of language-oriented AI

of $d = 1.60$. This suggests a high overall didactic effectiveness of the technical curriculum, taking into account that generalizability may be limited due to the relatively small convenience sample size.

The post-test questionnaire also included a retrospective question where participants were asked to reassess their technical knowledge of language-oriented AI at the beginning of the course.[16] This was done to account for a potential Dunning-Kruger effect at the onset of the study. The results of this retrospective self-assessment ($M = 2.93$, $SD = 2.19$) were notably lower than the initial self-assessment ($M = 3.72$, $SD = 2.13$), suggesting a response shift. As the curriculum introduced participants in detail to the technical foundations of language-oriented AI, they likely became aware of previous knowledge gaps that may have skewed their initial self-assessment. This indicates that the knowledge gain associated with the technical curriculum was even more pronounced than the initial pre-post comparison suggests. The response shift may also indicate an increase in participants' metacognitive awareness in terms of a more realistic picture of their own level of competence with regard to language-oriented AI.[17]

As mentioned above, the interpretation of results must take into account a potential selection effect (students with higher initial IT affinity completing the course and students with lower initial IT affinity leaving the course). However, both the observed response shift and the very large effect size of $d = 1.60$ (or $d = 2.07$ for the retrospective self-assessment of $M = 2.93$, $SD = 2.19$) provide evidence for a genuine treatment effect. These factors indicate that the self-reported knowledge gain indeed resulted from participants' engagement with the technical topics covered in the curriculum, rather than being merely an effect of participant self-selection.

3.2 Didactic potential of Jupyter notebooks

The post-test questionnaire also asked participants to assess the didactic potential of Jupyter notebooks in the context of the technical curriculum. When participants were asked whether these notebooks were a suitable didactic instrument for covering the technical course contents (item 2.1), 12 (80%) strongly agreed, 1 (6.7%) agreed and 2 (13.3%) neither agreed nor disagreed. This largely confirms the results reported in Krüger (2022, 519–520) concerning the didactic potential of Jupyter notebooks in translation technology teaching. When asked specifically whether the combination of multimodal documentation (text + figures + linked videos) and executable code cells helped participants to understand the technical contents of the curriculum (item 2.2), 11 (73.3%) strongly agreed, 3 (20%) agreed and 1 (6.7%) neither agreed nor disagreed, which can be taken as further evidence in favour of the notebooks' didactic suitability.

Participants in the post-test study could also provide open comments on the didactic potential of the Jupyter notebooks for covering the technical course contents (five participants provided such comments). Overall, feedback was positive. E.g., according to a MATS student, the notebooks *illustrated the content very clearly and generally helped me to review the lecture material at my leisure (the notebooks were also very detailed, which was also helpful)*.[18] A MAFKÜ student found *the structure and detail of the descriptions/explanations very helpful. Better than slides*

(*scipy.stats.ttest_ind*, *equal_var=False*).

[16] *Looking back at the beginning of the course, how would you now assess your knowledge of the technical foundations of language-oriented AI at that time?* (item 1.4, 11-point likert scale ranging from 0 = *Hardly any/No knowledge* to 10 = *Complete/Perfect knowledge*)

[17] Metacognition has also been discussed as a dimension of AI literacy acquisition in that it "[enables] individuals to reflect on their learning processes and understand how AI can enhance their cognitive abilities." (Tadimalla and Maher, 2024, 3)

[18] Participants provided all open comments in German. The comments were machine-translated into English using DeepL and post-edited by the author.

with bullet points, because it gives you the opportunity to read up on something that has already been explained in detail by the lecturer in the seminar. The same student reported having difficulties saving the notebooks locally (which could have been achieved through adequate lecturer support). Another MAFKÜ student made an interesting point regarding further didactic scaffolding in addition to the notebooks: *In the context of our course and your explanations of the content of the lines of code, I found these very helpful. However, I'm not so sure whether I—or other users who have no prior knowledge of coding—would have been able to learn well with these alone, especially since the relevant information was often hidden in long lines of code*. This indicates that, in the zone of proximal development, expert guidance cannot be fully embedded into the notebooks themselves and should also be provided by the lecturer (in this specific case by highlighting relevant portions of longer and potentially complex code cells). Future iterations of the AI course will also introduce students to using LLMs as coding assistants for producing, highlighting, explaining, changing or debugging computer code, as discussed in a translation context by Krüger et al. (forthcoming). Such LLM integration may offer further didactic potential, e.g., in terms of using these models as general learning assistants facilitating deliberate practice (Angelone, 2026) in addition to lecturer support or in self-learning scenarios where no such lecturer support is available.[19]

3.3 Further feedback on the technical curriculum

When asked whether the progression of the course contents from 1) vector embeddings via 2) the technical foundations of neural networks and 3) subword tokenization to 4) transformer neural networks was suitable from a didactic point of view (post-test questionnaire, item 3.1), 12 participants (80%) found the order of topics highly suitable and 3 (20%) found it rather suitable. In an open comment, one of the professional participants noted that they would have liked a final overview of how the different AI concepts covered in the course relate to each other. This will be implemented in a future version of the curriculum.

Finally, participants were asked to provide concluding feedback (in the form of open comments) on the AI course and the technical curriculum, e.g., in terms of the usefulness of the course in the context of their MA programme or with a view to participants' (future) careers, in terms of the appropriateness of the technical level at which the content was covered or in terms of any content that was missing or was covered too briefly or superficially (item 4.1). Seven participants provided concluding feedback.

The technical level was considered demanding (particularly in the context of transformer language models), but participants largely agreed that it could be mastered using the didactic resources provided. A MAFKÜ participant noted: *Although I consider the course to be challenging, I find that the content is explained very clearly and from a perspective that is easy for me to understand (as a translator)*. A professional translator participant commented that, at times, there was a risk of getting lost in too much technical detail. Another MAFKÜ participant responded that, although the lecturer always encouraged students to ask questions if they did not understand something, it was sometimes difficult to put such questions into words in the first place. These issues will have to be addressed in future iterations of the course, also taking into account the didactic potential of LLMs as coding or general learning assistants as discussed above.

Overall, participants acknowledged the relevance of the course and the curriculum for their (later) professional career. A MAFKÜ student noted that even if a technical understanding of language-oriented AI was not directly relevant in students' future careers, *I believe that this background knowledge gives participants a certain degree of confidence*. A professional translator noted that security restrictions currently prevented them from using language-oriented AI technologies extensively at their workplace but that *I can at least discuss with our IT department what is important for language service providers, make suggestions and discuss things on an equal footing*. These comments support the hypothesis stated in Section 1 that technical AI literacy as acquired through the technical curriculum includes an empowerment dimension and may strengthen L&T stakeholders'

[19] LLM's potential as learning assistants may be further reinforced by the possibility to prompt these models to tailor the "level of explanatory ambition" (Engberg, 2025, 82) to individual knowledge requirements (e.g., explaining the concept of a backward pass through a neural network to MSc students in information technology vs. MA students in translation/specialised communication).

(algorithmic) agency (e.g., by discussing AI 'on an equal footing' with IT experts and/or consulting on workflow design) and may contribute to their overall digital resilience (in terms of confidence) in highly automated environments.

4 Conclusion and outlook

This paper presented a technical curriculum on language-oriented AI in translation and specialised communication, which stands in the tradition of other L&T stakeholder-specific initiatives such as MultiTraiNMT, DataLit^{MT}, adaptMLLM and LT-LiDER. The overall didactic suitability of the curriculum was tentatively confirmed in an AI-focused MA course at TH Köln. Participants acknowledged the relevance of the course and the technical curriculum for their (later) professional careers, with individual participant feedback supporting the hypothesis that technical AI literacy may strengthen L&T stakeholders' (algorithmic) agency and digital resilience in professional high-automation contexts. Participants also confirmed the didactic potential of Jupyter notebooks for covering the technical course contents. However, participant feedback also indicated that the technical content could be overwhelming at times or was only manageable through additional lecturer support. This indicates that, in the zone of proximal development, where learners are challenged by the novelty and/or complexity of content, the required expert guidance cannot be fully embedded into the Jupyter notebooks and should also be provided through higher-level didactic scaffolding, ideally in terms of lecturer support. Participant feedback will be implemented into the curriculum on an ongoing basis in order to tailor it further to the specific knowledge requirements of the target audience. This is intended to strengthen the curriculum's *human-centered explainable AI* (HCXAI) dimension, which aims to place non-expert users at the center of AI explainability (Ridley, 2025, 98). Also, future work will investigate strategies for introducing LLMs as coding or general learning assistants into the curriculum, which could facilitate deliberate practice in self-learning scenarios.

Acknowledgements: The author would like to thank his colleagues Janiça Hackenbuchner and Erik Angelone for their valuable comments and suggestions on the manuscript version of this paper.

References

Angelone, Erik. 2026. Generative AI as a facilitator of deliberate practice in translator training. In Penet, JC, Joss Moorkens, and Masaru Yamada, editors, *Teaching Translation in the Age of Generative AI: New Paradigm, New Learning?*, page 27–47. Language Science Press, Berlin.

Barba, Lorena A., Lecia J. Barker, Douglas S. Blank, Jed Brown, Allen B. Downey, Timothy George, Lindsey J. Heagy, Kyle T. Mandli, Jason K. Moore, David Lippert, Kyle E. Niemeyer, Ryan R. Watkins, Richard H. West, Elizabeth Wickes, Carol Willing, and Michael Zingale. 2019. *Teaching and Learning with Jupyter*.

Bowker, Lynne. 2026. Teaching translation students about data in the age of generative AI. In Penet, JC, Joss Moorkens, and Masaru Yamada, editors, *Teaching Translation in the Age of Generative AI: New Paradigm, New Learning?*, pages 67–85. Language Science Press, Berlin.

Celik, Ismail. 2023. Exploring the determinants of artificial intelligence (AI) literacy: Digital divide, computational thinking, cognitive absorption. *Telematics and Informatics*, 83:1–11.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, Jill, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis. Association for Computational Linguistics.

Doherty, Stephen and Dorothy Kenny. 2014. The design and evaluation of a statistical machine translation syllabus for translation students. *The Interpreter and Translator Trainer*, 8(2):295–314.

Engberg, Jan. 2025. Specialised communication and cognition. In Roelcke, Thorsten, Ruth Breeze, and Jan Engberg, editors, *Specialized Communication. An International Handbook*, page 67–86. De Gruyter Mouton, Berlin/Boston.

Gage, Philip. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Gran, Anne-Britt, Peter Booth, and Taina Bucher. 2021. To be or not to be algorithm aware: A question of a new digital divide? *Information, Communication & Society*, 24(12):1779–1796.

Jiménez-Crespo, Miguel A. 2025. Human-centeredness in translation. Advancing translation studies in a human-centered AI era. *InContext*, 5(1):7–17.

Kenny, Dorothy. 2019. Machine translation. In Rawling, Piers and Philip Wilson, editors, *The Routledge*

Handbook of Translation and Philosophy, pages 428–445. Routledge, London/New York.

Kenny, Dorothy, editor. 2022. *Machine Translation for Everyone*. Language Science Press, Berlin.

Kornacki, Michał and Paulina Pietrzak. 2024. *Hybrid Workflows in Translation: Integrating GenAI into Translator Training*. Routledge, London/New York.

Krüger, Ralph and Janiça Hackenbuchner. 2024. A competence matrix for machine translation-oriented data literacy teaching. *Target*, 36(2):245–275.

Krüger, Ralph and Erik Angelone. forthcoming. The AI age in the language and translation industry – Opportunities for automation and literacy requirements. In Jiménez-Crespo, Miguel A. and Vanessa Enríquez Raído, editors, *The Routledge Handbook of Translation and Technology (2nd ed.)*. Routledge, London/New York.

Krüger, Ralph, Sergi Álvarez Vidal, and Janiça Hackenbuchner. forthcoming. Bridging the digital divide – Making Python and the Python ecosystem accessible to translators. In Castilho, Sheila, Dorothy Kenny, Joss Moorkens, and María Isabel Rivas Ginel, editors, *Developing AI Literacy for Translation*. Language Science Press, Berlin.

Krüger, Ralph. 2022. Using Jupyter notebooks as didactic instruments in translation technology teaching. *The Interpreter and Translator Trainer*, 16(4):503–523.

Krüger, Ralph. 2024. Outline of an artificial intelligence literacy framework for translation, interpreting and specialised communication. *Lublin Studies in Modern Languages and Literature*, 48(3):11–23.

Krüger, Ralph. forthcoming. Artificial intelligence literacy for the language and translation industry – Conceptual foundations, operationalisation, acquisition, measurement. In Castilho, Sheila, Dorothy Kenny, Joss Moorkens, and María Isabel Rivas Ginel, editors, *Developing AI Literacy for Translation*. Language Science Press, Berlin.

Kudo, Taku. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In Gurevych, Iryna and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne. Association for Computational Linguistics.

Lambson, Nick and Lintao Han. 2025. Localization in the AI era. In Sun, Sanjun, Kanglong Liu, and Riccardo Moratto, editors, *Translation studies in the age of artificial intelligence*, pages 62–84. Routledge, London/New York.

Lankford, Séamus, Haithem Afli, and Andy Way. 2023. adaptMLLM: Fine-tuning multilingual language models on low-resource languages with integrated LLM playgrounds. *Information*, 14(12):1–24.

Millman, K. Jarrod and Fernando Pérez. 2018. Developing open-source scientific practice. In Stodden, Victoria, Friedrich Leisch, and Roger D. Peng, editors, *Implementing Reproducible Research*, pages 149–184. CRC Press, Boca Raton.

Mills, Kathy A. and Amanda Gutierrez. 2026. *Critical Literacy in an AI World*. Routledge, London/New York.

Moorkens, Joss, Pilar Sánchez-Gijón, Esther Simon, Mireia Urpí, Nora Aranberri, Dragoș Ciobanu, Ana Guerberof-Arenas, Janiça Hackenbuchner, Dorothy Kenny, Ralph Krüger, Miguel Rios, Isabel Ginel, Caroline Rossi, Alina Secară, and Antonio Toral. 2024. Literacy in digital environments and resources (LT-LiDER). In Scarton, Carolina, Charlotte Prescott, Chris Bayliss, Chris Oakley, Joanna Wright, Stuart Wrigley, Xingyi Song, Edward Gow-Smith, Mikel Forcada, and Helena Moniz, editors, *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 55–56, Sheffield, UK. European Association for Machine Translation (EAMT).

Nitzke, Jean, Silvia Hansen-Schirra, and Carmen Canfora. 2019. Risk management and post-editing competence. *Journal of Specialised Translation*, 31:239–259.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI.

Raschka, Sebastian. 2025a. The big LLM architecture comparison. From DeepSeek-V3 to Kimi K2: A look at modern LLM architecture design. Ahead of AI Blog.

Raschka, Sebastian. 2025b. From GPT-2 to gpt-oss: Analyzing the architectural advances and how they stack up against Qwen3. Ahead of AI Blog.

Ridley, Michael. 2025. Human-centered explainable artificial intelligence: An annual review of information science and technology (ARIST) paper. *Journal of the Association for Information Science and Technology*, 76(1):98–120.

Rivas Ginel, María Isabel and Joss Moorkens. 2025. Translators' trust and distrust in times of GenAI. *Translation Studies*, 18(2):283–299.

Schuster, Mike and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152, Kyoto.

Shneiderman, Ben. 2020. Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction*, 12(3):109–124.

Tadimalla, Sri Yash and Mary Lou Maher. 2024. AI literacy for all: Adjustable interdisciplinary socio-technical curriculum. In Phillips, Margaret, Jason B. Reed, and Dave Zwicky, editors, *2024 IEEE Frontiers in Education Conference (FIE)*, pages 1–9, Los Alamitos. IEEE Computer Society.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Guyon, I., U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 1–11. Curran Associates, Inc.

Vig, Jesse. 2019. A multiscale visualization of attention in the transformer model. In Costa-jussà, M. R. and E. Alfonseca, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence. Association for Computational Linguistics.

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:

Proceedings of the 1st International Workshop on Teaching AI-Based Translation and Technologies (TAITT 2026), pages 12–18
June 15, 2026 ©2026 Association for Computational Linguistics

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.

Mimicking Neural Machine Translation History for Pedagogic Reasons

Vincent Vandeghinste
KU Leuven, Brussels, Belgium
Instituut voor de Nederlandse Taal, Leiden, the Netherlands
vincent@ccl.kuleuven.be

Abstract

In this paper we describe how we mimic the different steps in the historical development of NMT systems, all trained and evaluated on the same small data set. We do this for pedagogic reasons so students can see the effect of each of the steps on metrics like BLEU but also on qualitative examples, which the training scripts generate after each epoch. As MT paradigms, we discuss NMT training from scratch, fine-tuning pre-trained encoder-decoder models, and finally prompt engineering for decoder-only models. All models run in Kaggle sessions and all Python scripts and Jupyter notebooks are made available to the MT teaching community through GitHub and public Kaggle sessions.

1 Introduction

Neural machine translation (NMT) has undergone rapid architectural evolution during the past decade. Early sequence-to-sequence models based on recurrent neural networks (RNNs) were quickly followed by attention mechanisms, subword modeling, transformer architectures, multilingual models, pre-trained models, and more recently decoder-only large language models.

For translation students encountering machine translation (MT) for the first time, this rapid development can make the field difficult to understand. Modern state-of-the-art systems rely on highly complex pretrained architectures whose design motivations are not immediately obvious. Introducing such systems directly may therefore obscure the conceptual foundations of neural machine translation.

In this paper, we describe a pedagogical approach that addresses this challenge by mimicking the historical development of NMT. Rather than presenting current architectures directly, students learn to reproduce simplified versions of the models that marked major milestones in the development of the field, while also getting used to the MT experimental research paradigm.

This approach is implemented as the main part in the course *Machine Translation Advanced Topics* in the Translation Technology post-graduate programme for translators at KU Leuven. There are 27 contact hours and this counts for 3 ECTS points, so expecting a total average work load for students of 90 hours, including contact hours.

Until recently, we had been using OpenNMT (Klein et al., 2017), an open-source NMT platform that allowed easy training of NMT systems without much programming or other technical knowledge, making it an ideal tool to use in classes. But as development to OpenNMT was discontinued, it became very difficult to reconstruct the NMT history.

This is the main reason why we have developed a number of Python scripts with lots of command-line switches that mimic some of the functionality of OpenNMT. We have also created a series of Google Colab / Kaggle sessions in which these scripts are used.[1] Everything is made available via GitHub.[2]

[1] We use Kaggle as an alternative to the often used Google Colab environment, as Kaggle allows running Python notebooks while the user is offline and it has a fixed GPU quotum of currently 30 hours of GPU time per week.

[2] https://github.com/VincentCCL/MTAT/

Proceedings of the 1st International Workshop on Teaching AI-Based Translation and Technologies (TAITT 2026), pages 19–27
June 15, 2026 ©2026 Association for Computational Linguistics

In what follows, we show through evaluation with SacreBLEU (Post, 2018) (which was built-in in the scripts) and example translations how translation quality improves over the different inventions that have led to the current state-of-the-art in MT.

Section 2 discusses the dataset that was used. Section 3 discusses NMT training from scratch. Section 4 discusses building a multilingual NMT system and demonstrates a zero-shot system. Section 5 discusses fine-tuning and translation with pre-trained encoder-decoder models. Section 6 discusses prompt engineering with decoder-only models. And finally, Section 7 draws conclusions.

2 The dataset

All the Kaggle sessions make use of a dataset that was downloaded from OPUS (Tiedemann, 2012). We chose the Tatoeba (Tiedemann, 2020) English-Dutch dataset as it is rather small, in order to avoid long waiting times during class. Table 1 shows details about the dataset.

Tatoeba v2023-04-12 Properties	
nr of parallel sentences	79,541
nr of English tokens	579,984
nr of Dutch tokens	586,024

Table 1: Tatoeba v2023-04-12 properties

These data have been preprocessed similar as what is described in DataLit^{MT}'s (Hackenbuchner and Krüger, 2023) section on Data Preprocessing: entries with missing sentences are dropped, as well as duplicates and pairs where source and target sentence are equal, leaving us with 79,482 sentences, removing 59 sentence pairs (0.07%).

The next step consists of removing sentence pairs where either of the sentences contains more than 50 spaces, and those where the ratio between the source sentence number of words and the target sentence number of words (or vice versa) exceeds 2.5, removing another 101 sentences (another 0.12% of the initial set), leaving us with 79,381 sentences.

Then the data was tokenized with the SacreMoses tokenizer[3] and split into a training set, a development set and a test set. We limited the size of the development and test sets to 1000 parallel sentences each, again in order to not have to wait too long when running evaluations in a classroom setting.

[3] `https://pypi.org/project/sacremoses/`

In the remainder of this paper, we will not use or report about the test set, as this is kept as unseen data for students to perform a final evaluation once the best-performing systems have been selected based on the development set. Nevertheless, the cleaned data set with splits is made publicly available as a Kaggle dataset.

After each epoch, the training scripts show, depending on the options, the development set BLEU scores (Papineni et al., 2002) and a number of example sentences from the development set, in order to show students how results evolve during training, enabling them to see qualitative progress beside quantitative progress.

Table 2 shows 5 lowercased example sentences from the development set, with their lowercased reference translations, for which translation hypotheses are shown after each epoch.

Nr	**Src/Ref**	**Sentence** (lower cased)
1	SRC	nobody reads about my country .
	REF	niemand leest over mijn land .
2	SRC	i sleep in my car .
	REF	ik slaap in mijn auto .
3	SRC	there 're clean sheets under the bed .
	REF	er liggen schone lakens onder het bed .
4	SRC	i have a donkey .
	REF	ik heb een ezel .
5	SRC	betty drives fast .
	REF	betty rijdt snel .

Table 2: 5 example sentences from the development set

3 MT Models trained from scratch

We first discuss basic recurrent neural network models in Section 3.1, add subwording in Section 3.2, move to transformers in Section 3.3 and then present results on all the discussed models in Section 3.4.

3.1 Basic RNN models

After some explanation on how neural networks, word embeddings and plain RNN language modeling work, we start with training the RNN baseline, that is, a plain vanilla RNN encoder-decoder model (Cho et al., 2014; Sutskever et al., 2014). Table 3 shows the default hyperparameters for the RNN training script in the RNN column.

After the baseline, we have accordingly added gating to the RNN cells, comparing performance on Long Short-term Memory cells (LSTMs) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Units (GRU) (Cho et al., 2014), added

Setting	RNN	Transformer
Epochs	10	10
Batch size	32	32
Embedding size	64	512
Hidden size	128	2048
Encoder layers	1	6
Decoder layers	1	6
Number of heads	-	8
RNN type	RNN	-
Attention	none	self
Bidirectional encoder	no	-
Max sentence length	20	128
Learning rate	0.001	0.0005
Teacher forcing	0.7	1.0
Gradient clipping	1.0	1.0
Optimizer	Adam	AdamW
Lowercasing	yes	no
Subwording	none	SentPiece
Max source vocab	unlimited	8000
Max target vocab	unlimited	8000
Seed	42	42

Table 3: Default hyperparameters for the RNN and Transformer scripts. Some of these hyperparameters have been explained to the students, but for many this would lead too far as it would involve a technicality in maths which is not expected from translators. We provide them here mainly for reproducibility.

bidirectionality to LSTM (*biLSTM*), and added cross-lingual attention (*att*) (Luong et al., 2015).

3.2 Subwording with SentencePiece

We also test the effect of applying SentencePiece subword tokenization (Kudo and Richardson, 2018) with unigrams, in two variants: with separate sentence piece models for source and target languages (see +*sp* in Figure 1 and Table 4) and with a single joint model for source and target language (see +*spj*). We always set vocabulary size at 8000.

3.3 Transformer models

For transformers we made a script that uses the `transformers` library from Hugging Face. We ran two versions of the Transformer model, a small version using as much as possible the hyperparameters from the RNN network (i.e., changing the default settings from the Transformer column in Table 3 for embedding size to 64, hidden size to 128 and maximum generated length to 20. Then we also trained a larger transformer model with the settings given in the Transformer column in Table 3. Note that for Transformers we have used separate sentence piece models for source and target language. Testing the accuracy of the model with a joint sentence piece model is left as an exercise for the students.

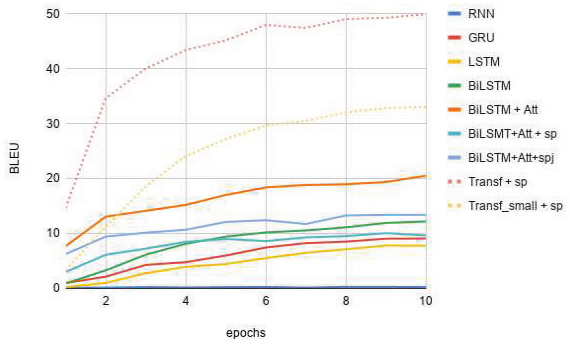

Figure 1: BLEU scores for the different models that were trained from scratch, over 10 epochs

3.4 Results

Figure 1 presents the learning curves for BLEU scores on the development set, clearly showing improvements with each newly introduced step. The difference betweeen LSTM and GRU is rather negligible.

Table 4 shows the hypotheses of the respective MT models after 10 epochs. As we can see while vanilla RNNs only produce some high-frequency words, later models already may get some words right. As soon as we start introducing bidirectionality, we see output increasingly resembling full sentences. When attention is added results are even better, but still far from decent translations.

As Figure 1 shows, using sub-wording (*sp* and *spj* conditions) in our case does not improve further BLEU scores, but there is clearly a higher BLEU score for the joint model over the separate models.

Both transformer models score substantially better than the RNN models, and the larger transformer model clearly outperforms the small transformer. The improved performance is also visible from the generated examples, where the small transformer already shows some correctly translated sentences, and for the larger transformer we see only a single mistake in the five sentences.

4 Multilingual NMT (MNMT) models: testing zero-shot translation

We also want to demonstrate how multilingual models work. We take the French to Dutch Tatoeba set as an additional dataset, which was processed in the same way as was done in Section 2.

We created a function that takes a source file, a target tag and an output filename as its arguments and we use it on the files as indicated in Table 5.

Model	Nr	Hypotheses
RNN	1	is een het . .
	2	is een het . .
	3	is niet het van een . .
	4	is is . .
	5	is is . .
GRU	1	lijkt dat mijn mijn land . .
	2	heb in mijn auto . .
	3	zijn onder bed de bed .
	4	ben een . .
	5	rijdt snel . .
LSTM	1	leest op mijn land .
	2	ben in mijn mijn .
	3	heeft onder onder die niet . .
	4	heb een groot .
	5	rijdt snel .
BiLSTM	1	verdedigde over land land .
	2	slaap in mijn auto .
	3	zijn veel onder bed de bed .
	4	heb een . .
	5	rijdt beter .
BiLSTM	1	leest op mijn land .
+ att	2	slaap in mijn auto .
	3	zijn schoon op onder bed bed .
	4	heb een ezel .
	5	rijdt snel .
BiLSTM	1	las over mijn land .
+ att	2	ik slaapt in mijn auto .
+ sp	3	hang zijn bed onder bed onder bed .
	4	ik heb een . .
	5	t rijdt snel .
BiLSTM	1	leest van mijn land .
+ att	2	ik slaap in mijn auto
+ spj	3	zijnnt laket onder onder bed bed
	4	ik heb een ezel .
	5	tty rijd snel snel
Transformer	1	Niemand leest over mijn land .
small	2	Ik slaapt in mijn auto .
	3	Er zijn ketchup onder het bed .
	4	Ik heb een boot .
	5	Bet niet snel rijden .
Transformer	1	Niemand leest over mijn land .
	2	Ik slaap in mijn auto .
	3	Er zit schone lakens onder het bed .
	4	Ik heb een ezel .
	5	Betty rijdt snel .

Table 4: MT from scratch examples after 10 epochs. The **Nr** column refers to the example Nr in Table 2.

Original File	Tag	Ouput File
tatoeba-en-nl/train.en	<2nl>	en-nl-train.en_2nl
tatoeba-en-nl/dev.en	<2nl>	en-nl-dev.en_2nl
tatoeba-fr-nl/train.fr	<2nl>	fr-nl-train.fr_2nl
tatoeba-fr-nl/dev.fr	<2nl>	fr-nl-dev.fr_2nl
tatoeba-en-nl/train.nl	<2en>	en-nl-train.nl_2en
tatoeba-en-nl/dev.nl	<2en>	en-nl-dev.nl_2en
tatoeba-fr-nl/train.nl	<2fr>	fr-nl-train.nl_2fr
tatoeba-fr-nl/dev.nl	<2fr>	fr-nl.dev.nl_2fr

Table 5: Adding target language tags for multilingual MT

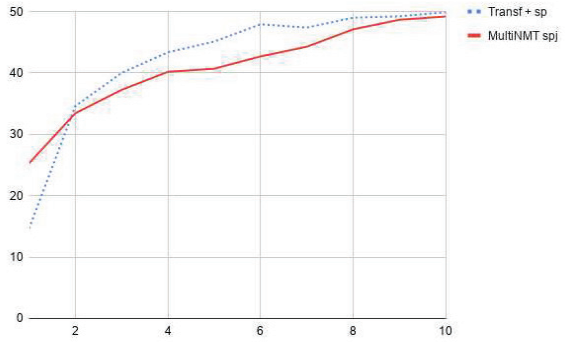

Figure 2: BLEU score for the multilingual NMT model. For ease of reference, the best-scoring system from Figure 1 is shown in a dotted line. Note that the two are not fully comparable, as the NMNT system is evaluated on a multilingual NMT file, not only on EN-NL.

Model	Nr	Hypotheses
MNMT	1	Niemand laira dans mon pays.
	2	Il y a du courant dans mon auto.
	3	Il y a des pteaux dans le lit.
	4	Il y a du cou.
	5	Betty rijdt snel.

Table 6: Translation hypotheses of the five example sentences in the zero-shot EN-FR multilingual setting

We create four datasets, in which the source sentences are prefixed with a target language tag. All training data was concatenated and a joint SentencePiece model was built. Then a transformer was trained with default settings (cf. Table 3) on these data. Results are presented in Figure 2, although the BLEU score is not fully comparable as the evaluation was done on the concatenation of the four development tests (See table 5).

Next, we tested the zero-shot capabilities of our model by translating the English development set in French. Table 6 shows example translations, from which we can see that the translations are not very good, as they contain a mixture of French and Dutch. We cannot evaluate this setting with metrics as we do not have the reference translations from English to French for these sentences.

5 Fine-tuning and using pre-trained Encoder-Decoders

For fine-tuning pre-trained encoder-decoders we have made separate scripts for T5 and mBART. For translation without fine-tuning, we also add M2M, NLLB and MADLAD-400.

Model	Nr	Hypotheses
T5	1	Niemand leest over mijn land.
	2	Ik slaapt in mijn auto.
	3	Er zijn snel bleeks op het bed.
	4	Ik heb een ees.
	5	Betty rijdt snel.
mT5	1	Niemand leest over mijn land .
	2	Ik slaap in mijn auto .
	3	Er zijn schoon kussens onder het bed .
	4	Ik heb een doek .
	5	Betty rijdt snel .
mBART (9 epochs)	1	Niemand leest over mijn land .
	2	Ik slaap in mijn auto .
	3	Onder het bed liggen schone lakens .
	4	< > Ik heb een ezel .
	5	Betty rijdt snel .
M2M (no finetuning)	1	Niemand leest over mijn land.
	2	Ik slapen in mijn auto.
	3	Er zijn schoon bladeren onder het bed.
	4	Ik heb een donkey.
	5	Betty rijdt snel.
NLLB (no finetuning)	1	Niemand leest over mijn land .
	2	Ik slaap in mijn auto .
	3	Er zijn schone lakens onder het bed .
	4	Ik heb een ezel .
	5	Betty rijdt snel .

Table 7: Encoder-Decoder examples after 10 epochs

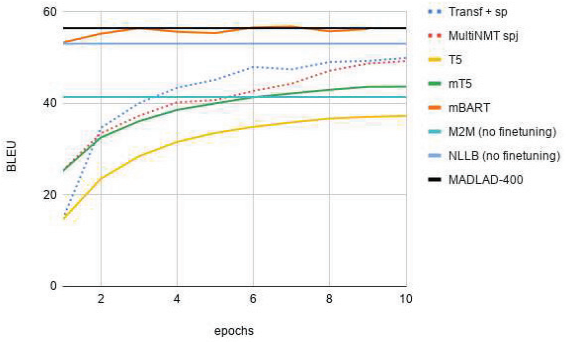

Figure 3: BLEU scores for the different pretrained encoder-decoder models over 10 epochs. The dotted line is the best scoring model as shown in Figure 1.

5.1 T5-variants (Text-to-Text Transfer Transformer)

T5 (Raffel et al., 2020) was one of the first large-scale models to systematically explore how transfer learning can be unified across many NLP tasks within a single encoder-decoder architecture.

Variant mT5 (multilingual T5) (Xue et al., 2021) extends the T5 framework to the multilingual setting. Like T5, it treats all tasks as text-to-text generation, but it does so using a model pretrained on many languages rather than mainly English.

5.2 mBART (Multilingual Bidirectional and Auto-Regressive Transformer)

mBART (Liu et al., 2020) extends BART to the multilingual setting and was one of the first models to show that multilingual denoising pretraining alone can substantially improve machine translation.

Fine-tuning it within Kaggle is possible, but our run was stopped after 9 epochs because it exceeded the max allowed execution duration.

5.3 M2M (Many-to-Many)

Unlike T5, mT5, and mBART, which are primarily pretrained using monolingual denoising objectives, M2M-100 (Fan et al., 2021) is trained directly on large-scale parallel translation data. This makes it fundamentally a multilingual translation model rather than a general-purpose text generation model adapted later to MT.

The M2M-100 models are too large to be fine-tuned comfortably in a standard Kaggle setup, so in this course we use the pretrained model directly for translation.

5.4 NLLB (No Language Left Behind)

The NLLB project (NLLB Team et al., 2022) extends massively multilingual machine translation to 200+ languages, with particular attention to low-resource and underrepresented languages. Its main goal is not only to scale translation to more languages, but also to improve quality for language communities that have historically received little support in MT.

Similar as for M2M, we do not fine-tune the model in Kaggle, but we can use the pretrained model directly for translation.

5.5 MADLAD-400 (Multilingual And Document-Level Large Audited Dataset)

MADLAD-400 (Kudugunta et al., 2023) represents a recent development in pretrained multilingual models that combines ideas from both encoder–decoder MT systems and text-to-text models such as T5.

Like for T5 and mT5, all tasks, including translation, are expressed as transformations from input text to output text. However, unlike earlier text-to-text models, it is trained explicitly for **massively multilingual translation**, covering hundreds of languages.

Similar as for M2M and NLLB, we do not fine-tune the model in Kaggle, but we can use the pretrained model directly for translation.

5.6 Results

As we can see in Figure 3, the T5 model scores a little better than the small transformer variant trained from scratch. The mT5 model scores a lot better, but still not as good as the large transformer model.

For mBART, we do not see a lot of improvement in fine-tuning, but nevertheless it is the best scoring model overall.

While we could not fine-tune M2M, NLLB nor MADLAD-400, due to hardware limitations in Kaggle, we show their performance in Figure 3 as straight lines, without fine-tuning. This shows that, even while M2M scores a BLEU score of 41, this is largely outperformed by NLLB, with a score of 53, which comes very near to the fine-tuned mBART score, so we can hypothesize that fine-tuning NLLB may beat mBART. MADLAD-400 scores 56.37 from the start, without fine-tuning, which is better than any of the fine-tuned systems.

Nevertheless, it should be noted that for these pretrained models data leakage between our development set and the data used for pretraining is rather likely, as the Tatoeba dataset is freely available so there is no reason why these systems would not have been trained on these data.

6 Prompt engineering with decoder-only models

We also introduce the students to prompt engineering with decoder-only models, and provide Kaggle sessions for working with GPT-2 (Radford et al., 2019), LLaMA (Touvron et al., 2023) and Mistral (Jiang et al., 2023), and we distinguish between different types of prompts, as shown in Table 8, such as, (1) zero-shot baseline, where we begin with a minimal instruction prompt; (2) format prompting – we make the structure of the task more explicit; (3) few-shot prompting – we provide one or more examples in the prompt; and (4) chat-style prompting. We also explore the effect of sampling and beam size.

Due to the limitations of Kaggle, we cannot download real size state-of-the-art models into the Kaggle environment. Therefore, for LLaMA, we also provide a Kaggle session with an API to the larger model running on a Hugging Face endpoint.

We also provide a Kaggle session which is using endpoints, such as Blablador (Strube, 2024) and Groq (Groq Inc., 2024), two services which offer access and compute for many different models.

Output of these decoder-only models shows that it is rather difficult and ad-hoc to make them output only the target sentence and nothing else, as they tend to overgenerate text. As such, overgeneration would largely affect the BLEU scores (not included here), but for the larger models, the output often contains nearly correct translations.

7 Conclusions

We have presented a number of scripts and notebooks that allow showing the impact of specific factors on NMT quality. A limitation of what we presented is that we have only trained / fine-tuned systems on a small dataset for a single language, but with these notebooks, it should be relatively easy for students in machine translation to train their own systems and test the effects of certain hyper-parameters and system architectures on the translation quality of their preferred language pair and domain.

While in this paper we have only used BLEU as an evaluation metric, calculated with SacreBleu, students have been taught how to use other metrics, be it through the use of MATEO (Vanroy et al., 2023) or through the use of scripts that perform evaluations with BERTScore (Zhang et al., 2020), BLEURT (Sellam et al., 2020) or COMET (Rei et al., 2020).

All the notebooks presented here should run in Kaggle, within the hard limitations set by the free version of Kaggle on GPU usage, processing time, disk space and RAM, these factors play. Nevertheless, if other GPU environments are available, training may go faster, fine-tuning of M2M and NLLB may become possible and working with larger open-weight decoder-only models may become feasible. Moreover, for the decoder models we have provided sessions that use free (for researchers) inference end-points for recent very large language models.

By sharing the scripts and making the notebooks public we hope to provide other teachers in MT to non-technical people with the right tooling to make students grasp the effect of the selection of training data.

It must be said that initially this course is very far outside the comfort zone of the students, of which only a minority has followed the first term course *Introduction to Machine Learning with Python*, but gradually most students get into the topic, as a lot of time is spent on making things

Prompt style	Example prompt
Zero-shot	Translate the following text from {source_lang} to {target_lang}. Only output the translation.\n\n{text}
Formatted	English: {text}\nDutch:
Few-shot	English: I eat apples.\nDutch: Ik eet appels.\n\nEnglish: She is at home.\nDutch: Zij is thuis.\n\nEnglish: {text}\nDutch:
Chat	You are a careful machine translation system. Translate faithfully and output only the translation. Translate this sentence from {source_lang} to {target_lang}. Only output the translation.\n\n{text}

Table 8: Example prompt styles and prompts

work on their own dataset. This involves quite a lot of individual help from the teacher, so the approach is probably only feasible for small groups. Nevertheless, during the course, students have some success experiences with the mechanics of MT training, and gradually they also see the MT quality improve. A downside is that training these models takes quite some time, which requires careful planning of when to explain MT theory versus when to do the hands-on work. All in all, most students that keep up the lessons till the end get a good grip on MT experimental methodology, the importance of data and finetuning and the general mechanisms of current-day language modeling.

Future work involves integrating the scripts as much as possible so it becomes even easier for students to play around with the hyper-parameters. We also want to investigate whether and how it would be possible to run these notebooks at the high performance computing facilities of the university instead of depending on commercial parties who may change running conditions as they please.

References

Cho, Kyunghyun, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics.

Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(1):4839–4886.

Groq Inc. 2024. Groq API documentation. `https://console.groq.com/`. Accessed: 2026-04-03.

Hackenbuchner, Janiça and Ralph Krüger. 2023. DataLitMT – Teaching data literacy in the context of machine translation literacy. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 285–293, Tampere. European Association for Machine Translation.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv*.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In Bansal, Mohit and Heng Ji, editors, *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver. Association for Computational Linguistics.

Kudo, Taku and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 66–71, Brussels. Association for Computational Linguistics.

Kudugunta, Sneha, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook. Curran Associates Inc.

Liu, Yinhan, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. In Johnson, Mark, Brian Roark, and Ani Nenkova, editors, *Transactions of the Association for Computational Linguistics, Volume 8*, pages 726–742, Cambridge. MIT Press.

Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In Isabelle, Pierre, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia. Association for Computational Linguistics.

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, Brussels. Association for Computational Linguistics.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI Blog.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1):5485–5551.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Strube, Alexandre. 2024. Helmholtz Blablador and the LLM models' ecosystem. EuroSciPy 2024, Szczecin.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS'14: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, pages 3104–3112, Cambridge, MA. Curran Associates, Inc.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul. European Language Resources Association (ELRA).

Tiedemann, Jörg. 2020. The Tatoeba translation challenge – Realistic data sets for low resource and multilingual MT. In Barrault, Loïc, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vanroy, Bram, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine Translation Evaluation Online. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 389–393, Tampere. European Association for Machine Translation.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR 2020)*.

Teaching Linguistic Prompt Control for LLM-Based Translation: A Classroom Approach to Developing Critical and Responsible AI Literacy

Katrin Menzel
Leipzig University / Institut für Angewandte Linguistik und Translatologie
katrin.menzel@uni-leipzig.de

Abstract

As Large Language Models (LLMs) are increasingly integrated into professional translation, there is a growing need for pedagogical frameworks that move beyond simple trial-and-error prompting across any available tools used for translation tasks. This paper presents a framework developed in a university MA seminar on translation to foster responsible AI literacy as well as linguistically grounded prompt control and structured prompting. The approach also emphasises that AI is understood as a tool for suggesting possible translation solutions, while human translators remain the final decision makers. Integrating translation-oriented text analysis and corpus-informed feature extraction from parallel and comparable data, the approach teaches students to develop register-specific and model-interpretable instructions when translating specialised texts with the assistance of generative AI tools. During a one-semester course, students familiarised with prompting strategies and different tools, including commercial, freely available models, GDPR-compliant institutional infrastructure and local models. These steps and the comparative evaluation of these tools allowed the students to identify optimal configurations that conform best to professional quality standards for specialised texts, and they ultimately led to highly improved translation output compared to unsteered model results.

1 Introduction

1.1 Overview of the growing role of LLMs in translation

As Large Language Models (LLMs) become increasingly common tools, professional translators and translation students are using them more frequently to support the translation process, e.g., by generating draft translations and revising target texts. The growing presence of these tools has important pedagogical implications. While LLMs can produce fluent translations, their output is shaped by statistical tendencies inherited from their training datasets. The output can vary across runs and lead to apparently random or inconsistent results for the same input. Moreover, particularly in unsteered baseline translations produced without specific instructions, LLM-generated translations may exhibit certain unwanted characteristics that set them apart from both traditional machine translation and human translation, including, for instance, differences in stylistic variation, lexical choice and degrees of explicitness (cf. Sizov et al., 2024).

1.2 Related research on AI-supported translation and pedagogical implications

Research on AI-supported translation tasks and translator education emphasises the importance of maintaining human-centered decision making and the development of the capacity to analyse and manage model behaviour rather than treating

Proceedings of the 1st International Workshop on Teaching AI-Based Translation and Technologies (TAITT 2026), pages 28–35
June 15, 2026 ©2026 Association for Computational Linguistics

LLMs as opaque drafting tools or black boxes. Recent publications in the field include work by Angelone (2026), Ayvazyan and colleagues (2024, 2025), Cui et al. (2025), Gao et al. (2023), He (2024), Ren and Wang (2025), Siu (2024), Yamada (2023, 2026), Yang et al. (2024) and Zhang et al. (2023). Existing studies highlight the need for structured training that equips translation students and professional translators with the skills to engage critically with LLM outputs and to control and interpret model behaviour in translation processes. Such training may draw on research on efficient prompt engineering and general prompting frameworks that provide systematic approaches for designing prompts. For instance, the CO-STAR framework structures prompts around the dimensions of context, objective, style, tone, audience and response (cf. Teo, 2023). The CLEAR framework emphasises prompts that are concise, logical, explicit, adaptive and reflective (Lo, 2023). Such frameworks offer general principles applicable to a wide range of LLM tasks. In addition to such general frameworks, various model- and task-specific prompt templates have been suggested. For example, TranslateGemma, an open variant of the Gemma 3 model specifically enhanced for translation has been reported to work well with a prompt that instructs the model to produce a faithful and culturally sensitive translation as follows: *"You are a professional {source_lang} ({src_lang_code}) to {target_lang} ({tgt_lang_code}) translator. Your goal is to accurately convey the meaning and nuances of the original {source_lang} text while adhering to {target_lang} grammar, vocabulary, and cultural sensitivities. Produce only the {target_lang} translation, without any additional explanations or commentary. Please translate the following {source_lang} text into {target_lang}:\n\n\n{text}"* (Finkelstein et al., 2026).

While the above-mentioned frameworks and such systematic translation-related prompts demonstrate that structured prompting can reduce unwanted variability in LLM outputs, they serve only as a general starting point for professional translation tasks due to their broad applicability. Translators and educators can further refine them to address the specific requirements of different registers and specialised domains when working with selected tools and language pairs.

1.3 Characteristics of LLM-generated translations

In specialised translation contexts, attentive readers and language experts, including professional translators, can identify recurring AI fingerprints in LLM translation outputs that deviate from established professional standards for the target texts. These tendencies may include repeated and unwanted syntactic structures – such as a disproportionate overuse of non-finite clauses in English texts – and a general inclination toward stylistically flattened, register-neutral formulations. Such patterns may be unproblematic in some contexts, but they can become limitations in domains where specific register features and specific lexico-grammatical and stylistic choices are essential for the text's communicative function. For instance, in an informational specialised text, an LLM-generated translation may fail to maintain the high density of nominal structures and impersonal expressions that would be typically desired by the producer of the target language text to convey institutional neutrality and compact informational content. It may tend to default to more general language or conversational patterns.

1.4 Considerations in model selection

The challenge for translators in navigating AI tools is exacerbated by the vast and heterogeneous landscape of available technologies and the rapid pace of ongoing developments in the field. AI systems differ in how well they handle multilingual tasks including translation. This performance gap is particularly evident when working with different language pairs. While models typically achieve higher accuracy for high-resource European languages that are typologically close, their effectiveness declines for low-resource languages and typologically distant combinations. To navigate this complexity, practitioners can rely on their own linguistic expertise to assess translation quality. They may also consult external evaluation frameworks from academic and industry contexts. Such frameworks provide relatively standardised comparisons of model performance and practice-oriented rankings of system capabilities, some focusing specifically on translation and others addressing translation in addition to other multilingual and creative language tasks (cf., for instance, Burkert, 2025 and Spiridonov, 2026). However, such benchmarks capture only limited aspects of translation quality, and it might be difficult for translators to interpret the scores or understand

how they were computed. Moreover, such rankings may not generalise to specialised translation contexts.

Furthermore, the use of AI in translation tasks and translator training is complicated by GDPR-related concerns, alongside practical limitations such as the requirement for specialised infrastructure and access to cost-effective AI services. The use of commercial closed models, i.e. proprietary systems that cannot be inspected or modified by users and are typically accessed via an API, has become increasingly common in translation workflows. In such setups, user input is processed on external servers, which raises concerns regarding data protection when handling confidential material. This contrasts with locally deployed models, which can be run entirely on the user's own hardware without transmitting data externally and may in some cases be adapted or fine-tuned by users.

These considerations have stimulated growing interest among translators (who are typically not trained in computer science) in local deployment solutions that do not require substantial hardware investment or advanced technical expertise. There is also increasing interest in secure, non-commercial, institutionally managed cloud infrastructures. However, such solutions are still under active development. Many tools remain experimental and sometimes provide less user-friendly interfaces and fewer features than the leading commercial systems, for example, limiting the amount of text that can be uploaded at one time. For enterprise users, individual cloud-based translation solutions are available, such as Language Weaver,[1] which is described by its developers as a secure, scalable AI translation platform. It combines neural machine translation and generative AI and was developed with confidentiality in mind.

1.5 AI-supported translator training in practice

Building on the observations outlined above and the prior research discussed, this paper proposes a pedagogical framework that integrates linguistic analysis, structured prompt design and reflective evaluation into a unified teaching sequence. It summarises tasks from a one-semester seminar at Leipzig University intended to guide MA translatology students toward becoming informed designers of AI-supported translation workflows. Integrating translation-oriented text analysis based on Systemic Functional Linguistics (SFL) and corpus-informed feature extraction from parallel and comparable data, the approach helped students to develop register-specific and model-interpretable instructions when translating specialised texts with the assistance of generative AI tools. The students utilised a diverse ecosystem over the course, including commercial and freely available models, GDPR-compliant institutional infrastructure and local models. The seminar also benefited from input by Martin Czygan from the Leipzig University Library, who shared his expertise on AI tools and workflows in a workshop held within the translation seminar.

One of the key outcomes of the seminar was a reduction in inefficient trial-and-error prompting and unnecessary iterations, as well as the mitigation of unacceptable translation outputs suggested by individual models. This was achieved by guiding students to employ a structured prompt architecture comprising a configuration prompt, a task prompt and a revision step and by supporting informed model selection based on quality considerations.

2 Pedagogical framework

2.1 Baseline analysis of LLM output

The pedagogical framework was implemented within an MA-level, cross-linguistic methodological seminar, open to students of translation and interpreting studies who work with a range of different language pairs in their respective specialisations. While most students in the seminar had substantial experience with English-German translation, some also had considerable experience with other language combinations involving, for instance, French, Spanish or Arabic as one of their main working languages.

The process began with students producing an unsteered baseline translation of an informational English text from an EU website on a topic of current relevance, such as antimicrobial resistance, using a minimal prompt (e.g., *"Translate this text into German"*) with a model of their choice. The text selected for this task[2] was approximately 1,200 words in length, with students initially working on shorter extracts before progressing to larger sections and, ultimately, the full text. It can be classified as a specialised text targeting a non-specialist but informed readership, including

[1] https://www.rws.com/language-weaver/

[2] https://health.ec.europa.eu/antimicrobial-resistance/eu-action-antimicrobial-resistance_en

stakeholders and interested members of the public. In this initial phase, students primarily relied on widely available commercial platforms that provide access to LLMs, such as those underlying OpenAI's ChatGPT, Microsoft Copilot (institutionally licensed) or Google Gemini. These tools produced translations that the students used to discuss default AI translation patterns and recurring linguistic features, based on their prior linguistic knowledge.

For the majority of the cohort, the English-German language pair served as the primary framework, with English-French provided as an option for those who were not native speakers of German. However, since several students primarily work with other language combinations in their daily practice and academic curriculum, e.g., involving Arabic, a specific challenge emerged in identifying institutional text resources with comparable stylistic consistency and parallel availability. While informational EU texts, which are available in various languages, served as suitable benchmarks for observing typical register-specific features and for comparing officially published versions with LLM-generated outputs, United Nations texts were proposed as alternative resources for students working with Arabic, given their availability as comparable institutional web texts on specialised topics in a public-facing communicative context.

2.2 Translation-oriented text analysis, corpus-based exploration and terminology work for prompt preparation

The next step involved deepening the students' understanding of the same informational texts they had used in the initial task by conducting a register analysis grounded in SFL (Halliday & Matthiessen 2013). The students examined how ideational, interpersonal, and textual meanings are realised, and they identified linguistic features in the texts that contribute to the thematic domain (Field), the construction of social roles and relations of expertise (Tenor) and the mode of communication in terms of medium, information structure and cohesive organisation (Mode).

The students then consulted large-scale corpus data in order to validate their observations derived from individual texts and to establish empirical frequencies of register-specific linguistic features. The corpora were selected for their register proximity to the analysed text type; however, no exact matching corpus of informational institutional website texts was available, so functionally and thematically comparable corpora were used instead. Europarl corpora, accessible for instance via Sketch Engine[3] and via OPUS,[4] were used to verify whether patterns identified in individual institutional web texts (such as high nominal density, frequent passive constructions and specific hedging strategies) constitute systematic features of the broader institutional register. The United Nations Parallel Corpus available via Sketch Engine was additionally used to ensure access to comparable institutional large datasets for the less commonly represented language combinations in the seminar.

The students observed that the analysed EU informational web texts and EU parliamentary debates in Europarl share core features of institutional discourse. These include a formal tone, the use of domain-specific and relatively standardised terminology, frequent quantification through measurable indicators and statistical data, explicit references to data sources, the use of comparative markers and references to procedural and institutional authority and to collective actors. Such characteristics provide important contextual information that can be integrated into translation prompts to guide AI tools toward more contextually appropriate and institutionally authentic outputs when processing similar texts.

The students further noted that the analysed EU web texts were typically more information-oriented and characterised by higher nominal density and a more impersonal style. In contrast, the Europarl data, which consist of edited and published plenary debates, naturally display more features of spoken political discourse, including increased personal reference and evaluative language. These characteristics are specific to debate-based data and should not be transferred to specialised informational web texts. Instead, translation prompts in this case should prioritise features that occur across different forms of EU discourse, together with the structural and stylistic conventions that are typical of institutional informational writing which is addressed directly to civic audiences.

Furthermore, the students utilised Sketch Engine's OneClick Terms tool[5] to identify terminology and multi-word expressions from existing comparable

[3] https://www.sketchengine.eu/
[4] https://opus.nlpl.eu/
[5] https://terms.sketchengine.eu/

institutional texts and parallel text pairs. In addition to EU and UN materials, they also worked with texts from other specialised domains, particularly in scientific and legal domains, to apply the tasks across different types of specialised discourse. This workflow enabled the students to perform structural alignment of existing reference texts and to create bilingual glossaries that could also serve as input for prompts. The students also learned to preprocess texts to improve structural alignment and to evaluate statistically derived term candidates before integrating them into their translation workflows.

2.3 Model selection and exploration of GDPR-compliant options

In early phases, the students experimented with commercial models, as outlined above, and compared their outputs in class discussions. After some practical experience in evaluating models from a qualitative linguistic standpoint, benchmarking approaches were additionally discussed in a dedicated workshop with Martin Czygan, a software engineer working at the Leipzig University Library (Digital Services). The workshop provided the students with further insights into evaluation criteria and comparative assessment strategies for model outputs in translation tasks. In this workshop, the seminar also expanded the previously used ecosystem of tools and platforms to include demonstrations of local models such as TranslateGemma, run autonomously using the Jan.AI tool,[6] as well as GDPR-compliant institutional infrastructure, in particular the Academic Cloud.[7]

The Academic Cloud was developed and coordinated within German higher education networks with significant involvement of the Göttingen State and University Library and is technically operated by the Gesellschaft für wissenschaftliche Datenverarbeitung mbH (GWDG). It was tried out by the students in the seminar as it provides a GDPR-compliant environment for research and teaching, including AI-based applications. Despite not being intended for professional translators as a target group, the Academic Cloud ChatAI service[8] has proven to be a valuable experimental tool in our academic translator-training context. It provides access to a curated selection of LLMs, including institutionally hosted open-source models such as Intel Neural Chat 7B, Mixtral 8x7B Instruct, Qwen1.5 72B Chat and Meta LLaMA 3 70B Instruct as well as externally hosted models such as OpenAI GPT-4. Internally hosted models are operated on infrastructure managed within the Academic Cloud/GWDG environment, whereas externally hosted models are accessed through integrated interfaces to third-party providers under GDPR-compliant data-processing arrangements. Such infrastructures are particularly useful in educational contexts, as they enable students and instructors to experiment with different LLMs through a unified academic platform.

2.4 Prompt architecture: configuration, task and revision prompts

Generally, after selecting an LLM interface and, where applicable, a specific model, the students structured their prompting into three components: a general configuration prompt, a more specific task prompt and a revision prompt.

Configuration prompt: This prompt involves a set of role-based instructions and general translational principles formulated in operational terms that was developed collaboratively with the students. This provides stable, high-level behavioural guidelines for the model and reduces the need to address individual translation issues in this step. In the present seminar context, this involved assigning the model a translator role for specialised texts, e.g., EU health-related informational websites translated from English into German or vice versa, and specifying consistent translational principles. These include fidelity to source meaning without unwarranted additions or omissions, register consistency in maintaining a formal institutional, information-oriented style for a non-expert but informed readership, and terminological precision aligned with EU domain-specific usage. In addition, this prompt may instruct the model to produce only the translation itself, as in the preferred TranslateGemma prompt discussed above, to ensure that no unintended content is added. Where supported by the chosen tool, the configuration prompt could be implemented as an assistant- or agent-style prompt.

Task prompt: The task prompt incorporates language-pair-specific considerations regarding systematic differences (e.g., between English and German) in the given register. As the students have identified a wide range of features in previous steps, their individual prompts will ultimately prioritise those aspects most relevant to them

[6] https://www.jan.ai/

[7] https://academiccloud.de/de/

[8] https://academiccloud.de/services/chatai/

based on their text analysis and corpus-driven findings. These considerations may include differences in word-formation preferences in institutional discourse, e.g., the greater tendency towards closed nominal compounds in German. Another example that some students may consider worth including, in light of insights from a specific text and corpus analysis, is the avoidance of premodifying hyphenated structures in German as translation equivalents for comparable constructions in English. For instance, where English may use forms such as "*bacterium-antibiotic resistance combinations"*, German institutional texts might prefer to encode relations more explicitly and in a different order through postmodifying prepositional phrases rather than premodifying attributive compound structures, e.g., *"Resistenzkombinationen zwischen Bakterien und Antibiotika"*, but LLM baseline outputs sometimes translate such structures too literally. The prompt can also address issues of grammatical gender representation, that is, the need to make gender inclusivity explicit in German where English uses gender-neutral expressions (e.g., *government experts* vs. *Regierungsexpertinnen und -experten*). Furthermore, the task prompt should address the handling of institutional terminology, including names, policy labels and abbreviations, e.g., in cases where full and short forms do not correspond directly across languages (e.g., *European Centre for Disease Prevention and Control (ECDC) / Europäisches Zentrum für die Prävention und die Kontrolle von Krankheiten (ECDC)*), since German institutional discourse may avoid repeated use of initialisms derived from foreign-language full forms.

A key pedagogical insight emerged during this step from the comparison between human-oriented and LLM-oriented instructions. When the students attempted to brief an LLM as they would a human translator, they observed that the model sometimes followed individual instructions inconsistently, even when these appeared straightforward and unambiguous. In such cases, the students attempted to include more examples and contextual information in their prompts in order to translate linguistic constraints into more explicit, model-friendly instructions. More explicit commands with examples usually worked better, especially when a revision step followed the task prompt.

However, rather than attempting to account for every possible linguistic construction, the task prompts generally combined rule-based guidance with some illustrative examples and provided guidance without aiming to be exhaustive. In this step, the students also need to consider model-specific capacities and limitations, including differences in context-window size, maximum prompt length and the number and types of possible supplementary files (e.g., glossaries, reference translations or other background documents) that can be integrated into the prompting workflow.

Revision prompt: This prompt is applied after the initial translation has been generated during the preceding steps. It instructs the model to assess its output relative to the specified constraints and produce a revised version accordingly. In practice, the model does not perform a human-like review; rather, it enters a second, constraint-aware generation phase in which it negotiates potentially competing requirements to improve alignment with the defined translation objectives. When multiple constraints conflict, the model may prioritise higher-level constraints, and it can be instructed in this step to flag any constraints that could not be fully satisfied and to provide a brief justification.

This refinement step enables iterative self-correction and reduces the need for repeated user-driven prompt modifications. To a certain extent, it enhances transparency in the model's decision-making process. Preliminary observations indicate that this self-evaluation mechanism improves adherence to register-specific constraints and leads to a higher quality of the translation output as well as improved compliance with the given instructions. It also helps the students to identify areas where LLMs systematically struggle to comply with or follow specific instructions.

Overall, this architecture provides the students with a reusable framework for controlling LLM output while keeping the prompting process structured and manageable. Over the course, the students worked with texts representing different specialised registers, primarily within the English-German language pair (in both directions). They were then free to choose their own case study and apply the methods and knowledge they had acquired during the semester. The students selected specialised texts for translation and carried out individual translation projects which concluded with a final presentation of their work and findings.

3 Discussion

In conclusion, the paper illustrated how a structured pedagogical framework combining linguistic analysis, corpus-informed insights and staged prompting can help translation students become

informed designers of AI-assisted translation workflows. The teaching framework that was presented in this paper also addresses a central challenge of LLM-assisted translation: the tension between linguistic specificity and prompt complexity. The approach encourages critical reflection on the limitations of different model types and their appropriate applications.

One limitation remained evident even when detailed and example-rich task prompts were provided: LLMs did not always consistently monitor or apply stylistic and linguistic constraints in the way human translators typically do. Human translators naturally monitor a wide range of features and vary their choices depending on the register; they rarely need to be told, for example, to avoid producing the same syntactic structure repeatedly. LLMs, however, sometimes exhibit recurrent patterns that would be unusual in human translation. Users may find themselves giving instructions that would be less necessary for human translators, such as *"Avoid using participle clauses repeatedly across consecutive sentences or stacked in a single sentence in this English translation"*, because such patterns can emerge as unwanted regularities in a model's output. Instructions that may seem straightforward for humans may not always produce predictable effects on model behaviour. In this sense, LLMs need to be guided to attend to aspects of the text differently than human translators and will often require additional concrete examples rather than relying solely on abstract linguistic instructions.

More explicit commands as the following worked much better, but were still sometimes followed inconsistently: *"Prefer full clauses with clear subjects and finite verbs. Participle constructions can be used where appropriate, but sparingly. Avoid using participle constructions such as 'emerging as…' or 'driven by…' repeatedly across consecutive sentences or stacking them in a single sentence as in the following non-preferred example: 'Antimicrobial resistance, emerging as a major public health concern and driven by the overuse of antibiotics, occurs when microorganisms, exposed to repeated treatments and adapting to survive hostile environments, evolve to resist the effects of medicines designed to kill them.'"*

Inconsistencies in following such prompt instructions with linguistic specifications do not necessarily arise from ambiguous instructions; rather, they reflect the fact that LLMs cannot reason about or systematically apply human-defined linguistic rules for translation purposes. Translation pedagogy can leverage this challenge by helping students learn to reframe linguistic constraints as prompts that LLMs are most likely to interpret reliably. This effect appears to be especially pronounced when the model is given a dedicated revision step after the initial task prompt. At the same time, instances where LLMs fail to follow instructions may serve as opportunities for cultivating critical AI literacy.

Overall, the proposed framework led to improved translation outputs, reduced reliance on trial-and-error prompting and heightened student awareness of model-specific strengths and limitations. It fosters more responsible and effective use of AI tools in translation tasks and can be adapted to similar translation contexts or other domains that require context-sensitive language production. Far from diminishing the role of linguistic expertise, the use of LLMs highlights the necessity of human knowledge and the capacity to craft precise, contextually informed instructions. Students learn to engage with LLMs as systems that require active, linguistically grounded guidance. More broadly, this approach provides a sustainable and pedagogically meaningful framework for integrating LLMs into translation training while reinforcing core disciplinary competencies and human decision making as central to the process.

4 Carbon impact statement

This study involved classroom-based translation tasks in which students used existing AI-powered tools to translate texts of different lengths to refine their prompting strategies. The computational resources required for the study were not directly measurable. The research also aimed to explore efficient uses of AI in translation workflows and training, for example by reducing unnecessary iterations and limiting the number of prompts needed to achieve satisfactory results, thereby contributing to more resource-efficient practices in the future.

Acknowledgements

The author would like to thank Martin Czygan from the Leipzig University Library for sharing his expertise in a workshop held within the discussed translation seminar. During the workshop, he introduced local LLMs, demonstrated the functions of, for instance, the Academic Cloud, Jan.AI and TranslateGemma, provided an overview of benchmarks and the

possibilities of different LLMs and supplied a repository of documentation.

References

Angelone, Erik. 2026. Generative AI as a facilitator of deliberate practice in translator training. In Penet, J. C., Joss Moorkens, and Masaru Yamada, editors, *Teaching Translation in the Age of Generative AI: New Paradigm, New Learning?*, pages 27–47. Language Science Press, Berlin.

Ayvazyan, Nune, Yu Hao, and Anthony Pym. 2024. Things to do in the translation class when technologies change: The case of generative AI. In Peng, Yuhong, Huihui Huang, and Defeng Li, editors, *New Advances in Translation Technology: Applications and Pedagogy*, pages 219–238. Springer, Singapore.

Ayvazyan, Nune. 2025. Human-centered pedagogies in the age of generative AI: Examining student perceptions of augmentation, progress and agency. *InContext: Studies in Translation and Interculturalism*, 5(1):167–193.

Burkert, Tomáš. 2025. LLM benchmarking leaderboard: Languages and creativity tasks. *RWS Blog.*

Cui, Feng, Defeng Li, and Chiyuan Zhuang. 2025. Introduction: Transforming translation education through Artificial Intelligence. *The Interpreter and Translator Trainer*, 19(3-4):227–233.

Finkelstein, Mara, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. 2026. *TranslateGemma technical report*.

Gao, Yuan, Ruili Wang, and Feng Hou. 2023. How to design translation prompts for ChatGPT: An empirical study. *arXiv*.

Halliday, Michael A. K. and Christian M. I. M. Matthiessen. 2013. *Halliday's Introduction to Functional Grammar* (4th ed.). Routledge, London/New York.

He, Sui. 2024. Prompting ChatGPT for translation: a comparative analysis of translation brief and persona prompts. *arXiv*.

Leo S. Lo. 2023. The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship,* 49(4):1–3.

Pym, Anthony and Yu Hao. 2024. *How to Augment Language Skills: Generative AI and Machine Translation in Language Learning and Translator Training*. Routledge, London/New York.

Ren Xiaobin and Wang Ruoxuan. 2025. Integrating human-AI collaboration into translation education: A comprehensive protocol for assessment, diagnosis and strategy development. *PLoS One*, 20(12):1–8.

Siu, Sai Cheong. 2024. Revolutionising translation with AI: Unravelling Neural Machine Translation and generative pre-trained Large Language Models. In Peng, Yuhong, Huihui Huang, and Defeng Li, editors, *New Advances in Translation Technology. Applications and Pedagogy*, pages 29–54. Springer, Singapore.

Sizov, Fedor, Cristina España-Bonet, Josef Van Genabith, Roy Xie, and Koel Dutta Chowdhury. 2024. Analysing translation artifacts: A comparative study of LLMs, NMTs, and human translations. In *Proceedings of the 9th Conference on Machine Translation*, pages 1183–1199, Miami. Association for Computational Linguistics.

Spiridonov, Ilya. 2026. Best LLM for translation in 2026: A data-driven engine scoreboard. *Alconost Blog.*

Teo, Sheila. 2023. How I won Singapore's GPT-4 prompt engineering competition. *Towards Data Science.*

Yamada, Masaru. 2023. Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 195–204, Macau. Asia-Pacific Association for Machine Translation.

Yamada, Masaru. 2026. Teaching translation with AI: Bridging theory and practice through prompt engineering. In Penet, J. C., Joss Moorkens, and Masaru Yamada, editors, *Teaching Translation in the Age of Generative AI: New Paradigm, New Learning?*, pages 87–104. Language Science Press, Berlin.

Yang, Zhishen, Tosho Hirasawa, Edison Marrese-Taylor, and Naoaki Okazaki. 2024. Large language models as manga translators: A case study. In *Proceedings of the 30th Annual Meeting of the Association for Natural Language Processing*, pages 2012–2017, Tokyo, Association for Natural Language Processing.

Zhang, Biao, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning*, pages 41092–41110, Honolulu. Journal of Machine Learning Research.

Evaluative Judgement in Teaching AI-based Translation: A Classroom Case Study of AI-Mediated Translation and Post-Editing

Gokhan Doğru
Universitat Pompeu Fabra
Barcelona, Spain
gokhan.dogru@upf.edu

Abstract

Drawing on 23 anonymized student projects from a fourth-year Machine Translation and Post-editing course in a BA-level translation programme, this paper examines how structured comparison of general-purpose LLMs and online MT systems can elicit evaluative judgement in AI-mediated translation. Students translated short specialised English Wikipedia texts into Catalan or Spanish, generated four system outputs, evaluated them using automatic metrics and human adequacy/fluency assessment, selected one output for post-editing, and justified their decision in written reports. Descriptive counts are reported for all 23 projects, while qualitative interpretation is based on the 22 cases accompanied by written reports. Results show that students did not treat automatic metrics as final authority: final post-editing selections often diverged from metric rankings and were justified through adequacy, fluency, terminology, naturalness, and expected post-editing effort. The study therefore does not benchmark systems under controlled conditions; it analyses how students justified system choice within an authentic classroom assignment.

1 Introduction

Translation technology pedagogy has long argued for evaluation, comparison, and post-editing as central elements of translator training, alongside tool operation. Earlier work in this field had already argued that students should learn not only how to operate translation tools and machine translation (MT) systems, but also how to evaluate them critically, compare technologies, and understand post-editing as a professional and judgement-based activity rather than a merely mechanical one (Ayvazyan et al., 2024). More recent studies build on this perspective by emphasizing that translator education must now engage explicitly with AI literacy, including issues such as human agency, prompting practices, data awareness, trust, and ethical decision-making (Freeth and Toto, 2026; Zhang and Doherty, 2025; European Commission, 2022; Massey and Ehrensberger-Dow, 2026).

The case study reported here comes from a fourth-year university course on Machine Translation and Post-editing. The aim is not to determine which system performs best in general terms, but to examine how a structured assignment can support students in developing "evaluative judgement" (Tai et al., 2018) in AI-mediated translation. The assignment had five stages. Students first produced a human translation, then generated four system outputs, evaluated them with automatic metrics and human adequacy/fluency criteria, selected one version for post-editing, and justified that decision in a report. This design follows recent calls for classroom experimentation in response to technological change, particularly tasks in which students compare human, machine-generated, and post-edited versions rather than treating AI/MT output as a finished product (Ayvazyan et al., 2024; Kenny, 2020; Guerberof Arenas and Moorkens, 2019).

The problem is especially relevant in a context where AI-generated translations may appear fluent, coherent, and convincing while still containing terminological, adequacy-related, or stylistic problems that demand careful human scrutiny (Ayvazyan, 2025). In line with recent higher education research, this study treats evaluative judgement as a core graduate

Proceedings of the 1st International Workshop on Teaching AI-Based Translation and Technologies (TAITT 2026), pages 36–48
June 15, 2026 ©2026 Association for Computational Linguistics

capability: the ability to make informed decisions about the quality of work and to justify those decisions using explicit criteria and evidence. In the context of GenAI, this becomes even more important, since students must remain the "arbiters of quality" (Bearman et al., 2024, 894) rather than uncritical accepters of machine output. The assignment is valuable because students had to leave a trace of their reasoning: metric interpretation, error comments, system selection, and post-editing choices.

The study therefore contributes an integrated classroom design for teaching critical AI use in translator education. More specifically, it examines how students compare translation outputs, interpret metric scores in relation to human judgement, identify recurring error patterns, and justify why one version is more suitable than others for post-editing. In doing so, the paper argues that structured comparison and evidence-based selection can serve as an effective pedagogical means of fostering critical, reflective, and professionally relevant judgement in the age of GenAI.

This study is guided by three research questions:

RQ1. How did students justify their final selection of one LLM/MT output for post-editing?

RQ2. To what extent did students' final post-editing choices converge with or diverge from automatic metric rankings?

RQ3. What kinds of evaluative judgement do students articulate when comparing LLM and NMT outputs within a translation/post-editing assignment?

The unit of analysis is therefore not the translation system as such, but the student evaluative decision: how students interpreted different forms of evidence and justified the choice of one output as a post-editing candidate.

2 Related Work

2.1 From MT literacy to AI literacy in translator education

Translation technology pedagogy has increasingly moved from procedural tool instruction toward literacy-oriented models of translator education including MT literacy, AI literacy and data literacy (Krüger and Hackenbuchner, 2024; Kenny, 2020; Rico and González Pastor, 2022; Moorkens et al., 2024). In these models, competence includes understanding how systems work, when they are appropriate, and what consequences follow from using them in specific communicative and professional contexts. O'Brien and Ehrensberger-Dow (2020), for example, conceptualize MT literacy as a contextualized and cognitively grounded ability to understand what MT does, what it does not do, and how its use affects translation processes and decisions. Ehrensberger-Dow et al. (2023) extend this perspective by proposing the role of translators and trainers as "MT literacy consultants," arguing that education should mitigate the risks of uninformed MT use by cultivating informed judgement and guidance.

Recent scholarship on GenAI broadens this literacy agenda further. Penet et al. (2026) argue that GenAI is not merely another translation tool, but a development that requires translator education to re-engage questions of human agency, ethics, and sustainability. Within that volume, AI literacy is framed in explicitly translation-specific terms through the notions of suitability, data literacy, prompting, and ethical reasoning (Bowker, 2026; Inglada, 2026; Moorkens and Doğru, 2026; Yamada, 2026). This translation-oriented work aligns with broader AI literacy research, where AI literacy is defined as a set of competencies that enable people to critically evaluate AI systems and interact with them responsibly (Long and Magerko, 2020). It also resonates with recent empirical evidence showing that novice translation students already use AI tools early in their training, often before they have developed robust critical or ethical routines for assessing output quality (Zhang and Doherty, 2025). As a whole, these studies suggest that translator education should not simply teach students how to obtain AI-generated output, but how to judge when, how, and why such output should be used, compared, and revised in context (Ayvazyan et al., 2024).

2.2 Evaluation and post-editing as pedagogical practices

The pedagogical value of evaluation and post-editing has a longer history in translator education. Early EAMT work already treated both MT evaluation and post-editing as legitimate objects of instruction and assessment (Belam, 2002; O'Brien, 2002). Later studies further embedded MT and post-editing into translation curricula through explicit learning objectives, professional workflows, and classroom tasks designed to connect academic training to industry practice (Doherty and Kenny, 2014; Guerberof Arenas and

Moorkens, 2019). Seen in this wider context, evaluation and post-editing pedagogy are not merely ways of familiarising students with professional workflows; they also cultivate the added value of human translators as adaptive and digitally literate experts whose role increasingly lies in reading, prompting, evaluating, editing, and making informed decisions about how AI-supported output should be used (Massey et al., 2023; Massey and Ehrensberger-Dow, 2026).

Within this tradition, evaluation is especially important because it shifts students from passive tool users to active judges of quality. Moorkens (2018) proposes a practical in-class NMT evaluation exercise that combines adequacy, post-editing productivity, and a simple error taxonomy in order to demystify technological hype and foster critical judgement. Öner and Öner Bulut (2021) similarly report that a post-editing-oriented NMT quality-evaluation and error-annotation task can support translator training by prompting students to identify, categorize, and reflect on translation problems more systematically, rather than relying only on general impressions of whether a translation "sounds good." At the same time, the post-editing literature warns against assuming that such practices are self-explanatory to novices. Flanagan and Christensen (2014) show that trainees do not always interpret post-editing guidelines consistently, which means that judgement criteria and revision expectations must be explicitly scaffolded. Nitzke et al. (2019) add a further dimension by framing post-editing competence as risk management: competent translators must judge not only how to post-edit, but also whether MT output is suitable for a given task, text type, and quality requirement. This perspective is particularly relevant for pedagogical designs in which students must compare several outputs and justify the selection of one version for post-editing.

2.3 Automatic metrics, human judgement, and evaluative judgement

A second core strand of related work concerns how quality is judged and how students learn to interpret different forms of evidence. In higher education research, evaluative judgement is understood as the capacity to make decisions about the quality of one's own work and that of others (Tai et al., 2018). In the context of GenAI, this capacity becomes even more important because fluent and persuasive outputs can mask deeper adequacy, factual, or domain-specific problems (Bearman et al., 2024). This insight is highly transferable to translation pedagogy, where students must learn to distinguish surface fluency from fuller notions of translation-oriented quality.

Automatic metrics are relevant here not because they replace human judgement, but because they provide one source of evidence that students must interpret critically. Rei et al. (2020) introduce COMET as a learned evaluation framework designed to correlate more strongly with human judgements than older lexical-overlap metrics. However, meta-evaluation work from WMT23 shows that metric–human correlation remains sensitive to evaluation conditions, including the quality of the reference translation itself (Freitag et al., 2023). For classroom practice, the implication is concrete: students need to learn why a high COMET, BLEU, or ChrF score may still coexist with terminology, adequacy, or naturalness problems. In this respect, structured error annotation provides an important complement to both metrics and impressionistic human evaluation. MQM offers an established framework for making error categories explicit and discussable, even when simplified for educational purposes (Lommel et al., 2014). When used together, metrics, adequacy/fluency ratings, and error analysis give students several kinds of evidence to reconcile.

2.4 Ethics, trust, and agency

Ethics is also part of this pedagogical problem. Ramírez-Polo and Vargas-Sierra (2023) show that ethical competence is often under-explicit in translation technology syllabi, despite the growing relevance of issues such as confidentiality, bias, accountability, and responsible use. Tipton (2024) similarly argues for integrating ethics into classroom tasks and reflective practice rather than treating it as detached policy knowledge. In translator education specifically, Moorkens and Doğru (2026) frame AI ethics as a practical discipline of deciding when GenAI should and should not be used, linking ethics directly to classroom reasoning and decision-making. This position is consistent with broader educational guidance that advocates human-centred and risk-aware approaches to GenAI in learning environments (Miao and Holmes, 2023).

The ethics dimension also intersects with questions of trust and professional agency. Rivas Ginel and Moorkens (2025) show that translators' trust and distrust in GenAI are shaped not only by perceived quality, but also by accountability, risk, and appropriate use. For translator education, this

suggests that students should not merely be encouraged to use AI tools, but should be trained to calibrate reliance on them and to justify that reliance explicitly. Within this framework, the pedagogical task is not to identify a universally best system, but to help students make accountable decisions about quality, intervention, and responsibility while retaining a sense of human control and agency in their interaction with AI tools.

Existing work has established the pedagogical value of MT evaluation, post-editing, and AI literacy in translator education. However, there is still limited empirical evidence on classroom designs that combine LLM and NMT outputs, automatic metrics, human adequacy/fluency assessment, error annotation, and a consequential post-editing decision within a single assignment. The present study addresses that integrated pedagogical configuration. The importance of combining these components in a single assignment is that the final post-editing choice becomes consequential: students cannot discuss metrics, errors, system affordances, and ethical reliance on AI as separate issues, but must reconcile them in one accountable decision about what to revise and why.

3 Methodology

3.1 Research design

The study adopts an exploratory qualitative case-study design with descriptive quantitative support, because the aim is not to benchmark systems across a controlled shared test set, but to analyse the reasoning observable in student decisions within an assessed final assignment completed as part of the course. More specifically, the study investigates how students compared multiple system outputs, justified quality judgements, and selected one version for post-editing on the basis of explicit criteria rather than on the basis of automatic metric scores alone. In this respect, the analysis draws on two concepts: evaluative judgement and AI literacy. Evaluative judgement is treated here as students' developing capacity to discern translation quality, justify preferences, and estimate the extent and type of human intervention required. AI literacy is addressed through students' critical engagement with the affordances and limitations of GenAI and NMT systems within a realistic translation workflow.

3.2 Course context

The study was conducted in the 4th year undergraduate course Traducció automàtica i postedició at Pompeu Fabra University, within a translation-focused degree programme, during the 2025–2026 academic year. The course drew on Kenny's textbook (Kenny, 2022) and accompanying activities for core MT and post-editing concepts, while newer topics such as large language models were addressed through additional materials and classroom activities. The assignment was situated in the later part of the course, after students had already received instruction on MT, post-editing, and translation quality evaluation during 10 weeks with both theoretical and practical weekly classes. In pedagogical terms, the task functioned as an integrative final assignment: rather than practising isolated techniques, students were asked to combine human translation, system comparison, metric interpretation, human quality assessment, and post-editing in a single workflow. This integrative design is broadly aligned with recent pedagogical proposals that respond to technological change through student-led experimentation, comparative evaluation, and explicit engagement with post-editing and AI literacy (Ayvazyan et al., 2024). The assignment therefore required students to explain why one output, with its particular strengths and weaknesses, was a better post-editing candidate than the alternatives.

3.3 Participants and materials

This study draws on 23 student projects produced for the final assignment in the course Traducció automàtica i postedició. All 23 cases are represented in the final evaluative-judgement master table linked below. Twenty-two cases also include a written report that supports qualitative interpretation, while one case lacks the final report and therefore contributes only limited structured information to the descriptive counts. Students worked into Catalan (n = 13) or Spanish (n = 10) and selected short and specialized English Wikipedia texts of approximately 300 words from several domains, most frequently zoology and translation technology. The task was designed to ensure variation in topic while maintaining a broadly comparable level of genre and text type, as students worked with short expository texts from a publicly available encyclopaedic source. The student reports confirm that the assignment was based on the comparison between a self-produced human translation and four machine-

generated versions, followed by the analysis of their relative strengths and weaknesses.

The study analyses student-produced coursework from a regular assessed module and reports the findings as an anonymized classroom case study. Formal ethics approval was not obtained before the assignment or the subsequent analysis, and the study therefore does not claim institutional ethics clearance. To minimize risk, all student work was anonymized before research reporting: students are identified only through case codes, no names or demographic identifiers are reported, and no direct quotations from student reports are used. Findings are presented in aggregate form or through paraphrased, non-identifying case summaries. The analysis focuses on patterns of evaluative reasoning in the assignment rather than on the assessment or profiling of individual students.

3.4 Assignment procedure

The assignment followed a multi-stage process. First, each student produced a human translation of the selected English source text into Catalan or Spanish. This stage served both as a reference for later evaluation and as a form of pedagogical preparation: by translating the text themselves before consulting AI or MT outputs, students had already made decisions about meaning, syntax, terminology, and style before comparing system outputs. This prior engagement helped them identify the main decision points in the source text and provided a stronger basis for the subsequent evaluation of machine-generated translations. Second, students generated four automatic translations: two with LLM-based systems and two with neural MT systems. In the case of the LLM systems, prompts and system versions were not fully standardized across all cases. While this reflects authentic classroom use of GenAI, it also limits strict technical reproducibility. Across the reports, the exact tools vary somewhat, but the pedagogical structure remained stable: students always compared two systems presented as GenAI and two systems presented as neural MT. Third, the four system outputs were evaluated automatically using the MATEO environment (Vanroy et al., 2023) and the metrics COMET, BLEU, TER, and ChrF2. Fourth, students carried out a human evaluation focused on adequacy and fluency on a scale of 4 at segment level, often accompanied by segment-level comments and error identification. Fifth, students selected the output they considered most suitable for post-editing and justified that choice. Finally, they performed post-editing on that selected version only, typically aiming at a publishable or fully revised target text. Student reports generally confirm this sequence of human translation, four-system comparison, automatic evaluation, human evaluation, system selection, and post-editing.

3.5 Data analysed

The analysis draws on several layers of student-produced material. These include the initial human translations, the four AI-generated outputs, the automatic metric scores, the human evaluation scores and comments, the error annotations recorded during evaluation, the post-edited final versions, and the written rationales included in the final reports. In analytical terms, the reports were analysed as evidence of students' reasoning, not just as records of system preference. Particular attention was therefore paid to the criteria students used when justifying preferences, especially where automatic and human evaluations diverged. Across the dataset, students frequently referred to adequacy, fluency, terminology, naturalness, structural calques, and expected post-editing effort when explaining their decisions. For this reason, the reports are central to the analysis: they show how students moved from scores and rankings to justified decisions about quality and revision.

In addition to the student reports themselves, the analysis relied on a final master coding table that consolidated, for each participant, the text domain, target language, systems compared, best automatic-evaluation system, best human-evaluation/final post-editing system, narrative evidence summary, reflection on AI tools, evaluative-judgement coding, and error notes. This master table made it possible to combine descriptive quantitative reporting with cross-case qualitative interpretation.

Descriptive counts based on system frequencies and broad coding distributions include all 23 projects represented in the master table. Interpretive claims about student reasoning, however, are grounded primarily in the 22 cases for which a written report was available.

3.6 Analytical procedure

The study was conducted by the course instructor, who designed the assignment and subsequently analysed the student submissions for research purposes. This dual role is methodologically relevant. The study does not claim the neutrality of an external observer; rather, it treats the reports as a pedagogically grounded classroom corpus

interpreted reflexively in relation to the course design.

The analysis combined descriptive quantitative aggregation with qualitative cross-case interpretation. Each report was first summarized using a common extraction template covering target language, systems compared, automatic-evaluation winner, student-selected system for post-editing, stated reasons for preference, recurring error types, post-editing effort, and explicit reflection on AI or MT tools. This template made it possible to compare cases systematically while preserving the contextual specificity of each student report.

The coding strategy was hybrid. Deductive categories were derived from the assignment design and the conceptual framing of the study, including adequacy, fluency, metric–human comparison, post-editing effort, and final system selection. Inductive subthemes were then identified through close reading of the reports, especially where students foregrounded issues such as terminology instability, acronym handling, scientific naming, syntactic naturalness, orthotypographic conventions, genre appropriateness, or the limitations of the human reference translation.

The reports were coded at case level rather than at segment level. The first layer of analysis recorded descriptive patterns, including whether the system favoured by automatic metrics was also selected by the student after human evaluation, or whether there was a divergence between metric-based and human judgement. Particular attention was paid to cases in which students explicitly rejected the metrically strongest output in favour of another system, since these moments offered especially clear evidence of evaluative judgement.

A second layer of coding focused on the criteria students used to justify their decisions. These included adequacy, fluency, terminology, naturalness, structural calques, formatting, orthotypographic conventions, grammatical accuracy, omissions, genre fit, and estimated post-editing effort. References to post-editing effort were coded when students commented on the extent of reformulation required, the number or severity of interventions needed, or whether the chosen output allowed for localised rather than extensive revision.

Finally, each case was assigned an evaluative-judgement rating on an ordinal scale: Low, Medium, Medium to High, or High. A report was coded as High when it triangulated several forms of evidence, such as automatic metrics, human adequacy and fluency evaluation, error analysis, and post-editing effort, and used them to justify the final system choice. Medium to High was used when the reasoning was explicit and multidimensional but less fully developed. Medium was assigned when the report included some qualitative justification but remained partly descriptive or metric-led. Low was used when the report mainly reported scores or outcomes without substantial evaluative interpretation.

AI-assisted drafting was used only for linguistic standardization of coding summaries after the researcher had reviewed the reports and assigned the categories; it was not used to assign codes, determine ratings, or generate analytical findings. All analytical decisions, category definitions, evaluative-judgement ratings, and final codes were checked and finalized by the researcher.

4 Results

4.1 Overview of student submissions

The results are organized in relation to the three research questions: first, the distribution of final selections and selection rationales; second, the relationship between automatic metric rankings and final human/post-editing choices; and third, the forms of evaluative judgement and critical AI use observable in the reports.

The final master table[1] contains 23 student projects, of which 22 are accompanied by written reports substantial enough for qualitative interpretation. The dataset includes 13 Catalan-target and 10 Spanish-target projects. Across the corpus, the most frequently used systems are ChatGPT (22 projects), Gemini (21), and DeepL (20), followed by Softcatalà (11), Bing Translator (8), Google Translate (5), Mistral (3), and eTranslation (2). This confirms that the assignment generated a broad but still comparable set of multi-system comparisons.

The distribution of "best" systems differs depending on whether one looks at automatic or

[1] Evaluative Judgement Master. The full anonymized coding matrix is available as supplementary material. It includes case identifier, target language, domain, systems compared, automatic-evaluation winner, final selected system, selection rationale, error categories, post-editing effort, reflection on AI tools, and evaluative-judgement level. https://docs.google.com/spreasheets/d/1BuKgQuMnQ2m8Bqo6XN6dO6KZQsJWtfmS7UckIuzGQ1U/edit?usp=sharing

human/final selection. In the automatic-evaluation column, ChatGPT is the most frequent best-ranked system (9 cases), followed by DeepL (6) and Gemini (3); there is also one ChatGPT/Gemini tie, plus 2 wins for Bing Translator, 1 for Softcatalà, and 1 for Google Translate. By contrast, in the final human/post-editing selection column, DeepL is chosen most often (10 cases), followed by Gemini (7), ChatGPT (5), and Bing Translator (1). At the family level, final choices are almost evenly split between LLM/GenAI systems (12 cases) and NMT systems (11 cases). These distributions suggest that final preferences were not dominated by a single tool family.

A further descriptive pattern appears at target-language level. In the Catalan-target projects, Gemini is the most frequent final choice (7 of 13), followed by DeepL (3), ChatGPT (2), and Bing Translator (1). In the Spanish-target projects, DeepL is the most frequent final choice (7 of 10), followed by ChatGPT (3). These counts should be interpreted cautiously because the source texts vary across projects, but they still show that student preferences were justified and text-dependent rather than uniform.

4.2 Evidence of evaluative judgement

The strongest result in the corpus is not that one system consistently "won," but that the assignment elicited criterion-based evaluative judgement. In the final coding, 14 reports are classified as High in evaluative judgement and 2 as Medium to High; 5 are coded as Medium, and only 1 as Low. The distribution indicates that most reports contained some form of explicit justification, usually involving more than one quality criterion. Table 1 summarizes the main forms of evaluative judgement identified in the report coding.

Manifestation of evaluative judgement	How it was evidenced in the coding	n/22	Typical form in the reports	Analytical significance
Evaluative judgement was substantively present in most reports	Reports coded High or Medium to High in the master table	**16**	Students justified preferences with explicit criteria rather than merely naming a preferred system	Shows that the task elicited genuine judgement, not simple tool preference
Students frequently compared automatic and human evaluation explicitly	Critical-judgement coding referred to direct comparison between metric results and human assessment	**15**	Students contrasted MATEO rankings with adequacy/fluency judgements and then explained why they converged or diverged	Supports the argument that metrics were used critically rather than accepted uncritically
Students often overrode the automatically best-ranked system	Best automatic system differed from the final post-editing choice	**11**	A system ranked best automatically was not always selected for post-editing	This is one of the clearest indicators of evaluative judgement in the dataset
Quality was usually treated as multidimensional	Reports used at least two quality dimensions beyond raw metric ranking	**20**	Students distinguished combinations such as adequacy, fluency, terminology, naturalness, grammar, formatting, coherence, or genre fit	Shows emerging translator reasoning about quality as composite and justified
Post-editing effort frequently entered into the final decision	Reasons for preference explicitly invoked fewer changes, easier revision, lower intervention burden, or faster post-editing	**14**	Students selected the version that would require fewer or more localized corrections, even when another system scored well automatically	Connects judgement to workflow awareness and professional usability

A smaller but important group adopted a metaevaluative stance	Reports explicitly questioned metric interpretation, reference-translation effects, domain limitations, or validated terminology externally	**6**	Students reflected on the limits of TER/COMET, the influence of their own reference translation, or the need to check domain terminology	Marks the strongest form of critical AI literacy in the dataset

Table 1. Evidence of critical/evaluative judgement across student reports (n = 22 analysable reports)

Across the stronger reports, students justified choices by weighing adequacy, fluency, terminology, naturalness, genre fit, semantic precision, and expected post-editing effort against one another. In other words, the assignment made traceable the kind of multidimensional reasoning that professional post-editing requires. This pattern is especially clear in cases where students explicitly rejected the metrically strongest system in favor of the output they judged more suitable for full revision.

At the same time, not all reports showed the same degree of analytical depth. A small minority remained largely metric-descriptive or offered limited qualitative justification. The weaker reports are also informative, because they show where additional scaffolding is needed: especially in connecting metric scores to error analysis and post-editing decisions.

4.3 Automatic metrics and human judgement

The clearest quantitative pattern concerns metric/human convergence. In the 22 report-backed cases, the best-ranked system in automatic evaluation and the final selected system matched in 12 cases and diverged in 10. The near-even split shows that metric rankings influenced the reports, but did not determine the final post-editing choice. Instead, students treated the metrics as one source of evidence among others. Table 2 compares the system most strongly favoured by automatic evaluation with the system ultimately selected by students for post-editing, making traceable the extent to which metric-based ranking and final human judgement converged or diverged across submissions. The student who did not submitted the final report is also included in the table. The best automatic system was coded as the system identified as strongest in the student's automatic-evaluation summary. Where students reported conflicting metric rankings, the coding followed COMET priority or the student's own stated automatic winner. Ties were retained where no single automatic winner was identifiable.

System	Best automatic system (n)	Final selected system (n)	System family
ChatGPT	9	5	GenAI
DeepL	6	10	NMT
Gemini	3	7	GenAI
Bing Translator	2	1	NMT
Google Translate	1	0	NMT
Softcatalà	1	0	NMT
ChatGPT=Gemini (tie)	1	0	GenAI tie

Table 2. Distribution of automatically best-ranked systems and final human/post-editing selections (n = 23)

This pattern is also visible in the contrast between system frequencies. ChatGPT appears most often as the automatic evaluation winner, but DeepL and Gemini together dominate the final human/post-editing choices. In practice, students often used automatic results as a starting point and then revised their conclusions when human evaluation surfaced issues of terminology, semantic fit, naturalness, or edit burden. Several reports make this tension explicit. In one case (C07), DeepL had stronger BLEU and TER values, but ChatGPT was selected because its Catalan syntax sounded more natural and its scientific terminology required fewer interventions. In another case (C10), ChatGPT was ranked first at the automatic evaluation step, but DeepL was chosen because it avoided a key semantic imprecision and preserved the source structure in a way that would reduce post-editing effort; the same report rejects Gemini partly because its unnecessary restructuring would create extra workload for a real post-editor.

4.4 Error annotation findings

The master table also makes it possible to report broader descriptive patterns in the error analyses.

A broad consolidation of the errors noticed field shows that terminology issues appear in 19 reports, accuracy in 15 reports, syntactic issues in 14, fluency/style issues in 14, and grammar issues in 13. These groupings are descriptive rather than absolute, but they indicate clearly where students' attention was concentrated. Table 3 summarizes the broad error categories most frequently identified in student reports, based on a cross-case coding of the error descriptions recorded in the master table.

Error category	Reports (n=22)	Typical problems observed
Terminology / lexis	19	unstable technical terms, lexical inaccuracies, false friends, untranslated items, acronym handling
Adequacy / meaning transfer	15	omissions, additions, reinterpretation, false sense, conceptual imprecision
Syntax / calques / structure	14	English-influenced structures, passive voice, literal syntax, awkward sentence structure
Naturalness / fluency / style	14	less idiomatic phrasing, awkward wording, low stylistic naturalness
Grammar / orthography / formatting	13	agreement, punctuation, spelling, apostrophation, numeric/notation conventions

Table 3. Broad error categories identified across student reports.

The error analysis shows that students' attention was concentrated not on fluency alone, but on specialized terminology, semantic precision, literal or English-influenced syntax, omissions/additions, and target-language conventions such as orthotypography, numeric formatting, and code or label preservation. One student report (C11), for example, values ChatGPT not only for adequacy and fluency, but also for preserving code, tags, and numbering, while still noting structural calques, anglicisms, and orthotypographic issues. A scientific Catalan report (C22) selects ChatGPT over DeepL because it offers a better overall balance, even though DeepL scores better on BLEU and TER metrics; the student highlights scientific terminology, conceptual precision, notation, and syntactic naturalness as decisive factors. Another report (C21) selects Bing Translator despite the student's initial expectation that the AI systems would perform better, because Bing Translator offered stronger terminological coherence and more natural syntax than the alternatives.

4.5 Selection and post-editing effort

The reports show that students usually selected not the output they considered best in the abstract, but the output they considered the best basis for full post-editing. In the final table, the chosen system is very often described as requiring low or low-to-moderate intervention, and students repeatedly explain their decisions in terms of fewer severe errors, less restructuring, better terminology, or more natural target-language syntax.

Many students evaluated outputs as post-editing candidates: the preferred version was often the one that seemed safest and most efficient to revise. That logic is explicit in several reports. One student (C10) prefers DeepL over ChatGPT partly because ChatGPT's errors would require heavier reformulation; the same report rejects Gemini because its restructuring would create unnecessary real-world workload. C09 chooses Gemini even though automatic evaluation leans toward ChatGPT, because Gemini, according to the report, offers the best final quality and needs only a few changes in specialized vocabulary.

4.6 AI literacy and critical use

Several reports show conditional use of AI/MT outputs: they accepted useful suggestions, rejected problematic ones, and justified where human revision remained necessary. Where explicit reflection is present, students do not treat AI-generated outputs as inherently superior, nor do they reject them categorically. Instead, they repeatedly evaluate them in relation to text type, terminology, target-language conventions, reference quality, and anticipated post-editing effort. This is consistent with the broader pattern in the master table, where reflections on AI tools tend to frame them as tools that can accelerate parts of the translation process but still require terminological checking, adequacy control, and stylistic revision by a human translator.

A further sign of developing AI literacy is that some students no longer frame the task as choosing a single "perfect" system. Instead, they begin to see different systems as offering partial strengths that can inform a better final result. One report (C08) explicitly comments that each tool contributed something useful to the final translation process, even though one system still had to

be selected for full post-editing. This is pedagogically important because it suggests that the student was not choosing a system blindly, but using differences between systems to improve the final translation.

When considered together, these reflections indicate that the assignment appears to have supported, and at least documented, not only system comparison, but also a more critical understanding of when AI output can be useful, where its risks lie, and why human judgement remains necessary. In that sense, the task appears to support AI literacy as a practical form of professional reasoning: students practice deciding when to rely on a system, when to question it, and what kind of human correction is needed.

5 Discussion

The results reposition system ranking as a means rather than an end: its main function was to prompt students to justify a post-editing decision. Students' final choices were distributed across both LLM/GenAI and neural MT systems, and the automatically best-ranked system did not always coincide with the version selected for post-editing.

The pattern fits the MT literacy and AI literacy literature, where technological competence is defined as contextual understanding rather than operation alone (O'Brien and Ehrensberger-Dow, 2020; Ehrensberger-Dow et al., 2023; Long and Magerko, 2020). In the present study, students used automatic metrics as one layer of evidence, but stronger reports did not treat metric scores as final judgements. Instead, students weighed them against adequacy, fluency, terminology, target-language naturalness, structural fit, and expected post-editing effort. The task positioned automatic evaluation as an object of interpretation, not an answer key.

The error patterns identified in the reports further confirm the value of combining automatic metrics with human assessment and error annotation. Students repeatedly focused on terminology, semantic precision, calques, passive structures, orthotypographic conventions, numerical formatting, omissions, additions, and target-language idiomaticity. Such issues may be underrepresented in purely score-based evaluation, especially when an output is fluent or close to the reference. In this respect, the assignment builds on earlier pedagogical work on MT evaluation and post-editing (Moorkens, 2018; Öner and Öner Bulut, 2021) by placing similar evaluative practices in a contemporary workflow that includes both LLM and NMT outputs.

The findings also caution against a simplified opposition between LLMs and NMT. The dataset does not support the claim that one system family is categorically superior. Instead, preferences were shaped by target language, domain, terminology, naturalness, and post-editing burden. For teaching, this variability is useful because it requires students to justify preferences in relation to a specific text, language, domain, and revision goal. It encourages students to treat AI and MT systems as resources with different affordances and risks, rather than as stable hierarchies of quality. Such a conclusion is consistent with recent translation-specific discussions of AI literacy, which emphasize suitability, data awareness, prompting, ethics, and contextual decision-making (Bowker, 2026; Inglada, 2026; Yamada, 2026; Moorkens and Doğru, 2026).

Finally, the study suggests one way of operationalizing evaluative judgement, AI literacy, and professional agency in a classroom assignment. Students were required to generate outputs, compare them, interpret metrics, perform human evaluation, identify errors, select one candidate, and post-edit it. This sequence left an analysable record of their reasoning. It also encouraged students to qualify their trust in the systems: outputs were useful, but still had to be checked against meaning, terminology, style, and post-editing effort. In this sense, the assignment operationalizes the idea that students must remain the "arbiters of quality" in AI-mediated translation (Bearman et al., 2024). Its contribution lies in showing that classroom tasks can move students beyond passive tool use toward critical comparison, accountable selection, and professionally grounded revision.

6 Conclusion

This case study has shown that the assignment's strongest contribution was to document how students justified translation-quality decisions within a realistic post-editing workflow. Across the 23 projects, preferences were neither uniform nor mechanically determined by automatic metrics. Instead, students justified their choices by weighing adequacy, fluency, terminology, naturalness, structural fit, and anticipated post-editing effort. The most important outcome was the reasoning itself: students connected quality, usability, and revision effort when deciding which output to post-edit.

A second key finding is that automatic metrics functioned as a useful starting point, but not as a sufficient basis for decision-making. In a substantial number of cases, the automatically best-ranked system was not the one ultimately selected for post-editing, because human evaluation brought into view issues that metrics alone could not capture adequately, especially semantic precision, terminology, target-language naturalness, restructuring, and local conventions. In this sense, the assignment helped students move beyond score comparison toward a more professionally relevant understanding of quality as multidimensional, contextual, and closely tied to revision effort.

The findings therefore support the use of multi-system comparison as a pedagogical design for AI literacy in translator education. The task encouraged a more conditional stance toward AI: students learned to use machine output when it was useful, question it when it was misleading, and justify the need for human intervention. In doing so, they learned to question fluent-looking output, interpret metrics critically, distinguish understandable from professionally acceptable translation, and justify why one version was preferable for a specific text and purpose.

This paper contributes: (1) an adaptable classroom workflow for comparing LLM and NMT outputs through automatic and human evaluation; (2) descriptive and qualitative evidence of how students justify post-editing selections; and (3) an argument for treating evaluative judgement as a practical component of AI literacy in translator education.

These conclusions should be read in light of the study's limits. The dataset is small, classroom-based, and built on text-specific comparisons using student-produced reference translations, so the paper does not support broad claims about whether LLMs or NMT systems are categorically superior. Its contribution is pedagogical: it shows how a comparison-and-post-editing task can generate evidence of students' evaluative reasoning. In this setting, students become less passive users of AI-based translation tools and more critical comparators, revisers, and arbiters of quality.

References

Ayvazyan, Nune, Yu Hao, and Antony Pym. 2024. Things to do in the translation class when technologies change: The case of generative AI. In Peng, Yuhong, Huihui Huang, and Defeng Li, editors, *New Advances in Translation Technology,* pages 219–238. Springer, Singapore.

Ayvazyan, Nune. 2025. Human-centered pedagogies in the age of generative AI: Examining student perceptions of augmentation, progress, and agency. *InContext. Studies in Translation and Interculturalism,* 5(1):167–193.

Bearman, Margaret, Joanna Tai, Philip Dawson, David Boud, and Rola Ajjawi .2024. Developing evaluative judgement for a time of generative artificial intelligence. *Assessment and Evaluation in Higher Education*, 49(6):893–905.

Belam, Judith. 2002. Teaching machine translation evaluation by assessed project work. In *Proceedings of the 6th EAMT Workshop: Teaching Machine Translation*, pages 131–136, Manchester. European Association for Machine Translation.

Bowker, Lynne. 2026. Teaching translation students about data in the age of generative AI. In Penet,, J. C., Joss Moorkens, and Masaru Yamada, editors, *Teaching Translation in the Age of Generative AI: New Paradigm, New Learning?*, pages 67–85, Language Science Press, Berlin.

Doherty, Stephen and Dorothy Kenny. 2014. The design and evaluation of a statistical machine translation syllabus for translation students. *The Interpreter and Translator Trainer*, 8(2):295–315.

Ehrensberger-Dow, Maureen., Alice Delorme Benites, , Caroline and Lehr. 2023. A new role for translators and trainers: MT literacy consultants. *The Interpreter and Translator Trainer*, 17(3):393–411.

European Commission. 2022. *European Master's in Translation: Competence Framework 2022*. Directorate-General for Translation.

Flanagan, Marian and Tina Paulsen Christensen. 2014. Testing post-editing guidelines: How translation trainees interpret them and how to tailor them for translator training purposes. *The Interpreter and Translator Trainer*, 8(2):257–275.

Freeth, Peter J. and Piero Toto. 2026. The impact of AI-in-the-loop pedagogy: An action research approach to integrating AI technologies in translator training. *The Interpreter and Translator Trainer*, 1–18.

Freitag, Markus, Nitika Mathur, Chi-kiu Lo, Elftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 Metrics Shared Task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Guerberof Arenas, Ana and Joss Moorkens. 2019. Machine translation and post-editing training as part of a master's programme. *The Journal of Specialised Translation*, 31:217–238.

Inglada, Ramon. 2026. AI literacy: The concept of suitability and core translation skills. In Penet, J. C., Joss Moorkens, and Masaru Yamada, editors, *Teaching Translation in the Age of Generative AI: New Paradigm, New Learning?*, pages 49–64. Language Science Press, Berlin.

Kenny, Dorothy. 2020. Technology in translator training. In O'Hagan, Minako, editor, *The Routledge Handbook of Translation and Technology*, pages 498–515. Routledge, London/New York.

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

Krüger, Ralph and Janiça Hackenbuchner. 2024. A competence matrix for machine translation-oriented data literacy teaching. *Target. International Journal of Translation Studies*, 36(2):245–275.

Lommel, Arle, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 12, 455–463.

Long, Duri and Brian Magerko. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–16, Honolulu. Association for Computing Machinery.

Massey, Gary, Elsa Huertas-Barros, and David Katan, editors. 2023. *The Human Translator in the 2020s*. Routledge, London.

Massey, Gary and Maureen Ehrensberger-Dow. 2026. Translation competence in the age of generative AI: Debates, dilemmas, directions. In Penet, J. C., Joss Moorkens, and Masaru Yamada, editors, *Teaching Translation in the Age of Generative AI: New Paradigm, New Learning?*, pages 3–26. Language Science Press, Berlin.

Miao, Fengchun and Wayne Holmes. 2023. *Guidance for Generative AI in Education and Research*. UNESCO.

Moorkens, Joss. 2018. What to expect from neural machine translation: A practical in-class translation evaluation exercise. *The Interpreter and Translator Trainer*, 12(4), 375–387.

Moorkens, Joss, Pilar Sánchez-Gijón, Esther Simon, Mireia Urpí, Nora Aranberri, Dragoș Ciobanu, Ana Guerberof-Arenas, Janiça Hackenbuchner, Dorothy Kenny, Ralph Krüger, Miguel Rios, María Isabel Rivas Ginel, Caroline Rossi, Alina Secară, and Antonio Toral. 2024. Literacy in digital environments and resources (LT-LiDER). In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)* , pages 55–56, Sheffield. European Association for Machine Translation.

Moorkens, Joss, and Gokhan Doğru. 2026. Teaching AI ethics for translation students. In Penet, J. C., Joss Moorkens, and Masaru Yamada, editors, *Teaching Translation in the Age of Generative AI: New Paradigm, New Learning?*, pages 105–122. Language Science Press, Berlin.

Nitzke, Jean, Silvia Hansen-Schirra, and Carmen Canfora. 2019. Risk management and post-editing competence. *The Journal of Specialised Translation*, 31:239–259.

O'Brien, Sharon. 2002. Teaching post-editing: A proposal for course content. In *Proceedings of the 6th EAMT Workshop: Teaching Machine Translation*, pages 99–106, Manchester. European Association for Machine Translation.

O'Brien, Sharon and Maureen Ehrensberger-Dow. 2020. MT literacy—A cognitive view. *Translation, Cognition and Behavior*, 3(2):145–164.

Öner, Işın and Senem Öner Bulut. 2021. Post-editing oriented human quality evaluation of neural machine translation in translator training: A study on perceived difficulties and benefits. *transLogos Translation Studies Journal*, 4(1):100–124.

Penet, J. C., Joss Moorkens, and Masaru Yamada, editors. 2026. *Teaching Translation in the Age of Generative AI: New Paradigm, New Learning?* Language Science Press, Berlin.

Ramírez-Polo, Laura and Chelo Vargas-Sierra. 2023. Translation technology and ethical competence: An analysis and proposal for translators' training. *Languages*, 8(2):1–22.

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, Online. Association for Computational Linguistics.

Rico, Celia and Diana González Pastor. 2022. The role of machine translation in translation education: A thematic analysis of translator educators' beliefs. *Translation & Interpreting: The International Journal of Translation and Interpreting Research*, 14(1):177–197.

Rivas Ginel, María Isabel and Joss Moorkens. 2025. Translators' trust and distrust in the times of GenAI. *Translation Studies*, 18(2), 283–299.

Tai, Joanna, Rola Ajjawi, David Boud, Philip Dawson, and Ernesto Panadero. 2018. Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education*, 76, 467–481.

Tipton, Rebecca. 2024. *The Routledge Guide to Teaching Ethics in Translation and Interpreting Education*. Routledge, London/New York.

Vanroy, Bram, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine Translation Evaluation Online. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 499–500, Tampere. European Association for Machine Translation.

Yamada, Masaru. 2026. Teaching translation with AI: Bridging theory and practice through prompt engineering. In Penet, J. C., Joss Moorkens, and Masaru Yamada, editors, *Teaching Translation in the Age of Generative AI: New Paradigm, New Learning?,* pages 87–104. Language Science Press, Berlin.

Zhang, Jia and Stephen Doherty. 2025. Investigating novice translation students' AI literacy in translation education. *The Interpreter and Translator Trainer*, 19(3–4), 234–253.

COPECO-Speech: Multimodal Post-Editing with Speech and LLMs for Translation Teaching

Jeevanthi Liyana Pathirana[1], **Pierrette Bouillon**[2], **Jonathan Mutal**[2], **Sabrina Girletti**[2], **Lise Volkart**[2]
University of Geneva
Jeevanthi.Liyana@etu.unige.ch
Pierrette.Bouillon@unige.ch, Jonathan.Mutal@unige.ch
Sabrina.Girletti@unige.ch, Lise.Volkart@unige.ch

Abstract

In this demo, we present COPECO-Speech, a multimodal post-editing workbench designed for translator training with AI-based translation technologies. It extends an existing pedagogical post-editing platform (COPECO) by integrating speech input and Large Language Model (LLM)-assisted editing. The workbench has different post-editing modalities and helps teachers annotate student tasks using either a shared or personalized annotation scheme. The system logs all interactions—including keystrokes, speech input, editing actions and LLM operations—enabling detailed analysis of post-editing processes.

1 Introduction

Machine translation (MT) and large language models (LLMs) are transforming translation workflows, requiring training approaches that emphasise interaction design, technical literacy and critical AI use (Bowker and Buitrago Ciro, 2019; Jiao et al., 2023). Recent LLM-assisted post-editing workflows exist (Castaldo et al., 2025), but CAT tools (e.g. MateCat[1]) are largely keyboard-based and only support basic speech interaction and no LLM-driven editing.

We introduce a multimodal PE workbench for classroom use. It lets students perform PE using different modalities (typing, respeaking, and LLM-assisted editing), while logging process data for instructor-led reflection and translation curricula integration.

2 System overview

Our work extends COPECO[2], a collaborative pedagogical post-editing platform (Mutal et al., 2020; Liyana Pathirana et al., 2024), by integrating multimodal interaction and LLM-assisted PE. Instructors upload source texts and MT output from any system, assign PE tasks, and configure system behaviour through prompts and settings. The system supports multiple languages and can be extended to additional language pairs. This work represents the first integration of LLM-based functionality within the COPECO platform. PE is performed through two interaction modes:

- **Typing:** users directly edit the MT output using a standard text interface.
- **Speech interaction:** users interact through speech for both navigation and editing. A small set of predefined commands (e.g. "Next/Previous/Save segment") supports navigation, while spoken editing instructions (e.g. insertions, deletions, or reformulations) are interpreted and executed by an LLM to modify or regenerate the translation.

Speech input is recognized using Azure Speech services[3]. COPECO-Speech can replace ASR and LLM components. The speech instruction, the source text and MT output are sent to an LLM (currently GPT-4.1-mini), guided by prompts. The LLM performs the requested edits and returns an updated translation. (Figure 1).

[1]https://guides.matecat.com/translate-1

[2]https://copeco.unige.ch

[3]https://learn.microsoft.com/en-us/azure/ai-services/speech-service/speech-to-text

Proceedings of the 1st International Workshop on Teaching AI-Based Translation and Technologies (TAITT 2026), pages 49–51
June 15, 2026 ©2026 Association for Computational Linguistics

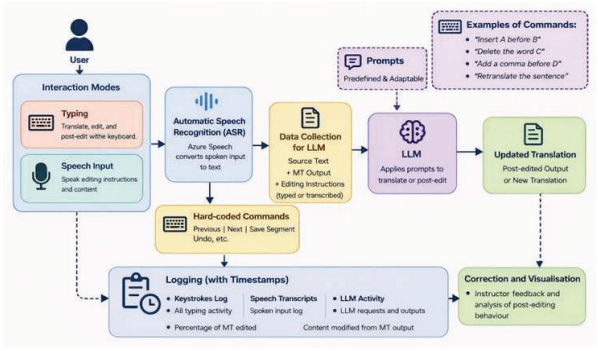

Figure 1: Workflow of COPECO-Speech, a multimodal speech-driven post-editing workbench

The prompt uses the input and guides the LLM in tasks (language detection, targeted PE or full re-translation, while ensuring consistency in the generated output). Users can switch between interaction modes. All actions—including keystrokes, speech input, and LLM activity—are logged for process-oriented analysis (Carl et al., 2016). The system currently supports English and French and can be extended for other language pairs.

3 Pedagogical use case

In a typical lab session, students post-edit texts using different interaction modes (e.g. typing,speech/LLM based editing). Instructors review, annotate and discuss the resulting translations and logs, providing targeted feedback and reflection on PE behavior. This supports discussion of PE strategies, efficiency, and use of AI (Koponen, 2016).

4 Demo description

In this demo, we will present the workbench and its main functionalities, focusing on speech-driven LLM-assisted post-editing. A live demo will illustrate how spoken commands can be used to modify and improve MT output, and how the system logs user interactions :

- **Loading:** Source text and MT output loaded into COPECO platform.
- **Post-editing via typing** (baseline): Direct keyboard-based corrections.
- **Speech interaction:**
  - **Navigation (hard-coded commands):** “Next”, “Previous”, “Save segment”, “Undo”.
  - **Editing via LLM:** “Insert A before B”, “delete second occurrence of C”, “re-translate the sentence”.
- **Correction and visualization:** Instructor feedback through annotation, comparison with reference translations, and reports with post-editing statistics.
- **Logs:** Keystrokes log, spoken input log, time taken, teacher’s corrections, and LLM requests/outputs, percentage of MT content edited.

5 Conclusion

This workbench offers a practical tool for teaching multimodal PE and human–AI interaction. By logging spoken editing instructions and enabling comparison across interaction modes, it supports reflective pedagogy and the empirical exploration of PE strategies in AI-assisted translation workflows within the classroom.

References

Bowker, Lynne and Jairo Buitrago Ciro. 2019. *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Emerald Publishing Limited, Bingley.

Carl, Michael, Moritz Schaeffer, and Srinivas Bangalore. 2016. The CRITT translation process research database. In *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*, pages 13–54. Springer, Cham.

Castaldo, Antonio, Sheila Castilho, Joss Moorkens, and Johanna Monti. 2025. Extending CREAMT: Leveraging large language models for literary translation post-editing. In *Proceedings of Machine Translation Summit XIX*, page 506–515. European Association for Machine Translation.

Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? A preliminary study. *arXiv preprint arXiv:2301.08745*.

Koponen, Maarit. 2016. Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *The Journal of Specialised Translation*, (25):131–148.

Liyana Pathirana, Jeevanthi, Perrine Schumacher, Pierrette Bouillon, and Jonathan David Mutal. 2024. Enhanced multilingual speech-based post-editing based on user feedback with COPECO-SPEECH. In *Proceedings of Translating and the Computer 46*, pages 161–167, Luxembourg. Tradulex.

Mutal, Jonathan, Pierrette Bouillon, Perrine Schumacher, and Johanna Gerlach. 2020. COPECO: a collaborative post-editing corpus in pedagogical context. In *Proceedings of the Workshop on Post-Editing Technology and Practice (PEMDT) at AMTA*, pages 61–78.

Beyond Post-Editing: A Project-Based Module on MT and LLM Integration for Trainee Translators

Alina Karakanta
Leiden University Centre for Linguistics, Leiden University
a.karakanta@leidenuniv.nl

Abstract

With the rapid technologisation of translation, skills beyond post-editing (PE), such as data literacy, technology evaluation, and critical engagement with AI-based tools are becoming essential competencies for trainee translators. This paper presents a syllabus for a translation technology module that equips MA Translation students with end-to-end MT evaluation skills, from engine selection and domain adaptation to automatic and human evaluation, post-editing, and results reporting. Large language models (LLMs) are integrated throughout, as translation engines, as a basis for prompting and domain adaptation strategies, and as a source of explainable quality estimation signals. Grounded in project-based learning, students apply these skills in a simulated client scenario. Trends on students' engine selection and customisation preferences observed across three years suggest a gradual shift towards LLM-based tools, but traditional engines and built-in options remain a firm presence in student practice.

1 Introduction

Translation is more technological than ever. The widespread adoption of neural machine translation (NMT) and, more recently, large language models (LLMs) has fundamentally altered how translation is produced, reviewed, and delivered. Translators increasingly operate within complex, multi-step workflows where MT output passes through quality estimation filters, automated post-editing suggestions, and human review before reaching its final form. CAT tools have evolved accordingly, now embedding quality signals, adaptive suggestions, and LLM-powered features directly into the translation interface.

This rapid shift has significant implications for translator training. The professional landscape is diversifying: while many translators continue to work with MT as end users, there is growing demand for adjacent roles such as MT specialist and language technology consultant. Even for those who remain primarily in translation practice, critical awareness of how MT systems work, how they are evaluated, and how they can be adapted is increasingly expected. In this age, critical skills are more relevant than ever. Skills such as data literacy, technology assessment, and evidence-based reasoning about MT quality are no longer the exclusive domain of computational linguists but are becoming relevant across a range of profiles that translation graduates may occupy.

Yet translator training has been slow to reflect this breadth. Existing syllabi have made significant progress in covering the theoretical underpinnings of MT, pre-editing, human and automatic evaluation, project management, and post-editing as a core professional skill (Doherty and Kenny, 2014; Moorkens, 2018; Guerberof Arenas and Moorkens, 2019; Mellinger, 2017; Pym and Hao, 2024). However, the pace of technological development and particularly the rise of LLMs means that the field is still figuring out how to incorporate these tools meaningfully into translator education. The role of LLMs within the MT pipeline, whether as translation engines, customisation tools, or sources of explainable quality signals, remains under-represented in published ped-

Proceedings of the 1st International Workshop on Teaching AI-Based Translation and Technologies (TAITT 2026), pages 52–62
June 15, 2026 ©2026 Association for Computational Linguistics

agogical proposals.

This paper addresses that gap by presenting a translation technology module grounded in project-based learning, in which students develop end-to-end MT evaluation competencies across seven sessions. The module combines commercial and open-source tools incorporating state-of-the-art MT engines and LLM-based tools. The approach is flexible in that its underlying framework is adaptable as new technologies emerge.

2 Research on teaching translation technologies

2.1 Skills

Early work on teaching translation technologies established a template that has proven durable. O'Brien (2002) presented one of the first systematic proposals for teaching MT, combining theoretical knowledge of MT, terminology, and pre-editing with practical elements such as PE with different engines and programming macros. Subsequent course designs followed this integrated spirit with topics rangin from MT evaluation, controlled language, PE of MT output (Doherty et al., 2012; Kenny and Doherty, 2014), ethics, payment, collaboration, the role of the human translator (Moorkens, 2022), and translation workflows, to PE effort and process research (Koponen, 2015). Doherty and Kenny (2014) offered a full course design incorporating a range of technology-related competencies alongside translation practice. Guerberof Arenas and Moorkens (2019) designed a machine translation and post-editing course, that includes two modules: one on PE and one on MT project management. TAUS PE guidelines and course.

Compared to PE skills, evaluation has only recently received increasing attention in translator training. With the advent of neural MT, Moorkens (2018) presented a proposal for a practical in-class translation evaluation exercise where students carried out evaluations of neural and statistical MT using translation quality assessment: adequacy, post-editing effort (time), and a simple error taxonomy (word order errors, mistranslations, omissions, and additions). Macken et al. (2023) proposed incorporating evaluation with automatic metrics into translator training, by developing an open-source, accessible evaluation platform (Vanroy et al., 2023).

Hands-on work with MT systems has also been supported by dedicated resources. Two seminal coursebooks on statistical and neural MT (Koehn, 2010; Koehn, 2020) have provided theoretical and practical grounding for courses engaging with engine-level knowledge, however these are often inaccessible to translators due to their lack of technical and programming knowledge. To address this this technical barrier, the MultiTraiNMT project[1] developed a coursebook (Kenny, 2022) and an online platform allowing students to train, fine-tune, and evaluate MT systems without programming skills. Krüger (2022) approached the same challenge through prepared Jupyter notebooks, enabling students to work with MT data and evaluation pipelines hands-on without requiring a programming background.

The concept of MT literacy (Bowker and Ciro, 2019), encompassing not just the ability to use MT tools but to understand, evaluate, and critically engage with them, has been influential in framing what translators need to know. Hackenbuchner and Krüger (2023) extend this to data literacy by proposing DataLit^{MT}. which includes a Professional MT Literacy Framework, an MT-specific data literacy framework and a competence framework for Machine Translation-Oriented Data Literacy Teaching. They argue that professional machine-translation literacy encompasses various skills and knowledge areas, including technical, linguistic, economic, societal, and cognitive aspects.

2.2 Pedagogical approaches

Research broadly agrees on the value of combining a theoretical component with hands-on practice. Lecture-based methods are typically used to convey background knowledge such as MT history and system types, while practical skills are developed through student-centred activities ranging from task-based group work to large-scale simulated projects (Hao and Pym, 2023). (Mellinger, 2017) making the case for embedding technology throughout the translation curriculum rather than treating it as a standalone topic. Project-based and authentic assessment approaches have also been advocated, with students working on realistic scenarios that mirror professional practice (Kiraly, 2005; Li et al., 2015; Mitchell-Schuitevoerder, 2020).

[1] https://www.multitrainmt.eu/

2.3 GenAI and LLMs

The integration of generative AI and LLMs into translation education is still emerging. Pym and Hao (2024) offers a guide to using generative AI for language learning activities, while Ayvazyan et al. (2024) provides a practical inventory of activities involving generative AI for translators. These contributions signal growing interest in the area, but systematic pedagogical proposals for integrating LLMs into translator training, particularly within an end-to-end MT evaluation framework, remain scarce.

3 Context

The module described in this paper forms the second block of Translator's Tools, a first-year course in the MA Translation programme at Leiden University. The first block introduces students to CAT tools, terminological databases, quality assurance, and cloud-based collaboration interfaces, assessed through an in-class CAT tool assignment. Students therefore enter the second block with a working knowledge of translation workflows, translation memories, terminology management, and cloud and collaborative tools. The second block, the focus of this paper, covers language technologies, machine translation, customisation, evaluation, post-editing, and quality estimation.

The module consists of seven weekly two-hour sessions, combining a theoretical and a practical component. Sessions take place in a translation lab with access to high-performing laptops. Licences for commercial MT engines and CAT tool integrations are obtained through the institution's resource infrastructure or through educational agreements with providers for the duration of the block. The module has been taught over three consecutive years to cohorts of approximately 25 students per year, allowing for iterative refinement based on student feedback and developments in the field.

4 Approach

The module is grounded in project-based learning, an approach that seeks to foster independent thinking, teamwork, and problem-solving in unfamiliar situations, with a view to acquiring transferable skills that remain relevant regardless of technological change (Kiraly, 2005). Rather than strictly following any single pedagogical model, the module draws on the spirit of Kiraly's empowerment-oriented approach (Kiraly, 2005), in which students take ownership of a complex, realistic task and develop competencies through doing. This is complemented by formative assessment: students submit weekly individual exercises based on the practical activities of each session, receiving written feedback or participating in class discussions that address common tendencies across the cohort. The summative assessment is a group project in which students propose and evaluate an MT solution for a simulated client scenario, described in detail in Section 4.3.

While the weekly exercises develop discrete skills progressively, the project brings these together into an end-to-end MT evaluation pipeline: students select engines, perform customisation, conduct automatic and human evaluation, carry out post-editing, and report their findings and recommendations. This integrated pipeline is the central pedagogical contribution of the module. Rather than treating each skill in isolation, students experience how the components of MT evaluation relate to and inform one another in practice.

Large language models are integrated throughout the module, serving three distinct functions. First, LLMs are included alongside conventional MT engines in the engine selection and comparison activities, giving students direct experience of their translation behaviour and characteristics. Second, LLMs are used as a basis for prompting strategies in domain adaptation, including retrieval-augmented generation approaches, allowing students to explore customisation methods that go beyond the translation memory and glossary integration available in traditional MT systems. Third, in the quality estimation session, LLMs provide explainable quality signals (word- and segment-level error highlights and automatic post-editing suggestions) that students learn to interpret and use to support their post-editing decisions. This integration reflects the current professional landscape, in which LLMs are increasingly embedded in MT workflows, and prepares students to engage with them critically rather than as passive end users.

The module uses a combination of commercial and open-source tools throughout, exposing students to the range of options available in professional practice. The underlying framework is designed to be flexible: as new engines, metrics, and LLM-based features emerge, individual components can be updated without disrupting the overall

structure.

4.1 Learning objectives

By the end of the module, students will have developed a comprehensive set of competencies spanning the full MT pipeline. They will understand how MT systems and LLMs work and be able to identify their role in contemporary translation workflows, as well as prepare and manage the data resources these systems rely on, including corpora, translation memories, and glossaries. Students will be able to customise MT systems using different types of data and prompting strategies, and assess the quality of MT output through both automatic metrics and human evaluation protocols. They will practise post-editing in accordance with professional guidelines, and learn to interpret quality estimation signals, such as explainable LLM-based outputs, to support post-editing decisions. Finally, students will gain experience in conducting, analysing, and reporting on an end-to-end MT evaluation project situated in a realistic professional scenario.

4.2 Session descriptions

The module is organised in 7 sessions. Each session includes a theoretical part, explaining basic notions and clarifying the readings, and a practical session with hands-on activities.

Session 1. Introduction to MT and language technologies

This session introduces machine translation by first eliciting students' own conceptualisations of the technology, before covering its evolution from rule-based and statistical approaches to neural MT and large language models. The theoretical component explains how NMT and LLMs work at a conceptual level, surveys language technologies relevant to the translator, such as optical character recognition and automatic speech recognition, and clarifies the distinction between translation memories and MT engines.

Practical activities:

- Explore a range of commercial and open-source MT engines, including DeepL[2], Google Translate[3], Lara[4], Widn[5], Apertium[6], and No Language Left Behind (Costa-Jussà et al., 2022), mapping each to its underlying architecture and noting available functionalities such as image-based translation, speech input, glossary integration, and quality or explanation features.
- Translate two short texts from different domains (institutional and audiovisual) using multiple engines and compare outputs based on translation theory knowledge, without formal evaluation criteria.
- Connect an MT engine to a CAT tool (e.g. Trados Studio) and practise working with MT suggestions within a translation interface.

Formative exercise: Students submit a short comparative reflection on the engine outputs for the two texts, discussing differences in translation behaviour across engines and domains based on their observations.

Session 2. Corpora and glossaries

This session covers the data that is used to build MT systems. The theoretical component explains the role of parallel corpora and glossaries in MT training and customisation, introducing students to the concept of in-domain and out-of-domain data and the principle that data quality directly affects translation quality. Students are introduced to existing parallel corpus resources such as OPUS (Tiedemann, 2009), and learn about common data quality issues including noise, misalignments, and domain mismatch. The session also covers how glossaries can be prepared and formatted for use in MT customisation.

Practical activities:

- Explore the OPUS parallel corpus repository and discuss the types of data available, their origins, and their suitability for different translation domains.
- Perform translation memory cleaning using a TM maintenance tool or Okapi Olifant,[7] identifying and removing noisy or misaligned segments.
- Prepare a domain-specific glossary in the format required for MT customisation, including tab-separated term pairs and any required metadata.

Formative exercise: Students submit a cleaned TM and a prepared glossary based on provided data, with a short reflection on the quality issues identified and the decisions made during cleaning.

[2] https://www.deepl.com/
[3] https://translate.google.com/
[4] https://laratranslate.com/
[5] https://www.widn.ai/
[6] https://apertium.org/
[7] https://okapiframework.org/wiki/index.php?title=Olifant

Session 3. **Machine Translation customisation**
This session covers the difference between generic and domain-adapted MT systems. The theoretical component explains what a domain is, why domain adaptation matters for translation quality, and the main strategies available for customising MT engines. Students learn about three approaches to customisation: fine-tuning with in-domain parallel data, glossary integration, and prompt-based adaptation. The latter is introduced through a demonstration of an API call to a large language model with few-shot translation examples and glossary terms integrated into the prompt, illustrating how LLMs can be steered towards domain-specific translation behaviour without retraining.

Practical activities:

- Customise a MT engine using a domain-specific glossary (e.g. via the DeepL glossary function), comparing output with and without customisation.
- Use an MT engine (e.g. Lara) to translate with an integrated translation memory, observing how the system incorporates existing translations into its output.
- Apply prompt-based customisation by constructing prompts with domain instructions, few-shot examples, and glossary terms, and evaluating the effect on translation output.

Formative exercise: Students submit a comparison of the three customisation approaches applied to a short text, discussing the effect of each strategy on translation quality based on their observations.

Session 4. **Human evaluation of MT**
This session introduces the methods used to evaluate MT quality through human judgement. The theoretical component covers the main human evaluation protocols: adequacy-fluency ratings, direct assessment (Graham et al., 2013), ranking, and error annotation based on Multidimensional Quality Metrics (MQM) (Lommel et al., 2014). The session features a collaborative in-class evaluation activity designed to make the evaluation process concrete and tangible. Large printed sheets are posted around the room, each displaying a source sentence, a human reference translation, and the output of three MT systems. Working together, students collect adequacy-fluency ratings, perform direct assessment, rank the outputs, and conduct MQM annotation, discussing their judgements and disagreements as a group. Students also learn how to select sentences for human evaluation, including strategies for sampling representative and informative segments from a test set.

Practical activities:

- Perform in-class human evaluation using adequacy-fluency ratings, direct assessment, and ranking across multiple MT outputs.
- Conduct MQM annotation, identifying and classifying errors in MT output using an MQM scorecard.
- Discuss inter-annotator agreement, the subjectivity of human judgements, and the implications for evaluation reliability.

Formative exercise: Students submit a completed MQM annotation of a short MT output, including error classification and a brief reflection on the challenges encountered and decisions made during annotation.

Session 5. **Automatic evaluation of MT**
This session introduces automatic MT evaluation as a complement to human judgement. The theoretical component motivates the need for automatic metrics: human evaluation is costly, time-consuming, and difficult to scale. It also covers the main families of metrics in use today. Students learn about surface-level metrics based on word and phrase overlap, i.e., BLEU (Papineni et al., 2002), chrF (Popović, 2015), and TER (Snover et al., 2006), embedding-based metrics that capture semantic similarity, i.e., BERTScore (Zhang et al., 2020), and BLEURT (Sellam et al., 2020), and learned metrics trained on human judgement data, i.e., COMET (Rei et al., 2020). The strengths and limitations of each family are discussed, including their varying degrees of correlation with human assessments.

Practical activities:

- Prepare a test set for automatic evaluation, ensuring correct formatting and alignment between source, hypothesis, and reference files.
- Evaluate the output of multiple MT systems using the MATEO platform (Vanroy et al., 2023), applying a range of metrics and comparing results across systems.
- Interpret metric scores, analyse agreements and disagreements across metrics, and discuss the use of statistical significance testing to validate comparisons.

- Reflect on the relationship between automatic scores and the human judgements collected in the previous session, identifying cases of agreement and divergence.

Formative exercise: Students submit an automatic evaluation of two MT systems using MATEO, including a written interpretation of the results across metrics and a discussion of the relationship of the scores obtained with the human evaluation scores from Session 4.

Session 6. Post-editing
This session introduces post-editing as a professional practice. The theoretical component covers what post-editing is and how it differs from traditional editing and reviewing, the distinction between light and full post-editing, and professional guidelines and standards. The session also addresses how professional translators experience and perceive post-editing, questions of cognitive effort, productivity, and the implications for professional practice and payment.

Practical activities:

- Discuss the differences between light and full post-editing and when each is appropriate in a professional context.
- Study and apply post-editing guidelines, including the TAUS post-editing guidelines, to a sample MT output.
- Conduct a timed productivity exercise in which students post-edit two short texts from two different MT engines, following a counterbalanced design with two groups. Students record their completion times, which are then aggregated and discussed as a class. The exercise illustrates variability across post-editors, the relationship between engine quality and post-editing effort, and the methodological challenges of designing valid productivity studies.
- Reflect on the types of errors encountered and the decisions made during post-editing, connecting observations to the error typologies covered in Session 4.

Formative exercise: Students submit a post-edited text with tracked changes, accompanied by a short reflection on the types and frequency of errors encountered, the effort required, and their experience of working with MT output in a CAT tool environment.

Session 7. Quality estimation and human-in-the-loop
This session introduces quality estimation (QE) as a method for predicting MT quality without a reference translation. The theoretical component explains the difference between quality evaluation, which requires human judgements or reference translations, and quality estimation, which operates on source and MT output alone. Students learn why QE is used in production workflows, how it can be used to triage segments for post-editing, and how recent developments in explainable QE provide word- and segment-level error signals that go beyond a single quality score. Multi-agent workflows are introduced, including LLMs for generating explainable QE outputs, automatic post-editing (APE) suggestions and for quality assurance.

Practical activities:

- Compare quality estimation and quality evaluation, discussing the scenarios in which each is appropriate and their respective limitations.
- Examine the output of state-of-the-art QE models, including xCOMET (Guerreiro et al., 2024) and xTower (Treviso et al., 2024), focusing on word-level error highlights and segment-level quality scores.
- Use SmartPE,[8] an error highlighting and post-editing tool, to practice post-editing with automatic error highlights and translation correction suggestions generated through xTower's automatic post-editing.
- Discuss the explainability functionalities enabled by LLMs in QE tools, critically evaluating the reliability and usefulness of the signals provided.

Formative exercise: Students submit a post-edited text completed using SmartPE, with a reflection on how the quality estimation signals influenced their post-editing decisions, which suggestions they accepted or rejected, and why.

4.3 Machine Translation project

The summative assessment for the module is a group project in which students of 3–4 propose and evaluate an MT solution for a simulated client scenario. The project is designed to bring together the skills developed across the seven sessions into an end-to-end evaluation pipeline, requiring students

[8] https://github.com/fatalinha/SmartPE

to make and justify decisions at each stage rather than simply applying procedures mechanically.

Students are provided with a dataset one month before the deadline, containing a test set of aligned source and target sentences, a translation memory in both .tmx and Trados formats, and a domain-specific glossary. The goal is to answer the research question: which is the best MT engine and setting for this content? Working from this question they proceed through four stages. First, they translate the test set using at least three MT engines of their choice. Second, they perform automatic evaluation using a range of metrics via the MATEO platform (Vanroy et al., 2023), selecting the best-performing baseline engine. Third, they customise one engine using the provided translation memory and glossary, and evaluate the effect of customisation through automatic metrics. Fourth, they select the two best-performing systems and conduct human evaluation on a sample of fourty sentences, choosing from adequacy-fluency rating, ranking, MQM annotation, and perform post-editing.

Project report Students report their findings in a research paper of 3000–4000 words, excluding references and appendices. The introduction situates the project within the relevant literature on MT, domain adaptation, and evaluation; the methods section describes the data, scenario, and evaluation procedures; the results section presents and interprets findings across automatic evaluation, customisation, and human evaluation; and the discussion reflects on the best MT solution for the given content, the reliability of evaluation methods used, and relevant ethical and professional considerations. Human evaluation data are included as an appendix. The project is completed collaboratively, with students dividing tasks and contributing jointly to the written report. Group-based assessment reflects the collaborative nature of MT evaluation work in professional settings, where tasks are typically distributed across team members with different areas of responsibility (Kiraly, 2005).

Assessment criteria Projects are assessed on five dimensions: i) academic content, including the quality of the literature review and theoretical grounding; ii) technical soundness, covering the suitability and correct application of evaluation methods and tools; iii) quality of analysis, including the interpretation and critical discussion of results; iv) structure and argumentation; and v) language quality. Together these criteria reward both technical competence and the ability to communicate findings clearly and critically in an academic register.

Scenario The simulated scenario changes each year, providing students with a realistic context that motivates the evaluation choices they make. In one iteration of the module, the scenario involved translating institutional health reports from Dutch into English for online publication, a domain that combines technical terminology with a requirement for high output quality. The scenario is designed to raise questions that go beyond metrics: is MT suitable for this content at all? Is the raw output publishable, or are further steps needed?

5 MT usage and customisation trends

How do student practices evolve alongside developments in the broader technology landscape? As LLMs are becoming dominant in life and market, are students adopting them more for translation, abandoning traditional NMT engines? This section examines student preferences in engine selection and customisation strategies, drawing on the submitted project reports from all three years of the module.

Engine	2023	2024	2025
Google Translate	6 (32%)	4 (13%)	6 (17%)
DeepL	5 (26%)	7 (23%)	5 (14%)
ModernMT	4 (21%)	5 (17%)	3 (9%)
NLLB	4 (21%)	1 (3%)	1 (3%)
ChatGPT/OpenAI	-	4 (13%)	6 (17%)
Widn	-	1 (3.%)	2 (6%)
Lara	-	-	3 (9%)
Others	-	2 (7%)	-
Total	19	30	35

Table 1: Distribution of MT engine selections by year

Table 1 shows the distribution of generic MT engine selections across the three years of the study. It should be noted that the available engine pool expanded over time. Widn and Lara, for instance, were only introduced in class from 2024 onwards. This means that the trends reflect both changes in the offering and shifts in student preferences. Google Translate and DeepL remained the most consistently chosen engines, with 6, 4, and 6 selections, and 5, 7, and 5 selections respectively. ModernMT was the next popular selection in 2023 and 2024, but declined in 2025. NLLB, the only open-

source NMT engine in the selection, was chosen in 4 projects in 2023 but only in one in subsequent years. This may be due to a better offering of engines for this high-resource language pair. The most notable trend is the rapid adoption of ChatGPT/OpenAI, which was absent in 2023 but rose to 4 selections in 2024 and 6 in 2025, matching Google Translate as the most selected engine. The initial absence is partly explained by a practical limitation: pasting the entire text in the ChatGPT interface did not preserve sentence boundaries, requiring manual realignment before automatic evaluation. This issue was subsequently resolved by accessing the model via the OpenAI API. Newer engines such as Widn and Lara also gained traction in the last year. These trends suggest that while traditional NMT providers are not being abandoned, students are increasingly experimenting with LLM-based solutions, reflecting the shifts in the professional translation technology market.

Method	2023	2024	2025
Glossary	4 (57%)	3 (50%)	7 (64%)
TM	3 (43%)	2 (33%)	2 (18%)
Prompting	-	1 (17%)	2 (18%)
Total	7	6	11

Table 2: Customisation options selected per year

Table 2 shows the customisation methods selected by students across the three years. Glossary-based customisation was the most popular method throughout, accounting for the majority of selections in all three years. Most providers have the option to directly upload a glossary, which students reported to be the easiest method for customisation. TM-based customisation was the second method preferred with 2 to 3 projects selecting it. Prompting is gradually emerging as a customisation strategy, with students including context information about the use case and/or the glossary entries in the translation prompt. This reflects the growing familiarity of students with LLM-based tools and their affordances for domain adaptation. Nevertheless, traditional customisation methods built into the engine interface, such as glossary upload and TM integration, remain preferred, likely because they are more straightforward to apply and easier to evaluate systematically than a custom prompt whose effect on output quality is harder to isolate.

6 Student evaluation and impressions

Student feedback was collected through anonymous voluntary surveys administered at the end of each iteration. The following reflections draw on three years of feedback and the instructor's own observations across the cohorts.

What worked well Students consistently recognised the relevance of the module to their professional development. The technological dimension was valued precisely because of its currency, with students recognising the importance of gaining experience with MT and AI. The variety of tools explored across the module (commercial, open-source, and LLM-based) was also appreciated. The combination of theory and hands-on practice was frequently highlighted as effective, with students appreciating that concepts introduced in the theoretical part were immediately applied in practical activities. The project-based assessment was particularly well received, with students valuing the experience of working on a realistic scenario with assignments for clients in the real world.

Points for improvement The most consistent criticism across all three years concerned the balance between theory and practice. Students frequently reported feeling overwhelmed by the theoretical content, particularly in the early sessions, and requested more time for hands-on activities. In response, the format was adjusted from the second year onwards: slide presentations were shortened, readings were assigned before class rather than presented in full during sessions, and contact time was reserved for practical work and discussion of questions arising from the readings.

The integration of LLM-based engines introduced a practical challenge around costs. Since students required access to LLM APIs for translation tasks, token usage had to be monitored and sometimes limited to avoid excessive costs. In practice, this was managed by pre-generating translations and sharing them with students, avoiding repeated generation of the same output. While functional, this solution reduced the immediacy of the hands-on experience. This remains an open challenge, particularly as LLM-based tools become more central to the module. Two tools were introduced incrementally and only reached their current form in the most recent iteration: translation-specific LLMs (Widn, Lara) were added gradually in the second and third years

as they became available, and the enhanced post-editing tool with explainable QE signals was piloted in the final iteration. Both additions were positively received but have not yet been evaluated across a full cohort in their current form.

Recommendations Based on three years of experience, several recommendations emerge for instructors designing similar modules. First, limiting in-class theory in favour of flipped classroom elements, for example assigning readings and explanatory materials before sessions, frees contact time for the practical work students find most valuable. Second, managing access to commercial and LLM-based tools requires planning: educational licences, pre-generated outputs, and token budgets should be arranged well in advance and are not always available. This is why investing on open-source tools is a valuable alternative. Finally, students expressed interest in hearing from working professionals: a guest lecture by an MT specialist or language technology consultant would complement the academic content and reinforce the professional relevance of the skills developed.

7 Conclusion

This paper has presented a project-based translation technology module that equips MA Translation students with end-to-end MT evaluation skills, integrating LLMs throughout as translation engines, customisation tools, and sources of explainable quality signals. The module has been taught over three consecutive years, allowing for iterative refinement in response to student feedback and developments in the field.

Analysis of student project reports over three consecutive years reveals a gradual shift in engine preferences, with LLM-based tools gaining ground while traditional engines remain in use. In terms of customisation, glossary-based methods continue to dominate due to their ease of implementation and perceived efficiency, though prompting is emerging as an alternative strategy.

Some limitations should be acknowledged. The module has been taught at a single institution with a single language pair (English-Dutch), and no formal measurement of learning outcomes has been conducted beyond student feedback. The evaluation of recently introduced components, such as translation-specific LLMs and the explainable QE tool, remains incomplete. Formal evaluation of the module, including self-efficacy measures and more detailed qualitative and quantitative feedback, will be conducted in future iterations. Furthermore, the module relies heavily on commercial engines and tools, a deliberate choice to maximise relevance to real-world professional practice, made possible through institutional resources and academic agreements with technology providers. However, it is recognised that not all institutions have access to such resources. A related challenge concerns instructor expertise: setting up open-source engines, hosting evaluation tools, or running models such as xCOMET requires technical knowledge that many translation technology instructors may not have. While commercial alternatives for enhanced post-editing exist, e.g. platforms such as Phrase and Bureauworks offer comparable functionality, the development of accessible, open-source post-editing and QE tools remains important for broadening access to this type of instruction (Lefer et al., 2024; Mutal et al., 2020).

More broadly, the challenge facing translator training programmes is not simply to incorporate specific tools, but to develop frameworks flexible enough to absorb continuous technological change without requiring wholesale curriculum redesign. The module presented here attempts to address this by grounding instruction in a stable evaluation pipeline that can accommodate new engines, metrics, and LLM-based features as they emerge. Future work includes the development of an online component to deliver theoretical content outside contact hours, freeing in-class time for the hands-on practice and critical discussion that students find most valuable.

References

Ayvazyan, Nune, Yu Hao, and Anthony Pym. 2024. Things to do in the translation class when technologies change: The case of generative AI. In Peng, Yuhong, Huihui Huang, and Defeng Li, editors, *New Advances in Translation Technology: Applications and Pedagogy*, pages 219–238. Springer, Singapore.

Bowker, Lynne and Jairo Buitrago Ciro. 2019. *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Emerald Group Publishing, Bingley.

Costa-Jussà, Marta R, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Doherty, Stephen and Dorothy Kenny. 2014. The design and evaluation of a statistical machine translation syllabus for translation students. *The Interpreter and Translator Trainer*, 8(2):295–315.

Doherty, Stephen, Dorothy Kenny, and Andy Way. 2012. Taking statistical machine translation to the student translator. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Commercial MT User Program*, San Diego. Association for Machine Translation in the Americas.

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In Pareja-Lora, Antonio, Maria Liakata, and Stefanie Dipper, editors, *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia. Association for Computational Linguistics.

Guerberof Arenas, Ana and Joss Moorkens. 2019. Machine translation and post-editing training as part of a master's programme. *The Journal of Specialized Translation*, 31:217–238.

Guerreiro, Nuno M, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Hackenbuchner, Janiça and Ralph Krüger. 2023. DataLitMT–Teaching data literacy in the context of machine translation literacy. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 285–293, Tampere. European Association for Machine Translation.

Hao, Yu and Anthony Pym. 2023. Choosing effective teaching methods for translation technology classrooms: Teachers' perspectives. *Forum*, 21(2):190–212.

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, editor. 2022. *Machine Translation for Everyone*. Language Science Press, Berlin.

Kiraly, Don. 2005. Project-based learning: A case for situated translation. *Meta*, 50(4):1098–1111.

Koehn, Philipp. 2010. *Statistical Machine Translation*. Cambridge University Press, Cambridge.

Koehn, Philipp. 2020. *Neural Machine Translation*. Cambridge University Press, Cambridge.

Koponen, Maarit. 2015. How to teach machine translation post-editing? Experiences from a post-editing course. In *Proceedings of the 4th Workshop on Post-editing Technology and Practice*.

Krüger, Ralph. 2022. Using Jupyter notebooks as didactic instruments in translation technology teaching. *The Interpreter and Translator Trainer*, 16(4):503–523.

Lefer, Marie-Aude, Romane Bodart, Justine Piette, and Adam Obrusník. 2024. MTPE quality evaluation in translator education: the postedit.me app. In Scarton, Carolina, Charlotte Prescott, Chris Bayliss, Chris Oakley, Joanna Wright, Stuart Wrigley, Xingyi Song, Edward Gow-Smith, Mikel Forcada, and Helena Moniz, editors, *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 23–24, Sheffield. European Association for Machine Translation.

Li, Defeng, Chunling Zhang, and Yuanjian He. 2015. Project-based learning in teaching translation: Students' perceptions. *The Interpreter and Translator Trainer*, 9(1):1–19.

Lommel, Arle, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):455–463.

Macken, Lieve, Bram Vanroy, and Arda Tezcan. 2023. Adapting machine translation education to the neural era: A case study of MT quality assessment. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 305–314, Tampere. European Association for Machine Translation.

Mellinger, Christopher D. 2017. Translators and machine translation: Knowledge and skills gaps in translator pedagogy. *The Interpreter and Translator Trainer*, 11(4):280–293.

Mitchell-Schuitevoerder, Rosemary. 2020. *A project-based approach to translation technology*. Routledge, London/New York.

Moorkens, Joss. 2018. What to expect from neural machine translation: A practical in-class translation evaluation exercise. *The Interpreter and Translator Trainer*, 12(4):375–387.

Moorkens, Joss. 2022. Ethics and machine translation. In Kenny, Dorothy, editor, *Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence*, pages 121–140. Language Science Press, Berlin.

Mutal, Jonathan, Pierrette Bouillon, Perrine Schumacher, and Johanna Gerlach. 2020. COPECO: a collaborative post-editing corpus in pedagogical context. In *Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation*, pages 61–78, Virtual. Association for Machine Translation in the Americas.

O'Brien, Sharon. 2002. Teaching post-editing: A proposal for course content. In *Proceedings of the 6th EAMT workshop: Teaching Machine Translation*, Manchester. European Association for Machine Translation.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In Isabelle, Pierre, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia. Association for Computational Linguistics.

Popović, Maja. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In Bojar, Ondřej, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon. Association for Computational Linguistics.

Pym, Anthony and Yu Hao. 2024. *How to Augment Language Skills: Generative AI and Machine Translation in Language Learning and Translator Training*. Routledge, London/New York.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts. Association for Machine Translation in the Americas.

Tiedemann, Jörg. 2009. News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. In Nicolov, N., K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing V: Selected Papers from RANLP 2007*, pages 237–248. Benjamins, Amsterdam.

Treviso, Marcos V, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan Van Stigt, and André FT Martins. 2024. xTower: A multilingual LLM for explaining and correcting translation errors. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15222–15239, Miami. Association for Computational Linguistics.

Vanroy, Bram, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine Translation Evaluation Online. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 499–500, Tampere. European Association for Machine Translation.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR 2020)*.

Integrating AI-Based Technologies into Translation Workflows through a Simulated Translation Bureau (STB)

Koen Kerremans
Vrije Universiteit Brussel, Brussels,
Belgium
koen.kerremans@vub.be

Abstract

This paper reports on an exploratory, qualitative study of technology use in a Simulated Translation Bureau (STB) in a master's programme in translation. The STB is a practice-oriented module in which students work on authentic translation projects and integrate a range of technologies, including AI-based tools, into their workflows. The paper addresses how students' technology-related decision-making can be made visible and assessable, including students' reflections on how they justify and verify the use of output-generating technologies, and how they describe perceived changes in their technology use over the course of the simulation. The study shows that the STB is an effective pedagogical model for teaching and assessing technology judgement in AI-enabled translation workflows, but also highlights the need to foreground broader ethical dimensions of AI literacy more explicitly in future iterations of the pedagogical design.

1 Introduction

AI-based technologies – from machine translation (MT) integrated into computer-assisted translation (CAT) tools to large language models (LLM)-based assistants – are increasingly embedded in professional translation workflows, as evidenced by recent market-based studies such as ELIS (2026). This evolution creates a concrete challenge for translation training programmes: if professional practice changes, translation education must integrate these technologies in ways that prepare students for professional practice. This challenge is not limited to teaching students how such technologies function at a technical level. It also involves helping them understand how these technologies are integrated in real translation workflows, while also developing the AI literacy required to use them responsibly, critically, and ethically (Krüger, 2024).

A Simulated Translation Bureau (STB) offers a particularly suitable pedagogical setting for such integration as it is explicitly practice-oriented (Buysschaert et al., 2018): student teams work as translation bureaus with different roles, carry out authentic translation projects, and deal with professional constraints such as deadlines and quality requirements (Van Egdom et al., 2020). This structure makes it possible to treat and assess technological tools as part of the workflow, rather than as separate topics taught in class (Zappatore, 2020, 2022).

This paper reports on an exploratory, qualitative study of how technology use becomes visible in an STB. It focuses on the data sources built into the STB module design of a master's programme in translation to monitor how students use technologies in their translation workflows. The starting point in the STB module is that students should have sufficient autonomy to choose and integrate any type of technology (including AI-based tools) in meaningful ways, while also being expected to explain how and why they used it. Importantly, the STB design does not isolate AI-based technologies from the wider set of technologies used in the translation workflow but examines how such technologies are used alongside other tools, such as CAT, project management, or collaboration tools.

Proceedings of the 1st International Workshop on Teaching AI-Based Translation and Technologies (TAITT 2026), pages 63–70
June 15, 2026 ©2026 Association for Computational Linguistics

In this respect, the study addresses three questions: How does a translation bureau simulation make students' technology-related decision-making visible and assessable (RQ1)? Which practices do students report and demonstrate when integrating AI-based technologies into translation workflows, and how do they justify and verify this use (RQ2)? To what extent do students report changes in their technology use during the simulation compared to prior practice (RQ3)?

In this study, technology-related decision-making is considered visible when students explicitly report where in the workflow a technology was used, for what purpose, and with what outcome. It is considered assessable when these reports provide evidence to evaluate how transparently students describe their technology use and justify their choices, and how far they demonstrate creative or innovative uses.

This assessability was also supported by the course design itself. Students were informed at the start of the STB module that technology use formed one of the dimensions on which group performance would be evaluated, and they received a rubric specifying how this aspect would be assessed. The rubric distinguished between limited use of technology, guided use, more creative integration into workflows, and highly proactive, innovative, and critically justified use of technology. This rubric clarifies the pedagogical criteria against which students' technology-related decision-making became assessable.

2 Background and pedagogical design

Research on translator training has increasingly treated technology as an integral part of contemporary translation practice, while also emphasising that students need to learn not just how to use tools, but when, why, and in which contexts they are best applied (Bowker, 2014; Kenny, 2020). Recent discussions on AI literacy in translation extend this argument by stressing that users need not only operational technological competence, but also the ability to evaluate, justify, and monitor the responsible use of AI-based tools (Bindels et al., 2024; Krüger, 2024).

Within this broader context, the STB model is especially relevant because it links technology learning to authentic professional workflows (Zappatore, 2022). Earlier work on simulated translation bureaus has shown their value for immersing students in realistic projects, external collaborations, entrepreneurship, and service provision (Buysschaert et al., 2018; Van Egdom et al., 2020; Konttinen & Salmi, 2025). The present study builds on that tradition by examining how students use technologies and report on their technology-related decisions.

The argument advanced here is that the use of technology in STB training should be studied at two levels. At the individual level, students need to be able to explain how and why they use particular technologies in specific tasks. At the group level, bureaus also need to decide collectively how they organise work, which tools they use, how they handle planning and communication, etc. The STB is therefore not only a teaching environment but also a useful research site for observing how technology judgement is negotiated in a practice-oriented setting.

3 Research design

3.1 Course context and participants

The study was conducted as part of a mandatory STB module in a master's programme in translation. Student teams worked in small simulated bureaus as either project managers or translators/revisers and carried out authentic projects for real and/or simulated clients. The module was explicitly designed to simulate professional practice by combining project management, client communication, translation, revision, external collaborations, and technology use. CAT tools were part of the working environment from the outset, but students retained meaningful autonomy in deciding how to integrate technologies into their workflows.

The present paper analyses a first-cycle corpus consisting of 17 artefacts produced by three simulated bureaus: 11 individual portfolios, three group technology reflection notes, and three final group presentations. The 11 portfolios were written by students performing different roles, including three project managers and eight translator/reviser profiles. The STB module was taught in the second semester, after a first-semester course on technologies for translators, and students had already gained insight into the technical dimensions of language and AI during their prior BA-level training. Within the STB, they were explicitly allowed to use AI-based technologies, provided this use was documented transparently in their reporting, in line with the university's broader policy on the responsible and transparent use of AI in edu-

cational contexts. To ensure anonymity, individual portfolios are cited as ID1–ID11 and the three bureaus are referred to as Bureau A, B, and C.

3.2 Data collection

Data collection was embedded in the mandatory course assignments to capture students' technology use in a naturalistic, practice-oriented way, without explicitly framing the course around AI-based technologies. This was important because the study sought to examine how students use AI-based tools alongside other technologies in the workflow, such as CAT, MT, terminology, project management, and collaboration tools. As noted above in the discussion of visibility and assessability (see Section 1), students were informed at the start of the STB module that group performance would be evaluated in part through specific indicators, including one focused on technology use, and that this aspect would be assessed using an explicit rubric (see Appendix 4). This was intended to encourage them to take an active role in identifying technological solutions and integrating them into their projects, while still leaving them enough autonomy to decide what to incorporate and how to incorporate it.

At group level, each bureau submitted a brief technology reflection note of approx. 500–800 words (See Appendix 2). In this document, groups had to report at least three concrete examples of technology use in a fixed format including: a description of the specific tool that was used for a specific purpose, task or process, a reflection on the outcome or effect of its use and a link to supporting evidence. If the example involved an AI generating output, groups were additionally required to indicate how the output was verified. These notes provide direct evidence for RQ1 by showing where in the workflow technologies were used and for RQ2 by making verification explicit.

A second group-level data source was the final presentation (see Appendix 3), in which each bureau reflected on projects, challenges, SWOT factors, technology choices, and lessons learned. Although this presentation partly overlapped with the separate technology reflection note, it served a distinct pedagogical and analytical purpose. Whereas the written reflection note required groups to document technology use in a structured way, the presentation invited them to synthesise these experiences for an audience of peers, allowing other bureaus to learn from their choices, challenges, and strategies. The oral format therefore added an additional pedagogical and research dimension to the written reflection by showing how groups collectively framed their workflow, justified their technology use, and highlighted what they considered most relevant.

At individual level, each student submitted a portfolio describing their involvement in projects, their specific role(s), learning reflections, and a separate section on technology use. This individual perspective was important because it added a more fine-grained and role-specific dimension to the two group-level sources. Whereas the technology reflection note and the presentation mainly captured group-level technology practices and shared strategies, the individual portfolios made it possible to examine how students themselves experienced and interpreted these practices from their own position within the bureau (see Appendix 1). They were therefore particularly valuable for identifying differences between project managers and translators/revisers. Students were also explicitly asked to reflect on whether and how their technology use had changed during the STB module compared to before the module and, if relevant, to provide examples of output-generating technology use. This makes the individual portfolios especially important for RQ3, which concerns perceived change, while also providing role-specific evidence for RQ1 and RQ2.

Taken together, the three types of data sources provide a basis for triangulating reported practices, verification strategies, and perceived changes in technology use across roles and workflow stages.

3.3 Data analysis

The analysis was based on qualitative content analysis of the collected group- and individual-level materials, combining deductive and inductive thematic procedures. At a deductive level, it was organised around three main categories derived from the research questions. In relation to RQ1, the analysis will identify patterns of technology use across workflow stages. For RQ2, the focus was on how students report checking and validating AI technology use. In relation to RQ3, the analysis examined students' reported changes in technology use over the course of the simulation, based on individual reflections and supporting examples. At an inductive level, recurrent themes were identified within each category through close reading of the data sources.

Meaningful reflection segments were used as coding units. After the initial coding, the group- and individual-level materials were compared in order to identify convergences and divergences between collective and personal reporting on technology use.

4 Results

4.1 Making technology-related decision-making visible and assessable (RQ1)

In the individual portfolios, translators and revisers most often connect technologies to translation, terminology, revision, and quality control, while project managers connect them to planning, communication, task allocation, and document management. These materials show that the STB does not merely reveal what technologies students use but also how they locate them within their specific workflows. This also appears clearly in the group materials. In Bureau A's presentation, problems are categorised as: organisation efficiency, technology/workflow, and communication; the accompanying technology slide refers to the use of a CAT tool, language supporting tools (e.g. electronic dictionaries, translation memories, spelling checkers) as distinct resources within the workflow. Bureau B explicitly separates a specific tool for communication and collaboration (e.g. a shared agenda for planning), a CAT tool for translation and revision, and a MT tool for translation support. Bureau C likewise connects technology to their workflow, linking a collaboration and communication tool, a project management tool, a CAT tool and a MT tool to successive phases in their workflow, such as task distribution, translation, and revision.

Assessability, however, depends on the degree of structured documentation. The strongest portfolios and reflection notes make decision-making assessable by reporting a tool, a purpose, an outcome, and a quality-control procedure. ID5, ID8 and ID9, for instance, explicitly describe what a tool produced and what they did with that output afterwards. By contrast, some documents mention technology in more general terms ("it was efficient", "it helped us") without further specifications on the effect of the outcome and how the output was verified. The group reflection notes tend to be more consistent than some of the individual portfolios because they use a more standardised reporting template. Taken together, the materials suggest that the STB makes students' technology-related decisions easier to assess when the reflection task clearly asks them to explain why they used a tool, what effect it had, and how they controlled or checked its use. When the task instructions do not follow such a strict reporting format, it also becomes harder to assess those decisions.

4.2 Practices of integrating and verifying AI-based technologies (RQ2)

The group- and individual-level data consistently show that AI-based technologies are used as part of a broader technological workflow. Most reported uses concern CAT-integrated MT, AI-based writing assistants, and, in one bureau, the use of a generative AI chatbot for suggesting a company name and logo rather than for translation itself. Apart from that, the use of a standalone generative AI chatbot for translation drafting is almost absent from the research materials. Bureau B reports having tested MT in the CAT tool they used but having rejected it because the quality was insufficient; Bureau C states that MT suggestions were only used in one project and were always revised; Bureau A also highlights CAT-integrated MT, but also foregrounds electronic dictionaries, language assistants and spell-checking tools in its workflow. The overall pattern is therefore one of AI-based tools embedded within a larger set of technological tools.

A second pattern is that output-generating tools are mostly treated as supportive rather than substitutive. Students typically describe using them to generate a first draft, explore alternative formulations, find lexical options, or speed up (repetitive) tasks. In their individual portfolios, ID8 reports using MT integrated in a CAT tool to support the translation work but then switching off the MT function when post-editing became less efficient than translating independently, especially for the specific project on which they had been working (i.e. the translation of historically sensitive farewell letters written by people during the Second World War shortly before their execution). ID9 describes using MT as a starting point for translating a specialised text on biodigesters, but adapting the output almost systematically, especially terminological suggestions. ID6 similarly describes consulting the output of a built-in MT feature in a CAT tool mainly to support the translation of difficult grammatical constructions or single lexical items rather than adopting full segments. At the group level, Bureau B describes MT use primarily as a way of finding more idiomatic alternatives and states explicitly that it was sometimes used

only when no other satisfactory formulation could be found. These accounts suggest a pragmatic use of AI-based tools rather than any attempt to outsource translation completely.

Verification emerges as the clearest and most consistent marker of responsible AI-related practice. Students repeatedly describe post-editing, comparing generated output with the source text, consulting dictionaries, checking terminology via IATE, EurLex, or parallel texts, relying on peer revision, and carrying out final quality checks. Bureau A's technology reflection note, for instance, explicitly states that MT suggestions in their CAT tool were treated only as "a first idea or example", after which the text was post-edited, revised several times by the translator, and then revised either internally or by an external partner. Bureau C's presentation similarly stresses that MT output was never adopted without further scrutiny, but checked against the source text, dictionaries, and peer revision. The individual portfolios confirm the same pattern. ID9, for example, reports that MT output for a specialised article was verified against the source text and professional reference sources, while suggestions from an AI-based writing assistant were only accepted after checking tone and coherence in the full text. Taken together, these findings suggest that students did not place unconditional trust in AI-based tools, but subjected their output to explicit verification.

At the same time, the data shows that students' understanding of responsible use is still mostly operational. They focus on revision, source checking, and not accepting output blindly. Much less is said about privacy, bias, intellectual property, or environmental impact. One project manager (ID1) states that technologies should be used in the "right, ethical and efficient way", but such broader ethical reflection remains relatively rare. The present materials therefore capture verification and transparency very well, but only partially capture broader ethical AI literacy.

4.3 Reported change in technology use (RQ3)

The clearest pattern in the individual portfolios is a reported shift from reluctance towards CAT tools to greater familiarity and acceptance. Several students state that before the STB they preferred to work in Word, used CAT tools only when required, or perceived them as too complex. Through repeated use in authentic projects, they became more aware of the practical benefits of CAT tools for workflow support and collaboration. ID8, for example, reports a fundamental change in attitude after working with the chosen CAT tool, which came to be seen as a useful and supportive part of the translation process rather than as an unnecessarily complex tool. ID6 similarly describes a better understanding of why CAT tools are efficient and professionally relevant. ID3 emphasises their value for the translation workflow, for monitoring deadlines, and for maintaining a clear overview of translation projects within the team. This shift is also reflected at group level: Bureau B lists learning to work with CAT tools among its strengths, while Bureau C presents its CAT environment as useful not only for translation itself, but also for organising the workflow.

Reported change is not limited to the tools students use, but also concerns how these tools help students to organise the translation process itself. Several students describe a move away from a more linear way of translating a text towards a more structured workflow divided into phases such as project setup, drafting, revision or quality control. Others emphasise that they now see technologies as integral components of the overall process. This is particularly visible in the portfolios of the project managers. ID1 and ID4, for instance, describe learning to use communication and planning tools, shared agendas, and CAT tools more purposefully in order to monitor deadlines, distribute work, and maintain a clear overview of tasks in their projects. Bureau C's presentation illustrates this shift clearly: the group explains that it initially worked in a largely ad hoc manner, but gradually developed a more structured workflow supported by several dedicated technologies, including the selective use of MT. The STB thus appears to foster change not only in tool use, but also in how students conceptualise workflow, coordination, and the place of technology, including AI-based tools, within professional translation practice.

A third pattern is that students report using MT- and AI-supported tools in a selective and cautious way. As discussed in Section 4.2, generated output is mainly used as support, for example as inspiration or a first suggestion, and rejected when correcting it takes more time than translating directly. This pattern also appears at group level. Overall, the data point less to a broad enthusiasm for AI as such than to a more selective and workflow-dependent use of these technologies.

5 Discussion

Taken together, the findings show that the STB makes it possible to analyse technology use at both individual and group level. The individual portfolios provide insight into how students describe and justify their own technology-related decisions, whereas the reflection notes and presentations show how each bureau organised its workflow, distributed roles, and combined different tools. This matters because the findings show that technology use in the STB depends both on students' own choices and on how the group organises its workflow and divides roles. Project managers mainly report technology use in relation to planning, communication, deadline monitoring, and task distribution, while translators/revisers focus more on translation, revision, terminology work, and quality control.

This broader role- and workflow-based view of technology use is in line with earlier work on simulated translation bureaus and translator training, which stresses that technology is embedded in collaborative, practice-oriented environments rather than in isolated tool exercises (Buysschaert et al., 2018; Van Egdom et al., 2020).

Decision-making became most visible when students were required to specify where in the workflow a technology was used, for what purpose, what resulted from that use, and, in the case of AI-generated output, how the result was checked. In this respect, the richest evidence came from those data sources that followed a fixed reporting structure. In such cases, technology use could be analysed in a more systematic way. By contrast, where reporting was more open or less systematically structured, statements about technology use often remained broad.

AI-based tools were typically used in a supportive way, alongside CAT tools, dictionaries, terminology tools/resources, project management tools, and collaboration platforms. The most common pattern was to use AI-generated output for inspiration, a first draft, or lexical support, followed by revision, com-parison with the source text, and other forms of verification. This means that, in the present dataset, AI use appears less as a disruptive force than as one resource within a broader workflow in which people still play the main role in checking quality and making decisions. This aligns with work on MT literacy, which emphasises that students and translators need to understand when and how MT can best be used, and that human intervention remains central to translation work (Kenny, 2020).

At the same time, the findings suggest that the current design captures some dimensions of AI literacy more clearly than others. Students frequently describe selective uptake, post-editing, terminology checking, source comparison, and peer revision. In other words, they provide evidence of operationally responsible use. By contrast, they much less often reflect on broader ethical issues such as confidentiality, bias, intellectual property, or environmental impact. What emerges from the present design is therefore mainly a picture of responsible use in terms of verification and quality control. A useful next step would be to retain the current workflow-oriented setup, but to add one or two explicit prompts that invite students to reflect on risks, limitations, or ethical considerations when using AI-based tools in their projects. This distinction is also consistent with recent work on AI literacy and translation technology ethics, which argues that responsible use involves not only functional and evaluative skills, but also attention to wider ethical and professional implications (see e.g. Moorkens, 2022).

The study also has some limitations. First, it is based on one course run and a relatively small dataset. Second, the data come from assessed course assignments, so they are partly self-reported. However, combining individual and group materials helps reduce this limitation by allowing personal reflections to be compared with group reflections and concrete examples. The paper therefore does not claim to prove that students developed these competences because of the STB design. Instead, it shows that the design makes important aspects of technology-related learning visible, including decision-making, the checking of AI- and MT-supported output, and reported changes in students' use of translation technologies. Further research is needed to examine more directly to what extent the STB supports the development of these competences over time.

6 Conclusion

This study has shown that the STB makes technology-related decision-making visible through workflow-specific reporting at both the individual and group levels. It has also shown that AI-based technologies are mainly used as supportive tools within a broader workflow, and that students mainly report change in three areas: greater familiarity with CAT tools, a more explicit

integration of technology into the workflow, and a more critical and selective use of MT- and AI-supported tools.

The paper's broader contribution is both pedagogical and methodological. Pedagogically, it suggests that a practice-oriented STB offers a useful setting for teaching and assessing technology judgement in AI-enabled translation workflows. Methodologically, it shows the value of combining individual and group-level data in order to capture both personal experiences and shared ways of organising and reflecting on technology use. Future research can build on this first cycle by expanding the dataset and refining the task instructions so that broader ethical dimensions of AI literacy are addressed more explicitly, alongside workflow judgement and quality control.

References

Bindels, Joop, Aletta G. Dorst, and Mark Pluymaekers. 2024. Machine translation in the workplace: Deciding on the whether, why and how. *Global Advances in Business Communication*, 11(1):1–15.

Bowker, Lynne. 2014. Computer-aided translation: Translator training. In Chan Sin-wai, editor, *Routledge Encyclopedia of Translation Technology*, pages 88–104. Routledge, London/New York.

Buysschaert, Joost, Maria Fernández-Parra, Koen Kerremans, Maarit Koponen, and Gys-Walt van Egdom. 2018. Embracing digital disruption in translator training: Technology Immersion in Simulated Translation Bureaus. *Revista Tradumàtica. Tecnologies de La Traducció*, 16, 125–133.

ELIS Research. 2026. *European Language Industry Survey 2026. Trends, expectations and concerns of the European language industry.*

Kenny, Dorothy. 2020. Technology in translator training. In O'Hagan, Minako, editor, *The Routledge Handbook of Translation and Technology*, pages 498–515. Routledge, London/New York.

Konttinen, Kale and Leena Salmi. 2025. Simulated Translation Bureaus. In Walker, Callum and Joseph Lambert, editors, *The Routledge Handbook of the Translation Industry*, pages 511–525. Routledge, London/New York.

Krüger, Ralph. 2024. Outline of an artificial intelligence literacy framework for translation, interpreting and specialised communication. *Lublin Studies in Modern Languages and Literature*, 48(3):11–23.

Moorkens, Joss .2022. Ethics and machine translation. In Kenny, Dorothy, editor, *Machine Translation for Everyone. Empowering Users in the Age of Artifical Intelligence*, pages 121–140. Language Science Press, Berlin.

Van Egdom, Gys-Walt, Kale Konttinen, Sonia Vandepitte, Maria Fernández-Parra, Rudy Loock, Rudy, and Joop Bindels. 2020. Empowering translators through entrepreneurship in Simulated Translation Bureaus. *HERMES - Journal of Language and Communication in Business*, 60:81–95.

Zappatore, Marco. 2020. Lifecyle design and effectiveness evaluation for simulated translation agencies. *Current Trends in Translation Teaching and Learning E*, *7*:118–166.

Zappatore, Marco. 2022. A design methodology for a computer-supported collaborative skills lab in technical translation teaching. *Journal of Information Technology Education: Research*, 21:137–167.

Appendix 1. Instructions for the individual technology reflection

Reflect on your use of technology within the STB. Technology is understood broadly here and may include CAT/MT, terminology and QA tools, corpora and search tools, collaboration tools, project management tools, automation, writing or language assistants, and similar tools.

A) Change in technology use through the STB

Answer the following question explicitly: Do you use technology differently since participating in the STB than you did before?

- If yes, explain what has changed and why. For example, you may refer to: different tools, a different way of working, more or less trust in certain technologies, different quality-control practices, different ways of collaborating or planning. Illustrate your answer with at least one concrete example (for example a screenshot, a workflow step, or a short description of a procedure).
- If no, explain why your use of technology has not changed.

B) Use of output-generating technology (only if applicable)

If you used AI-technology that generates output (for example translations, reformulations, summaries, or terminology proposals), report a maximum of two different concrete examples using the following format:

- Tool(s): [name of tool]
- Category (choose one): CAT/MT – terminology/QA – corpus/search tool – collaboration/project management – automation – writing/language assistant – other

- Task/purpose (what was it used for?):
- Input/instruction (e.g. prompt):
- Output (brief excerpt or short summary in 2–3 sentences):
- What did you do with the output? (accept / adapt / reject + explain why)
- Quality control (how did you check it?)

Appendix 2. Instructions for the group technology reflection note

Submit a technology reflection note of 500–800 words in which you explain how technology was used within your bureau in a creative and well-considered way. The note must include at least three concrete examples of technology use. For each example, use the following reporting format (maximum 5–8 lines per example):

- Task/process step: (e.g. planning, terminology work, translation, revision, quality assurance, communication, delivery)
- Tool/technology: [name of tool]
- Purpose: (e.g. speed/efficiency, quality, consistency, collaboration, risk management)
- What came out of it / effect: (briefly)
- Quality control: (how did you check this?)
- (If the tool generated output, also indicate briefly what input/instruction you gave and what you did with the output.)
- Evidence: [link to a concrete file or screenshot]

Appendix 3. Instructions for the group technology reflection in the group presentation

Include at least one slide explaining: which technologies were used and what they were used for. If AI-based technologies that generate output were used (for example suggestions or reformulations), include one short example showing:

- tool,
- task,
- input (briefly),
- output (briefly),
- what was done with it,
- how it was checked.

Appendix 4. Technology adoption assessment rubric

The following rubric was used as part of the group assessment to evaluate how student bureaus integrated technology (including AI-based technologies) into their workflow in a deliberate, creative, and critically justified way.

- Insufficient: The bureau uses technology only to a limited extent or without a clear approach. Students do not follow a consistent workflow and show little initiative. They use only whatever happens to be available. There is no visible creativity in the way technology is used. Generative AI is either not used at all or used in a random and unstructured way.
- Basic: The bureau uses the technologies introduced in the course in a correct way. These technologies support the basic functioning of the bureau, for example in translation, communication, and planning, but students mainly follow instructions. There is little initiative or creative application.
- Good: The bureau makes deliberate choices in its use of technology and adapts tools to its own way of working. Students show initiative and use technology creatively to improve collaboration and organisation.
- Excellent: The bureau actively looks for new tools or applications and uses technology in a creative and well-considered way. Students experiment and make conscious decisions about what works best for their bureau. Generative AI, if used, is employed critically and responsibly.

Teaching Data Management to Translation Students: From Documentation Practices to Data Literacy

Pilar Sánchez-Gijón
Universitat Autònoma de Barcelona
Spain
pilar.sanchez.gijon@uab.cat

Abstract

This paper proposes an approach to teaching data management in translation programmes, framing it as an extension of established documentation practices. It argues that data governance, sourcing, and processing are core competences for translators working with NMT and LLMs, and outlines a training framework that supports responsible data reuse, quality optimisation, and professional agency.

1 Introduction

The rapid integration of generative AI, neural machine translation (NMT), and large language models (LLMs) into translation workflows has repositioned linguistic data from a secondary by-product of translation projects to a strategic asset that directly affects translation quality, system performance, and ethical compliance. In NMT- and LLM-based workflows, system behaviour and output quality are critically conditioned by the data used for configuration, adaptation, and reuse, making data management a prerequisite for informed and responsible use of machine translation technologies in both professional practice and translator training. As a consequence, data management has emerged as a core competence for future translators, yet it remains underrepresented and insufficiently systematised in translator training (Bowker, 2026; DePalma & Lommel, 2025; Mellinger & Hanson, 2025; Sánchez-Gijón & Torres Simón, forthcoming).

This shift is explicitly reflected in recent competence frameworks. Data management is identified as a transversal competence in the LT-LiDER Language Technology Map, which outlines the skills required for participation in language-centric AI workflows (Secară et al., 2024). This recognition marks a move away from tool-centred training towards competence-based profiles, in which the ability to understand how data are created, described, governed, and reused becomes central to professional agency. While currently in the pilot phase, the proposed framework is robustly grounded in FAIR data principles (Findable, Accessible, Interoperable, and Reusable) and established documentation standards, ensuring its alignment with both academic rigor and industry requirements for data readiness.

2 Industry drivers: data readiness and data spaces

The growing relevance of data management is further reinforced by industry developments. According to the GALA Business Barometer Report (2025), only around 30% of business-generated language data are currently suitable for reuse in AI-driven workflows, revealing a substantial gap between data availability and data readiness (GALA, 2025). In this context, data readiness is defined as the state in which linguistic resources are structurally, qualitatively, and legally prepared for immediate integration into AI-driven workflows. This concept is operationalised through three primary dimensions: technical quality (including cleaning, alignment, and proper formatting), linguistic adequacy (ensuring domain relevance, terminological consistency, and genre conformity), and legal-ethical compliance (encompassing clear licensing, traceable provenance, and data privacy).

Proceedings of the 1st International Workshop on Teaching AI-Based Translation and Technologies (TAITT 2026), pages 71–75
June 15, 2026 ©2026 Association for Computational Linguistics

At the same time, large-scale initiatives such as the European Language Data Space (https://language-data-space.ec.europa.eu/index_en) foreground data governance, interoperability, licensing, and traceability as foundational elements of the emerging language technology ecosystem. These developments signal a shift towards regulated and value-driven data reuse, in which language professionals are expected to interact critically with datasets rather than merely consume technology outputs (Secară et al., 2024).

3 Teaching data management in translation programmes: a proposal

This paper argues that data management should be explicitly taught in translation programmes, not as a form of technical specialisation, but as a natural extension of documentation practices already familiar to translation students. Translators have long worked with parallel texts, corpora, translation memories, and termbases to ensure adequacy, genre conformity, and functional fitness (Bowker, 2026). From a Translation Studies perspective, data management can therefore be framed as an expansion of documentary and corpus competence adapted to AI-mediated reuse.

In line with Bowker's proposal to teach data through a science-communication approach, abstract notions such as data governance, sourcing, and processing are introduced using familiar concepts, analogies, and visualisation techniques. This approach helps non-technical audiences engage with data-centric concepts while reinforcing the translator's role as an expert decision-maker rather than a passive user of AI (i.e. NMT or LLM) systems.

3.1 Linking data management to Translation Studies theory

The proposed framework explicitly connects data management to established Translation Studies concepts, particularly functionalist approaches and corpus-based translation pedagogy. From this perspective, decisions about data selection, reuse, and annotation are closely aligned with core notions such as *skopos*, communicative purpose, genre conventions, register, and audience expectations (Nord, 2013). Linguistic data are thus understood not as neutral or interchangeable inputs, but as situated resources whose value depends on their adequacy for a specific translation task and context (Sánchez-Gijón & Torres Simón, forthcoming).

This view builds on traditional documentary practices in translator training, where parallel texts, corpora, translation memories, and terminological resources are used to observe how specialised discourse communities communicate and to model target-language production. In AI-mediated translation workflows, these same resources increasingly function as datasets used to customise and optimise NMT systems and LLMs. Framing data management within Translation Studies theory therefore allows students to recognise continuity between long-standing documentation practices and contemporary data-driven technologies, rather than perceiving AI as a conceptual rupture.

In line with Bowker's work (2026) on data literacy for translators, this framework also draws on a science-communication approach to make data-centric concepts accessible to non-technical audiences. By linking abstract notions such as governance, sourcing, and processing to familiar translational decisions, students are encouraged to reflect critically on how theoretical principles inform practical data management choices. This integration reinforces the translator's role as an informed decision-maker who applies theoretical knowledge to the management and reuse of linguistic data in AI-supported translation environments (Bowker, ibid.).

3.2 Core aspects to be taught

This proposal conceptualises data management in translation education as a structured set of three core dimensions that together support the responsible, effective, and sustainable use of linguistic data throughout the translation process. These dimensions reflect key decision-making areas that translators already engage with implicitly, and which require explicit articulation in AI-mediated workflows:

Data governance refers to how linguistic resources are documented, versioned, traced, and made reusable over time. It includes decisions about transparency, responsibility, access conditions, and long-term maintenance of translation data. By addressing transparency and responsibility, this dimension directly operationalises ethical compliance, ensuring that the management of linguistic assets meets both legal standards and professional moral obligations.

Data sourcing concerns the identification and selection of appropriate linguistic resources, taking

into account provenance, relevance, licensing conditions, representativeness, and alignment with the translation brief.

Data processing involves the preparation of selected resources for use, including cleaning, filtering, normalisation, annotation, and other transformations required to ensure that data are fit for a specific translation purpose or for optimising NMT and LLM performance.

From a pedagogical perspective, these aspects can be introduced progressively within translation or specialised translation courses. They can be aligned with familiar stages of the translation process: sourcing decisions are addressed when analysing the brief and selecting documentation; processing decisions are explored during the preparation of translation resources; and governance issues are foregrounded at delivery and closure stages.

Crucially, these dimensions must be considered both at the start and at the end of a translation project. Initial decisions shape how data are assembled and optimised to support the translation task, while closing-phase decisions ensure traceability, ethical compliance, and future reuse. Teaching this lifecycle-oriented perspective reinforces data management as an integral component of professional project preparation and closure.

3.3 Data processing for translation purposes

In the context of AI-mediated translation workflows, data processing constitutes a pivotal set of activities that directly mediate between the translation brief and the behaviour of NMT systems and LLMs. Drawing on Krüger's characterisation of data processing (2024), this section conceptualises data processing not as a neutral technical pipeline, but as a series of context-dependent decisions embedded in professional translation work.

Datasets serve as 'instructional material' to fine-tune NMT and LLM systems. By curating high-quality examples, students 'teach' models to follow specific terminology or styles. This is essentially a documentary task where students select and clean resources to optimize system performance for a given task. From a translation perspective, data processing tasks take place primarily during the project preparation phase, when linguistic resources are identified, assessed, and transformed to ensure alignment with the communicative purpose of the assignment. Typical processing activities include cleaning and filtering translation memories, selecting or discarding segments based on relevance and quality, normalising terminology and style, deduplicating heterogeneous content, and annotating data with minimal but meaningful metadata. Basic annotation entails adding essential metadata (such as domain, genre, context or client labels) to help NMT and LLM systems prioritize relevant linguistic patterns and adapt to specific contextual requirements. These decisions directly influence how NMT systems and LLMs prioritise patterns, handle terminology, and adapt to genre, register, or client-specific constraints once deployed.

Importantly, data processing practices in translation mirror long-standing quality-oriented activities such as revision, validation, and documentary selection, but acquire renewed significance in AI-driven contexts due to the increased sensitivity of models to noisy, heterogeneous, or poorly documented data. Framing processing as a translation-relevant activity enables students to understand how preparatory data work conditions translation outcomes before generation takes place, rather than treating quality issues solely as post-editing problems.

Pedagogically, data processing can be addressed within integrated translation and specialised translation courses that cover the full project lifecycle. By embedding processing decisions alongside brief analysis, documentation, and delivery tasks, students are encouraged to reflect on how theory-informed choices about data preparation contribute to fit-for-purpose MT and LLM performance. This approach reinforces the translator's role as a strategic agent who actively configures linguistic resources to achieve adequacy, consistency, and contextual appropriateness.

3.4 Proposal for a data management training framework

Rather than isolating data-related skills in stand-alone technology courses, the framework is conceived as a transversal structure that can be embedded across translation and specialised translation modules, mirroring the full lifecycle of translation projects. This orientation is also consistent with current discussions around the revision of the European Master's in Translation (EMT) Competence Framework, which increasingly foregrounds data-related competences, including the management, documentation, and responsible reuse of linguistic data in technology-enhanced translation workflows.

The framework is organised around three complementary moments. At the project initiation stage, students learn to assess the translation brief and to identify, source, and preliminarily evaluate linguistic resources with explicit attention to provenance, relevance, and licensing. During the project preparation and execution stage, students engage in data processing activities, such as filtering translation memories, compiling or refining small corpora, and performing basic annotation, to support fit-for-purpose NMT or LLM use. Finally, at the project closure stage, students are explicitly trained to consolidate, document, and govern the linguistic data that have been used or produced, ensuring traceability, ethical compliance, and conditions for future reuse.

Within this framework, two examples of viable pedagogical activities illustrate how data management competences can be operationalised and will be presented and discussed during the workshop presentation:

1. **From project initiation to delivery**: students carry out a complete translation project, from analysing the brief and sourcing data to processing selected resources and delivering the translation with NMT or LLM support. Throughout the project, students document data-related decisions and link them systematically to observed effects on output quality, adequacy, and consistency.
2. **Project closure and documentation reuse**: after project completion, students focus explicitly on the closing phase. They consolidate validated resources, enrich them with minimal metadata (e.g. domain, genre, client, revision status), define reuse or restriction conditions, and reflect on ethical implications.

This strategic approach also integrates broader critical concerns, including the ethical and social implications of data provenance and the ecological sustainability of data-intensive translation workflows. The aim of this framework is not to train translators as data engineers, but to enable them to conceptualise, justify, and document data management decisions as an integral part of translation expertise and professional practice in AI-mediated environments. This activity foregrounds sustainability, efficiency, and responsibility in data reuse (Moorkens, 2022; Weber, 2021).

4 Final remarks

Training in data management represents a significant professional opportunity for translators in an increasingly data-driven language industry, particularly in contexts where NMT and LLMs are integrated into everyday translation workflows. As the performance, reliability, and limitations of these technologies are largely conditioned by the data on which they rely, data management competences enable translators not only to use MT systems, but also to understand, configure, and critically assess them. By developing competences in the governance, processing, and reuse of linguistic data, translators can move beyond the role of end-users of machine translation technologies and position themselves as knowledgeable managers of linguistic assets who actively shape AI-mediated workflows. More broadly, data management training empowers translators to exert greater control over their own production and over the linguistic resources generated through translation work. This competence supports more efficient project preparation, facilitates the responsible reuse of validated documentation, and enhances long-term value creation from translation data. In this sense, data management is not merely a technical add-on, but a strategic skill that strengthens the translator's professional agency and relevance in AI-mediated translation environments.

5 Acknowledgements

The LT-LiDER (ref. 2023-1-ES01-KA220-HED-000161531) has been funded with support from the European Commission. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

Bowker, Lynne. 2026. Teaching translation students about data in the age of generative AI. In Penet, J. C., Joss Moorkens, and Masaru Yamada, editors, *Teaching Translation in the Age of Generative AI: New Paradigm, New Learning?*, pages 67–86. Language Science Press, Berlin.

DePalma, Donald A. and Arle Lommel, 2025. Transforming translation: The evolution and impact of AI on language transfer and communication. In Sun, Sanjun, Kanglong Liu, and Riccardo Moratto, editors, *Translation Studies in the Age of Artificial Intelligence*, pages 18–41. Routledge, London.

GALA. 2025. *Business Barometer Report – Technology, AI, and Automation. Globalization and Localization Association*.

Krüger, Ralph. 2024. Some reflections on the interface between professional machine translation literacy and data literacy. *Journal of Data Mining and Digital Humanities*. Special Issue: Towards a Robotics of Translation, 1–11.

Mellinger, Christopher D. and Thomas A. Hanson. 2025. Data sources and issues of optimisation in translation and interpreting technologies. In Baumgarten, Stefan and Michael Tieber, editors, *The Routledge Handbook of Translation Technology and Society*, pages 300–312. Routledge, London.

Moorkens, Joss. 2022. Ethics and machine translation. In Kenny, Dorothy, editor, *Machine Translation for Everyone. Empowering Users in the Age of Artificial Intelligence*, pages 121–138. Language Science Press, Berlin.

Nord, Christiane. 2013. Functionalism in translation studies. In Millán, Carmen and Francesca Bartrina, editors, *The Routledge Handbook of Translation Studies*, pages 201–212. Routledge, London.

Secară, Alina, María Isabel Rivas Ginel, Antonio Toral, Ana Guerberof Arenas, Dragoș Ciobanu, Justus Brockmann, Claudia Plieseis, Raluca-Maria Chereji, and Caroline Rossi. 2024. *LT-LiDER Language Technology Map – Technologies in Translation Practice and their Impact on the Skills Needed*.

Sánchez-Gijón, Pilar and Ester Torres-Simon, forthcoming. Data management planning. In Castilho, Sheila, Dorothy Kenny, Joss Moorkens, and María Isabel Rivas Ginel, editors, *Developing AI Literacy for Translation*. Language Science Press, Berlin.

Weber, Tobias. 2021. The curation of language data as a distinct academic activity: A call to action for researchers, educators, funders, and policymakers. *Journal of Open Humanities Data*, 7:1–10.

Facilitating Interaction-Oriented AI Literacy in Translator Training: A Process-Oriented Approach

Erik Angelone
Institute of Translation and Multilingual Communication
TH Köln – University of Applied Sciences Cologne, Germany
erik.angelone@th-koeln.de

Abstract

This paper proposes a process-oriented approach using screen recording and think-aloud output for facilitating the AI literacy of translation students in the domain of interaction with generative AI. It outlines potential indicators of intelligence augmentation and impairment, as documented in process protocols of student performance, which students can use as a point of departure for self-reflecting on how generative AI utilization impacts their performance in terms of quality and efficiency.

1 Introduction

As generative AI continues to redefine translation workflows, the translator training community is responding with a range of learning activities to familiarize students with this assistive technology (see Penet et al., 2026; Pym and Hao, 2025). Translation-specific AI literacy (Krüger, 2005) is now being prioritized, informing competence frameworks and updates in translator training curricula. Thanks to recent surveys, we have a better understanding of *why* translators use generative AI. However, there is relatively little empirical research on *how* the utilization of generative AI impacts student translation performance in both positive and negative ways, from the perspective of quality and efficiency. This paper proposes a process-oriented approach that uses screen recording and think-aloud output to enable student self-reflection on their own performance along these lines when using generative AI. This approach is intended to foster AI literacy in the domain of interaction, where metacognitive awareness plays a pivotal role.

2 Conceptual framework

Within the interaction domain of his translation-specific AI literacy framework, Krüger (2025, 18) foregrounds a cognitive level, with corresponding effects traced to interaction with generative AI. Conducive cognitive effects, such as efficiency gains and beneficial cognitive offloading, are regarded as forms of intelligence augmentation. Detrimental cognitive effects, such as AI priming and various forms of deskilling, are regarded as forms of intelligence impairment.

An inherent danger of deskilling lies in the fact that is often perceptual in nature (Appiah, 2025), and might go unnoticed. The same could hold true for perception of intelligence augmentation. In the context of process-oriented translator training, a combination of screen recording and think-aloud output enables the creation of translation process protocols documenting problems and corresponding problem-solving. These protocols can then serve as pedagogical framework for retrospective self-analysis, with students and instructors reflecting on salient indicators of performance. This approach can readily extend to explorations of intelligence augmentation and impairment in conjunction with generative AI interaction.

3 Pedagogical design and methodology

Initial instructor input on how to create a screen recording and the task of thinking aloud would scaffold the approach, which is otherwise largely rooted in self-regulated, discovery-based learning. If training is deliberately deductive, central ideas

Proceedings of the 1st International Workshop on Teaching AI-Based Translation and Technologies (TAITT 2026), pages 76–77
June 15, 2026 ©2026 Association for Computational Linguistics

pertaining to intelligence augmentation and impairment could be defined and contextualized upfront so that students know what to look for when reflecting on their performance. This starts with working definitions of each overarching concept and its respective forms. The instructor may also consider introducing concrete manifestations of augmentation and impairment found in each modality. Pedagogically speaking, it would be advisable to do so after the task has been completed in order to avoid priming student articulations and actions. Indicators introduced after task completion and prior to retrospective reflection might take the following forms.

Indicators of intelligence augmentation in screen recordings:

- Precision prompting in optimizing output
- Using AI as a successful control mechanism rather than a stand-alone authority
- Tool orchestration of gen AI with other resources to facilitate efficiency without sacrificing quality

Indicators of intelligence impairment in screen recordings:

- Single-shot acceptance of AI output without revision or cross-checking other resources
- Copying and pasting of content across interfaces without revision (when needed)
- Shallow initial prompts yielding problems or errors not addressed through follow-up prompting

Indicators of intelligence augmentation in think-aloud protocols:

- Articulations signaling direct engagement and critical reflection with gen AI in addressing problems
- Articulations predicting how gen AI might help prior to actual use
- Articulations involving rationales for why gen AI is assistive

Indicators of intelligence impairment in think-aloud protocols:

- Articulations reflecting superficial actions ("I'm entering a prompt") rather than critical reflection
- Complete absence of articulation in conjunction with not detecting or addressing problems
- Articulations reflecting a loss of anchoring and detachment from the translation task and source content

Given the cognitively demanding nature of thinking out loud and subsequent manual retrospective analysis of on-screen phenomena and articulations, the translation task should be limited in text length and duration. Previous studies using this approach have used source content that is 100–150 words in length, yielding screen recordings that are approximately 20–30 minutes in duration.

In terms of concrete learning activities, obtained process protocols can be used for meeting various objectives. One activity could have students gauge their performance in conjunction with the aforementioned indicators to discern what augmentation and impairment look like in their particular case. A related learning activity could involve reverse engineering errors in the translation product in relation to what was articulated (or not articulated) and transpiring on-screen in conjunction with them. Students could be asked to reflect and report on the following: Was generative AI being used at the time the error occurred? If so, why did AI fall short and what might you have done differently to prevent the error? Do errors correlate with indicators of intelligence augmentation or intelligence impairment? What might have transpired differently had generative AI not been used in these instances?

References

Appiah, Kwame Anthony. 2025. The age of de-skilling: Will AI stretch our minds—or stunt them? *The Atlantic*.

Krüger, Ralph. 2025. Implementing generative artificial intelligence technologies in language industry workflows – A competence perspective. *Lebende Sprachen*, 70(1):11–38.

Penet, JC, Joss Moorkens, and Masaru Yamada, Masaru, editors. 2026. *Teaching Translation in the Age of Generative AI: New Paradigm, New Learning?* Language Science Press, Berlin.

Pym, Anthony, and Yu Hao. 2024. *How to Augment Language Skills: Generative AI and Machine Translation in Language Learning and Translator Training*. Routledge, London/New York.

Translator Competence in the Age of Agentic AI Orchestration: A "Backcasting" Perspective

Yu Hao
University of Melbourne, Melbourne, Australia
haoyh2@unimelb.edu.au

Elise (Qing) Wu
University of Melbourne, Melbourne, Australia
qing.wu1@student.unimelb.edu.au

Ester Leung
University of Melbourne, Melbourne, Australia
ether.leung@unimelb.edu.au

Abstract

As an orchestration infrastructure, agentic AI systems now can plan and decompose the pre-defined goals into a sequence of steps, decide on external function calls, and coordinate one LLM or multiple LLMs with specialised roles. In this context, this position paper adopts a future studies "backcasting" approach that starts with a desirable future, envisioned as one in which AI-integrated translation workflows are transparent, accountable, and aligned with human values; it then works backwards to examine how translator expertise should be reconceptualised to sustain meaningful human-in-the-loop participation. In this sense, the study first conceptualises the current translation-service provision as a system structured around managerial, mediation, and authorising roles. It then analyses how these roles may be changed and augmented within the agentic AI-orchestrated workflows. Building on the analysis, we propose a series of competences that should be cultivated to achieve the envisioned future: 1) evaluation grounded in advanced language competence; 2) situated and context-sensitive judgement informed by cultural and experiential knowledge; and 3) strategic procedural planning in the design and oversight of agentic AI-orchestration workflows. The paper concludes with recommendations for future pedagogical development and empirical research.

1 Introduction

The ever-evolving technology has profoundly reshaped the landscape of the translation industry (e.g., ELIA, 2025; Translators Association of China, 2025) and changed the way we translate and teach translation. More recently, the surge of agentic AI marks a new phase of the AI boom. Unlike earlier generative AI systems, such as large language models (LLMs), which produce content in response to human prompts in each iteration, agentic AI now operates as an orchestration layer and can assume greater autonomy over the entire workflow (Acharya et al., 2025; Gridach et al., 2025; Sapkota et al., 2025; Schneider, 2025). In highly autonomous settings, agentic AI systems can now plan and decompose the pre-defined goals into a sequence of steps (i.e., based on a goal input by humans, the system can autonomously break the task down into a series of logical steps to achieve that goal), decide on external function calls, and coordinate one LLM iteratively or multiple LLMs with specialised roles (some as planners, while others as executors or evaluators), steps traditionally taken by humans. In this sense, their role shifts from task-level execution to higher-level control, authority, and governance. At the other end of the spectrum, the autonomy of agentic AI systems can be limited to text generation, as in earlier models, whereas all workflow steps and external tool use are predefined or hard coded by humans.

Although AI agents and agentic AI orchestration are still very new in the field of translation studies, some pioneering research has begun to explore how automated translation could potentially "go agentic." For instance, Wu et al. (2025) introduced a multi-agent LLM framework that mimics a real commercial translation company. The framework

Proceedings of the 1st International Workshop on Teaching AI-Based Translation and Technologies (TAITT 2026), pages 78–84
June 15, 2026 ©2026 Association for Computational Linguistics

involves several autonomous agents, including a CEO, Senior Editor, Junior Editor, Translator, Localization Specialist, and Proofreader. The system was used to translate a set of Chinese literary texts and showed great potential; compared with baseline translations (generated by ChatGPT-4), the agentic AI translations were preferred in both monolingual human evaluation and bilingual automated evaluation. In parallel, Briva-Iglesias (2025) conceptualized AI agents as configurable units that enable both single-agent and multi-agent workflows, mirroring professional translation processes to various degrees. In this work, four essential characteristics of AI agents and agentic workflows are discussed, including "autonomy" (the extent to which the system is trusted to make decisions independently), "tool use" (the system's capacity to integrate external resources, such as translation memories, glossaries, and domain-specific databases), "memory" (the capacity to recall clients' stylistic preferences and previous corrections from human revisors), and workflow customization (agentic AI workflows are not rigid and one-size-fits-all pipelines, but can be dynamically adapted on the fly to introduce specialized sub-agents or pause for human authorization when escalation is triggered). More recently, Briva-Iglesias and O'Brien (2026) have advocated for a human-centered language technology framework in increasingly autonomous environments, emphasizing reliability, safety, and trustworthiness across translation workflows.

A recent hype around "raising lobsters" (i.e., running OpenClaw AI agents on users' personal devices) shows how widely accessible and popular open-source AI agents have become, while also raising serious concerns. It is reported that such highly autonomous agents have deleted data, exposed sensitive information, or initiated unauthorised financial actions (New York Times, 2026). As agentic AI functions as an orchestration framework rather than a collection of isolated tasks, errors in these systems can have broader systemic implications. When such systems misinterpret predefined goals or act beyond user intent, as observed in cases of OpenClaw, errors can escalate into large-scale failures and are very unlikely to be traced back to a specific step.

So what might translation look like in the era of agentic AI? And what implications does this have for translator education? This position paper approaches these questions from a "backcasting" perspective, a future studies method that begins with a desirable future state and works backwards to identify the conditions necessary to achieve it (Dreborg, 1996; Holmberg and Robèrt, 2000). In this lens, this study envisions a future in which agentic AI is used responsibly and ethically to achieve shared and optimal outcomes for all stakeholders who would expect value from translation services. In such a future, human-AI collaboration is both effective and accountable. The paper then examines the conditions required to bring about such a future. It explores how professional translation workflows may evolve in ways that allow AI to augment, rather than replace, human judgment. Pedagogically, it also seeks to explore the kinds of human expertise, curricular changes, and institutional safeguards are needed to ensure that humans continue to remain meaningfully in the loop.

2 Human actors in contemporary workflows and their expertise

Before addressing the potential impacts of agentic AI, it is pertinent to first review the human actors involved in contemporary, commercialised translation workflows. This is conceptualised in our working model as a multi-actor framework structured around managerial, mediation, and authorising roles. Each category of actors contributes heterogeneous expertise, with different responsibilities distributed across these roles. We argue that understanding the current distribution of roles is essential for assessing how emerging technologies such as agentic AI may replace, reallocate, or augment human expertise.

In this study, account managers (AMs) and project managers (PMs) are both grouped under the category with managerial priorities. Compared with other role groups, managerial actors exert oversight and more constantly engage in interpersonal interactions, as they function as coordination nodes that connect clients indirectly with all the other roles involved. Account managers are involved from the very beginning as the primary point of contact with the client, in which they negotiate project scope, cost, quality and delivery time. Their work is strategic but also highly interpretive. They are often responsible for translating ambiguous client needs into structured guidelines (commonly referred to as a "translation brief") to initiate the project. Project managers are then brought on board to coordinate the workflow. They analyse source texts, plan or select appropriate workflows, and allocate tasks to relevant roles;

throughout the process, they constantly calibrate the shared goals, oversee timelines, and manage any uncertainties that might happen in real-world situations; and finally, they are responsible for the final delivery of the product. Project managers may not be familiar with all the languages involved in a project (particularly in large-scale multilingual settings); however, they have to rely on knowledge of translation technologies and genre-specific requirements to inform their decision-making. In this sense, managerial roles require strong interpersonal and communication skills, as well as the strategic skillsets to align expectations, advise on appropriate procedural solutions, and mitigate potential risks.

Next, language and cultural mediation roles constitute the core personnel of what is traditionally understood as "translation", namely text processing and production. These roles include, among others, from-scratch translators, pre-editors engaged in language control, post-editors who correct machine-generated outputs, and in some cases cultural consultants who address cultural intricacies in communication. Note that these roles are not mutually exclusive in real-world scenarios; practitioners may assume multiple roles sequentially in a workflow or concurrently across different projects. In today's language industry, these actors increasingly work within hybrid environments in which automated outputs are explicitly integrated into or at least consulted during the translation process (Krüger, 2025; Shormani and Al-Sohbani, 2025). Alongside from-scratch translation, they are increasingly required to review and refine AI-generated outputs, exercising situated judgment in each iteration. Across the board, these roles require AI literacy as well as high levels of language competence, particularly as AI models can generate increasingly fluent sounding but distorted translations.

The final category is associated with authorising responsibilities. In this group, text revisers and quality assurance (QA) specialists provide an essential layer of quality validation and accountability. These authorising roles function as crucial checkpoints within the workflow. During text production, once a draft is complete, revisers critically assess the translated text for accuracy and stylistic appropriateness; this work requires advanced bilingual and bicultural expertise. Moving towards the final stages of text production, QA specialists focus on procedural verification, ensuring consistency and formatting in the final product, as well as compliance with the translation brief.

3 A conceptualisation of an agentic AI translation workflow

Having outlined the human roles and their managerial, mediation, and authorising expertise, this section explores which aspects of these roles could potentially be augmented within agentic AI orchestration. Recent studies have suggested that agentic AI is particular suitable for work characterised by existing, clear, and structured workflows (Hughes et al., 2025; Sapkota et al., 2025), as is the case for many commercialised translation services. So, in what follows, we conceptualise the potentials and associated risks of agentic AI systems across the three role groups in our working model.

Agentic AI systems now work on an infrastructural layer that coordinates activities across agents in an AI orchestration. In this framework, the systems can decompose tasks, assign roles and resources, monitor progress, and even adjust workflows on the fly, functions traditionally central to managerial tasks (Hughes et al., 2025; Sapkota et al., 2025). If routine management can potentially be automated, what then becomes of managerial roles? We argue that the social-interactive dimension of management is far more resistant to automation. This includes interpreting ambiguous client needs, negotiating value, and framing project purpose, all activities that involve open-ended sensemaking and judgment anchored in situational awareness. At the same time, we hasten to add that it remains unclear to what extent such systems can autonomously reflect on errors and call on the corresponding external functions or additional LLMs to address risks and unexpected situations.

For language and cultural mediation roles, core "translation" activities in routine or low-risk domains have long been automated, to various extents, by a range of technologies including translation memories and machine translation. Standard LLMs can also be prompted to undertake automated revision and language control on textual outputs. However, recent studies have reported that these systems continue to struggle when meaning is ambiguous, culturally embedded, or contextually high stakes (Hajek et al., 2025; Hao and Pym, forthcoming in 2026). It indicates the need for competent language and cultural experts

to remain involved for iterative revision and editing. In an orchestrated AI workflow, these human mediation actors should not only be involved in routine evals, i.e., output-based evaluation, but also be engaged when uncertainty thresholds or escalation triggers are reached.

Authorising roles present a mixed picture. These roles seem to shift from undertaking text-level changes to auditing system-level decisions. For instance, while some work of QA specialists that align closely with algorithmic validation can easily be made (semi-)automated, including consistency and formatting checks, the situated interpretive aspects of their work remain human-centred. Similarly, although generative AI can be curated to undertake editing at scale and agentic systems are able to call on "third-party" LLMs as an evaluator, human authorisation remains essential for accountability purposes when "escalation alarm" is triggered or before the output is made available to wider audiences.

In a nutshell, as AI increasingly performs coordination, decision-routing, and governance functions, human expertise in the loop should centre on situated judgement (for language and cultural mediators), edge-case handling (for managers, also across roles), and system-level accountability (co-supervised by managerial and authorising roles). However, recent research on human–machine cooperation (HMC) suggests that, paradoxically, as automation increases, human actors may experience reduced situational awareness, becoming less able to maintain an "eagle-eye" view of system errors (e.g., Hao and Pym, forthcoming in 2026). One possible explanation is automation bias: systems perceived as reliable may lead human actors to over-trust AI outputs and under-scrutinise their limitations (Tatasciore and Loft, 2024; Romeo and Conti, 2025). This out-of-the-loop condition can also occur in translation workflows across roles, where managers, mediators, and revisers fail to intervene when problems arise. This creates a central tension: while human expertise is increasingly expected to centre on situated judgement, cultural-, linguistic, and managerial-wise, highly competent AI systems may compromise that judgement by encouraging over-reliance on their outputs. So what role could institutional training play to address this tension?

4 Cultivating competence for future translator education

A backcasting approach offers a useful forward-looking conceptual framework. Backcasting is theoretically grounded in a normative and problem-oriented perspective in futures studies. Dreborg (1996) conceptualizes backcasting as being underpinned by an understanding of social change in which human agency, strategic choice, and evolving knowledge all shape future outcomes. Additionally, backcasting is less concerned with predicting probable futures than with exploring feasible and desirable alternatives to inform long-term decision-making and development. Importantly, this perspective does not assume the future to be an extension of present trends; rather, it emphasizes transformative change, particularly in contexts where existing initiatives and trajectories may no longer remain sustainable.

Adopting this approach, we begin with a desired future, one in which agentic AI-integrated translation workflows are transparent, accountable, and aligned with human values, and work backwards to identify the translation competences required to sustain meaningful human participation. In such a future, the key question is no longer whether humans remain involved (a concern more closely associated with predicting probable futures), but how their roles are reconfigured and what forms of expertise would enable them to contribute within increasingly autonomous workflows. To bring about this envisioned future, translator education should prepare students for sustained human-AI collaboration and cultivate competences that are not easily replicated by automation. We therefore propose a list of interconnected competence for rethinking translator education in the agentic AI era.

First, fundamental language competence may become more important than ever. L1 and L2 proficiencies have long been viewed as foundational skills for translators and are often listed as admission prerequisites in translation programmes. Their importance in translator education is also reflected in the ways language skills are acknowledged and integrated into most translation competence models (e.g., PACTE Group, 2003; EMT, 2017) and previous surveys on translation competence (e.g., Schmitt et al., 2016; Toudic, 2017; Hao and Pym, 2021). Traditionally, language competence has been understood as the ability to comprehend and produce content in one's first and

second languages, as well as to transfer meaning between them (particularly in models where "translation" is not treated as a stand-alone competence). However, as agentic AI systems can autonomously generate and revise texts, language competence may shift from what is required to translate from-scratch towards the ability to critically comprehend source-language content, detect errors in AI-generated output as well as propose appropriate alternatives in the target language. In this sense, language competence currently forms the foundation of post-editing and revision AI output, which is required across mediating and authorising roles. Although AI systems continue to evolve rapidly, they may increasingly generate fluent-sounding yet distorted outputs. Thus, the ability to detect, diagnose, and iteratively refine such errors may become critical, positioning translators as quality gatekeepers of AI-generated content.

At the curricular level, this technology-driven change reopens longstanding debates with respect to the role of advanced language training in translator education (Schaffner, 2000; Pym, 2018), particularly whether standalone courses or modules in students' L1 and L2 should be incorporated into a translation curriculum. Programmes may therefore need to reinforce advanced proficiency in both languages, potentially through rigorous assessment formats such as in-person examinations, particularly during earlier stages of training. At the same time, language competence should be developed alongside revision and evaluation practices so that students are prepared to assume quality-authorisation responsibilities, including reviewing, validating, and authorising AI-generated translations in relation to communicative briefs and ethical standards.

Second, situated judgement of automated translation grounded in contextual and cultural awareness becomes increasingly important. This is somehow related to what O'Brien (2002) and a few other visionaries previously proposed with respect to the ability to post-edit machine translation; there has long been explicit recognition of post-editing abilities in L1 and L2 in the EMT translation competence model (EMT, 2009; 2017). However, judgement and decision-making in traditional post-editing were primarily retrospective and focused on improving the one-shot machine output, while in agentic AI environments they become more recursive and increasingly process-oriented. In other words, translators in the loop are increasingly required to exercise contextual and cultural judgement, in order to iteratively prompt, evaluate, and authorise the (semi-)automated translation process itself and its respective AI outputs. At the same time, such judgement that underpins high-quality translation should be situated, embodied, and context-driven. Research in cognitive science suggests that language understanding is not purely abstract but deeply rooted in embodied, affective, and cultural experience (Hauk et al., 2004; Pulvermüller, 2018). Meaning-making therefore depends not only on linguistic knowledge but also on human experiential insight and contextual sensitivities (Tian, 2024; Zhou and Wang, 2025). That is, while AI systems often lack such grounding in factual and embodied experience, human translators can be trained to navigate ambiguities, interpret culturally embedded connotations, and mediate for marginalised communities where AI systems may exhibit bias or insufficient coverage.

Perhaps rather than requiring large-scale curricular restructuring, the development of this competence may be achieved through targeted pedagogical interventions. Classroom practices such as close readings, interpretive discussions, and comparative analysis of multiple valid translations (of the same start text) can encourage students to articulate and justify interpretive choices. Role-play, scenario-based tasks, and audience adaptation exercises, such as those aligned with Skopos theory modules, can further enhance learners' awareness of register, tone, and communicative purpose.

Third, an emerging competence foregrounds strategic procedural planning, which we argue differs from traditional project management skills. In conventional translation or localisation contexts, a project manager traditionally functions as the managerial and logistical centre for the clients and all professionals involved in pre-defined translation workflows. According to the EMT translation competence framework (EMT 2017), project management skills refer to the ability to plan, coordinate, and oversee language or translation services in a professional context, including managing client relationships, ensuring quality standards are met, and organising budgets, timelines, as well as translators and other service providers involved in the process. However, as agentic AI systems can coordinate multiple AI agents to automate a wide range of tasks (workflow planning, translation, terminology, editing, quality assurance, invoicing) as well as assume routine managerial functions, a project manager may stop manually coordinating

translators and other professionals but instead manage a semi-autonomous ecosystem of AI agents and humans who remain in the loop. Human oversight remains critical. Managers' expertise should shift to decision-making with respect to 1) risk anticipation and assigning the appropriate degree of autonomy to agentic AI systems, i.e., the extent to which systems are trusted to make autonomous decisions; 2) escalation pathways for ambiguous or high-stakes cases, i.e., when and under what conditions human intervention should occur in a semi-autonomous workflow; and 3) human accountability, i.e., how quality assurance mechanisms should be embedded into autonomous orchestration systems; and 4) dynamic workflow adjustment, i.e., how workflows should be adaptively restructured and adjusted when conditions change. All of these require the human manager not only to initially match the nature of the project with a suitable agentic AI-assisted workflow, but also to dynamically adjust such workflows to call on (non)human agents and external resources when errors occur. Human actors are therefore likely to move away from direct task execution towards higher-level workflow design, supervision, and rapid procedural adaptation in response to evaluation outcomes. That said, as mentioned in Section 3, the effective communication skills required in client relationships may remain unchanged and could be automation resistant.

Translator education could address this shift by incorporating training in strategic workflow design in their current technology or project management modules, and when possible, expose students to authentic agentic AI systems for translation purposes. For instance, students might engage in agentic-AI powered localisation projects in which they design workflows, allocate tasks to AI agents and human in the loop, and ensure that outputs meet professional standards.

Last but not least, a baseline level of AI literacy is necessary to sustain meaningful human-in-the-loop participation under conditions of increasing automation. This includes not only technical skills like prompt engineering and system interaction, but ethical awareness, and perhaps it is equally important to stay technological curious when technology is evolving so quickly.

In short, note that this work remains exploratory and "hypothesis-generating"; and the arguments presented here are conceptual and should all be further examined empirically. We also note that this position paper focuses specifically on commercialised and team-based translation workflows and their implications for institutional training. Additionally, our primary goal is not to provide specific pedagogical solutions for all translation programs before we have more empirical justifications, as we believe there is no one-size-fits-all solution: programmes guided by various initiatives will need to accommodate and adapt in terms of their institutional contexts and available resources. A broader range of issues associated with agentic AI remains beyond the scope of the present study but warrants further investigation. These include risks such as task mis-decomposition by orchestrating systems, loss of contextual coherence across distributed agents, financial costs associated with large-scale AI use, and environmental sustainability. These challenges underscore the need for continued critical engagement with agentic AI beyond its immediate functional advantages. While this paper foregrounds the reconfiguration of human expertise and corresponding pedagogical responses, future research must also address the systemic, technical, and ethical complexities inherent in multi-agent environments.

References

Acharya, Deepak Bhaskar, Karthigeyan Kuppan, and B. Divya. 2025. Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey." *IEEE Access* 13:18912–18936.

Briva-Iglesias, Vicent. 2025. Are AI agents the new machine translation frontier? Challenges and opportunities of single- and multi-agent systems for multilingual digital communication. In *Proceedings of Machine Translation Summit XX*, pages 365–377, Geneva. European Association for Machine Translation.

Briva-Iglesias, Vicent and Sharon O'Brien. 2026. Human-centered AI language technology (HCAILT): An empathetic design framework for reliable, safe and trustworthy multilingual communication. *International Journal of Human–Computer Interaction*, 1–15.

Dreborg, Karl H. 1996. Essence of backcasting. *Futures*, 28(9):813–828.

ELIA. 2025. *European Language Industry Survey*.

EMT. 2017. *European Master's in Translation Competence Framework 2017*.

EMT Expert Group. 2009. *EMT Competences for Professional Translators, Experts in Multilingual and Multimedia Communication*.

Gridach, Mourad, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes, and Christina Mack. 2025. Agentic AI for scientific discovery: A survey

of progress, challenges, and future directions. *arXiv* preprint arXiv:2503.08979.

Hajek, John, Anthony Pym, Yu Hao, Maria Karidakis, Ambrin Hasnain, Anila Hasnain, Ke Hu, and Juerong Qiu. 2024. *Understanding and Improving Machine Translations for Emergency Communications*. The University of Melbourne, Melbourne.

Hao, Yu and Anthony Pym. 2021. Translation skills required by Master's graduates for employment: which are needed, and which are not? *Across Languages and Cultures*, 22(2):158–175.

Hao, Yu, and Anthony Pym. 2026. Can generative AI help students revise translations? Yes, with low, vigilant trust. In Feinauer, Ilse and Amanda Marais, editors, *The Reality of Revision*. Benjamins, Amsterdam.

Hauk, Olaf, Ingrid Johnsrude, and Friedemann Pulvermüller. 2004. Somatotopic representation of action words in human Motor and premotor cortex. *Neuron*, 41:301–307.

Holmberg, John and Karl-Henrik Robèrt. 2000. Backcasting—A framework for strategic planning. *International Journal of Sustainable Development & World Ecology*, 7(4):291–308.

Hughes, Laurie, Yogesh K. Dwivedi, Tegwen Malik, Mazen Shawosh, Mousa Ahmed Albashrawi, Il Jeon, Vincent Dutot, et al. 2025. AI agents and agentic systems: A multi-expert analysis. *Journal of Computer Information Systems*, 65(4):489–517.

Krüger, Ralph. 2025. Implementing generative artificial intelligence technologies in language industry workflows—A competence perspective. *Lebende Sprachen*, 70(1):11–38.

New York Times. 2026. China is embracing OpenClaw, a new A.I. agent, and the government is wary. March 17, 2026.

O'Brien, Sharon. 2002. Teaching post-editing: A proposal for course content. In *Proceedings of the 6th EAMT Workshop: Teaching Machine Translation*, pages 99–106. Manchester, European Association for Machine Translation.

PACTE Group. 2003. Building a translation competence model. In Alves, Fabio, editor, *Triangulating Translation: Perspectives in Process-Oriented Research*, pages 43–66. Benjamins, Amsterdam/Philadelphia.

Pulvermüller, Friedemann. 2018. Neural reuse of action perception circuits for language, concepts and communication. *Progress in Neurobiology*, 160:1–44.

Pym, Anthony. 2018. Where translation studies lost the plot: Relations with language teaching. *Translation and Translanguaging in Multilingual Contexts*, 4 (2):203–222.

Romeo, Giulio and Dario Conti. 2025. Exploring automation bias in human–AI collaboration: A review and implications for explainable AI. *AI & Society*, 41:259–278.

Sapkota, Ranjan, Konstantinos I. Roumeliotis, and Manoj Karkee. 2025. AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges. Information Fusion, 126.

Schäffner, Christina. 2000. Running before walking? Designing a translation programme at undergraduate level. In Schäffner, Christina and Beverly Adab, editors, *Developing Translation Competence*, pages 143–156. Benjamins, Amsterdam.

Schmitt, Peter A., Lina Gernstmeyer, and Sarah Müller. 2016. *Übersetzer und Dolmetscher – Eine Internationale Umfrage zur Berufspraxis*. BDÜ Fachverlag, Berlin.

Schneider, Johannes. 2025. Generative to agentic AI: Survey, conceptualization, and challenges. *arXiv preprint* arXiv:2504.18875.

Shormani, Mohammed Q. and Yehia A. Al-Sohbani. 2025. Artificial intelligence contribution to translation industry: Looking back and forward. *Discover Artificial Intelligence*, 5:1–29.

Tatasciore, Michael and Shayne Loft. 2024. Can increased automation transparency mitigate the effects of time pressure on automation use? *Applied Ergonomics*, 114:1–8.

Tian, Xujun. 2024. Personalized translator training in the era of digital intelligence: Opportunities, challenges, and prospects. *Heliyon*, 10:1–13.

Toudic, Daniel. 2012. *Employer Consultation Synthesis Report*. OPTIMALE Academic Network Project on Translator Education and Training, Rennes.

Translators Association of China. 2025. *2025 China Language Service Industry Development Report*. Beijing.

Wu, Minghao, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. (Perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *arXiv preprint* arXiv:2405.11804.

Zhou, Tong and Jinghui Wang. 2025. Embodied empathy in translation studies: Enhancing global readers' cognitive and emotional engagement with translations of traditional Chinese medicine terminology. *Frontiers in Psychology*, 16:1–10.

Author Index