Proceedings of the

26th Annual Conference of the European Association for Machine Translation

Volume 2: Products and Projects, Implementations and Case Studies

16–18 June 2026
Tilburg, The Netherlands

Edited by

Dimitar Shterionov, Eva Vanmassenhove, Mirella De Sisto, Fred Blain, Javad Pourmostafa Roshan Sharami, Lisa Lepp, Chiara Manna, Argentina Anna Rescigno, Alina Karakanta, Ayla Rigouts Terryn, Manuel Lardelli, Natalia Resende, Elena Murgolo, Janica Hackenbuchner, Anna Zaretskaya, Miquel Esplà-Gomis, Thierry Etchegoyhen, Dagmar Gromann, Rachel Bawden, Barry Haddow, Sara Szoc, Mikel Forcada, Helena Moniz

ISBN: 9789403901404
DOI: 10.26116/9789403901404

Foreword from the General Chair

As president of the European Association for Machine Translation (EAMT) and General Chair of the 26th annual conference of EAMT, it is my utmost pleasure to write these opening words (the last time for me as your president!). Be most welcome to our EAMT 2026!

As usual, the EAMT Executive Committee (EC) has been very busy. Mikel Forcada (treasurer) and Sara Szoc (secretary) have been tirelessly supporting all initiatives. Carolina Scarton and Sara Szoc took great care of our bursaries. Patrick Cadwell, André Martins, Dimitar Shterionov, and Manuel Lardelli were our chairs for the Research Projects. Our very own Mary Nurminen, chair of the bid proposals for our next events, has been busy selecting our next venue! EAMT 2027 venue will be disclosed in our closing ceremony at Tilburg!

One of our core initiatives, the best thesis award – Rachel Bawden and Barry Haddow, chairs of the Best Thesis Award, had a very difficult time selecting a candidate, since the submissions were of very high quality. Our congratulations to Gabriele Sarti (PhD carried out at the University of Groningen, currently at Northeastern University), "From Insight to Impact: Actionable Interpretability for Neural Machine Translation", supervised by Arianna Bisazza, Malvina Nissim and Grzegorz Chrupała. In addition, the committee judged that the thesis of David Stap (PhD carried out at the University of Amsterdam, currently at NXAI) entitled "Analyzing and Improving Cross-lingual Knowledge Transfer for Machine Translation", supervised by Christof Monz and Vlad Niculae was "highly commended".

EAMT has been sponsoring the MT Marathon for several years. We would also like to thank the University of Helsinki, Jörg Tiedmann, for organizing the 18th MT Marathon. The event included MT lectures and labs, covering the basics and tutorials; keynote talks from experienced researchers and practitioners; presentations of research and open source tools related to MT; and hacking projects to advance tools or research in one week or start new collaborations.

Our EAMT 2026 in Tilburg will have a four-day intense program, put together by our chairs: Alina Karakanta and Ayla Rigouts Terryn (research: technical track chairs); Manuel Lardelli and Natalia Resende (research: translators & users track chairs); Elena Murgolo and Janica Hackenbuchner (implementations & case studies track chairs); Miguel Esplà-Gomis and Anna Zaretskaya (products & projects track chairs) and Thierry Etchegoyhen and Dagmar Gromann (workshop and tutorial chairs). Our deepest gratitude for being our filters of quality! We will also have a one-day workshops and tutorials event. Join us! So much to discuss.

Our gratitude to our keynote speakers, Rachel Bawden, fellow at PR[AI]RIE-PSAI research institution, and Antonio Toral, Distinguished Researcher in Machine Translation at Universitat

d'Alacant, Spain. Thoughtful voices in our MT community covering low-resourced and unseen languages (Rachel) and alternative translation pipelines that flip the roles, placing the translator before the machine (Antonio). I am sure that we will have a lot of food for thought with our outstanding speakers.

EAMT 2026 would not be possible without a fantastic local organizing team! A very energetic, engaged, proactive, and hard working local organising team! It has been a pleasure working with you! Our local organizers, Dimitar Shterionov, Eva Vanmassenhove, Mirella de Sisto, Fred Blain, Javad Pourmostafa, Lisa Lepp and Chiara Manna from Tilburg University, and Argentina Anna Rescigno from the University of Pisa, did a great job and put a lot of love in hosting you!

EAMT has been supported by generous sponsors in its initiatives along the years, long time friends and new ones, to all of you, thank you! This year is again a very exceptional year in terms of sponsoring activities. Our gratitude to our Platinum sponsor, who will also be giving a research oral presentation, AppTek. Our golden sponsor – the Sectorplan for the Humanities, "Humane AI" theme, funded by the Dutch Ministry of Education, Culture and Science. Our Silver sponsors: BIG Language Solutions, Cohere, Powerling, STAR, Translated, Transperfect, Tilburg.AI and the CSAI research centre (our sustainability sponsor). To all our Supporter sponsors: Apertium, Prompsit, Springer Nature (our Supporter sponsor for the Best Paper award) and Open Press Tilburg University (responsible for the publishing of our proceedings and booklet). Finally, to our Media sponsors, GALA, MultiLingual and Slator. Your support is vital in our efforts to give back to our community through grants and other initiatives. To the University of Tilburg, our sincere thank you!

A special thank you to all our members and community! Without you no effort would make sense! Let us take this opportunity to create scientific collaboration and give constructive feedback. To fully enjoy the conference, please check our Code of Conduct. I am looking forward to seeing you all and celebrating our community gathering!

It is our organisation's greatest wish to continue giving back to our community and to drive and be driven by our community's energy and enthusiasm. Reach out to us if you have new ideas or suggestions you would like to implement. We will try hard to accomplish it with you. Learn more about us.

Finally, my thank you for being your President! It has been a pleasure and an honor to serve this community! I have learned so much, I grew as a person and I have definitely pushed my boundaries! I wish the very best to our future President and to our beautiful community! Thank you, EC, for your amazing support along the years! You are definitely a "group of very nice people"!

Helena Moniz
President of the EAMT
General Chair of EAMT 2026
University of Lisbon, Portugal

Message from the Organising Committee

Welcome to Tilburg – the city of Schrobbelèr, Willem II (the football team), Dutch textile,... a social and experimental city; the city of Tilburg University, Schouwburg Concertzaal and MindLabs; the city of EAMT 2026. We are delighted to host the 26th edition[1] of the Annual Conference of the European Association for Machine Translation which gathers every year researchers, practitioners, students, and industry professionals from around the world to exchange ideas, present advances, and discuss the future of machine translation and multilingual language technologies.

Machine translation is evolving at a rapid pace thanks to the remarkable developments in generative artificial intelligence, foundational and large language models and the advancements in computing infrastructures, to reach a state where translation technologies are no longer confined to specialized tools. They are embedded across various digital infrastructures, mediating evergrowing multilingual communication in governmental, medical, commercial or educational settings. MT and LLM-based language technologies are playing an increasingly prominent role in everyday life while ethical and responsible deployment remains understudied.

In this context, the 2026 meeting of the European Association for Machine Translation (EAMT) provides a timely gathering that critically examines the future for MT research, education and deployment. It is unique in its interdisciplinary nature, intersecting between computer science, computational linguistics, translations studies and translation professionals. It thereby bridges and supports the needed coordinated dialogue across these communities that often operate separately while integration is essential for the coherent, socially responsible development of multilingual technologies.

The EAMT 2026 conference encompasses a diverse range of contributions, including two keynote and one industry presentations, 87 publications including research papers, project demos and descriptions, three workshops and three tutorials. Reflecting on theoretical advances, innovative applications, societal implications and economical insights, these constitute the new state-of-the-art in Europe and globally.

Presented in a four-day programme (a full day of workshops and tutorials, followed by the three-day main conference) we hope that these activities foster inspiring discussions, stimulate

[1] While there is a debate on the edition number and perhaps we should count this as the 27th EAMT conference, for now we stick to 26 and leave the discussion for the next edition.

new collaborations, and strengthen connections across academia, industry, and public-sector organizations. This year we received 184 submissions, spread over 4 tracks and 3 workshops. We are proud to note that the 133 submissions to the Research – Technical, Research – Translators and Users, Products and Projects, and Implementation and Case Studies tracks as well as the 7 workshop and tutorial submissions have been a record number for the EAMT conference. This incremental trend is indicative not only of the rapid evolution of MT and the related fields, but also of the importance of translation and language technologies in the global scheme of science and society.

Organizing EAMT 2026 has been a collective effort and wouldn't have been possible without the support of our sponsors, who ensured we could provide a top-level experience spread over two amazing venues, publish our proceedings and disseminate our event; the two venues – MindLabs (workshops and tutorials) and Schouwburg (main conference), which provided the facilities to host this conference; the EAMT executive committee for their guidance, and the CSAI research centre and Tilburg University for the institutional support. Furthermore, we would like to express our sincere gratitude to the authors, for their work and the reviewers and programme committee members for their critical assessment; to the track chairs for their oversight on submissions, reviews and decisions; to the workshop and tutorial organisers for bringing such interesting and important topics to the scene and to our volunteers for their on-site assistance. Their dedication and hard work have been instrumental to the EAMT 2026 conference.

We hope that you find the two volumes of the EAMT 2026 proceedings engaging, motivating and contributory to your work, and that you have a productive, inspiring and memorable conference.

The EAMT 2026 Organizing Committee
Tilburg, The Netherlands
June 2026

Preface by the Programme Chairs

Research – Technical Track

The Research-Technical track invited submissions on significant results in any aspect of MT and related areas, including multilingual technologies. As in previous years, this track proved the most popular of the four tracks, receiving a total of 55 submissions, higher than previous years. With five desk rejections, 31 papers were accepted, resulting in an acceptance rate of 56%, slightly above the previous year. Nine of the accepted papers will be presented orally and the remaining 22 will be presented as posters.

Following current practices in the field, many papers focus on large language models (LLMs) for translation. A number of contributions address corpora and resources (Dias et al., Ciesiółka et al., Hingrajiya et al., Mash et al.) while the largest group of the accepted papers focus on evaluation (Shterionov et al., Wiśniewski and Czudy, Yanishevsky and Norris, Miró Maestre and Martínez-Murillo, Nishimwe et al., Iakivchuk, Hauhio et al., Dahan et al.), reflecting the field's growing emphasis on assessing LLM-based translation quality. On the methodological side, papers explore prompt engineering (Sánchez-Torrón et al.), retrieval-augmented generation (Zafar et al., Bouthors et al.), quality estimation (Guttmann et al.), data selection and augmentation (Aulamo et al., Kalikman et al.), and multi-agent workflows (Shen et al.). Fine-tuning and domain adaptation remain active areas of inquiry, with work spanning legal (Di Natale et al., Lorini et al.), literary (Yirmibeşoğlu Balal and Güngör), e-commerce (Zhang et al.), and oil & gas (Yang et al.) domains, as well as automatic post-editing (Deoghare et al.). Low-resource translation continues to attract attention this year (Fishel and Yankovskaya, Qian and Scherrer). The programme also features contributions addressing ethical and societal dimensions, including gender bias (Hackenbuchner et al., Ivanovs et al.) and MT for crisis communication (Castaldo et al., Moerman et al.).

We would like to give our thanks to all the authors who submitted to the track and to the 67 reviewers, who provided feedback and insightful comments for the submissions received. We are particularly grateful to the emergency reviewers who agreed to review papers at the last minute, allowing decision notifications to be sent out on time.

Research – Translators and Users Track

The Translators and Users Track is dedicated to the human-centric aspects of machine translation, exploring how translation technologies are applied, experienced, and evaluated by professionals and end-users in real-world contexts. This year, the track received a total of 34 submissions. Following six desk rejections and a rigorous review phase, 24 high-quality papers were accepted, resulting in an acceptance rate of 70.6%. Nine of these will be presented orally, while the remaining 15 will be featured as poster presentations. Unsurprisingly, a substantial portion of this year's program reflects the field's rapid pivot toward Large Language Models (LLMs) and generative AI, investigating how these tools reshape workflows through prompt engineering, safety protocols, and domain adaptation (Lai and Li, Pandeiro et al., Rios Gaona et al.). The track also highlights the widespread adoption of MT beyond traditional localization, featuring insights into user experiences across parliaments, public sectors, journalism, and social media (Leal-Wyss et al., Ilkılıç et al., Nurminen and Havumetsä). Furthermore, the intersection of automation and creativity remains a focal point, with papers exploring the evaluation of literary and poetic translations (Gerrits et al., Resende and Hadley), multimodal applications like subtitles (Schlüter et al.), and the evolving realities of post-editing workflows and the translation profession (van Tellingen et al., Bowker and Rodrigues). Looking at this year's program, it is clear that our community is moving far beyond simply asking whether to use machine translation. The focus has decisively shifted to how we interact with these increasingly capable systems, navigating prompt languages, safeguarding confidentiality, managing cognitive loads during post-editing, and preserving human creativity. The 2026 Translators and Users Track serves as a testament to the resilience and adaptability of language professionals as they shape the future of human-AI collaboration. We would like to extend our deepest gratitude to all the authors who submitted their work and shared their valuable insights. We are equally grateful to our 34 dedicated reviewers, who worked diligently to provide constructive feedback and ensure a rigorous, fair, and timely review process for all submissions. A special note of appreciation goes to those who stepped in as emergency reviewers on short notice, whose vital assistance ensured we could finalize decisions and notify authors without delay.

Implementations and Case Studies Track

This year, the Implementations and Case Studies Track received 16 submissions, of which 9 high-quality papers were accepted. Submissions stem from industry members, practitioners and academia alike. The topics are multifold with emphasis on information extraction, Large Language Model (LLM) enhancement, specialised Machine Translation (MT) systems, multimodal context, low-resource languages and post editing. Half of the papers will be presented as oral presentations, half as poster presentations. We expect interesting, fruitful and interactive sessions. We are grateful for the 22 committed reviewers, who have submitted their work on time without the need for emergency reviewers. We would like to deeply thank them for their work to ensure a fair reviewing process.

Products and Projects Track

For this year's Products and Projects track, 23 papers were accepted from 26 submissions, showcasing a rich variety of community-driven projects and products. The program highlights EAMT-sponsored, European, and regional initiatives, alongside cutting-edge work from top industry and research institutions. More than just a showcase, this track offers a vital platform for participants to share their initiatives, gather constructive feedback, and unlock new opportunities for collaboration and community engagement. We anticipate a lively session, anchored by our traditional "poster boasters" and dedicated poster presentations. We extend our deepest thanks to the 33 reviewers who worked diligently under tight deadlines to ensure a rigorous and fair review process.

Workshops and Tutorials

This year we received three workshop proposals, all of which were accepted. These workshops are the fourth edition of the Workshop on Gender-Inclusive Translation Technologies (GITT), the first edition of the Workshop on Teaching AI-based Translation and Technologies (TAITT), and the first edition of the Workshop on Style in GenAI-Translated Content (StyGenAI). All three will provide lively forums for the exchange of ideas on highly relevant and recent topics in the field of (machine) translation, from inclusiveness to style in translation technology. We also received four tutorial proposals, three of which were accepted. The tutorials cover topics in line with the typically diverse audience at EAMT, including human evaluation methods of ever increasing relevance in current research, core translation evaluation methods and tools for translation freelancers and LSPs, and MT integration into CAT tools for modern translation practice. The EAMT 2026 workshops and tutorials aim to provide a rich environment for all participants of the conference, and we would like to thank all organizers, authors, and presenters for what will certainly be lively and fruitful sessions.

EAMT 2025 Best Thesis Award (Anthony C Clarke Award)

Four PhD theses defended in 2025 were received as candidates for the 2025 year edition of the EAMT Best Thesis Award, all of which were eligible. Four external reviewers were recruited to examine and score the theses alongside seven EAMT executive committee members. Each thesis was evaluated according to predefined criteria: how challenging the topic was, how relevant the results were to the MT field and the strength of its impact in terms of scientific publications. 2025 was yet again a strong year for PhD theses in machine translation, and the decision was not easy.

All PhD theses were of good quality, focused on interesting topics and were all highly appreciated by reviewers. A panel of two EAMT Executive Committee members (Barry Haddow and Rachel Bawden) was assembled to process the reviews and select a winner that was later ratified by the EAMT executive committee.

We are pleased to announce that the winner of the 2025 edition of the EAMT Best Thesis Award is Gabriele Sarti (PhD carried out at the University of Groningen, currently at Northeastern University), "From Insight to Impact: Actionable Interpretability for Neural Machine Translation", supervised by Arianna Bisazza, Malvina Nissim and Grzegorz Chrupała.

In addition, the committee judged that the thesis of David Stap (PhD carried out at the University of Amsterdam, currently at NXAI) entitled "Analyzing and Improving Cross-lingual Knowledge Transfer for Machine Translation", supervised by Christof Monz and Vlad Niculae was "highly commended".

The winner will receive a prize of €500, together with an inscribed certificate. In addition, Dr. Sarti will present a summary of their thesis at the EAMT 2026 in Tilburg, the Netherlands, receive complimentary membership to the EAMT in 2027 and will receive a travel bursary of €200.

Chairs of the Best Thesis Award 2025
Rachel Bawden, Inria, Paris, France
Barry Haddow, Aveni

Programme Committee

Research – Technical Track

Aleš Tamchyna, Ana Guerberof Arenas, Andrea Piergentili, Andrei Popescu-Belis, Antonio Castaldo, Antonio Valerio Miceli Barone, Artur Nowakowski, Atul Kr Ojha, Beatrice Savoldi, Benyamin Ahmadnia, Chiara Manna, Christophe Declercq, Daniel Ortiz-Martínez, Delu Kong, Ekaterina Lapshinova-Koltunski, Eleni Gkovedarou, Esther Ploeger, Fan Zhou, Felipe Sánchez-Martínez, Fred Blain, Guillaume Wisniewski, Haiyue Song, Hiroshi Echizenya, Jasmijn Bastings, Javier Iranzo-Sánchez, Jinan Xu, John Ortega, Jonathan Mutal, Josef Jon, Luisa Coheur, Marco Gaido, Marco Turchi, Maria Kunilovskaya, Marina Sánchez-Torrón, Masao Utiyama, Mattia Antonino Di Gangi, Mayra Nas, Michael Carl, Miguel Menezes, Miquel Esplà-Gomis, Mirella De Sisto, Patrick Simianer, Rejwanul Haque, Rik van Noord, Rodolfo Joel Zevallos Salazar, Rui Sousa-Silva, Sai Koneru, Senyu Li, Sergi Alvarez-Vidal, Siddharth Divi, Siqi Liu, Sokratis Sofianopoulos, Taro Watanabe, Thomas Moerman, Tong Xiao, Vera Cabarrão, Vicent Briva-Iglesias, Vilém Zouhar, Vincent Vandeghinste, Yasmin Moslem, Yves Lepage

Research – Translators and Users Track

João Pedro Campos, Ana Guerberof Arenas, Callum Walker, Nora Aranberri, Aletta G. Dorst, Federico Gaspari, Bettina Hiebl, Maarit Koponen, Ekaterina Lapshinova-Koltunski, Antoni Oliver, Constantin Orasan, Frederike Schierl, Federica Vezzani, Suad Al Rahbi, Sergi Alvarez-Vidal, Vicent Briva-Iglesias, Ines Buchegger, Helle Dam Jensen, Silvana Deilen, Maria Fernandez-Parra, Sabrina Girletti, Sari Hokkanen, Dorothy Kenny, Rudy Loock, Joss Moorkens, Masaaki Nagata, Mary Nurminen, David Orrego-Carmona, Celia Rico, Akiko Sakamoto, Maria del Mar Sánchez Ramos, Susana Valdez, Kirti Vashee, Patrick Cadwell

Products and Projects Track

Miquel Esplà-Gomis, Pierrette Bouillon, Sabrina Girletti, Marie-Aude Lefer, Chiara Manna, Antoni Oliver, Juan Antonio Pérez-Ortiz, Raphael Rubino, Víctor M. Sánchez-Cartagena, Arda Tezcan, Antonio Toral, Daniel Torregrosa, Sergi Alvarez-Vidal, Giuseppe Attanasio, Pavel Doronin, Pedro Luis Díez-Orzas, Johanna Gerlach, Judith Klein, Rebecca Knowles, Varun Kotte, Ekaterina Lapshinova-Koltunski, Manuel Lardelli, Lisa Lepp, Helena Moniz, Mara Nun-

ziatini, Maja Popovic, Randy Scansani, Dimitar Shterionov, Felipe Sánchez-Martínez, Rik van Noord, Tom Vanallemeersch, Anna Zaretskaya, Eleftherios Avramidis

Implementation and Case Studies Track

Vicent Briva-Iglesias, Gokhan Dogru, Ana Guerberof Arenas, Ann Huehls, Maria Illescas, Miguel Ángel Jiménez Crespo, Martin Kappus, Rebecca Knowles, Maarit Koponen, Marie-Aude Lefer, Joss Moorkens, Matiss Rikters, Caroline Rossi, Dierk Runne, Florian Schottmann, Maria del Mar Sánchez Ramos, Pilar Sánchez-Gijón, Marina Sánchez-Torrón, Daniel Torregrosa, Mireia Vargas-Urpí, Sergi Alvarez-Vidal, Silvia Hansen-Schirra

Welcome to EAMT 2026

EAMT 2026 is organized in cooperation with Tilburg University (TiU)'s Inclusive and Sustainable Machine Translation (ISMT) Research Line led by Dimitar Shterionov with staff members from the Department of Intelligent Systems (DIS) and the Department of Computational Cognitive Science (DCS) within the Tilburg School of Humanities and Digital Sciences. Check the website of the ISMT group for more details: https://csai-ismt.github.io

Local Organizers

Dimitar Shterionov	Associate Professor, ISMT Group, DIS, TiU
Eva Vanmassenhove	Assistant Professor, ISMT Group, DCS, TiU
Mirella De Sisto	Assistant Professor, ISMT Group, DCS, TiU
Fred Blain	Assistant Professor, ISMT Group, DIS, TiU
Javad Pourmostafa	Lecturer, ISMT Group, DIS, TiU
Chiara Manna	PhD Student, ISMT Group, DCS, TiU
Lisa Lepp	PhD Student, ISMT Group, DIS, TiU
Argentina Anna Rescigno	PhD Student, University of Naples
Karin Berkhout	Management Coordinator, DCS, TiU
Eva Verschoor-Suitela	Management Coordinator, DIS, TiU
Sacha Elzinga	Management/Office Assistant, DCS/DIS, TiU
Sarah Blain	Management/Office Assistant, DCS/DIS, TiU

Student Volunteers

Luuk van Elewout, Tilburg University; Marie Dewulf, University of Antwerp; Aurora Trapella, University of Turin and Ghent University; Gül Karabaş, Tilburg University; Xiaoxiao Yang, Tilburg University; Isabelle van Stiphout, Tilburg University; Ozan Safak Kocak, Tilburg University; Claudia Yanez, Tilburg University; Adelina Violeta Sandu, Tilburg University; Melat Assefa, Tilburg University; Yuhuan Lee, Tilburg University; Nilsu Tari, Tilburg University; Mihaela Petrova, Tilburg University; Noa Muste, Tilburg University; Mathis van der Steen, Tilburg University; Olena Pazyuk, Tilburg University; Zahra Vafadar Nikjoo, Tilburg University; Nityaa Kalra, Tilburg University; Sumbul Syed, Tilburg University; Florence Bellemont, Leiden University; Rastislav Hronský, Tilburg University; Martijn van Leeuwen, Tilburg University; Maximos Christodoulou, Tilburg University; Alexandros Gkritzelis, Tilburg University

General Chair

Helena Moniz University of Lisbon / FLUL; INESC-ID, Portugal

Track Chairs

Research – Technical

Alina Karakanta	Leiden University
Ayla Rigouts Terryn	Université de Montréal / Mila

Research – Translators and Users

Manuel Lardelli	University of Padua
Natalia Resende	Trinity College Dublin

Implementations and Case Studies

Elena Murgolo	Custom.MT
Janiça Hackenbuchner	Ghent University

Products and Projects

Anna Zaretskaya	TransPerfect
Miquel Esplà-Gomis	University of Alicante

Workshops and Tutorials

Thierry Etchegoyhen	Vicomtech
Dagmar Gromann	University of Vienna

Contents

Keynote and sponsor presentations and tutorials

Large Language Models and Machine Translation: From Low-Resource to Unseen Languages

Dr. Rachel Bawden ALMAnaCH project-team, Inria Paris, France
Currently a fellow at PR[AI]RIE-PSAI research institution

Speaker Bio Dr. Rachel Bawden is a researcher in the ALMAnaCH project-team at Inria Paris, France. She is a specialist of Machine Translation (MT), having worked on contextual MT during her PhD at the LIMSI laboratory in France and MT for low-resource languages in her post-doc at the University of Edinburgh. She is currently working on a range of topics in MT and multilingual NLP, focusing mainly on language variation, both for historical and contemporary texts (for example user-generated content, dialectal variation), evaluation and resource creation. She is currently a fellow in the PR[AI]RIE-PSAI research institution.

Abstract Large language models (LLMs) have been offering new approaches to machine translation (MT). Much of today's research involves trying to tease out the underlying knowledge of the LLMs to improve translation quality, especially in scenarios where standard prompting does not lead to good results. For many of the world's low-resource languages, LLMs have not been the magic solution for translation, with new problems arising such as failure to translate in the right language and uncontrolled hallucination, and there remain significant challenges.

In this talk, I will be discussing several research directions in low-resource MT with LLMs that I recently published with colleagues. These include (i) the decomposition of sentences into simpler components to aid the search for useful few-shot examples, (ii) the creation of high quality synthetic parallel data for under-resourced languages and finally (iii) the explicit learning of translation from grammar descriptions, tested with encrypted and therefore unseen languages.

Flipping the Script: The Case for a Human-Initiated, AI-Augmented Translation Pipeline

Dr. Antonio Toral Universitat d'Alacant, Spain
Distinguished Researcher in Machine Translation

Speaker Bio Antonio Toral works as Distinguished Researcher in Machine Translation at the Universitat d'Alacant. Previously, he was an Associate Professor in Language Technology at the University of Groningen, where he coordinated the Computational Linguistics research group. Prior to that, he served as a postdoctoral researcher and research fellow at Dublin City University. He completed his PhD studies at the Universitat d'Alacant and the Istituto di Linguistica Computazionale.

His research interests include the application of machine translation (MT) to literary texts, MT for under-resourced languages and the computational analysis of translations produced by machines and humans. He coordinated the Abu-MaTran project, which was flagged by the European Commission as a success story and won the best paper award at MT Summit 2019 for his work on post-editese.

Abstract Over the last two decades the translation profession has witnessed a dramatic increase in the use of technology. Primary examples include translation memories and machine translation post-editing (MTPE), whose adoption has been primarily driven by productivity.

However, while MTPE is well-established and widely used, it presents important issues that affect both translators and the quality of the resulting translations.

In this talk, I will examine the main issues inherent in MTPE and propose an alternative translation pipeline that flips the roles, placing the translator before the machine. I will argue that, in such a setting, multi-agent AI can foster more informed translation decisions while safeguarding the translator's creative agency.

Finally, I will discuss why I think this approach is particularly suited for creative texts and peripheral languages, and also why it is not a far-fetched utopia, given current socioeconomic trends and developments.

Machine Translation in the Context of Subtitling and Dubbing: Challenges and Solutions

Dr. Evgeny Matusov (AppTek) This talk focuses on machine translation (MT) for audio-visual localization, and on subtitling and dubbing workflows in particular. We discuss practical challenges using examples from AppTek's real-life use cases. For dubbing, we examine MT in human-in-the-loop settings, highlighting the importance of fine-grained control mechanisms such as length constraints, pause transfer, and gender and formality customization. We also address practical challenges in text normalization for text-to-speech (TTS), which remain a bottleneck for high-quality synthetic dubbing; here, we show that MT training methods can be leveraged for this task. In the subtitling domain, we discuss MT integration with intelligent line segmentation (ILS) and compare large language model (LLM)-based post-editing with direct translation approaches, arguing for a smart selection of processing units. We further cover methods for learning from post-editing (PE) data to continuously improve system performance. Finally, we turn to live speech translation, contrasting traditional cascade systems with end-to-end speech LLMs. Key challenges such as handling multilingual speech on a single channel, preservation of instruction-following abilities in speech LLMs, and managing latency-quality trade-offs are analyzed. The talk provides an overview of current capabilities and open problems, emphasizing the role of controllability, efficiency, and human feedback in next-generation audiovisual MT systems.

Tutorial: Integrating Free NMT and LLMs into CAT Tools with MTUOC

Sergi Alvarez-Vidal and Antoni Oliver The landscape of machine translation (MT) has evolved dramatically since the advent of neural MT (NMT), which marked a breakthrough in translation quality and fluency. More recently, the rise of large language models (LLMs) has reshaped this landscape once again, introducing a new paradigm that merges translation, adaptation, and post-editing within a unified framework of multilingual text generation. These advances are expanding the possibilities for translators and language professionals, offering tools that can be tailored to domain-specific needs and local workflows. While commercial systems such as DeepL, Google Translate, ChatGPT, or Gemini dominate public attention, a vibrant ecosystem of free and open-source NMT and LLM resources has emerged. Projects like OPUS-MT, NLLB, and translation-oriented open LLMs such as Tower and Salamandra make it increasingly feasible to build and adapt high-quality MT pipelines for specific languages, domains, or institutional contexts. Yet, integrating these tools—each with its own dependencies and APIs—into professional computer-assisted translation (CAT) environments remains a technical challenge.

The MTUOC project addresses this gap by providing a comprehensive open-source framework that simplifies deployment and integration. This hands-on tutorial will guide participants in building and customizing their own tailored MT ecosystems using fully open and free technologies. Attendees will learn how to (1) set up the MTUOC-server, (2) deploy leading open models such as OpusMT and NLLB, (3) integrate translation-specialized LLMs (Tower, Salamandra) through MTUOC components, and (4) connect all these tools seamlessly within OmegaT, a widely used open-source CAT platform. By the end of the session, participants will have a fully operational and reproducible open-source translation workflow capable of combining neural MT and LLM-based translation within a professional environment. Since both MTUOC and OmegaT are distributed under the GNU-GPL license, the entire solution remains free, extensible, and adaptable to the needs of individual translators, research groups, and institutions.

Tutorial on Human Evaluation of Translation and Multilingual Tasks

Vilém Zouhar, Maike Züfle and Dominik Macháček Human evaluation is the gold standard for multilingual NLP but is frequently omitted due to operational complexity. This tutorial demonstrates how to design and execute rigorous human evaluation campaigns focusing on multilingual tasks (e.g. translation, multilingual, or multimodal evaluation), covering the full lifecycle: data selection, protocol selection, setting up the evaluation campaign, annotator management, and analysis of results. The practical focus will be on setting up the evaluation campaign with examples, while the theoretical part will be devoted to modern statistical techniques, such as turning pairwise preferences into absolute scores, or modelling benchmarking competitions. At the end, participants will have detailed knowledge of how to design, implement, and run high-quality human evaluation in their scientific and industry applications.

Translation Evaluation Tools for Everyone: A Hands-On Tutorial for Freelancers and Small LSPs

Yuri Balashov Tutorial materials: https://github.com/YuriBalashov/eamt2026-eval-tutorial

Before the tutorial, please complete the quick start steps outlined here: https://github.com/YuriBalashov/eamt2026-eval-tutorial#quick-start-before-the-tutorial

A half-day hands-on tutorial which introduces automatic translation quality evaluation methods and tools to an audience that has not traditionally used them: freelance translators, small language service providers (LSPs), translation project managers, and translation studies students with little or no programming experience. Evaluation techniques long reserved for MT research and large-scale industry workflows are now within reach of individual language professionals, thanks to two converging developments: user-friendly no-code web toolkits such as MATEO (Vanroy et al., 2023), and modern large language models (LLMs) that can serve as on-demand coding partners. Building on the emerging concept of Translation Analytics, the tutorial unfolds in four parts. Part 1 surveys manual and automatic evaluation, from MQM and direct assessment to BLEU, chrF, TER, COMET, BLEURT, BERTScore, and current developments (xCOMET, MetricX, LLM-based metrics). Part 2 walks participants through MATEO, where they run BLEU, chrF, TER, and COMET on multilingual evaluation sets in EN–DE, EN–RU, EN–JA, or EN–ZH. Part 3 interprets the outputs: score tables, confidence intervals, and sentence-level COMET in Excel. Part 4 introduces lightweight statistics (means, variance, p-values; Pearson, Spearman, and Kendall correlations) using Excel and LLM-assisted Python. All materials are openly available in a GitHub repository.

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, Finland. European Association for Machine Translation.

Products and Projects

Literacy-Grounded and Industry-Oriented Translation Training with LT-LiDER

Janiça Hackenbuchner[1], María Isabel Rivas Ginel[2], Joss Moorkens[2], Sheila Castilho[2], Nora Aranberri[3], Sergi Álvarez Vidal[4], María do Campo Bayón[4], Ralph Krüger[5]

[1]Ghent University [2]Dublin City University [3]University of the Basque Country UPV/EHU
[4]Universitat Autònoma de Barcelona [5]TH Köln

janica.hackenbuchner@ugent.be, nora.aranberri@ehu.eus, ralph.krueger@th-koeln.de
{joss.moorkens, sheila.castilho, isabel.rivasginel}@dcu.ie
{Sergi.Alvarez, maria.docampo}@uab.cat

Abstract

The Erasmus+-funded international research consortium LT-LiDER develops a range of digital training resources which are grounded in the overarching frameworks of digital and AI literacy and oriented towards practical application contexts in the language and translation industry. These resources can be implemented on a component basis or as a complete curriculum in higher-education language and translation classrooms.

1 Introduction

The language and translation (L&T) industry is currently experiencing another automation cycle fuelled by general-purpose artificial intelligence (GPAI) technologies in the form of large language models (LLMs). Given the wide scope of application of these GPAI models, the automation pressure felt in the industry may be higher than in previous automation cycles. According to L&T industry surveys, about 50 per cent of language service providers (LSPs) are already implementing GPAI LLMs for achieving efficiency gains or cost savings and for providing new capabilities or service offerings (ELIS Research, 2025). Naturally, L&T industry stakeholders are required to adapt to these changes (Rivas-Ginel et al., 2026). However, retraining and upskilling in terms of AI skills often remains an item on stakeholders' to-do lists, which underscores the need for initiatives fostering the development of translation-oriented digital and AI literacy (Krüger, 2025).

In late 2023, the Erasmus+-funded international research consortium LT-LiDER (Language and Translation: Literacy in Digital Environments and Resources)[1] launched such an initiative aimed at developing digital training resources grounded in digital and AI literacy and oriented towards the needs of the digitalised, datafied and increasingly AI-saturated L&T industry. These resources address "the many technical and ethical questions about use and reuse of data, appropriate and risky uses of technology, and positive and negative impacts on the many stakeholders in a translation process" (Moorkens et al., 2024, 54). This paper provides an overview of the current outputs produced by LT-LiDER, which stands in a tradition of related initiatives such as MultiTraiNMT[2], DataLitMT[3], adaptMLLM[4] and UPSKILLS[5].

2 Implementations and Outcomes

The current outputs of LT-LiDER are outlined below.

- **Interviews:** To obtain a multifaceted expert perspective that could serve as a baseline for developing the training resources, LT-LiDER interviewed 29 L&T industry stakeholders, including in-house and freelance translators and interpreters, LSP managers, heads of research and academics involved in translator education. Interviews were conducted in six different languages and extracts are available online.[6]

[1] https://lt-lider.eu/
[2] https://www.multitrainmt.eu/
[3] https://itmk.github.io/The-DataLitMT-Project/
[4] https://github.com/adaptNMT
[5] https://upskillsproject.eu/
[6] https://lt-lider.eu/interviews-general/

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 7–9
Tilburg, The Netherlands, June 2026.

- **Language and Technology Map:** Based on a synthesis of these expert interviews, LT-LiDER developed a comprehensive Language and Technology Map[7], which identifies technologies employed in the different phases of the translation process and maps these to relevant skills and competences (Secară et al., 2025).
- **ProMut** is a didactics-focused machine translation (MT) platform[8] for teaching users the essentials of MT, including managing training corpus data as well as training, using and evaluating MT engines. The platform is an evolution of MultiTraiNMT's MutNMT platform. ProMut is built on MarianNMT and provides an interface to the OPUS corpus collection to facilitate access to MT training corpora. Evaluation can be performed via BLEU, chrF3, TER and COMET (Sánchez-Gijón and Ramírez-Sánchez, 2025).
- **LT-LiDER book:** One of the main outputs of LT-LiDER is a comprehensive textbook comprising 18 chapters on diverse topics such as data, MT and AI literacy, data management, basic programming skills, implementing advanced NMT/LLM tools, speech technology, best practices for prompting, evaluation and fine-tuning. Chapters are grounded in relevant theory and current practices in translation technology and include specific hands-on exercises. Following an open peer-review phase, the book is scheduled to be published in late 2026.
- **Hands-on exercises:** The hands-on exercises mentioned above are compiled in a central GitHub repository.[9] This repository provides an easy overview of all activities, which can be accessed readily for learning and teaching translation-oriented digital and AI literacy.
- **Training events:** LT-LiDER regularly holds training events to inform interested audiences about current project activities. For example, in 2025 LT-LiDER held a "Train the Trainer Workshop"[10] in Vienna and a Translating Europe Workshop (TEW) in Grenoble. These will be followed by future TEWs and related events, as well as by the first International Workshop on Teaching AI-based Translation and Technologies[11] (TAITT) held at EAMT 2026.

3 Finalising LT-LiDER

In its final year, LT-LiDER will focus on publishing the textbook and organising further workshops and events. Crucially, the project invites critical stakeholder feedback on its outputs in the spirit of L&T industry stakeholder collaboration and in order to continually fine-tune its learning resources.

4 Acknowledgements

This project is funded from 2023-26 by Erasmus+ as a cooperation partnership in higher education, grant number KA220-HED-15E72916. All members of the consortium are listed on the LT-LiDER website.

References

ELIS Research. 2025. European language industry survey 2025.

Krüger, R. 2025. Artificial intelligence literacy for the language and translation industry - conceptual foundations, operationalisation, acquisition, measurement.

Moorkens, J., P. Sánchez-Gijón, E. Simon, M. Urpí, N. Aranberri, D. Ciobanu, A. Guerberof-Arenas, J. Hackenbuchner, D. Kenny, R. Krüger, M. Rios, I. Ginel, C. Rossi, A. Secară, and A. 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, UK, June. European Association for Machine Translation (EAMT).

Rivas-Ginel, I., J. Hackenbuchner, A. Secară, C. Rossi, and R. Krüger. 2026. A technology-oriented mapping of the language and translation industry – analysing stakeholder values and their potential implication for translation pedagogy.

Sánchez-Gijón, P. and G. Ramírez-Sánchez. 2025. ProMut: The evolution of NMT didactic tools. In *Proceedings of Machine Translation Summit XX: Volume 2*, pages 91–92, Geneva, Switzerland, June. European Association for Machine Translation.

Secară, A., I. Rivas Ginel, A. Toral, A. Guerberof, D. Ciobanu, J. Brockmann, C. Plieseis, R.-M. Chereji, and C. Rossi. 2025. LT-LiDER Language

[7] https://lt-lider.eu/mapa-de-las-tecnologias-de-la-lengua/
[8] https://promut.uab.cat/
[9] https://github.com/LT-LIDER/LT-LIDER
[10] https://haitrans.univie.ac.at/research/lt-lider-train-the-trainer-workshop/
[11] https://sites.google.com/view/taitt2026/home

Technology Map – Technologies in translation practice and their impact on the skills needed. Technical report, Universität Wien.

MaTIAS – Machine Translation to Inform Asylum Seekers: final results

July De Wilde, Anaïs Wouters, Arda Tezcan, Simon Van den Meersschaut, Katrijn Maryns and Lieve Macken
Department of Translation, Interpreting and Communication
Ghent University
Belgium
{firstname.lastname}@ugent.be

Abstract

This paper reports on the final stages of the MaTIAS project. A functional prototype of the multilingual notification tool was deployed across seven Belgian reception centres, accompanied by training and support. Feedback was gathered through interviews and surveys. Two rounds of machine translation evaluation revealed considerable differences in quality across languages. The translation quality of Tigrinya in particular was deemed too low to be usable.

1 Context and project overview

Asylum reception centres typically accommodate a highly diverse population with a variety of linguistic backgrounds, inevitably yielding communication barriers. Staff members frequently have to disseminate time-sensitive information ranging from administrative appointments and house rules to maintenance updates and transport disruptions, but they often face language obstacles. As a result, residents may misunderstand expectations or miss essential updates, leading to communication gaps.

To address this issue, a research team at Ghent University has developed MaTIAS (Machine Translation to Inform Asylum Seekers), a multilingual notification tool which enables staff to send messages via WhatsApp to residents in their native languages. Designed for one-way communication, the tool ensures the rapid delivery of simple informational messages (such as updates about activities or technical issues) to groups of residents, while more complex and sensitive interactions requiring resident input are handled through other language support channels.

Messages are composed in Dutch, French or English via a user-friendly web interface and automatically translated by ModernMT, customised with a context-specific translation memory.

The 2023–2025 project has been carried out in close collaboration with Fedasil, the federal agency for the reception of asylum seekers in Belgium. It has been co-financed by the European Commission under the Asylum, Migration and Integration Fund (AMIF 093-133). This contribution focuses on the final phases of the project: the deployment and evaluation of a functional prototype in several asylum centres, and the usability of the automatic translations.

2 Prototype deployment and evaluation

The MaTIAS prototype consists of a web platform that enables Fedasil staff to translate messages from Dutch, French, or English into 14[1] additional target languages, including several low-resource languages. The platform supports resident registration, message composition, and communication history review through an intuitive interface designed to minimize staff workload. A key feature is its integration with Fedasil's database, automatically synchronizing resident data every 24 hours to eliminate manual updates.

For the evaluation phase, MaTIAS was deployed on Fedasil servers to assess its usability in a live operational environment. We adopted a phased rollout strategy across seven reception centers in Flanders, Brussels, and Wallonia, to ensure technical stability and manageable user support. This phased approach integrated demos for each

[1] These 14 languages were selected by Fedasil based on current needs.

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 11–12
Tilburg, The Netherlands, June 2026.

new center added to the test pool, allowing the research team to provide tailored training. The demo sessions took place between October and November 2025 and were either given on-site or online via Microsoft Teams. A separate Microsoft Teams channel per center provided direct feedback and support, especially for initial technical issues with server and database connections.

Employee feedback was gathered through online interviews and a Qualtrics survey, involving 13 and 9 participants respectively. Staff members rated MaTIAS as user-friendly, highlighting the ease of resident registration, recipient selection, and message sending. Additional features like message reuse and delayed sending were also well utilized. The tool was primarily used for practical and organizational communication. Analysis showed that most translations were into French, English, Tigrinya, Turkish, Arabic, and Somali. Feedback from residents, collected by centre staff, was very positive: most residents found the system clear, user-friendly and accessible. However, analysis of the data revealed that not all registered residents read the messages.

3 Machine translation quality

The quality and usefulness of the automatically translated messages were assessed in two evaluations. The first, conducted in May–June 2025, covered all 14 target languages and involved 28 language experts, using 90 messages collected during the project's preparatory ethnographic phase. Results showed substantial variation in quality across languages, with translations into Tigrinya, Somali, and Persian remaining suboptimal. For detailed results, we refer to the publication of Macken et al. (2025). The second took place in March 2026 and focused on five target languages: Tigrinya, Arabic, Somali, Pashto, and Persian. For each language, a set of 40 actual messages collected during the test phase was assessed by one language expert. Although the Arabic, Persian and Somali translations were acceptable, they sometimes required effort to understand. The Pashto messages, on the other hand, were overly formal and poorly aligned with spoken language. The Tigrinya results were particularly poor, often losing the core meaning.

4 Discussion and recommendations

The results of the test phase demonstrate the potential of MaTIAS to enhance multilingual communication in reception centres. While the tool has demonstrated its potential to improve communication and reduce language barriers, several key considerations must be addressed to ensure successful integration and scalability: (1) Tools must be embedded in existing workflows to ensure a manageable workload for staff. Automating resident registration via database synchronisation proved essential for maintaining data consistency and minimising manual administrative effort. (2) GDPR and data protection compliance are paramount. Clear communication regarding data storage and protection is essential for maintaining trust and upholding ethical standards when working with vulnerable populations. The paid version of ModernMT was found to offer sufficient privacy safeguards[2], and staff were explicitly instructed through training and the user manual never to include personal data in messages. (3) As the system scales, it is vital to define clear roles and communication protocols. Specific guidelines on message frequency and thematic relevance must be established to prevent information overload and ensure the tool remains effective. (4) While ModernMT provides a strong baseline, achieving high-quality translation for certain low-resource languages remains challenging. Recent research (Moerman et al., 2026) has explored whether the in-context learning capabilities of large language models (LLMs) could offer a solution, though results have been variable.

Of course, the full impact of MaTIAS on efficiency and time savings, and areas for improvement, will only become clear after full implementation across the entire Fedasil reception network.

References

Macken, Lieve, Margot Fonteyne, Arda Tezcan, Ella van Hest, Katrijn Maryns, and July De Wilde. 2025. Machine translation in asylum reception centres : system selection and multilingual quality evaluation. *Revista Tradumàtica. Tecnologies de la Traducció*, (23):326–349.

Moerman, Thomas, Arda Tezcan, and Lieve Macken. 2026. Multilingual Communication in the Asylum Context: Evaluating LLM-Based Machine Translation with Fuzzy-Match Augmentation and Adaptive NMT across Resource Conditions under Low-Data Constraints. In *Proceedings of the 26th Annual Conference of the European Association for Machine Translation*, Tilburg, The Netherlands.

[2]See ModernMT's privacy policy: `https://www.modernmt.com/privacy`

OSCAIL - Open Science Communication through AI in EU Languages

Sheila Castilho,[1] **Susanna Fiorini,**[2] **Lynne Bowker,**[3] **Petr Motlicek,**[4] **Joss Moorkens,**[1] **Lieve Macken,**[5] **Dairazalia Sanchez-Cortes,**[4] **Janne Pölönen,**[6] **Sami Syrjämäki,**[6] **Mikael Laakso,**[6] **Mark Fishel,**[7] **Anastasia Stasenko**[8]
[1]Dublin City University, [2]OPERAS, [3]Université Laval, [4]Idiap Research Institute, [5]Ghent University, [6]Federation of Finnish Learned Societies, [7]University of Tartu, [8]pleias

Abstract

The Anglocentric nature of scholarly communication has many implications, such as limiting publication, discoverability and access from other language communities (even for major languages); putting minoritized languages at risk in the academic domain; and excluding many from peer review. The OSCAIL project addresses these challenges by exploring how machine translation (MT) enhanced by large language model (LLM)–based technologies can support access to scientific knowledge. Outputs will include evaluation datasets, protocols and best practices for MT in scholarly communication, and a prototype integration of MT tools into Open Journal Systems, the world's most widely used open-source scholarly publishing platform.

1 Introduction

The dominance of English in scholarly communication creates structural barriers to both participation in research and access to scientific knowledge. Researchers working in other languages often face limitations in publishing and disseminating their work internationally (Amano et al., 2023), while important research produced in languages other than English may remain less visible or accessible (Hannah et al., 2024). This linguistic imbalance also affects peer review processes and the discoverability of scientific outputs, and limits the accessibility of research for non-specialist audiences.

MT is already widely used informally by researchers, reviewers, and readers to access scientific content across languages (Fiorini, 2024). However, the effective and responsible integration of MT into scholarly communication workflows remains insufficiently explored (Ayeni et al., 2025). Previous studies and initiatives have shown that MT can help users understand the general meaning of scientific texts, but challenges remain regarding terminology, domain-specific accuracy, and the preservation of meaning in specialised contexts (Macken et al., 2024). In addition, ethical considerations related to transparency, governance, and the responsible use of artificial intelligence (AI) technologies (Moorkens et al., 2024) remain central to the adoption of automated translation tools in scholarly publishing.

The OSCAIL project (Open Science Communication through AI in EU Languages) aims to address these challenges by exploring how advances in MT and LLMs can support more inclusive multilingual communication in science. The project is funded under the CHIST-ERA ERA-NET Call 2025 "Science in Your Own Language", with the following funding agencies: Research Ireland, Swiss National Science Foundation (SNSF), The Academy of Finland (AKA), French National Research Agency (ANR), Estonian Research Council (ETAg). OSCAIL will run for 36 months (1 May 2026–30 April 2029) and will bring together an interdisciplinary and multilingual consortium including:

- Dublin City University (Ireland, coordinator)
- OPERAS Research Infrastructure (France)
- Idiap Research Institute (Switzerland)
- Pleias (France)
- University of Ghent (Belgium)
- Université Laval (Canada)

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 13–14
Tilburg, The Netherlands, June 2026.

- University of Tartu (Estonia)
- Federation of Finnish Learned Societies (Finland)

The consortium combines expertise in MT, natural language processing, LLM fine-tuning, translation evaluation, scholarly publishing infrastructures, and research policy.

Central to the project is the investigation of MT in three practical use cases inspired by real scholarly communication workflows. First, the project will explore the use of MT to support multilingual peer review, enabling reviewers to engage with manuscripts and author responses in their preferred working language while allowing authors to write and respond in the language with which they are most comfortable. Second, OSCAIL will investigate how MT can improve the discoverability of scientific publications, for example through the translation of metadata and key descriptive elements to support multilingual indexing and search. Third, the project will examine how MT and automatic summarisation can improve the accessibility of research outputs, including the translation and plain-language summarisation of scientific articles for broader audiences.

To support these goals, OSCAIL will gather domain-specific datasets (still under discussion) in project languages representing different levels of technological support according to the Digital Language Equality framework. These data will be used to develop, adapt, and evaluate MT systems, including approaches based on large language models, for the translation of scientific texts.

2 OSCAIL Objectives

OSCAIL pursues five main objectives:
Advance MT for scientific content by developing and benchmarking MT systems adapted to the linguistic and structural characteristics of scholarly texts.
- Support high-, mid-, and low-resource European languages through domain-adapted models and evaluation across diverse language pairs.
- Enable multilingual peer review workflows allowing authors and reviewers to work in their preferred languages without compromising the rigour of the review process. - Develop evidence-based evaluation frameworks combining automatic and human evaluation tailored to different user roles in scholarly communication.
- Promote responsible and inclusive scholarly AI, addressing ethical, legal, and governance considerations in the deployment of AI-based translation technologies.

3 Expected Results

As the project has recently been funded and just started, results are expected during the implementation phase. Planned outcomes include the creation of multilingual evaluation datasets and corpora for scientific MT, reproducible evaluation protocols and best-practice guidelines for MT deployment in scholarly communication, comparative analyses of MT systems across different scholarly use cases, and a prototype integration of MT functionalities into Open Journal Systems. The project will also produce ethical guidelines and recommendations to support responsible AI-driven translation in scholarly publishing and to promote more inclusive and multilingual access to scientific knowledge.

References

Amano, Tatsuya, Valeria Ramírez-Castañeda, Violeta Berdejo-Espinola, Israel Borokini, Shawan Chowdhury, Marina Golivets, Juan David González-Trujillo, Flavia Montaño-Centellas, Kumar Paudel, Rachel Louise White, and Diogo Veríssimo. 2023. The manifold costs of being a non-native english speaker in science. *PLOS Biology*, 21(7):e3002184.

Ayeni, Philips, Emanuel Kulczycki, and Lynne Bowker. 2025. Machine translation and scholarly publishing: A scoping review. *Canadian Journal of Information and Library Science*, 48(1):123–145.

Fiorini, Susanna. 2024. Exploratory studies for the creation of a technology-aided collaborative translation service in open scholarly communication.

Hannah, Kelsey, Neal R. Haddaway, Richard A. Fuller, and Tatsuya Amano. 2024. Language inclusion in ecological systematic reviews and maps: Barriers and perspectives. *Research Synthesis Methods*, 15(3):466–482.

Macken, Lieve, Vanessa De Wilde, and Arda Tezcan. 2024. Machine translation for open scholarly communication: Examining the relationship between translation quality and reading effort. *Information*, 15(8):427.

Moorkens, Joss, Sheila Castilho, Federico Gaspari, Antonio Toral, and Maja Popović. 2024. Proposal for a triple bottom line for translation automation and sustainability. *The Journal of Specialised Translation*, (41):2–25.

AIDA Agents: A Multi-Agent Translation Platform with Context-Aware Quality Control

Emanuele Di Rosa **Piotr Peszynski**
DATAmundi.ai
{emanuele.dirosa,piotr.peszynski}@datamundi.ai

Abstract

We present AIDA Agents, a multi-agent translation platform that orchestrates LLM-based agents—for translation, rating, post-editing, and re-rating—delivering context-aware translations without model fine-tuning. Optional retrieval-augmented generation (RAG) injects translation memories, terminology, and style guidelines at every pipeline stage. On WMT24++ (Deutsch et al.(2025)) (11 languages), AIDA Agents outperforms all systems on 10 of 11 pairs. On an industrial benchmark, 70–98% of segments are publication-ready without human post-editing. The platform is deployed with native XLIFF integration.

1 Introduction

Enterprise localization teams face a persistent trade-off: NMT engines require expensive fine-tuning and substantial post-editing, while LLMs show quality gains but lack workflow integration. Recent research on multi-agent translation systems (Wang et al.(2025)) demonstrates the potential of agent-based pipelines but lack production integration. AIDA Agents bridges this gap by orchestrating specialized LLM agents in a pipeline with native XLIFF integration. The platform is currently deployed internally at DATAmundi to support production localization for enterprise clients, offered as a proprietary service through client engagements, with potential, public commercial availability in the upcoming months. This paper describes the system, deployment configurations, and preliminary evaluation on general-purpose, WMT24++ translation benchmark, and an industrial one.

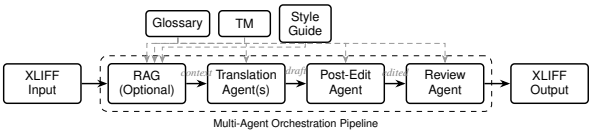


Figure 1: AIDA Agents pipeline. Dashed arrows: context injection via prompt refining and optional RAG.

2 System Description

Figure 1 illustrates the pipeline. Source content arrives from the integrated CAT tool; an optional RAG stage retrieves TM matches (via multilingual embeddings), glossary entries with approved/forbidden terms, and style instructions—enabling domain adaptation without fine-tuning. This context is injected at every pipeline stage, ensuring consistent application of client requirements. Up to three parallel translator agents produce candidates; an Arbiter/Rater, not represented in figure 1, enabled when multiple parallel translations are generated, evaluates each on MQM-inspired criteria (accuracy, fluency, terminology, style) with customizable weighted scoring, producing an explainable quality report, and selects the best one. A Post-Editor Agent refines it using the report, all candidates, and guidelines. A second Arbiter/Rater re-evaluates and produces the final output, returned to the CAT tool with XLIFF structure preserved.

Three product tiers address different quality-cost trade-offs: **AIDA LLM** (3 calls/segment): 1 translator + 1 post-editor + 1 rater; best quality-cost balance. **AIDA MultiLLM** (up to 6 calls): up to 3 parallel translators + rater + post-editor + re-rater; maximum quality through diversity. **AIDA NMT** (3 calls): 1 NMT engine + 1 LLM post-editor + 1 LLM rater; lowest cost, terminology can be injected at the post-editor step. The platform supports XLIFF format and CAT tool connectors.

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 15–16
Tilburg, The Netherlands, June 2026.

System	JA	ZH	DE	FR	ES	IT	IN	NO	VI	SW	PT
AIDA MultiLLM	27.9	47.2	31.5	**40.4**	**46.1**	43.6	34.8	45.1	37.0	41.5	42.0
AIDA LLM	29.6	41.5	33.7	40.1	44.5	**44.1**	35.9	45.5	37.1	43.6	41.7
AIDA NMT	**30.2**	**48.1**	**34.7**	38.7	44.8	43.2	**36.6**	**47.1**	**37.3**	42.8	**42.6**
GPT-4o	25.7	42.8	33.8	39.2	41.3	40.5	33.2	45.1	34.3	**44.9**	41.6
DeepL	26.3	39.2	33.4	38.2	45.5	42.9	32.5	44.1	–	44.1	41.3
Google Tr.	25.3	41.5	34.5	37.2	43.0	41.5	33.5	43.5	34.9	41.7	42.2
Gemini-1.5-Pro	23.7	45.7	31.8	36.0	39.4	37.7	31.4	40.9	29.7	38.0	39.5
Claude-3.5	23.3	38.6	30.8	37.7	39.0	38.8	31.4	40.5	30.8	41.9	40.6
Microsoft Tr.	22.1	35.8	31.2	31.7	41.2	41.2	30.9	46.1	30.9	42.8	40.9
OpenAI o1	23.6	38.7	31.8	33.3	41.2	36.2	30.0	43.0	31.0	40.9	39.0
Unbabel Tower	23.5	40.6	32.0	37.4	41.3	41.0	–	–	–	–	41.2
Δ AIDA vs. Best	**+3.9**	**+2.4**	**+0.2**	**+1.2**	**+0.6**	**+1.2**	**+3.1**	**+1.0**	**+2.4**	–1.3	**+0.4**

Table 1: WMT24++ BLEU (EN→target). AIDA Agents best on 10/11 languages. No TM/terminology available. Results for all 16 systems available upon request.

3 Experimental Results

3.1 WMT24++ Benchmark (11 Languages)

Table 1 reports BLEU on WMT24++ (Deutsch et al.(2025))—960 segments, 4 domains, no TM/terminology—isolating the multi-agent orchestration effect. AIDA Agents achieves the top score on 10/11 languages (up to +3.9 on Japanese). With RAG-injected resources, gains are larger (§3.2).

3.2 Industrial Benchmarks

Table 2 reports results on 96 segments per language (technical IT documentation, full TM/terminology/style guides, 3 linguists per language, blind 5-point scale) previously translated in a project by a competitor company involving human professional post-editors. Our goal is to show preliminary results about the added value of our automated translation system compared to a full human-in-the-loop post-editing process.

	BLEU			Good/Exc. (%)		
Lang.	MS	Comp.	AIDA	MS	Comp.	AIDA
JA	24.0	52.8	**59.3**	12.2	85.7	69.4
ZH	24.9	43.1	**45.2**	62.2	82.7	77.6
FR	42.2	47.6	**48.8**	60.2	94.9	90.8
KO	48.1	54.2	**56.9**	73.4	82.9	81.6
DE	55.1	56.1	**57.0**	82.7	91.9	**98.0**
ES	58.5	64.4	**65.3**	92.9	95.9	94.9

Table 2: Industrial benchmark. MS = Microsoft NMT; Comp. = commercial system + human post-editing; AIDA = AIDA MultiLLM (automated).

AIDA MultiLLM achieves the highest BLEU on all six languages; 70–98% of segments are publication-ready without human intervention, even surpassing the human-post-edited competitor for German (98.0% vs. 91.9%). Even though AIDA Agents workflow is automated suggesting a potential 50–70% reduction in post-editing effort, professional human post-editing is fully integrable. The experimental evaluation reported here focuses on general-purpose translation; an extended evaluation on terminology-constrained tasks and a specialized architecture have been presented in (Di Rosa(2026)).

4 Conclusion

Unlike standalone NMT engines or research multi-agent systems, AIDA Agents is a deployed orchestration platform with native CAT/TMS integration requiring no fine-tuning. It can incorporate any NMT or LLM backend and enables immediate deployment on new domains and languages. It consistently outperforms state-of-the-art MT across 11 languages on WMT24++ and delivers 70–98% publication-ready output on industrial data, even surpassing human-post-edited baselines for several language pairs. A live demo can be presented at the conference.

References

Daniel Deutsch et al. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of ACL 2025*, pages 12257–12284.

Emanuele Di Rosa. 2026. Multi-agent orchestration for terminology-constrained machine translation in industrial localization. In *Proc. of ACL 2026: Industry Track*. To appear.

George Wang et al. 2025. MAATS: A multi-agent automated translation system based on MQM evaluation. arXiv:2505.14848.

VERA: A Platform for Automatic and Human Evaluation of Machine Translation

Sofía García González[1 2], **Inés Quintana Raña**[1], **Jorge N. Afonso Cabido**[1], **Alberto Hernández Lado**[1], **German Rigau Claramunt**[2], **Sheila Castilho**[3]

[1]imaxin, Santiago de Compostela, Spain
{firstName.lastName}@imaxin.com
[2]University of the Basque Country (EHU), Donostia, Spain
german.rigau@ehu.eus
[3]SALIS/ADAPT Centre, Dublin City University, Dublin, Ireland
sheila.castilho@dcu.ie

Abstract

We present VERA, an easy-to-use platform for machine translation (MT) evaluation, combining both automatic metrics and the Multidimensional Quality Metrics (MQM) Core human evaluation framework in a single web environment. It supports reference-based metrics, multi-user annotation, corpus export, and PDF reports with automatic and human evaluation results, including their correlations.

1 Introduction

The rapid evolution of MT has led to an increasing need for robust and comprehensive evaluation platforms. While automatic metrics offer scalability and efficiency, human evaluation remains the gold standard for assessing MT quality. However, existing tools address these approaches separately. Notable examples include MATEO[1] (Vanroy et al., 2023) which focuses on automatic metrics via a web interface (BLEU, chrF, TER, COMET, BERTScore, and BLEURT); MT–LENS[2] (Gilabert et al., 2025) which extends to bias and robustness analysis beyond quality scores, and Pearmut[3] (Zouhar and Kocmi, 2026) which is specialised in human evaluation frameworks such as MQM, DA, and ESA. This fragmentation results in disconnected workflows. In this paper, we present VERA, a novel evaluation platform that seamlessly integrates automatic and human evaluation frameworks within a unified web environment. Uniquely, VERA further assesses the correlation between automatic and human evaluation results, enabling meta-evaluation and deeper analysis of MT evaluation methodologies. By bridging these perspectives, VERA provides more holistic and reliable evaluations of MT systems, supporting both research and industry needs.

VERA is being developed through an industrial PhD collaboration between imaxin, the University of the Basque Country (EHU), and the ADAPT Centre as a proprietary internal tool.[4]

2 Tool Description

VERA is a web-based platform composed of a Next.js[5] frontend and a FastAPI[6] backend. It provides multilingual support through next-intl[7] and integrates with Authentik[8] as an external identity provider, while NextAuth[9] manages session handling within the application. This architecture enables secure authentication and reliable access control for multiple concurrent users.

Automatic Evaluation MT outputs are evaluated using standard reference-based metrics—BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), TER (Snover et al., 2006) via SacreBLEU (Post, 2018)—, and COMET (Rei et al., 2020). Statistical significance between models is measured using bootstrap resampling (Koehn, 2004) for each metric.

[1]Source: https://mateo.ivdnt.org/
[2]Source: https://github.com/langtech-bsc/mt-evaluation
[3]Source: https://github.com/zouharvi/pearmut

[4]Although VERA is an internal tool, it is available via institutional licensing to universities and research groups for a fee. All the information is available at: https://imaxin.com/en/vera
[5]Source: https://github.com/vercel/next.js/
[6]Source: https://fastapi.tiangolo.com/
[7]Source: https://next-intl.dev/
[8]Source: https://goauthentik.io/
[9]Source: https://next-auth.js.org/

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 17–18
Tilburg, The Netherlands, June 2026.

Human Evaluation VERA supports multiuser annotation following the MQM Core framework.[10] To make the human evaluation process more efficient, the platform integrates sampling strategies from Zouhar et al. (2025) to select informative evaluation subsets. Users can specify the proportion of the corpus to evaluate, after which the platform automatically selects the most representative segments.

Within the same interface, VERA also allows the creation and refinement of reference corpora. Existing references can be directly edited, and when no references are available, new ones can be generated either from scratch or by editing MT outputs.

Results Deployment Evaluation results (both automatic and human) are accessible directly within the platform, offering an interactive overview of MT model performance and metric correlations calculated using Pearson (Pearson, 1896) and Spearman (Spearman, 1904). A comprehensive PDF report can also be generated, and metric scores, reference datasets, and annotated corpora can be exported in CSV and JSON formats, enabling advanced analysis and seamless integration with external processing pipelines.

References

Gilabert, Javier García, Carlos Escolano, Audrey Mash, Xixian Liao, and Maite Melero. 2025. MT-LENS: An all-in-one Toolkit for Better Machine Translation Evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 51–60.

Koehn, Philipp. 2004. Statistical Significance Tests for Machine Translation Evaluation. In Lin, Dekang and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

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, Pennsylvania, USA, July. Association for Computational Linguistics.

[10]Source: `https://themqm.org/the-mqm-typology/`

Pearson, Karl. 1896. VII. Mathematical Contributions to the Theory of Evolution.—III. Regression, Heredity, and Panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, (187):253–318.

Popović, Maja. 2017. chrF++: words helping character n-grams. In Bojar, Ondřej, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, editors, *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark, September. Association for Computational Linguistics.

Post, Matt. 2018. A Call for Clarity in Reporting BLEU Scores. In Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, et al., editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

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, November. 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, USA, August 8-12. Association for Machine Translation in the Americas.

Spearman, Charles. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72–101.

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.

Zouhar, Vilém and Tom Kocmi. 2026. Pearmut: Human Evaluation of Translation Made Trivial. *arXiv preprint arXiv:2601.02933*.

Zouhar, Vilém, Peng Cui, and Mrinmaya Sachan. 2025. How to Select Datapoints for Efficient Human Evaluation of NLG Models? *Transactions of the Association for Computational Linguistics*, 13:1789–1811.

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

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

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

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

Abstract

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

1 Project Overview

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

Project Reference: ANR-25-CHR4-0003

Project Duration: January 2026 – December 2028

Partners and Funding Agencies:

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

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

2 Project Objectives

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

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

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

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

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 19–21
Tilburg, The Netherlands, June 2026.

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

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

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

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

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

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

3 Project Languages

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

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

4 Expected results

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

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

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

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

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

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

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

References

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

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

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

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

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

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

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

ARTICULATE: Science in your Own Language

Yolanda Vazquez-Alvarez[1], Matthew P. Aylett[1], Benjamin R. Cowan[2], Justin Edwards[3], Sanna Järvelä[3], Ioannis Konstas[1], Madeleine Steeds[2]
Heriot-Watt University, Edinburgh[1] / University College Dublin[2] / University of Oulu[3]
[1] {Y.Vazquez_Alvarez, M.Aylett, I.Konstas}@hw.ac.uk
[2] {benjamin.cowan, madeleine.steeds}@ucd.ie
[3] {justin.edwards, sanna.jarvela}@oulu.fi

Abstract

The ARTICULATE project is an ambitious and interdisciplinary initiative funded by the CHIST-ERA call 2025. Its vision is to revolutionize science education and democratize scientific knowledge beyond academia and English-speaking audiences through the integration of AI with self-regulated learning. The aim is to translate science not just across language but across language style, to create engaging spoken digital experiences. We present an introduction to this project, an overview of the consortium and research approach, and a number of expected impacts.

1 Introduction

A major barrier to disseminating and democratising scientific advances is access to languages that dominate scientific content but also the mismatch between the formal language of scientists and the everyday expressions understood by the general public. The ARTICULATE project aims to translate science, not just across language but across language style, not just to the written form but to an engaging spoken form. We have made enormous advances in machine translation (MT), Large Language Models (LLMs) and Speech technology. The current challenge for the research community is to integrate these technologies with other innovative techniques and research to produce engaging and compelling use cases (Aylett and Romeo, 2023). In this project, we regard a text translation of science material as a starting point. Our challenge is reframing this material into speech and an engaging dialog, producing a delightful multilingual solution which can translate science into everyday language that can be more accessible.

Fictive dialogs are an approach where an idea is communicated by turning it into a conversation. Previous work has shown that dialogs can communicate information more effectively and be more persuasive (Aylett et al., 2024). We aim to generate interactive fictive dialogue across multiple multi-lingual science texts, in order to make a concrete contribution to educational services, as well as to increase engagement from groups who may be marginalised by background or less mainstream native languages. Just as Plato used fictive dialogs to bring to life the work and philosophy of Socrates, we aim to bring modern science to life for citizens, undergraduates and postgraduates across Europe.

2 Project Description

2.1 Duration & Partners

The ARTICULATE project will be completed over a 24-month period, from March 2026 to 2028. The Partners involved are the following:

- Heriot-Watt University, Edinburgh, UK
- University College Dublin, Ireland (UCD)
- University of Oulu, Finland

2.2 Approach and Research Method

Our primary goal is to develop ARTICULATE, a platform that supports Self-Regulated Learning through the use of an ensemble of Conversational AI Agents speaking the language of science learners. We will investigate the use of English, Finnish and dialectal Finnish. Figure 1 outlines a typical fictive dialogue scenario with ARTICULATE. Two Conversational AI agents (Expert and Peer) engage in a dialogue over a scientific topic, while the human Learner is actively listening and participating. The tone and dialect is dynamically adjusted to accommodate for different learning profiles and engagement of the Learner.

Following Vygotsky's concept of the Zone of Proximal Development (ZPD) (Vygotsky and

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 23–24
Tilburg, The Netherlands, June 2026.

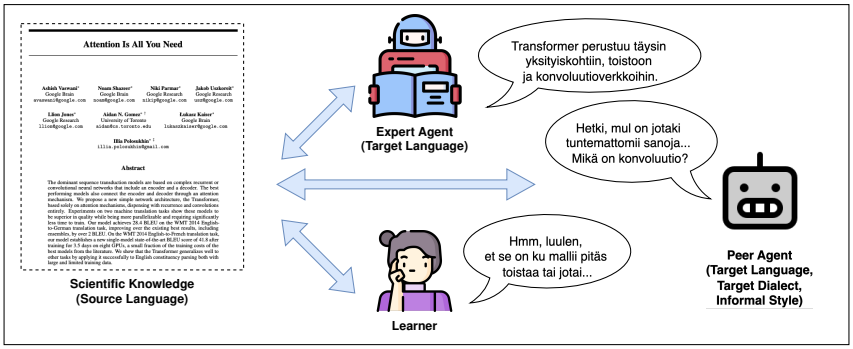


Figure 1: Typical ARTICULATE fictive dialogue in Finnish. The Expert Agent (translation in English: "Transformer is based solely on attention mechanisms, abandoning recurrence and convolutions entirely…") discusses with the Peer Agent (translation in English: "Wait, I have some unknown words… What is recurrence?") in their target native language about an input scientific domain, in this case a scientific paper on Natural Language Processing (Vaswani et al., 2017). The Expert Agent uses more sophisticated language, whereas the Peer Agent uses dialectical language. The Learner (translation in English: "Hm... I think it's when the model has to repeat or something..."), who studies the paper, interjects with their view also in the target language.

Cole, 1978), which suggests that *learning must be facilitated by support that is tailored to the learner's current abilities*, the interaction with the Learner is not driven by the AI Expert agent, but is offered as guidance in the form of a dialogue with the less knowledgeable Peer agent. The aim is to create a learning context in which a learner can co-create scientific knowledge by joining the conversation with their own questions, remarks, and ideas.

2.3 Project Objectives

There are four key project objectives:
1) Enable multilingual and inclusive science learning through high-quality document-level machine translation and develop accurate and context-sensitive translations of scientific materials into low-resource and dialectal languages (e.g. the Oulu dialect of Finnish).
2) Create two role-based AI agents (Expert and Peer) that engage learners in authentic, co-constructive scientific dialogues by grounding their responses in translated scientific documents and adapting language tone and style to suit learners' profiles.
3) Implement robust safety mechanisms to prevent misinformation and harmful content while maintaining factual scientific accuracy.
4) Build animated, speech-capable avatars that communicate in the target dialect, integrating advanced Automatic Speech Recognition (ASR), Text-to-Speech (TTS), and avatar.

3 Expected Impacts

The expected impacts of the ARTICULATE project are both educational and technological:

- Enhanced Access to Science Education in Low-Resource and Dialectal Languages (Educational, Societal).
- Empowered, Active Science Learners through Conversational AI (Educational, Technological).
- Advancement of Safe and Responsible AI in Education (Technological).
- Technological Breakthroughs in Multimodal, Multilingual AI Systems (Technological).
- Cross-Disciplinary Impact and Model for Collaborative AI Research.

Acknowledgements: This work is supported by the grant CHIST-ERA-25-SOL-01, and by the UK Research and Innovation UKRI3908.

References

[Aylett and Romeo2023] Aylett, Matthew Peter and Marta Romeo. 2023. You don't need to speak, you need to listen: Robot interaction and human-like turn-taking. In *ACM CUI '23*, pages 1–5.

[Aylett et al.2024] Aylett, Matthew Peter, Shiyi Tang, Xuanchen Li, Xinyang Liu, Chengcheng Liu, Ruiqing Li, Chenxi Meng, Zewen Qu, Sirui Wang, and Zechen Yang. 2024. Developing fictive dialogs for a classroom language learning conversational interface. In *ACM HAI '24*, pages 323–325.

[Vaswani et al.2017] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

[Vygotsky and Cole1978] Vygotsky, Lev Semenovich and Michael Cole. 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.

HERMeS:
Human Evaluation & Ranking of Multiple Systems

Rex VanHorn
Institute for Artificial Intelligence,
University of Georgia, Athens, Georgia, USA
ORCID ID: 0000-0002-3792-427
rex.vanhorn@uga.edu

Abstract

HERMeS is a lightweight human evaluation platform designed to streamline human evaluation and comparison of multiple MT systems across large translation sets. As a complement to existing evaluation tools, HERMeS focuses specifically on scalable comparison of many anonymized systems through a hybrid ranking and direct assessment workflow, using a practical approach that reduces cognitive load while maintaining data quality, security, and integrity.

1 Introduction

Reliable evaluation remains one of the central challenges in machine translation research. While automatic metrics provide convenient approximations of translation quality, human evaluation remains the gold standard. However, prior work has highlighted ongoing challenges in human MT evaluation, including annotator effort, consistency, and the difficulty of defining translation quality (Castilho, 2021). On the other hand, modern research settings frequently require comparisons across many systems, including commercial translation engines and large language models. In such scenarios, evaluation tools must balance methodological rigor with practical usability for human annotators.

This paper presents a human evaluation platform, HERMeS, designed to support scalable and reproducible, sentence-level evaluation of multiple MT systems, combining two complementary evaluation approaches: categorical ranking through bucket assignments and fine-grained direct-assessment scoring, thereby reducing evaluator fatigue while keeping evaluation quality consistent.

2 System Design

HERMeS follows a two-step workflow consisting of bucket assignment (coarse ranking) and direct assessment (numeric scoring). It was developed in support of our ongoing study and seeks to reduce the overhead of human evaluation while maintaining quality and integrity. The system presents source sentences alongside MT-/LLM-generated translations to efficiently collect human judgments.

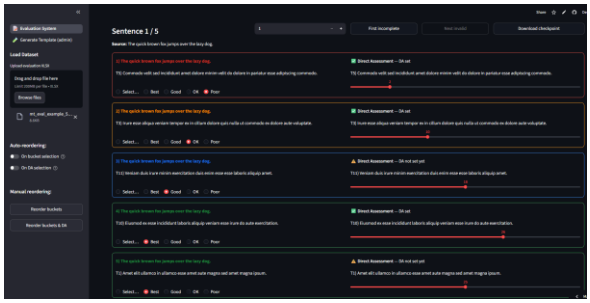


Figure 1: Screenshot of HERMeS

2.1 Evaluation Interface

The evaluators load the evaluation spreadsheet into HERMeS' user interface, through which one source sentence is displayed at a time as a translation card. The translation card shows the source sentence, translation, an evaluation bucket on the left, and a direct-assessment slider on the right. Each translation card allows the evaluator to provide two types of judgments (required).

Bucket Assignment

Evaluators first assign translations to one of four relative quality categories: Best, Good, OK, or Poor (these names are configurable). Bucketed ranking provides a low-effort entry point for evaluation by allowing annotators to make coarse

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 25–26
Tilburg, The Netherlands, June 2026.

comparative judgments before assigning precise scores. This reduces early-stage decision complexity and helps establish a stable internal ranking prior to fine-grained assessment. HERMeS can automatically group all translations by bucket upon ranking, or upon completion, such that all translations assigned to a common quality category are displayed together.

Direct Assessment (DA)

Evaluators then provide a numerical quality score on a fine-grained scale of the research team's choice; the range is configurable. In our current research, each bucket corresponds to a fixed scoring range of size 7 (e.g., 1–7 for Poor, 8–14 for OK, etc.). This score represents the perceived adequacy and fluency of the translation relative to the source sentence. The system can order the DA scores within a bucket upon assessment, and upon the evaluator's button click. Additionally, evaluators can make notes about the source sentence, its translations, or the evaluation itself, per source sentence.

While adequacy and fluency are distinct dimensions, prior work has shown they are often correlated in practice. In lieu of separate dimension scoring, HERMeS offers a hybrid workflow that integrates bucketed ranking and direct assessment, capturing both relative and absolute quality while reducing annotator burden, in a format that easily lends itself to final review. Note that scores can be reassigned at any time, and the system is not constrained by dataset size or text length.

2.2 Data Integrity

Source data are randomized, anonymized, and integrity-checked prior to evaluation. Using hash-based methods to validate the inputs and outputs enables the detection of tampering and the reliable merging of distributed annotations.

3 Evaluation Workflow

HERMeS' annotation workflow is designed to minimize evaluator friction while ensuring coherent data collection. Evaluators proceed through the evaluation sentences sequentially, completing rankings for each source sentence's translations. The interface supports automatic reordering of translations as bucket assignments are made, or upon completion, which helps maintain coherent rankings. Evaluators then provide direct assessment scores for translations in each of the buckets, with automatic or on-demand ordering, if desired. Evaluators then have the option to review all assessments together.

HERMeS is a stateless application that requires no login, and is implemented as a web-based interface built using the Streamlit framework,[2] and runs on the Streamlit Community Cloud platform.[3] Accordingly, the results are not saved by the system, but rather evaluators are reminded and encouraged to save their work by downloading 'checkpoints', which can be exported and saved at any time in spreadsheet format, thereby allowing annotation sessions to be paused and resumed without data loss, and with no cost to the researchers or evaluators. The software is open-source and available on GitHub.[4]

Upon completion of their annotations, evaluators save the final checkpoint spreadsheet and return it to the research team, which can then de-anonymize and de-randomize the results for analysis.

Appraise (Federmann, 2012) established a strong general-purpose foundation for manual MT assessment. HERMeS is designed as a complement to fine-grained annotation systems like Appraise, targeting large-scale, direct-assessment comparison of many anonymized systems in a single workflow. It was designed to reduce evaluator fatigue, provide distributed, spreadsheet-based portability, and offer built-in, hash-based integrity checks, without the need for coding or technical expertise. It was not designed to capture fine-grained error categories and is not intended for detailed error analysis. In an evaluation study using HERMeS (13 systems, 100 sentences; 1,300 translations), preliminary qualitative feedback suggests that a bucket-first workflow may reduce perceived effort and improve alignment and consistency.

4 References

Castilho, S. (2021, April). Towards document-level human MT evaluation: On the issues of annotator agreement, effort and misevaluation. In Proceedings of the workshop on human evaluation of NLP systems (HumEval) (pp. 34–45).

Federmann, C. (2012). Appraise: an Open-Source Toolkit for Manual Evaluation of MT Output. Prague Bull. Math. Linguistics, 98, 25–36.

[2] https://streamlit.io

[3] https://mt-eval.streamlit.app/?type=admin

[4] https://github.com/RexVH/St-MT-Evaluation-app

The MULTI-TRAD Project: Parallel Corpora and Multidimensional Analysis of Human, Machine and Post-Edited Translation in the Third Social Sector

M.M. Sánchez Ramos[1], D. Biber[2], C. Cano Fernández[1], I. Fuentes Pérez[1], D. González Pastor[3], L. Goulart da Silva[4], M. Yuri Himoro[5], D. Kenny[6], L. M. Mónaco[7], M. T. Ortego Antón[8], I. Peñuelas Gil[8], C. Plaza Lara[9], V. Redondo Astilleros[10], C. Rico Pérez[11], T. Salvador Blázquez[12], M. Shakir[13], F. J. Vigier Moreno[14], M. Aenlle Curras[15]

[1]Universidad de Alcalá, [2]Northern Arizona University, [3]Universitat de València, [4]Montclair State University, [5]UNED, [6]Dublin City University, [7]Universidade da Coruña, [8]Universidad de Valladolid, [9]Universidad de Málaga, [10]Universitat Oberta de Catalunya, [11]Universidad Complutense de Madrid, [12]UDIMA, [13]University of Münster, [14]Universidad Pablo de Olavide, [15]Independent Consultant

Abstract

Domain adaptation remains a major challenge for machine translation, particularly in institutional communication. This paper presents the MULTI-TRAD project, which develops English–Spanish parallel corpora for the Third Social Sector communication. The project integrates three complementary objectives: (i) the compilation of a domain-specific parallel corpus, (ii) the analysis of linguistic variation across human translation (HT), machine translation (MT), and post-edited (PE) texts using Multidimensional Analysis (Biber, 1988), and (iii) the development of a domain-adapted neural machine translation system. In particular, the project investigates how different translation processes give rise to distinct functional profiles, related to phenomena such as translationese and post-editese. This paper presents the project design and initial progress.

1 Introduction

Communication within the Third Social Sector plays a central role in global public discourse. Organizations in this sector regularly produce multilingual content aimed at informing the public, mobilizing support and documenting institutional activities (Tesseur, 2017). Despite their importance, these texts remain underrepresented in corpus-based studies of translation and multilingual communication. In the field of machine translation, large-scale English–Spanish parallel corpora such as Europarl or ParaCrawl have played a central role in training neural systems. However, these resources mainly reflect general or institutional domains and do not adequately capture the communicative practices of the Third Social Sector, limiting both domain adaptation and evaluation in this context.

The MULTI-TRAD project addresses this gap through the development of English–Spanish parallel corpora designed to investigate linguistic variation across translation modes and to support domain-adapted machine translation. In addition to its relevance for translation studies, the project contributes to ongoing efforts in machine translation to move beyond purely metric-based evaluation by incorporating linguistically informed approaches to analyzing MT output.

2 Project overview

MULTI-TRAD is a research project funded by the Spanish Ministry of Science, Innovation and Universities (Grant PID2024-157849OB-I00) running from 2025 to 2028. The project involves a consortium of academic institutions and collaborating Third Social Sector organizations (e.g., Fundación Abrazando Ilusiones, Caritas, and Plataforma del Voluntariado en España).

The project integrates three complementary objectives. The first objective is the compilation of an English–Spanish parallel corpus. Texts are collected in collaboration with Third Social Sector organizations and include institutional, advocacy, and web-based communication texts (Tesseur, 2017). The corpus is designed following a protocol that

Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 27–28
Tilburg, The Netherlands, June 2026.

aims to capture a wide range of communicative practices within the Third Social Sector. The corpus includes texts originally produced in both English and Spanish. The dataset is designed for Multidimensional Analysis (MDA), including aligned human translation (HT), machine translation (MT), and post-edited (PE) versions. The MT texts will be generated using the domain-adapted neural machine translation system developed within the project. The second objective is the linguistic characterization of translation processes using MDA (Biber, 1988). This approach enables the investigation of functional variation across HT, MT, and PE outputs, including phenomena such as translationese and post-editese. The third objective is the development of a domain-adapted neural machine translation system tailored to Third Social Sector discourse. The system is trained on in-domain parallel data compiled from institutional materials using the MTUOC framework and the Marian NMT toolkit (Junczys-Dowmunt et al., 2018).

3 Multidimensional Analysis

The analysis is based on MDA (Biber, 1988), a framework grounded in register theory. MDA identifies co-occurring linguistic features and groups them into functional dimensions of variation. The corpus is tagged using the Multi-Feature Tagger of English (MFTE, Le Foll and Shakir, 2023), an open-source tool for multivariate analysis of English corpora. For the Spanish sub-corpus, we are currently evaluating state-of-the-art NLP frameworks (such as spaCy) to build a custom pipeline capable of extracting the equivalent lexico-grammatical features.

4 Preliminary results

The project is currently in the corpus compilation and preprocessing stage. On the one hand, an English–Spanish parallel corpus is being compiled for multidimensional linguistic analysis, including texts representing different domains and their human translations (HT). The corpus design follows established practices in MDA, with an initial target of approximately 25 texts per domain across four domains (Tesseur, 2017) and three translation modes (HT, MT, and PE), resulting in a balanced dataset of around 300 texts. Each text contains around 1,000 words to ensure reliable analysis. The corresponding MT and PE versions will be generated once the domain-adapted neural machine translation system has been trained. On the other hand, in-domain parallel data are being collected and curated for the training of the neural machine translation system. Although both strands rely on data from the same domain, the datasets are designed for different purposes: the corpus used for MDA requires carefully controlled and aligned HT, MT, and PE versions, whereas the training data consist of a larger and more heterogeneous in-domain collection.

References

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Junczys-Dowmunt, Marcin et al. 2018. Marian: Cost-effective high-quality neural machine translation in C++. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation (WNMT)*, 129–135.

Le Foll, Elen and Shakir, Muhammad. 2023. MFTE Python (Version 1.0) [Software]. Available at: https://github.com/mshakirDr/MFTE

Tesseur, Wine. 2017. The translation challenges of NGOs: Professional and non-professional translation at Amnesty International. *Translation Spaces*, 6(2), 209–229.

Parallel Corpus Development Toolkit (*PCDT*): A Web-Based Platform for Multilingual Parallel Data Creation

Praveen Acharya
Dublin City University
acharyaprvn@gmail.com

Rupak Raj Ghimire
Kathmandu University
rughimire@gmail.com

Bipesh Subedi
Kathmandu University
bipeshrajsubedi@gmail.com

Prakash Poudyal
Kathmandu University
prakash@ku.edu.np

Balaram Prasain
Tribhuvan University
prasainbalaram@gmail.com

Bal Krishna Bal
Kathmandu University
bal@ku.edu.np

Abstract

This paper presents *PCDT*, a web-based platform for collecting sentence-aligned parallel corpora through a community-driven approach to support machine translation for under-resourced languages. The tool decentralizes the translation task to the target community and is subsequently reviewed by language experts.

1 Introduction

The purpose of the Parallel Corpus Development Toolkit (*PCDT*) is to create a web-based platform that streamlines the development of multilingual parallel corpora for machine translation systems, with Nepali as the primary language. The platform enables *administrators* to manage tasks and text corpora, *translators* to provide multilingual translations, and *reviewers* to verify translation quality, while also supporting the addition of new languages for future scalability. There are few systems (Zanata,[1] TEITOK,[2] Pootle,[3] WebAnno (Yimam et al., 2013)) popular for the translation task. These are mostly designed for application localization and linguistic annotation, and hence are not necessarily suitable for sentence-aligned corpus translation. Furthermore, these systems lack the multiversioning of the edits and recoverability, which is essential for mass deployment.

The scope of the application is to serve as a centralized platform for parallel corpus development in different languages, facilitating the creation of high-quality multilingual parallel corpora required for training a multilingual Neural Machine Translation (NMT) system, including Nepali, English, and Tamang, with potential expansion to other languages. The objective is to support community-driven parallel corpus development, a promising strategy for developing corpora in low-resource languages (Bal et al., 2024).

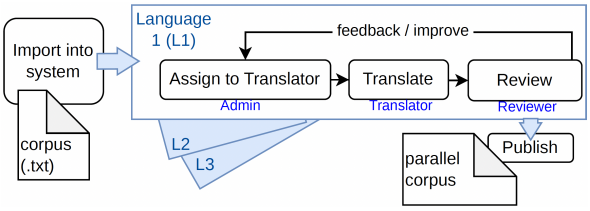


Figure 1: Workflow used in Parallel Corpus Development Toolkit.

2 System Overview

PCDT is a web-based application with three main user roles: 1) *Admin*: Manages users, uploads text corpora, assigns tasks, and oversees the corpus development process. 2) *Translator*: Translates the text/corpus into other specified languages. 3) *Reviewer*: Reviews and validates the translations provided by translators.

The system manages the structured data, including metadata related to corpus files such as domain or category, source, and timestamps. The overall workflow of the application has been outlined in Figure 1.

3 Functionalities

3.1 Admin Features

The admin module supports role-based user management under a single super admin, enabling account creation, updates, deletion, password resets,

[1]http://zanata.org/
[2]http://www.teitok.org/
[3]https://pootle.translatehouse.org/

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 29–30
Tilburg, The Netherlands, June 2026.

and blocking for Translator, Reviewer, and Admin level users. All the actions are logged with a timestamp, actor, and remarks. Corpus datasets are uploaded as `.txt` files (one sentence per line), auto-assigned unique IDs per sentence and organized into traceable batches, with full view, update and delete support, bulk/individual operations, versioned history, and timeline-based recovery. Admin workflows allow marking data ready for translation and finalizing reviewed data into the main corpus. The system supports dynamic language addition and a dashboard providing metadata, lexical density, per-user performance reports, workflow oversight, and progress monitoring.

3.2 Translator Features

There are three types of translators: 1) *Crowd*, 2) *Translator*, and 3) *Machine*. *Crowd* translators are registered users who sign up through the platform and contribute translations voluntarily. *Translator* users are professional translators added and managed by the system administrator. *Machine* translators represent automated translation services, such as machine learning models or online translation APIs, which are used when available for a language pair (e.g., Nepali $\leftrightarrow$ English).Users can translate each sentence available in their preferred language, with the preference set in the translator's profile. The system supports tracking work progress, provides reviewer status on their own translation, and maintains a history of edits of translations with version numbers. The edits can be recovered as needed.

3.3 Reviewer Features

Reviewers review the translations provided by the translators, make corrections if needed while maintaining the version of edits, allow the option to skip the review, maintain flags such as "Doubtful", "Language", and "Gender", where the flag can be dynamic and is maintained in the database so that group-based reporting can be done, and track the work progress.

4 Technologies Used

The system is implemented as a client/server web-based application that can be hosted on a web server (local or cloud) and accessible via standard web browsers. We used a modular architecture with a Python-based (Flask) backend, and a PostgreSQL database to store corpus sentences, corpus metadata, translations, and user information.

5 Application Availability and Licensing

The *PCDT* is developed as part of an effort to build a machine translation system for the Nepali-Tamang language (Ghimire et al., 2026). The source code[4] and documentation are available for research purposes.[5]

6 Conclusion

The Parallel Corpus Development Toolkit (*PCDT*) provides a robust, scalable solution for building a trilingual (and potentially multilingual) parallel corpus with Nepali as the primary language. By supporting batch uploads, dynamic language additions, and seamless collaboration between *Admins*, *Translators*, and *Reviewers*, the system will play a vital role in developing the parallel corpus required for training high-quality Neural Machine Translation.

Acknowledgement: This research was funded by Google through the 2024 Google Academic Research Award (GARA) under the Society-Centered AI initiative and Taighde Éireann – Research Ireland under Grant No. 18/CRT/6223.

References

Bal, Bal Krishna, Balaram Prasain, Rupak Raj Ghimire, and Praveen Acharya. 2024. Strategies for corpus development for low-resource languages: Insights from nepal. *Automatic Speech Recognition and Translation for Low Resource Languages*, pages 297–330.

Ghimire, Rupak Raj, Bipesh Subedi, Balaram Prasain, Prakash Poudyal, Praveen Acharya, Nischal Karki, Rupak Tiwari, Rishikesh Kumar Sharma, Jenny Poudel, and Bal Krishna Bal. 2026. NepTam: A Nepali-Tamang Parallel Corpus and Baseline Machine Translation Experiments. *arXiv preprint arXiv:2603.14053*.

Yimam, Seid Muhie, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A flexible, web-based and visually supported system for distributed annotations. In Butt, Miriam and Sarmad Hussain, editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6, Sofia, Bulgaria, August. Association for Computational Linguistics.

[4]Sources and documentations: https://github.com/ilprl/pcdt

[5]The tool is currently in production and hosted at https://cat.ilprl.ku.edu.np/

English–Nepali–Tamang: A Trilingual Parallel Corpus and Benchmark for Low-Resource Machine Translation

Praveen Acharya
Dublin City University
acharyaprvn@gmail.com

Rupak Raj Ghimire
Kathmandu University
rughimire@gmail.com

Prakash Poudyal
Kathmandu University
prakash@ku.edu.np

Balaram Prasain
Tribhuvan University
prasainbalaram@gmail.com

Bal Krishna Bal
Kathmandu University
bal@ku.edu.np

Abstract

This article describes the research project aimed at developing a Trilingual Machine translation System for English, Nepali, and Tamang language pairs.

1 Introduction

The predominance of English as the primary language of digital content creates a significant barrier for speakers of Nepali and minority languages such as Tamang. This linguistic divide restricts equitable access to information, knowledge resources, and digital services across sectors, including education, governance, healthcare, and commerce. Machine Translation (MT) provides a scalable solution to bridge this gap. However, MT systems for low-resource languages remain under-developed due to the scarcity of high-quality parallel corpora. Research on MT for English–Nepali–Tamang language pairs has been limited. Early work on English–Nepali MT, including rule-based (Bista et al., 2005) and lexicon-driven approaches (Dahal, 2011), showed limited scalability. Subsequent studies comparing Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) found that SMT performs better for English–Nepali, mainly due to the lack of sufficient parallel data for training NMT models (Acharya and Bal, 2018). While some domain-specific datasets exist for English–Nepali (Poudel et al., 2024), resources for Nepali–Tamang remain scarce. A small parallel corpus of about 15,000 sentence pairs is available (Chaudhary et al., 2020), and there is currently no parallel corpus for the English–Tamang language pair. This project addresses these challenges by developing a community-driven trilingual parallel corpus (Bal et al., 2024) and an extensible MT system for English, Nepali, and Tamang, with a strong focus on inclusivity, scalability, and real-world impact.

1.1 Objectives

The key objectives of this research project are:

- ***Corpus Development Framework:*** Design and implement a sustainable, community-driven mechanism for collecting and validating parallel corpora for low-resource languages.
- ***Community Engagement:*** Ensure active participation of native speakers, translators, and local stakeholders throughout to promote ownership and inclusivity.
- ***Extensible MT System:*** Develop a scalable multilingual MT system that is adaptable to additional languages in Nepal.

1.2 Funding Details

- *Project Title:* Empowering Information Access Rights: Developing a Trilingual Machine Translation System for English, Nepali, and Tamang.
- *Funding Agency:* Google Academic Research Award (GARA) Society-Centered AI
- *Award Year:* 2024 A.D.
- *Partner Institutions:* Kathmandu University and Tribhuvan University
- *Principal Investigator:* Bal Krishna Bal
- *Co-Principal Investigator:* Balaram Prasain

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 31–32
Tilburg, The Netherlands, June 2026.

- *Contributors:* Bipesh Subedi, Jenny Poudel, Nischal Karki, Prakash Poudyal, Praveen Acharya, Rishikesh Kumar Sharma, Rupak Raj Ghimire, Rupak Tiwari.

2 Expected Outcomes

The project is expected to deliver the following:

- ***Parallel Corpora:*** Approximately 100,000 sentence pairs for each language pair.
- ***MT System Framework:*** A scalable and extensible multilingual translation system.
- ***Community Awareness:*** Increased awareness of language technologies for preservation and digital inclusion.
- ***Adoption and Impact:*** Enhanced use in local and provincial governments, community schools, and public service delivery.
- ***Language Preservation:*** Strengthening the digital presence of Tamang and other underrepresented languages.

3 Progress and Current Outcomes

MT System: A beta version[1] of the system has been released for community testing.

Dataset and MT models: A large-scale Nepali–Tamang parallel dataset: *(1)* ***NepTam20K***, a corpus of 20,000 sentence pairs translated by experts, and *(2)* ***NepTam80K***, a synthetic corpus of 80,000 sentence pairs.

Both datasets are used to fine-tune state-of-the-art multilingual MT models (mBART, M2M-100, and NLLB-200). Among the evaluated models, the fine-tuned NLLB-200 model outperformed all other models across metrics in both translation directions for the Tamang–Nepali language pair. Details of the corpus creation pipeline, data composition, methodology, and experimental evaluation are presented in (Ghimire et al., 2026).[2]

4 Next Steps

The next phase of the project will focus on:

- Extending the corpus to the target of 100,000 sentence pairs by June 2026.
- Enhancing MT system performance through iterative refinement.
- Promoting adoption through collaboration with local and provincial governments.
- Increasing community engagement to ensure sustained usage and impact.

Acknowledgement: This research was funded by Google through the 2024 Google Academic Research Award (GARA) under the Society-Centered AI initiative and Taighde Éireann – Research Ireland under Grant No. 18/CRT/6223.

[1] https://tmt.ilprl.ku.edu.np/

[2] https://github.com/ilprl/NepTam-A-Nepali-Tamang-Parallel-Corpus-and-Baseline-Machine-Translation-Experiments

References

Acharya, Praveen and Bal Krishna Bal. 2018. A Comparative Study of SMT and NMT: Case Study of English-Nepali Language Pair. In *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 90–93.

Bal, Bal Krishna, Balaram Prasain, Rupak Raj Ghimire, and Praveen Acharya. 2024. Strategies for corpus development for low-resource languages: Insights from nepal. *Automatic Speech Recognition and Translation for Low Resource Languages*, pages 297–330.

Bista, S, B Keshari, J Bhatta, and K Parajuli. 2005. Dobhase: online English to Nepali machine translation system. In *The proceedings of the 26th Annual conference of the Linguistic Society of Nepal*.

Chaudhary, Binaya Kumar, Bal Krishna Bal, and Rasil Baidar. 2020. Efforts towards developing a Tamang Nepali machine translation system. In *Proceedings of the 17th international conference on natural language processing (ICON)*, pages 281–286.

Dahal, AR. 2011. Development of a Nepali-English MT system using the apertium MT platform. *The Language Technology Kendra*.

Ghimire, Rupak Raj, Bipesh Subedi, Balaram Prasain, Prakash Poudyal, Praveen Acharya, Nischal Karki, Rupak Tiwari, Rishikesh Kumar Sharma, Jenny Poudel, and Bal Krishna Bal. 2026. NepTam: A Nepali-Tamang Parallel Corpus and Baseline Machine Translation Experiments. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 8850–8861, Palma, Mallorca, Spain, May. European Language Resources Association (ELRA).

Poudel, Shabdapurush, Bal Krishna Bal, and Praveen Acharya. 2024. Bidirectional English-Nepali machine translation (MT) system for legal domain. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*, pages 53–58.

TELÓ: AI-Driven Automatic Subtitling for the Promotion of the Performing Arts

Antoni Oliver, Silvia Rodríguez Vázquez, Manel Jiménez
Universitat Oberta de Catalunya (UOC)
{aoliverg, srodriguezvaz, manel.jimenez}@uoc.edu

Abstract

The TELÓ project provides an open-source framework for automated subtitling in the performing arts. Integrating state-of-the-art ASR and NMT, the system enables bidirectional translation between Catalan, Spanish, English and French. Designed for live performances, it provides synchronized captions for multiple devices, enhancing cultural internationalization and accessibility.

1 Project information

- Acronym: TELÓ
- Funding Agency: Generalitat de Catalunya
- Project Reference: PLG002/25/000052
- Duration: 6 months (02/02/2026 – 01/08/2026)
- Lead Institution: UOC

2 Introduction and Motivation

The performing arts in Catalonia represent a consolidated industry of high international value. However, the limited number of Catalan speakers among visiting populations and newcomers poses a significant barrier to the full dissemination and internationalization of theatrical works. While subtitling technologies have existed for years, they typically rely on manual synchronization during live performances and require labor-intensive prior translation.

TELÓ (Catalan for "curtain") leverages state-of-the-art Artificial Intelligence (AI) to automate this workflow, providing an open-source tool accessible to professional companies, small theaters, and amateur groups alike. The system supports bidirectional, automated subtitling between Catalan, Spanish, French and English, with an architecture designed for easy adaptation to additional language pairs.

2.1 Context and Market Potential

The scale of the performing arts sector in Catalonia (787 spaces and 13,647 annual performances) highlights the language barrier: while non-verbal shows represent 8.33% of local performances, they account for 60.59% of international tours. Conversely, spoken Catalan performances drop from 60% locally to 12.38% abroad. TELÓ aims to reverse this trend by utilizing AI to make spoken-language drama viable for global audiences.

While Catalan is the primary focus, the system's multi-language capability is driven by immediate industry needs. At the time of writing, the first real-world implementation is scheduled for a French-language production to be subtitled simultaneously into Catalan and Spanish. Although the full calendar of performances for the pilot phase is still being finalized, this initial use case confirms the necessity of bidirectional support for the selected language pairs.

Beyond traditional theater, the modular architecture of TELÓ offers significant potential for adaptation to other live cultural contexts, such as audiovisual festivals, film screenings with live Q&A sessions, and international conferences. These environments share similar synchronization and translation challenges, making the TELÓ framework a versatile solution for the broader creative and cultural industries.

3 Project Objectives

The primary objective is to develop a AI-based system distributed under a free license to facilitate:

- Automatic Script Alignment: Synchronizing original scripts with existing translations.
- Domain-Adapted NMT: Generating high-quality translations tailored to the literary and dramatic genre.
- Real-time Subtitling: Delivering titles to projectors, mobile devices, and smart glasses.
- Accessibility: Providing same-language cap-

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 33–34
Tilburg, The Netherlands, June 2026.

tions for the Deaf and Hard of Hearing (DHH) community.

4 Technical Framework

The project utilizes a robust stack of open-source technologies and research from the Aina Project and Grial Research Group:

- **Transcriber**: Employs speech-to-text models to process live audio. It can run local models, such as Whisper (Radford et al., 2023) and the Catalan-optimized Whisper (Hernández Mena et al., 2024), or commercial systems via API.
- **Synchronizer**: This module acts as the system's core, aligning live transcripts with the pre-existing script and managing subtitle timing. It is capable of detecting actor improvisations; if a deviation from the script is identified, it automatically triggers the translator module to process the spontaneous dialogue in real time.
- **Translator**: Integrates the MTUOC framework (Oliver, 2025), supporting NMT and LLMs to translate both scripts and improvised dialogue. Future iterations of the project will focus on training translation systems specifically adapted to the linguistic registers and nuances of the performing arts.
- **Emissor**: Aggregates subtitles and generates web interfaces for broadcasting. Users can access independent language streams via any device (mobile phones, tablets, smart glasses). The system also supports direct projection onto physical screens.

This modular architecture facilitates adaptation to different user needs and viewing conditions. The system supports defining visual parameters like font size and color contrast depending on the device. To address theater policies, the interface includes a 'Theater Mode' with low-emission color schemes to minimize light pollution. Technical efficiency and low latency are ensured by the system's ability to run on a local server via a dedicated Wi-Fi intranet, though it can also operate via standard web services. The use of a local dedicated network not only guarantees sub-second latency but also isolates the system from external interference, providing a robust environment for high-stakes live performances where timing is critical for the dramatic effect.

While the current prototypes use a cascaded ASR+NMT pipeline to leverage existing language-specific models, we intend to explore the integration of these components into a single end-to-end speech-to-text translation framework in future iterations.

5 Conclusions and future work

The TELÓ project provides an open-source, AI-driven solution for the performing arts. At the time of writing, the core architecture is implemented and preliminary tests have been conducted, with full-scale pilot trials scheduled for the final stage (concluding August 2026) to validate the system in professional settings.

Beyond its technical framework, TELÓ is a vital resource for cultural accessibility. Upcoming trials will prioritize feedback from actors, theater professionals, and the Deaf and Hard of Hearing (DHH) community to ensure the technology integrates seamlessly into the artistic flow. These insights are essential for refining the 'Theater Mode', ensuring the system acts as a non-intrusive bridge that enhances spectator immersion without distracting from the performance.

Future efforts will focus on enhancing robustness against acoustic unpredictability and improvisations. Designed for long-term evolution, the platform will be distributed under a GNU-GPL v.3 license on GitHub[1] to encourage community contributions. While currently focused on four languages, the methodology is designed for extrapolation to further language pairs, promoting social inclusion and the global positioning of local culture.

References

Hernández Mena, Carlos Daniel, Carme Armentano Oller, Sarah Solito, and Baybars Külebi. 2024. 3catparla: A new open-source corpus of broadcast TV in Catalan for automatic speech recognition. In *Proc. IberSPEECH 2024*, pages 176–180.

Oliver, Antoni. 2025. MTUOC server: integrating several NMT and LLMs into professional translation workflows. In *Proceedings of Machine Translation Summit XX: Volume 2*, pages 81–82.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518.

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

DA + Criteria: A New Quality Assessment Method for Bridging the Gap Between Human and Machine Translation

Bettina Hiebl
University of Vienna
bettina.hiebl@univie.ac.at

Abstract

Direct Assessment (DA) + Criteria is a translation quality assessment method proposed based on a comprehensive systematic literature review on the concepts of quality in machine translation and translation studies. In the presented project the method was tested alongside Multidimensional Quality Metrics (MQM) on the German translations by humans, DeepL and ChatGPT of English non-fiction texts, using the results of the study as well as the participants' answers to further refine the method.

1 Introduction

Quality assurance is a central component of human translation (HT) and machine translation (MT). However, quality is conceptualized differently in the humanities-based field of Translation Studies (TS) and the AI-based field of MT: while MT focuses mostly on developing metrics (increasingly including concepts put forward by TS scholars, e.g. evaluation at document level and limitations of relying on reference translations), TS still focuses on time-consuming and costly manual evaluation.

The idea of linking the theoretical foundations of translation studies with the practical quality frameworks used in MT is not new (Čulo, 2014). However, previous surveys tend to concentrate on only one aspect of the field, such as post-editing (Koponen, 2016) or the perspective of translation studies (Koby et al., 2014). Additionally, the Multidimensional Quality Metrics (Lommel et al., 2014) proposes a comprehensive catalog of quality issues that can be used to derive evaluation scores for translations.

Focusing on bridging the gap between TS and MT in translation quality assessment (TQA), this project proposes and tests a unified assessment method (DA + Criteria) based on the results of a comprehensive systematic literature review on the concepts of quality in the two fields.

2 Project Overview

The overall project is a dissertation project at the University of Vienna comprising a systematic literature review on human and machine TQA based on the PRISMA method (Page et al., 2021). This review consists of 250 publications assessed focusing on both the authors' background and the main topics (preliminary results published in Hiebl and Gromann (2023b; 2023a)) building the basis for creating a criteria catalog with numerous criteria for TS and MT with relevant scientific references. Based on this a unified assessment method is proposed, i.e. DA + Criteria, combining Direct Assessment (Graham et al., 2013) with criteria from the catalog, i.e. in addition to rating a translation with a slider from 0 to 100, participants can choose criteria as reason for the rating.

The one-year project of testing the unified assessment method has been sponsored by the EAMT Sponsorship of Activities, Students' Edition 2025 and concerns the validation of the proposed unified assessment method by asking translation professionals to perform TQA of human and machine-translated texts with MQM and DA + Criteria. 12 professional translators have been asked to assess translation quality with both methods (each of the methods was used by 6 participants on each of the texts, i.e. all 12 participants assessed all

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 35–36
Tilburg, The Netherlands, June 2026.

texts). They evaluated 6 parts of English language texts out of non-fiction books on different topics (politics, history, and biology) and their officially published human translations as well as MT output by DeepL and ChatGPT into German and answered questions on both methods. For assessment with DA + Criteria they could choose up to four criteria (Accuracy, Fluency, Cultural Context, Terminology). While for MQM the assessment was only done sentence by sentence, for DA + Criteria the rating was done both sentence by sentence and overall for the texts.

3 Results

Table 1 shows the type of translation per text, the number of segments and the overall assessments per text and method. For MQM the best-rated text according to the total number of errors (30) and the total error points (37.1)[1] is biologyA translated by a human. Whereas this text also received good results by those participants who used DA + Criteria for its assessment, the best-rated text with this method was politicsB translated using DeepL. Both methods indicated that the worst translation was the one of biologyB by ChatGPT.

4 Conclusions

The main conclusions to be drawn by the participants' answers regarding both methods and the evaluation of the results are that the slider from 0 to 100 is considered too fine-grained, and having a criteria catalog to choose from is seen as useful, but not being able to indicate whether the ticked criteria had a positive or negative influence was considered not ideal. Therefore, the method DA + Criteria will be updated: instead of using just one set of criteria without specifying whether the influence is positive or negative, there will be two sets of the same criteria used for indicating the type of influence, and the slider will be changed to 0 to 10.

[1] 5 error points for each major error, 0.1 error points for each minor Fluency/Punctuation error, 1 error point each for all other minor errors

References

Čulo, Oliver. 2014. Approaching machine translation from translation studies–a perspective on commonalities, potentials, differences. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 199–206, Dubrovnik, Croatia.

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41.

Hiebl, Bettina and Dagmar Gromann. 2023a. Human & machine translation quality: comparing & contrasting concepts. In *Transl. Comput 45*, pages 108–128.

Hiebl, Bettina and Dagmar Gromann. 2023b. Quality in human and machine translation: An interdisciplinary survey. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 375–384.

Koby, Geoffrey S, Paul Fields, Daryl R Hague, Arle Lommel, and Alan Melby. 2014. Defining translation quality. *Tradumàtica*, 12:0413–420.

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(2).

Lommel, Arle, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 12:0455–463.

Page, Matthew J., Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, et al. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *International journal of Surgery*, 88:105906.

Info	**biologyA**	**biologyB**	**historyA**	**historyB**	**politicsA**	**politicsB**
translated by	HT	ChatGPT	DeepL pro	ChatGPT	HT	DeepL pro
segments	10	9	7	9	10	10
MQM total errors	**30**	**79**	62	51	68	38
MQM total error points	**37.1**	**234.1**	200	72.1	161.1	48
DA+Crit mean scale	76.72	**64.81**	68.14	70.87	76.98	**78.27**
DA-Crit Doc mean scale	70.50	**53.83**	57.50	71.17	75.83	**79.17**

Table 1: Main Results; mean values are per participant

Adaptive CAT-embedded MT for low-memory, low-compute end-user devices

Marek Sabo
STAR AG
Wiesholz 35
CH-8262 Ramsen
marek.sabo@star-group.net

Abstract

We present ACATMT, a compact bilingual encoder-decoder NMT system for English and Swedish, designed for professional computer-assisted translation (CAT) tools. It runs on-device in ONNX format, under 1 GB of RAM with no GPU needed, and features real-time post-edit based terminology adaptation. It also supports translation memory conditioning via decoder prefilling. Evaluation on 5,021 technical segments unseen during training shows significant improvements in COMET and BLEU when using glossaries.

1 Introduction

Adapting MT in realtime by updating model weights after each segment is impractical on average translator's hardware. Constrained decoding and source augmentation still require explicit glossary management from the translator. ACATMT instead detects terminology changes automatically from post-edits, storing updated pairs in a glossary that is automatically applied to subsequent translations.

2 System Description

2.1 Architecture

ACATMT is an encoder-decoder model initialized from mT5-base (Xue et al., 2020) with tied embeddings. A unigram tokenizer trained on English and Swedish was used to reduce the vocabulary to 50,000 tokens and the original embeddings were aligned to the new tokenizer where possible. The model was trained on bilingual masked span prediction, then fine-tuned on translations with and without bilingual term entries. If glossaries are provided, special tokens signal the start of a term pair and the start of the target term. In training, the glossaries were generated from parallel corpora using the Stanza NLP pipeline's syntactic parsing and our own rules to extract noun phrase candidates. Target term candidates were aligned using LaBSE. To better handle incomplete input from human's typing during post-editing, a later training stage introduced more varied unigram sampling. Glossary extraction uses another special token. For this task we used a few thousand bilingual texts from which DeepSeek API extracted term pairs. Translation memories are supported via decoder prefilling with TM translations provided as decoder context before generation and stripped from the displayed output. Exporting and quantizing in ONNX enabled a faster generation (tokens/second): 16 with max battery saving, 21 with a balanced setting, 33 with max performance on battery, and 46 when plugged in. Tested on a laptop with Intel Core i5-1135G7 (no GPU) and 16 GB RAM.

2.2 Interactive Post-Editing

Our localhost browser demo interface streams output token by token from the on-device Python server. The translator can click any word mid-generation to interrupt and request alternatives from that position, optionally typing the first characters of a desired word to filter suggestions. Pressing *Enter* accepts the typed input and the model regenerates the translation from that point; while *Escape* accepts the edit locally without regeneration, suitable for minor corrections.

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 37–38
Tilburg, The Netherlands, June 2026.

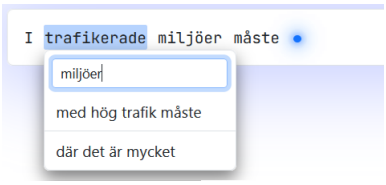


Figure 1: ACATMT interrupted when a user clicks on a word to rephrase it, types the beginning of an alternative, and ACATMT suggests completions in a dropdown.

2.3 Implicit Terminology Adaptation

The system's most distinctive feature operates without any deliberate user action. After translating a segment, the model extracts a bilingual glossary from its own output. If the translator post-edits the translation, the extraction is re-run. By comparing the two extraction results, the system detects changes in terminology and saves updated term pairs into a post-edit-based glossary, which is separate from any human-curated glossaries. If a new source text contains any captured term pairs, the system supplies them to the model. Term extraction is a trained behaviour, which works by providing both the source and the target to the encoder with a dedicated token for glossary generation.

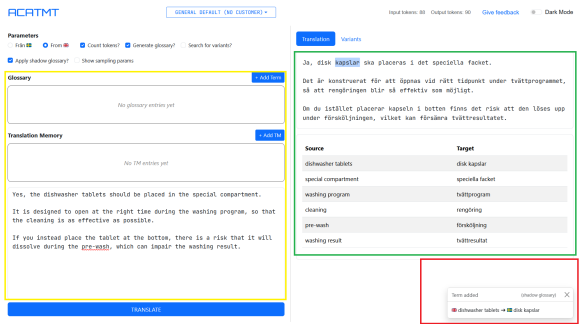


Figure 2: Source text (yellow) with optional user provided terms or translation memory, ACATMT output post-edited by the user (green) and the automatically extracted glossary. Terms automatically captured into the glossary appear in a toast notification (red)

3 Evaluation

Glossary usage is automatically evaluated on 5,021 segments from a customer's reference translation in medical technology. No human post-editors were involved. Segments where the source consisted solely of glossary terms were excluded from the evaluation. The customer data was not seen during training. We compare four ACATMT conditions against Helsinki-NLP's opus-mt-en-sv as a strong bilingual baseline. Terms used in the evaluation were derived from the same reference material using ACATMT's glossary extraction mode. We compare MT without glossary (vanilla), with the glossary, and with decoder prefilling using terms, and report the respective COMET (wmt22-comet-da) and sacreBLEU scores. When terminology is provided, ACATMT gains significant improvements in BLEU and COMET scores and outperforms Helsinki using decoder prefilling with terms.

Condition	COMET	BLEU
Helsinki opus-mt-en-sv		
Vanilla	0.8368	35.68
+ Decoder pref. with terms	0.8436	37.29
ACATMT		
Vanilla	0.8195	35.23
+ Glossary mode	**0.8719**	**45.19**
+ Glossary & pref.	0.8675	45.96
+ Decoder pref. with terms	**0.8733**	**47.57**

Table 1: Mean COMET and BLEU on 5,021 customer segments.

A formal evaluation of the implicit adaptation loop, measuring term capture rate and apply rate across post-edits in real human-facing interactions, remains as future work.

4 Pricing, licensing, and availability

ACATMT is intended for integration into STAR Transit, a commercially available CAT tool by STAR AG. Licensing and pricing are pending; the codebase is proprietary and cannot be released.

5 Discussion and Future Work

When the user clicks on a word to display alternatives, the system rarely surfaces source-language suggestions. If the model doesn't apply the supplied glossary term, the term often appears among the suggested alternatives.

Some terms are used very infrequently in customer texts. Even in the model variants that have been fine-tuned on specific customer's data, glossary input can occasionally further help the model.

We think that small, adaptive, privacy-preserving models represent a compelling direction for professional post-editing workflows, and hope ACATMT serves as a useful reference for the community.

References

Xue, Linting and others. 2020. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. *arXiv preprint arXiv:2010.11934*.

Scalable Video-Based Search in the VGT Dictionary

Toon Vandendriessche,
Mathieu De Coster, Joni Dambre
Ghent University – imec – IDLab AIRO,
Ghent, Belgium
toon.vandendriessche@ugent.be
mathieu.decoster@ugent.be
joni.dambre@ugent.be

Caro Brosens,
Hannes De Durpel
Vlaams Gebarentaalcentrum
(VGTC),
Ghent, Belgium
caro.brosens@vgtc.be
hannes.dedurpel@vgtc.be

Abstract

Video-based sign language dictionary search – in which a user records a sign to retrieve its translation – has been increasingly studied, yet never deployed in a large-vocabulary setting. We present the first such deployment: a fully scalable video-based search system integrated into the Flemish Sign Language (VGT) Dictionary, comprising over 11,000 signs. The system, released on November 28th, 2025, requires no retraining as new signs are added, and was validated on data collected *in the wild*. It was developed through an equal partnership between the deaf-led Flemish Sign Language Centre (VGTC) and AI researchers from Ghent University, and shows that closing the gap between sign language research and community impact is both achievable and essential.

1 The Flemish Sign language dictionary

The Flemish Sign Language (VGT) dictionary (Brosens, 2022) is the primary lexicographic resource for Flemish Sign Language (Link: woordenboek.vlaamsegebarentaal.be). Through its continual development in close collaboration with the Flemish signing community, it plays a central role in preserving, documenting, and teaching the language, while reflecting its contemporary lexical and regional diversity. Currently, the VGT dictionary comprises 11,248 unique signs, but this number is rapidly expanding through several curation mechanisms that rely on community involvement.

The VGT dictionary was among the first to support a form of sign-to-text search (Vermeerbergen and Van Herreweghe, 2018). Initially, users could search for sign translations using SignWriting (Sutton, 2022). Later, this was replaced with a system based on sign parameters, in which users search by selecting handshape(s) from a list or indicating the articulatory location on a pictogram of the human body. Similar (or even more extensive) search functionality is also available in several other sign language dictionaries.

However, parameter-based approaches fall short of enabling true sign-to-text search, where a user performs a sign in front of a camera and retrieves its relevant Dutch translations. Building a system capable of searching through large dictionaries poses several non-trivial challenges. It must handle any arbitrary sign mapped to a unique gloss, scale efficiently and robustly to very large vocabularies, and remain flexible enough to accommodate new entries without requiring retraining. Through our collaboration, video-based search has been added alongside existing parameter-based methods, making the VGT dictionary the first to support this at such a vocabulary scale. While Vandendriessche et al. (2026) detail the co-creation process that brought initial research results into a validated application, this paper describes the final system, which has been publicly available as part of the VGT dictionary since November 28th, 2025.

Before our addition, the dictionary was used primarily by hearing VGT learners looking up signs for Dutch words. Video-based search makes the VGT dictionary truly bidirectional: enabling reliable sign-to-text search opens the dictionary to DHH users as well – particularly those with lower Dutch proficiency – as it allows them to consult it in their native language. Beyond expanding the

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 39–40
Tilburg, The Netherlands, June 2026.

user base, this functionality also unlocks new applications, such as lexical research and the development of VGT learner support tools driven by the DHH community itself. While the full community impact remains to be seen, realising this functionality required concrete technical advances, which we describe below.

2 Product properties and usage statistics

To ensure frictionless access, we deliberately embedded this functionality into the trusted VGT dictionary rather than branding it as a standalone product. Technically, our solution is based on similarity search, using vector representations that were not specifically tailored to VGT. As a result, it is not affected by dictionary expansion. Most technical aspects described in (Vandendriessche, 2026) are retained in the current implementation. One key design choice was the privacy-oriented decision to extract keypoints locally on the user's device, ensuring that no personal data is ever transmitted over the internet. In the original implementation, however, real-time keypoint extraction using MediaPipe[1] introduced frame drops on less performant hardware, reducing retrieval accuracy. We therefore reimplemented the recording workflow to enforce the processing of all frames, at the cost of some processing time. Additionally, in-frame detection of the user's elbows and shoulders reduces keypoint jitter caused by out-of-frame body parts. Together, these changes significantly improved retrieval performance.

Fig. 1 shows steady platform usage since release. To support long-term monitoring we invite user-driven labelling: after each query, users can indicate whether the correct sign appeared among the retrieved results. Since this feedback is optional, annotated samples remain a minority (at the time of writing, only 172 labelled queries have been logged), which limits statistically significant conclusions. However, we expect that meaningful evaluation data will accumulate over time.

3 Conclusions and call to action

Bringing research technology to a real-world, large-vocabulary dictionary required steps that are rarely reported in academic research – and even more rarely rewarded by it. Validating performance *in the wild*, adapting to real hardware constraints, and iterating based on genuine user behaviour are not linear processes, yet they are essential to ensure that research outcomes actually reach the communities they are intended to serve. We argue that the field of sign language technology would benefit from more explicitly valuing these steps, both in how research is conducted and how it is evaluated. We therefore encourage sign language technology researchers to engage more openly with the practical challenges of real-world deployment, and to treat the gap between laboratory results and genuine community use not as an implementation detail, but as a research problem in its own right. In the end, this engagement is what transforms research outcomes into community-driven products with real short-term impact (Bragg, 2019) – and what gives co-creation its meaning beyond a buzzword.

[1]Link: github.com/google-ai-edge/mediapipe

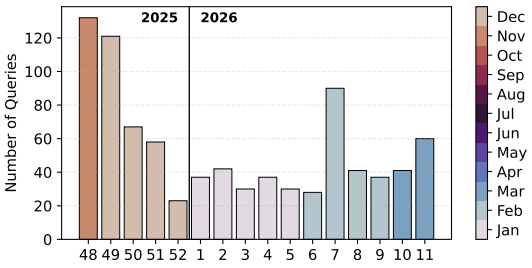


Figure 1: Weekly usage of video based lookup in the VGT online dictionary since release on Nov. 28th, 2025.

References

Bragg, D. et al. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st international ACM SIGACCESS conference on computers and accessibility*, pages 16–31.

Brosens, C. et al. 2022. Moving towards a functional approach in the flemish sign language dictionary making process. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 24–28.

Sutton, V. 2022. *Lessons in SignWriting*. SignWriting Press.

Vandendriessche, T. et al. 2026. Signbuddy: from sign language research to scalable co-created solutions. *Universal Access in the Information Society*, 25(2):45.

Vermeerbergen, M. and M. Van Herreweghe. 2018. Looking back while moving forward: The impact of societal and technological developments on flemish sign language lexicographic practices. *International Journal of Lexicography*, 31(2):167–195.

TaMTAS: Terminology-Aware Machine Translation for Accessible Science

Corpus compilation, terminology extraction and data augmentation

Antoni Oliver
Universitat Oberta de Catalunya
aoliverg@uoc.edu

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

Abstract

This paper presents the TaMTAS project (Terminology-Aware Machine Translation for Accessible Science), a research project coordinated by the Universitat Oberta de Catalunya (UOC) to develop an open-source translation ecosystem for the Life Sciences. While we provide a general overview of the project's organization into seven Work Packages (WPs) and its collaborative consortium, this article focuses specifically on the work of WP2. Led by the UOC, this package is responsible for the parallel corpus compilation for five languages (English, Spanish, Catalan, Estonian, and Irish), the enhancement of TBX-Tools for terminology extraction, and the development of synthetic data augmentation strategies. These linguistic assets are essential to power the downstream Large Reasoning Models (LRMs) and Automatic Post-Editing (APE) modules, ensuring terminological consistency in highly specialized scientific domains.

1 Introduction

The dominance of English in scientific dissemination creates a significant barrier to knowledge access, affecting the ability of researchers, students, and the general public to access global research, while also limiting the capacity of researchers to disseminate their own findings in their native languages. To address this challenge, the TaMTAS project emerges as a 36-month initiative (15 December 2025 – 14 December 2028). Currently in its initial phase, the work focuses on the foundational data architecture and the functional setup of WP2 tasks. This collaborative effort is led by the Universitat Oberta de Catalunya (UOC), acting as project coordinator, in partnership with the Barcelona Supercomputing Center (BSC), the University of Surrey (UoS), Dublin City University (DCU), and the University of Tartu (UT). Focusing on the highly specialized Life Sciences domain, the project supports five languages with varying resource availability: English, Spanish, Catalan, Estonian, and Irish.

Unlike traditional neural machine translation (NMT) approaches, which often struggle with terminological consistency in long documents, TaMTAS introduces a paradigm shift by utilizing Large Reasoning Models (LRMs). Within this framework, translation is treated as a multi-step reasoning task. By leveraging chain-of-thought prompting, glossary-guided constraints, and self-correction mechanisms, the system ensures that specialized terminology is rendered accurately and consistently throughout entire texts.

2 Architecture and Work Packages

TaMTAS is structured as a comprehensive software and data pipeline, transitioning from raw data acquisition to end-user application. The ecosystem is built through collaborative Work Packages (WPs):

- **WP1 (UOC):** *Project Management and Coordination.* Administrative governance and inter-partner synergy.
- **WP2 (UOC):** *Corpus compilation, terminology extraction and data augmentation with terminology.* Compilation and curation of Life Sciences corpora and functional enhancement of TBX-Tools (Oliver and Vàzquez, 2015).
- **WP3 (BSC):** *Terminology-aware MT.* Development of Large Reasoning Models (LRMs) for domain-specific translation.
- **WP4 (UoS):** *Terminology-aware Quality Estimation and Automatic Post-Editing.* Implementation of QE and APE modules to ensure terminological integrity.
- **WP5 (UT):** *Post-Translation Text Augmentation.* Content adaptation through simplification

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 41–42
Tilburg, The Netherlands, June 2026.

and explanatory enrichment for diverse audiences.

- **WP6 (DCU):** *Evaluation.* End-to-end validation with real-world stakeholders and scientific partners.
- **WP7 (DCU):** *Dissemination, Outreach and Impact.* Management of FAIR-compliant data release and strategic project communication.

3 WP2: Corpus Compilation, Terminology Extraction and Data Augmentation

Work Package 2 (WP2), led by the Universitat Oberta de Catalunya (UOC), plays a crucial role in the success of the TaMTAS project. Its main objective is to create and curate high-quality linguistic resources that are essential for training and validating terminology-aware machine translation models. These resources are vital for downstream components, such as the Large Reasoning Models (LRMs) and Automatic Post-Editing (APE) modules, ensuring terminological consistency and overall quality in highly specialized scientific domains. This work package focuses on three strategic areas:

- **Large-scale Corpus Curation:** The project involves compiling parallel and comparable corpora within the Life Sciences domain, focusing on high-quality alignment at the sentence and paragraph levels. Texts will be gathered from open-access academic journals, public health guidelines, and regulatory documents. The corpus will cover various text types, including research papers and medical guidelines, to create robust training material for the Large Reasoning Models (LRMs). Regular updates to the corpus will maintain its relevance and support the evolving needs of the field.
- **Functional Optimization of TBXTools:** WP2 will enhance TBXTools (Oliver and Vàzquez, 2015), UOC's open-source terminology software, to significantly improve the efficiency and accuracy of terminology extraction. The optimized TBXTools will be capable of automatically generating TBX-formatted termbases that are enriched with both morphological and syntactic metadata, allowing for a deeper understanding of the terms within their specific contexts. This enhancement will ensure that terminology remains consistent across translations, particularly in the specialized domain of Life Sciences. Moreover, the improved TBXTools will be able to handle complex multi-word expressions and highly specialized domain-specific terms, which are common in scientific texts. By enabling precise extraction and management of these terms, the system will contribute to higher accuracy and reliability in translations, ensuring that the final outputs are both technically correct and contextually appropriate.
- **Innovative Data Augmentation:** WP2 introduces a synthetic data augmentation strategy specifically designed to address the challenge of rare or out-of-vocabulary (OOV) terms. It leverages intelligent term substitution techniques within existing parallel segments of the corpus. By identifying similar, contextually appropriate terms from the corpus, the system can generate new training instances that expand the vocabulary without the need for manual data collection. This approach will prove particularly useful in low-resource languages such as Estonian and Irish, where obtaining extensive linguistic data can be challenging. As a result, the system's ability to handle specialized scientific terminology will be significantly enhanced.

4 Conclusion

The TaMTAS project represents a significant step forward in specialized machine translation. By framing translation as a reasoning task and backing it with the rigorous, high-quality data infrastructure developed by the UOC, the project will deliver a suite of practical products ready for integration into modern translation workflows. Consistent with our commitment to open science, all linguistic resources and software developed will be released under open-source licenses through the project website and FAIR-compliant repositories.

Acknowledgements

This work has been supported by project PCI2025-167063-2, funded by MICIU/AEI/10.13039/501100011033 and by the European Union (CHIST-ERA).

References

Oliver, Antoni and Mercè Vàzquez. 2015. TBXTools: a free, fast and flexible tool for automatic terminology extraction. In *Proceedings of the international conference recent advances in natural language processing*, pages 473–479.

Advancing CAT Tool Features to enhance Consistency in MT and Generative AI Outputs

Judith Klein
STAR AG
Wiesholz 35
CH-8262 Ramsen
judith.klein@star-group.net

Abstract

Recently, generative artificial intelligence (GenAI) has been perceived as a "silver bullet" for achieving faster, cheaper, and better translation production. However, in professional localisation, AI capabilities alone are not enough, as the still time-consuming post-editing (PE) of machine translation (MT) and GenAI output proves. The features and processes presented in this work aim to reduce these efforts by enhancing terminological control and translation consistency within the CAT environment STAR Transit[1].

1 Inconsistency in AI-driven Output

Despite high quality MT and GenAI outputs, inconsistencies and hallucinations remain a key challenge, often causing repeated corrections of similar error types. This highlights the need for advanced CAT tools features to improve consistency, particularly in domains such as technical documentation. To address this, four key features are proposed that leverage terminology integration, contextual processing, and MT reuse.

2 Advanced Features for Consistency

For terminological consistency, we introduce features for engine-level glossary integration and segment-specific term-pair constraints. For translation consistency, we present context-aware processing and systematic reuse of MT outputs.

2.1 Automated MT Glossary Synchronisation

As manual synchronising of glossaries of multiple language combinations is not feasible at scale, we implemented an automated workflow for consistent MT and GenAI output. Using the standard export functionality of TermStar[2] via its proprietary API, multilingual glossaries are extracted with a customised script, converted into the required format of the MT platform and uploaded through the platform's standard import interface.

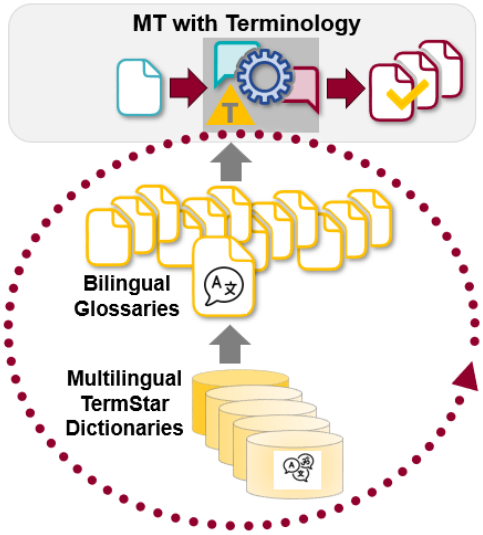


Figure 1: Automated workflow from glossary extraction to synchronisation with DeepL.

In the current use case, this scalable workflow synchronises over 650 domain-specific glossaries on a weekly basis, ensuring that the most recent dictionary updates are quickly available for translation workflows, and can be adapted to different requirements.

When multiple translation candidates are available, the process selects the first entry based on dictionary or entry order, without applying additional selection criteria such as contextual fields.

[1] Pricing/licensing upon request (star-group.net/en/products)

[2] STAR's Terminology Management System

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 43–44
Tilburg, The Netherlands, June 2026.

2.2 Term-pair Constraints for GenAI

Linking terminology directly to the source segment enables consistent use of validated customer terminology. Originally developed for NMT systems with integrated terminology constraints[3], the feature is now extended to GenAI engines. If a match is found, the corresponding source term and its approved translation are retrieved from the dictionary and automatically supplied to the GenAI engine with the source segment, thereby ensuring terminological consistency during generation.

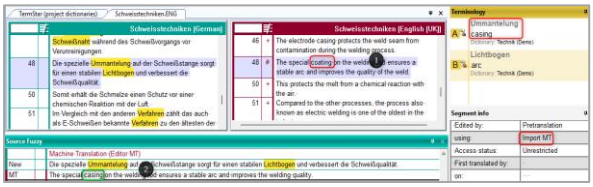


Figure 2: Translation output comparison: (1) Standard MT result vs. (2) MT with applied terminology constraint

The approach corresponds to recent AI-assisted CAT workflows[4] with terminology-aware generation. First results show improved terminology adherence, although occasional hallucinations confirm that careful verification remains necessary.

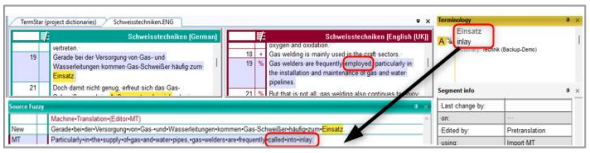


Figure 3: Hallucination produced by term constraint

The feature is activated automatically when a dictionary is assigned to a project. If multiple candidates are available, the system selects the first term pair. The feature will be included in the next release as part of the optional AI feature package.

2.3 Context-aware Segment Translation

Context-aware translation has become established through the integration of GenAI in CAT tools, helping to overcome the limitations of isolated segment processing.[5]

In STAR Transit, the GenAI engine receives not only the current source segment but also a configurable number of preceding and following sentences taken from the same or adjacent paragraphs, thereby providing paragraph-level context during generation. The context range is defined at project level and applied consistently across all documents in the project while preserving the segment-based structure required for TM efficiency. As illustrated in Figure 4, this improves the handling of pronouns and context-specific terminology and particularly resolves ambiguities and semantics.

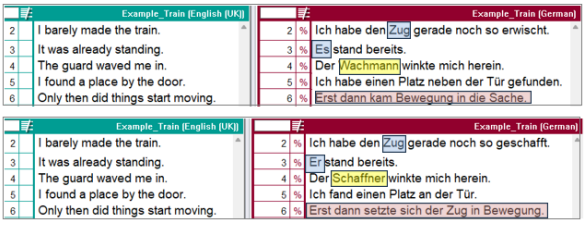


Figure 4: GenAI output comparison between segment-based (top) and context-aware (bottom) translation in STAR Transit

2.4 MT Fuzzy for MT Output Reuse

Existing approaches such as adaptive MT or features like "MatchPatch"[6] attempt to improve consistency by adapting translations from similar segments previously processed by MT. In contrast, the new standard feature already available in STAR Transit presents previously generated MT output for similar source segments as an "MT fuzzy" without automatically modifying the translation. This reuse of existing MT output provides substantial consistency benefits across related segments, while visually highlighted source-segment differences support efficient PE.

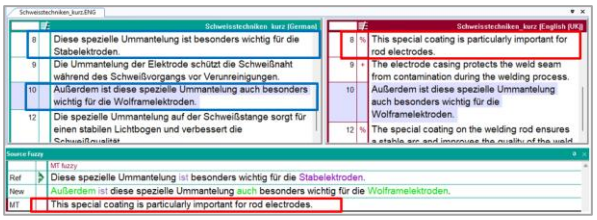


Figure 5: MT fuzzy: reuse of existing MT for fuzzy matches

3 Summary and Outlook

These features demonstrate how CAT tools can evolve to improve the quality of MT and GenAI output within professional translation workflows.

However, further refinements are still needed to enable more controlled and guided terminology support, for example through interactive prompting. In addition, future work will integrate AI-based quality assurance features for improved inconsistency detection and resolution.

[3] E.g. TextShuttle (Supertext)
[4] E.g. SDL Trados Studio (RWS), memoQ (Kilgray), and Phrase Platform.
[5] Cf. GenAI-based *Copilots* in market-leading CAT tools
[6] *MatchPatch* is a feature in memoQ.

Translation 2.0: Equipping linguists for the machine translation future

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

Vasilis Kalogiannis
Leiden University Centre for Linguistics
v.p.kalogiannis@umail.leidenuniv.nl

Abstract

Translation 2.0 addresses a critical gap in accessible, up-to-date educational resources on recent developments in Machine Translation and Large Language Models for students of linguistics and translation. It develops an online module with open-access learning materials, including knowledge clips, a workbook with incremental exercises to consolidate conceptual understanding, practical coding guides, and videos featuring industry professionals. The module aims to build both subject knowledge and computational literacy, freeing up contact hours for deeper engagement and critical discussions on practical, professional and ethical aspects. Translation 2.0 is funded through an Educational Innovation grant by the Faculty of Humanities at Leiden University and ECOLe (Expert Centre for Education and Learning) and runs from February to December 2026.

1 Introduction

Translators and linguists increasingly require training in language technologies and machine translation (MT) to meet market demands (ELIS, 2025), yet accessible educational resources remain scarce. Academic literature on MT tends to presuppose a computational background that many linguistics and translation students lack, creating a significant gap between available materials and the needs of this audience. Notable exceptions include the textbooks by Koehn (2020) and Kenny (2022), as well as collaborative projects such as MultiNMT[1] and LT-LiDER[2], which made NMT more approachable for translators.

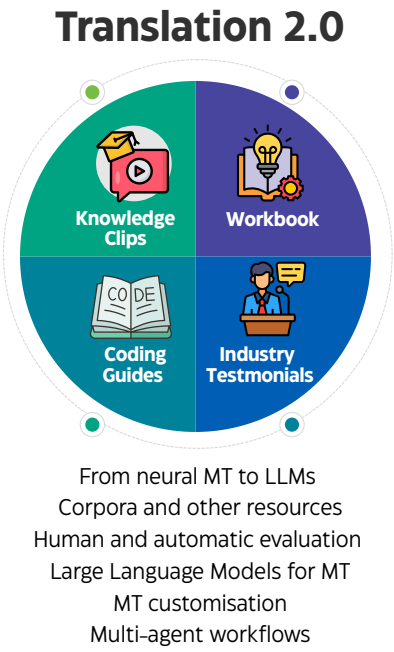


Figure 1: Structure of the Translation 2.0 module

The advent of Large Language Models (LLMs) has further widened this gap. LLMs have fundamentally transformed how MT systems are built and deployed, rendering much of the existing accessible literature obsolete. This leaves linguistics and translation programmes without adequate teaching materials on the current state of the field, and places teachers in a difficult position, as they are expected to deliver up-to-date instruction on a rapidly evolving technology while keeping pace

[1] https://www.multitrainmt.eu/index.php/en/
[2] https://lt-lider.eu/

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 45–46
Tilburg, The Netherlands, June 2026.

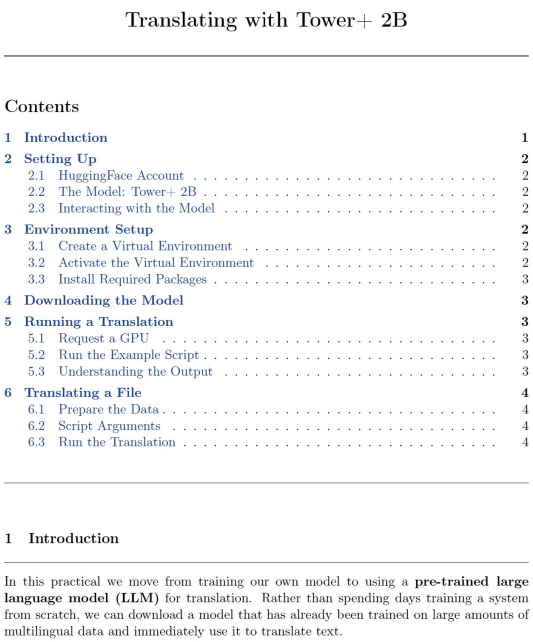
with those advancements. Compounding the problem, previous course materials have tended to prioritise declarative content over active learning: they offer limited incremental exercises to consolidate conceptual understanding, little opportunity for students to reason critically about MT implementation, and few practical guides that would equip them to work with open-source tools.

2 Project objectives

This project aims to develop a set of open-access learning materials for courses in translation technology and machine translation, addressed at students of both linguistics and translation programmes. The materials are designed to be flexible across curricula and thus intended for BA and MA students alike. Their goal is not only to convey subject knowledge, but also to build towards digital literacy and internalisation of essential computational principles, e.g. understanding where data is stored, how to estimate computational requirements, and how to structure computational steps and processes. Rather than relying solely on cloud services such as Google Colab or Kaggle, students will also learn to work on local infrastructure and on high performance computing (HPC) clusters[3], in order to develop foundational skills necessary for careers as computational linguists or translation technologists.

The structure of the module is shown in Figure 1. Each unit will include 2-4 knowledge clips (5-10 min duration each) covering core topics in MT and LLMs. All clips will be designed based on Web Content Accessibility Guidelines (WCAG)[4]. A workbook with incremental exercises serves improving conceptual understanding and developing hands-on procedural knowledge. Alongside these, step-by-step practical guides will walk students through the use of open-source tools to train and use MT systems in the terminal. An example of a practical coding guide is shown in Figure 2. Lastly, videos featuring industry professionals will discuss real-world applications of MT and language technology, bridging the gap between academic training and professional practice.

All materials will be housed in a module on an online e-learning platform (Articulate Rise 360[5]). The underlying structure is flexible: as new engines, metrics, and LLM-based features emerge, individual components can be updated without disrupting the overall structure. Students can work on the platform independently before class. This flipped classroom approach aims to free up contact hours for substantive discussion, assignment review, and individual feedback. The guides and clips will be piloted iteratively with students during development to ensure clarity and usability. The evaluation of the final module includes feedback by students and translator trainers. After completion, the module will be made publicly available as open educational resources.



Translating with Tower+ 2B

Contents

1 Introduction

In this practical we move from training our own model to using a **pre-trained large language model (LLM)** for translation. Rather than spending days training a system from scratch, we can download a model that has already been trained on large amounts of multilingual data and immediately use it to translate text.

Figure 2: Example taken from the practical guide to inference with a pre-trained LLM on the University HPC Cluster.

References

ELIS, Research. 2025. European language industry survey 2025: Trends, expectations and concerns of the european language industry. https://elis-survey.org/wp-content/uploads/2025/03/ELIS-2025_Report.pdf.

Kenny, Dorothy, editor. 2022. *Machine translation for everyone*. Number 18 in Translation and Multilingual Natural Language Processing. Language Science Press, Berlin.

Koehn, Philipp. 2020. *Neural machine translation*. Cambridge University Press.

[3]For Leiden University, students have access to the HPC ALICE computer cluster for running computationally heavy experiments. We acknowledge this may not be the case with many other programs.

[4]https://www.w3.org/WAI/standards-guidelines/wcag/

[5]https://360.articulate.com/rise

Making Jobs Accessible through AI-supported Easy Language Translation

Fabian Merkel[1], Marco Baumgartner[2], Athanasios Breskas[1], Silke Gutermuth[1], Silvia Hansen-Schirra[1], Elena Kick[2], Vanessa König[3], Tobias Kopp[2], Natalie Martin[2], Miriam Spieß[2]

[1]Johannes Gutenberg University Mainz
[2]Karlsruhe University of Applied Sciences
[3]SUMM AI

Abstract

Access to the primary labor market for people with cognitive impairments is hampered by barriers, notably the lack of workplace information in Easy Language (EL). Producing such texts is time- and cost-intensive and requires specialized translators. Project STARK-LS (Strengthening participation in the primary labor market through AI-generated Easy Language) addresses this gap by using an AI-translation tool to translate workplace materials into EL and integrating the approach into internships for people with cognitive impairments. An interdisciplinary team conducts mixed-methods evaluations by testing the EL translations for applicability, comprehensibility, and acceptance using lab-based eye-tracking and questionnaire studies, qualitative interviews with interns with cognitive impairments and experts for EL, and a quantitative survey with company representatives. The findings will inform best-practice recommendations for companies and rehabilitation agencies. The project advances scientific understanding of the perceived usefulness and potential barriers of EL in organizational contexts, while evaluating AI's influence on the diffusion of high-quality EL texts in companies.

1 Introduction

STARK-LS aims to strengthen participation (and acceptance) of people with cognitive impairments in the primary labor market through workplace internships investigating the role of providing AI-supported and post-edited Easy Language (EL) texts in authentic inclusive employment settings.

Existing projects highlight both the potential and the gaps: iDEM (Saggion et al., 2024) develops AI for accessibility but does not account for German. TOP.KI (2025) uses Plain Language instead of EL. Findings from the LeiSA project (Bock, 2019) show that EL texts often do not adhere to guidelines, have limited practical application, and are rarely used by companies. STARK-LS addresses these gaps.

The project is funded by the German *Federal Ministry of Labour and Social Affairs*.

2 Project Timeline and Partners

The STARK-LS project consists of four phases running from 2025 to 2029. First, internships for people with cognitive impairments are organized by the *Institute for Learning and Innovation in Networks* (ILIN) at Karlsruhe University of Applied Sciences (HKA) in collaboration with the *District Administration of Germersheim*, employers, rehabilitation stakeholders, and sheltered workshops. Second, relevant workplace materials are translated into EL using *SUMM AI*'s translation tool and post-edited by the *Tra&Co Center* at Mainz University to ensure compliance with EL standards and comprehensibility. Third, the internships are accompanied by qualitative and quantitative research conducted by ILIN and Tra&Co investigating the organizational use of EL, the role of AI-supported accessibility, and the readability and comprehensibility of AI-generated and post-edited EL. Finally, the project develops practical guidelines, actionable models, and market analyses to support inclusive workplace communication and participation in the primary labor market.

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 47–48
Tilburg, The Netherlands, June 2026.

3 Project Objectives

STARK-LS pursues a set of objectives aimed at advancing inclusive labor market participation through AI-supported EL.

EL translations will be empirically evaluated for their applicability, comprehensibility, and user acceptance. We thereby seek to assess the potential of AI-powered translation tools to assist inclusion processes. Furthermore, both interns and employers will be accompanied throughout the internship phase to document and analyze their experiences, challenges, and learning outcomes. Based on these findings, we will develop actionable models and practical guidelines for companies, rehabilitation providers, and other stakeholders in the long term. These guidelines aim to support the integration of EL into workplace communication and identify strategies for overcoming organizational acceptance barriers.

4 Planned Methods of Evaluation

To achieve the objectives mentioned above, we are planning several methods of evaluation, such as structured interviews with both the interns and the employers following the completion of the internships. Additionally, a quantitative study involving more employers is planned to identify barriers to disseminating workplace information in EL and evaluate the perceived usefulness of AI-supported EL. An error annotation of the output following the guidelines of the DIN SPEC 33429 (2025) to evaluate the quality of AI-generated vs. post-edited texts is in progress. In terms of the technical effort required to ensure high quality, a study will be conducted using keystroke logging to compare the post-editing of AI-generated EL with human translations.

Furthermore, studies will be conducted investigating the usability of EL texts in workplace scenarios involving participants with cognitive impairments. Adopting a mixed-methods approach, data collection includes technologies such as eye-tracking and augmented reality applications, alongside interviews, comprehension tasks, and questionnaires. The studies address readability, comprehensibility, usability, and user experience, as well as the acceptability of AI-generated compared to post-edited EL texts.

Results are pending, as the internships are still ongoing, and some are yet to be planned and organized.

References

Bock, Bettina M. 2019. *„Leichte Sprache“ – Kein Regelwerk. Sprachwissenschaftliche Ergebnisse und Praxisempfehlungen aus dem LeiSA-Projekt.* Korrigierte Druckfassung (= Kommunikation – Partizipation – Inklusion, Band 5). Berlin. Online (Fassung 2018): http://ul.qucosa.de/api/qucosa%3A31959/attachment/ATT-0/

DIN SPEC 33429. 2025. *Empfehlungen für Deutsche Leichte Sprache*. Berlin. Beuth. https://dx.doi.org/10.31030/3594547

Saggion, Horacio, S. Markus Bott, J. LLUÍS MARTÍN BERBOIS, S. Marcela Szasz, S. Sharoff, J. O'Flaherty, T. Blanchet, Volkan Sayman, M. Gollegger, A. Rascón Alcaina and S. Sanfilippo y L. Muñoz. 2024. *The iDEM Project: Addressing Linguistic Barriers in Deliberative Processes*.

TOP.KI. 2026. *Inklusive berufliche Prüfungen ohne Sprachbarrieren durch Textoptimierung mit Hilfe von Künstlicher Intelligenz.* https://top-ki.info

Prompsit's API and CLI: planet-friendly, privacy-first, open-source translation services for everyone

Lev Nikolaevich Berezhnoy, Gema Ramírez Sánchez
Sergio Ortiz Rojas, Mikel L. Forcada
Prompsit Language Engineering
Edif. Quorum III, Avinguda de la Universitat, s/n, E-03202 Elx
levnikolaevich,gramirez,sergio,mlf@prompsit.com

Abstract

Prompsit is launching an updated API and CLI for its open-source, planet-friendly machine translation services. Operating on a freemium model, the tools offer free limited access alongside tiered pricing for advanced features like MT evaluation, quality estimation, corpus scoring, and multilingual dataset annotation.

1 A classical MT service in 2026?

While large language models (LLMs) offer impressive linguistic nuance, traditional neural machine translation (NMT) remains the proven backbone for professional workflows. NMT is significantly faster —often outperforming LLMs by a large margin— and utilises a transparent, character-based pricing model. By removing the token overhead associated with prompting, NMT provides a more stable and cost-effective solution for both small and large-scale projects. This efficient approach inspires Prompsit's translation services, offering significant sustainability advantages by using purpose-built NMT engines that require a fraction of the computational power and energy needed by general-purpose LLMs. Alongside lean open-source NMT models built from well-curated corpora provided by OPUS (github.com/Helsinki-NLP/Opus-MT) and Mozilla (github.com/mozilla/firefox-translations-models), we provide high-quality Apertium machine translation (apertium.org) and AltLang language variety converters (altlang.net), offering the stable, predictable behaviour of rule-based systems (RBMT) that are even faster and more energy-efficient. Furthermore, we complement these translation engines with automatic evaluation and annotation services.

2 Services available

Translation We offer high-performance NMT and RBMT specializing in low-resource languages and regional variants to ensure contextually accurate output. The API supports text snippet and document translation across a wide range of languages and formats, including robust tag handling, optional quality estimation and leverage from a hierarchy of user's translation memories.

Evaluation Our tools measure translation quality using industry-standard automated metrics, allowing users to audit engines by analysing parallel corpora and model performance. This helps maintain professional standards and linguistic consistency across supported language pairs.

Scoring Parallel segments can be scored for translation likelihood using Prompsit's widely-adopted Bicleaner multilingual models (github.com/bitextor/bicleaner-ai). These scores are used to identify and filter low-quality translations, to help select higher-quality parallel data for model fine-tuning.

Annotation The API provides sophisticated data processing to deduplicate, label, and score multilingual datasets. Documents are enriched with language identification, personally identifiable information (PII) and adult content flagging, encoding fixes, and quality scores. This metadata enrichment is essential for top document selection in model refinement tasks.

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 49–50
Tilburg, The Netherlands, June 2026.

3 The API and CLI

Access to Prompsit's API is via an access token available to registered users. Here's a curl request to translate a short string:

```
curl -X 'POST' \
'https://edge.prompsit.com/v1/translation?enable
_qe=false' \
  -H 'accept: application/json' \
  -H 'Authorization: Bearer ...auth_token... \
  -H 'Content-Type: application/json' \
  -d '{
  "source_lang": "en",
  "target_lang": "es",
  "texts": [
    "Hello world"
  ]
}'
```

To translate a file, one would use a similar call which would return the URLs needed to check status and to download the result file. Currently customers can directly invoke the API from their internal tools and platforms with simple integration steps, or use the CLI described below. Prompsit plans to offer new CAT connectors to implement the translation services offered in the new API.

The CLI provides easier API access for human users but may also be used to easily script complex translation- related tasks. For instance, the command for the translation query inside the CLI above would simply be `translate "Hello world" -s "en" -t "es"`. From outside the CLI, a script could send `prompsit translate "Hello world" -s "en" -t "es"` and capture the result for further processing. The CLI is available under the Apache 2.0 licence at `github.com/Prompsit/prompsit-cli` for our customers to install locally. The underlying engines and most models are also open-source.

4 A bit of technological detail

Built as a REST application on top of Python 3.13 and FastAPI (`fastapi.tiangolo.com`), our API utilizes a microservice architecture to orchestrate 12 containerised modules that power several translation engines such as Apertium, AltLang, and CTranslate2 (`github.com/opennmt/ctranslate2`). An 8-step pipeline manages tag extraction, 5-level caching, and neural word alignment while a specialized formatting stack (Docling, Okapi, and Tikal) handles over 25 binary and text formats. MetricX (`github.com/google-research/metricx`) and COMET (`unbabel.github.io/COMET`) GPU-based estimation are used to ensure quality while Bicleaner-AI and Monotextor (`github.com/bitextor/monotextor`) provide respectively advanced parallel (sentence pairs) and monolingual corpus (documents) scoring and annotation. Asynchronous job progress is streamed in real-time via server-sent events, all accessible through an open-source CLI.

5 A summary of features

Energy efficiency: quantized NMT engines and microservices save GPU and power usage.

Data privacy: in-memory processing ensures no data storage or use for model training.

Latency: intelligent caching allows for millisecond responses and real-time document progress via streaming.

Language coverage: a selection of NMT and RBMT engines for 17 major and 3 low-resource languages (ca, gl, nn), 11 language varieties for 5 of them (such as fr-CA and fr-FR) in 52 language pairs as of May 2026.

Format support: tag-aware translation for 30 different formats, including Office, PDF, and localization file formats (such as TMX or PO).

Transparency: transparent commands for usage and health monitoring.

6 Use via an AI agent

The CLI repository includes machine-readable skill descriptions that enable most popular AI coding assistants to assist the human user to interact with the CLI programmatically to perform translation, evaluation, scoring, annotation, and initial setup. Skills are bundled with the CLI package and deployed automatically on first launch. This integration is a thin interface layer: the AI assistant interprets user intent and invokes CLI commands. The computational cost of translation services is the same regardless of whether the request originates from a human or an AI assistant.

7 Access and pricing

Visit `prompsit.com/en/contact` for free API access. A secret key will be sent to your email. Install the CLI with `npm install -g prompsit-cli` and authenticate with the provided login, or using your Google address. We offer a freemium pricing model, mostly free (with limits) for MT and paid for additional services.

Advancing Medical Communication: Multilingual, Multicultural, and Multimodal Processing for Translation and Simplification

Maria Pia di Buono
University of Naples "L'Orientale"
Via Duomo, 219 - Napoli (Italy)
mpdibuono@unior.it

Abstract

This paper presents the Multilingual, Multicultural, and Multimodal Medical Language Processing (4MLP) Project, funded through a competitive call of the University of Naples "L'Orientale" (Italy).

1 Introduction

Multilingual, Multicultural, and Multimodal Medical Language Processing[1] (4MLP) is a project started in December 2025 and funded through a competitive call of the University of Naples "L'Orientale" (Italy) under the University Research Projects funding scheme (PRA 2025). In addition to the principal investigator (PI), the team consists of one full professor and three doctoral students.

Although the project does not involve a formal consortium, external partners will be engaged through framework agreements. Among those, a recent agreement has been established with Senaso srl, a local social enterprise that provides support to immigrants, including medical escort services.

4MLP advances linguistic research and language technologies in healthcare by developing multilingual, multicultural, and multimodal resources, alongside specialized medical language models. It aims to improve healthcare communication by enhancing patients' access to, understanding of, and engagement with medical information and treatments, ultimately supporting more efficient and patient-centered care.

[1] https://www.unior.it/en/node/3574
https://sites.google.com/view/4mlp-project/home

Effective communication is crucial for successful treatment outcomes, yet misunderstandings, medical jargon, and fragmented information often create barriers. Moreover, as cultural and linguistic factors critically shape doctor–patient communication (Schouten and Meeuwesen, 2006), including differences in explanatory models of health and illness, cultural values, patient preferences regarding doctor–patient relationships, racism and perceptual biases, and linguistic barriers, medical language processing requires approaches that account for multilingual, multicultural, and multimodal dimensions.

Several studies have explored the use of Natural Language Processing (NLP) and large language models (LLMs) to improve healthcare communication, medical information retrieval, and patient comprehension. Reddy (2023) proposed a framework for assessing the translational value of LLMs in healthcare, while Sakakini *et al.* (2020) investigated context-aware simplification of health materials in low-resource settings. Other works focused on specialized medical resources and models, such as LeMe-PT (Simões and Gamallo, 2021), ChatDoctor (Li et al., 2023), and PharmMT (Li et al., 2020), aimed respectively at improving medical language understanding, medical advice generation, and the simplification of prescription instructions.

2 Objectives and Methodology

Building on previous research by the PI and co-authors, including studies on the use of LLMs as virtual assistants for patients in the context of drug administration (Giordano and di Buono, 2024), 4MLP develops responsible NLP-based approaches which combine symbolic knowledge and neural methods to provide clear, reliable, and ac-

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 51–52
Tilburg, The Netherlands, June 2026.

cessible healthcare communication.

In line with current research in the field, 4MLP leverages language models for medical information retrieval and text simplification, and is structured around a four-step workflow: (i) development of domain-specific enriched and linked resources, integrating cultural and multilingual dimensions to ensure interpretability and cultural grounding; (ii) subsymbolic modeling including adaptation and fine-tuning of language models and integration of multimodal data for robust representation and flexible processing; (iii) neurosymbolic integration to combine symbolic transparency with subsymbolic adaptability; (iv) evaluation and application to test developed tools in different healthcare communication scenarios, including iterative refinement of models and resources.

3 Preliminary Results

The project is currently in its initial phase, and the results presented here represent a first step in its development. The research initially focuses on Italian and will progressively extend to multilingual and multimodal settings involving languages spoken by migrant communities. Particular attention is devoted to contexts in which institutional languages (e.g., French or Portuguese) coexist with widely used vehicular languages such as Wolof (Minerba, 2021) or Portuguese Creole.

A first outcome of the project is the Italian Medical Term Simplification (I-MTS) resource, comprising 1,356 Italian term–simplification pairs enriched with semantic and lexical information (di Buono, 2026). Future work will expand both the linguistic coverage and the range of medical documents considered, including heterogeneous healthcare materials and multilingual resources, with the aim of supporting AI-assisted translation into lower-resourced languages and accessible multilingual healthcare applications.

Ethics Statement

The project addresses key ethical considerations, with particular focus on fairness and the mitigation of socio-demographic, religious, and ethnic biases through rigorous evaluation, careful data curation, and synthetic data generation. Human oversight by domain experts supports the prevention of inappropriate outputs and unintended outcomes.

Acknowledgements

This research is funded by the University of Naples "L'Orientale" under the University Research Projects funding scheme (PRA 2025) - CUP C66125002860005.

References

di Buono, Maria Pia. 2026. Italian medical term simplification: From patient information leaflets to simplified language resources. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC 2026*.

Giordano, Luca and Maria Pia di Buono. 2024. Large language models as drug information providers for patients. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024*, pages 54–63.

Li, Jiazhao, Corey Lester, Xinyan Zhao, Yuting Ding, Yun Jiang, and VG Vinod Vydiswaran. 2020. Pharmmt: a neural machine translation approach to simplify prescription directions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2785–2796.

Li, Yunxiang, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Minerba, Emiliano. 2021. Wolof language and literature: an introduction. In *Language and Identity Theories and experiences in lexicography and linguistic policies in a global world*, pages 202–216. EUT Edizioni Università di Trieste.

Reddy, Sandeep. 2023. Evaluating large language models for use in healthcare: A framework for translational value assessment. *Informatics in Medicine Unlocked*, 41:101304.

Sakakini, Tarek, Jong Yoon Lee, Aditya Duri, Renato FL Azevedo, Victor Sadauskas, Kuangxiao Gu, Suma Bhat, Dan Morrow, James Graumlich, Saqib Walayat, et al. 2020. Context-aware automatic text simplification of health materials in low-resource domains. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 115–126.

Schouten, Barbara C and Ludwien Meeuwesen. 2006. Cultural differences in medical communication: a review of the literature. *Patient education and counseling*, 64(1-3):21–34.

Simões, Alberto and Pablo Gamallo. 2021. Lemept: A medical package leaflet corpus for portuguese. In *10th Symposium on Languages, Applications and Technologies (SLATE 2021)*, pages 10–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Does Speech Translation Meet Users' Needs? An English to Portuguese Study Across Demographics

Giuseppe Attanasio[*Λ], **Beatrice Savoldi**[Φ], **Daniel Chechelnitsky**[Π], **Matteo Negri**[Φ], **Marine Carpuat**[ϒ], **André F.T. Martins**[ΣΛT]

[Λ] Instituto de Telecomunicações, Lisbon, Portugal
[Φ] Fondazione Bruno Kessler, Trento, Italy
[Σ] Instituto Superior Técnico, Lisbon, Portugal
[Π] Carnegie Mellon University, Pittsburgh, USA
[ϒ] University of Maryland, College Park, USA
[T] TransPerfect

Abstract

This paper introduces Ouvia, a research project to assess user-perceived usability and reliability of modern speech translation tools in En→Pt scenarios. The project centers on a user study in which we simulate real-life daily interactions by recruiting crowdworkers online from different sociodemographic groups. We collect their spoken requests and self-assessments about quality, satisfaction, and reliability. Here, we describe the project's motivation and objectives, the study design, and the expected outcomes we will provide to speech translation practitioners.

1 Introduction

While machine translation is consolidating its popularity among laypeople, scholarly accounts have advocated rethinking a new way to evaluate this technology—one centered on people, their communication needs, and real-world use contexts (Savoldi et al., 2025; Carpuat et al., 2025).

This project investigates **whether modern AI-based speech translation systems can meet people's communication needs** across diverse real-life scenarios. It centers on an online user study and data collection that mimics a one-to-one interaction: a person starts a conversation with a request in English, and an AI-based system translates it for a Portuguese listener. The project's goal is to measure success along two axes. First, we center the evaluation on user-perceived aspects—"I trust the AI translation to convey my message" or "I would use this AI translation in a real-world situation" are two of the statements we ask participants to rate their level of agreement with. Second, we replicate the study over different demographic groups, stratifying our speakers on first language (native English or not), gender, and ethnic group. This dimension aligns with recent calls for AI-enabled speech technologies to be equitable across different demographics (Attanasio et al., 2024).

Objectives. This project has three main objectives. We will (**O1**) establish whether current speech translation AIs provide outputs that end users would be willing to rely on in real-world En→Pt scenarios, (**O2**) measure the extent to which language variants and demographic factors affect user-perceived reliability, and (**O3**) assess whether current translation quality metrics capture such user-centered perceptions.

We will compile all collected data, including speaker recordings, annotations, and self-reported assessments, in a new speech recognition and translation benchmark. We plan to recruit participants (speakers) from three language groups (native and non-native English), three broad ethnic groups, and three gender identities. We will release all artifacts (data, annotations, code) in a GDPR-compliant bundle, licensed under CC BY 4.0.

Funding Agency and Partners. The project is a joint effort from Instituto de Telecomunicações (IT; leading), Instituto Superior Técnico (IST), Fondazione Bruno Kessler, Carnegie Mellon University, and the University of Maryland. The duration is 12 months, starting May 2025. Funds for the entire data collection are provided by the European Association for Machine Translation, awarded through the *2025 EAMT Sponsorship of Activities* call. IT provides computational and logistic support. The user study obtained ethical ap-

[*]Correspondence to giuseppe.attanasio@lx.it.pt.

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 53–54
Tilburg, The Netherlands, June 2026.

proval from IST's Ethics Committee.

2 Study Design

We conduct a four-stage online study to mimic a one-to-one exchange in which an English *speaker* interacts with a Portuguese *receiver*, with the speaker's voice translated by an AI.

Round 1. The speaker reads aloud and records a passage we provide them with. This *conversation starter* is 40-60 words long and contains key information, such as named entities or quantities, that a system should translate correctly. An example is "Hi, I've had a persistent rash on my arms and torso for five days. [...]" **Round 2.** We translate the speaker's recording automatically and provide the receiver with the outcome. Then, we ask the receiver to reply to a set of questions grounded in the original passage, e.g., "For how many days has the person had symptoms?" This approach borrows from QA-based machine translation evaluation protocols. **Round 3.** We recruit a third participant, who speaks both English and Portuguese fluently, to validate the translation in two ways: A scalar quality score and a binary judgment of which question was answered correctly by the receiver. We use several heuristics to check these assessments. **Round 4.** The initial speaker receives the validation outcome and answers a qualitative survey designed to capture self-assessed satisfaction, effectiveness, and reliability.

3 Preliminary Results and Conclusions

At the time of writing, our main outcomes are (**i**) a set of conversation starters and associated questions used in Rounds 1 and 2, (**ii**) a web-based platform to handle data collection and user participation, and (**iii**) data from 1200+ interactions.

We compiled (**i**) with healthcare-related and mundane (e.g., traveling) interactions to cover both high- and low-stakes scenarios, sourcing most starters from existing dialogue benchmarks. To increase linguistic coverage, we developed an AI-assisted pipeline in which we first prompt a flagship language model to generate synthetic starters, then manually validate and apply heuristics to quality-check them. All participants interacted through a custom-developed web app (**ii**), which streamlines users and recordings management and admin tooling. We recruited all crowdworkers through `prolific.com`. We collected scripted recordings from more than 100 US residents who self-identified as female or male and as white or black/African American (self-declared attributes), totaling (**iii**) 1,240 unique interactions (speaker recordings, receiver answers, validator quality assessments, and speaker final self-assessment).

Our data, platform, and cross-demographic findings will help identify blind spots in current speech translation systems and inform future development priorities. They will foster research across speech and demographic groups, laying the groundwork for focused inquiries into fairness and equity, as well as studies on how automatic metrics relate to user-defined reception and usability.

4 Acknowledgments

GA is supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI) and by cofunded EU funds under UID/50008: Instituto de Telecomunicações; AM by the project DECOLLAGE (ERC-2022-CoG 101088763). BS is funded by the Horizon Europe research and innovation programme, under grant agreement No 101135798, project Meetween.

References

[Attanasio et al.2024] Attanasio, Giuseppe, Beatrice Savoldi, Dennis Fucci, and Dirk Hovy. 2024. Twists, humps, and pebbles: Multilingual speech recognition models exhibit gender performance gaps. In Al-Onaizan, Yaser, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of EMNLP 2024*, Miami, Florida, USA, November. Association for Computational Linguistics.

[Carpuat et al.2025] Carpuat, Marine, Omri Asscher, Kalika Bali, Luisa Bentivogli, Frédéric Blain, Lynne Bowker, Monojit Choudhury, Hal Daumé III, Kevin Duh, Ge Gao, Alvin Grissom II, Marzena Karpinska, Elaine C. Khoong, William D. Lewis, André F. T. Martins, Mary Nurminen, Douglas W. Oard, Maja Popovic, Michel Simard, and François Yvon. 2025. An interdisciplinary approach to human-centered machine translation. In Christodoulopoulos, Christos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of EMNLP 2025*, Suzhou, China, November. Association for Computational Linguistics.

[Savoldi et al.2025] Savoldi, Beatrice, Alan Ramponi, Matteo Negri, and Luisa Bentivogli. 2025. Translation in the hands of many: Centering lay users in machine translation interactions. In Christodoulopoulos, Christos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of EMNLP 2025*, Suzhou, China, November. Association for Computational Linguistics.

CRITICS: Critical Science Without Borders by Translation of Scientific Knowledge

Rodrigo Agerri[♠], Itziar Aldabe[♠], Elena Cabrio[♣], Mark Cieliebak [◇], Jan Deriu [◇], Mariana Flores[♠], Jurgita Kapočiūtė-Dzikienė [Φ], Dovilė Kuizinienė [Φ], Arantza Rico[Ψ], Aritz Ruiz-González[Ψ], Aitor Soroa[♠], Mantas Vaškevičius [Φ], Serena Villata[♣]

[♠]HiTZ Center - Ixa, University of the Basque Country EHU

[♣]Université Côte d'Azur, Inria, CNRS, I3S, France

[◇]Centre for Artificial Intelligence, ZHAW School of Engineering

[Φ]Faculty of Informatics, Vytautas Magnus University, Lithuania

[Ψ]Department of Mathematics, Experimental and Social Sciences Education, University of the Basque Country EHU

rodrigo.agerri@ehu.eus

Abstract

The CRITICS project addresses science accessibility and literacy through the convergence of advanced Machine Translation (MT) based on Large Language Models (LLMs) and educational technology. By leveraging MT systems specifically optimized for scientific content, educational institutions can provide accurate, culturally relevant translations of scientific materials in students' native languages, ensuring that complex scientific concepts are comprehensible while maintaining technical accuracy. Novel research on MT for scientific documents aims to break down language barriers in accessing cutting-edge research and educational materials currently only available in high-resourced languages, thereby facilitating the democratization of scientific knowledge.

1 Introduction

CRITICS is a three-year CHIST-ERA IV Cofund 2025 project funded within the topic "Science in your own language"[1]. This call addresses the translation of scientific knowledge to bridge linguistic and cultural gaps for those who must disseminate and access scientific knowledge beyond their linguistic scope.

The project is coordinated by the HiTZ Center from the University of the Basque Country EHU[2], funded by MICIU/AEI /10.13039/501100011033 and by the European Union (PCI2025-167239-2). In addition to HiTZ, the EHU team includes researchers from the Department of Mathematics, Experimental and Social Sciences Education. The consortium is also formed by the CNRS/Université Côte d'Azur in France (grant no ANR-25-CHR4-0002-02), Vytautas Magnus University (VMU) (funded by the Research Council of Lithuania, agreement No. S-CHIST-ERA-26-1), and the ZHAW Center for Artificial Intelligence in Switzerland (grant no. 20CH-1_238349).

The accessibility to scientific content in our own languages through advanced MT naturally connects to the automated generation of science education materials, where LLMs can be applied to synthesize and adapt complex scientific concepts into level-appropriate and pedagogically sound resources. Thus, CRITICS will develop and adapt LLMs to facilitate the generation of customized science education materials (Méheut and Psillos, 2004). CRITICS will mostly follow recent science education research focused on Design-Based Research (Ruiz-González et al., 2025) and consider education materials to specify the *driving problem/questions*, the learning objectives focused on competency acquisition, the scientific practices including scientific argumentation and critical thinking, and the activities to be made by the science students. The availability of machine-translated scientific knowledge will be crucial to investigating and developing LLMs for the automatic generation of appropriate education materials in the students' native languages that relate to local students' experiences (Ruiz-González et al., 2025).

2 Objectives and Work Plan

Although CRITICS' vision applies across disciplines, the project focuses on two areas: (i) nat-

[1]https://www.chistera.eu/projects-call-2025

[2]https://hitz.eus

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 55–56
Tilburg, The Netherlands, June 2026.

ural sciences (biology, chemistry, and physics) and (ii) Artificial Intelligence. The former is key to competency-based assessments, while the latter poses specific translation challenges due to the continuous introduction of new terminology that may lack established equivalents in less-resourced languages (Kleidermacher and Zou, 2026; Zhang et al., 2024). CRITICS will target a diverse spectrum of languages, namely, Basque (agglutinative language isolate), Lithuanian (East Baltic, inflectional), German (West Germanic, inflectional), and French (Romance, synthetic-fusional).

Objective 1. LLM-based Machine Translation of Scientific Documents: Focused on developing and adapting open-weight LLMs for high-quality, document-level MT of scientific texts, particularly in low- and medium-resource language pairs (e.g., English–Lithuanian/Basque/French/German).

Objective 2. Argumentation serves as a fundamental mechanism in scientific discourse, facilitating the process of reaching conclusions and facilitating science literacy and critical thinking. This objective will focus on training LLMs to recognize evidence-based argumentation and identify fallacies and scientific misconceptions.

Objective 3. Automatic Assessment and Critical Thinking: Rather than merely identifying gaps in the scientific discourse, automatic assessment will also involve the generation of Critical Questions and Detailed Feedback in competency-based assessment settings (PISA style - Programme for International Student Assessment) where cross-linguistic comparability is essential.

Objective 4. Evaluation: qualitatively evaluate the generation of critical questions and automatic assessment feedback using LLMs in a way that can be compared with human-generated judgments (Calvo Figueras and Agerri, 2025).

3 Future Work

LLM-based Translation of Scientific Documents: Terminology-aware prompting and term injection have received growing attention as key techniques for addressing an open challenge in domain-specific MT: terminological inconsistency (Sabo et al., 2024; Kim et al., 2024). This is critical in scholarly translation, where inaccurate terminology can distort meaning or undermine academic precision (Kleidermacher and Zou, 2026). It should be noted that less-resourced languages have received limited attention in LLM-based MT research (Kapočiūtė-Dzikienė et al., 2025).

The use of LLM-based evaluation for text generation tasks related to MT and Critical Questions and Feedback generation in the scientific domain remains an open research problem (Calvo Figueras and Agerri, 2025). CRITICS will provide new LLM-based evaluation methods specifically tailored to the relevant features of translating MT of scientific documents and to the science teaching-related criteria of the automatically generated Critical Questions and Assessment/Feedback.

References

Calvo Figueras, Blanca and Rodrigo Agerri. 2025. Benchmarking Critical Questions Generation: A Challenging Reasoning Task for Large Language Models. In *Findings of the EMNLP 2025*, pages 5635–5652.

Kapočiūtė-Dzikienė, Jurgita, Toms Bergmanis, and Mārcis Pinnis. 2025. Localizing AI: Evaluating Open-Weight Language Models for Languages of Baltic States. In *NoDaLiDa/Baltic-HLT*, pages 287–295.

Kim, Sejoon, Mingi Sung, Jeonghwan Lee, Hyunkuk Lim, and Jorge Gimenez Perez. 2024. Efficient terminology integration for LLM-based translation in specialized domains. In *Proceedings of the Ninth Conference on Machine Translation*, pages 636–642.

Kleidermacher, Hannah Calzi and James Zou. 2026. Science across languages: assessing LLM multilingual translation of scientific papers. In *Findings of the EACL 2026*, pages 3932–3947.

Méheut, Martine and Dimitris Psillos. 2004. Teaching–learning sequences: aims and tools for science education research. *International Journal of Science Education*, 26(5):515–535.

Ruiz-González, Aritz, Arantza Rico, and Jenaro Guisasola. 2025. Learning About Sound in Initial Teacher Training: Evaluation and Redesign of a Teaching-Learning Sequence. In *Connecting Science Education with Cultural Heritage: Selected Papers from the ESERA 2023 Conference*, pages 157–171. Springer.

Sabo, Marek, Judith Klein, and Giorgio Bernardinello. 2024. Boosting machine translation with AI-powered terminology features. In *EAMT*, pages 25–26.

Zhang, Dan, Ziniu Hu, Sining Zhoubian, Zhengxiao Du, Kaiyu Yang, Zihan Wang, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024. Sciinstruct: a self-reflective instruction annotated dataset for training scientific language models. *Advances in Neural Information Processing Systems*, 37:1443–1473.

Implementations and Case Studies

AI Post-Editing in Production: A 71,262-Segment Evaluation Across Five Domains, Ten Languages and Five Systems

Mara Nunziatini
Welocalize
mara.nunziatini@welocalize.com

Mercedes Speroni
Welocalize
mercedes.speroni@welocalize.com

Abstract

This study evaluates an AI post-editing (AIPE) system in a professional translation setting, covering translation from English into ten target languages across five domains. We evaluate the system using automatic metrics on 71,262 production segments and human evaluation on a stratified sample of 6,618 segments (approximately 600 segments per target language) assessed by 60 professional translators. AIPE refines machine translation output using a secure publicly available LLM, retrieving language-specific style guides and high-quality bilingual examples to guide edits. We compare it with direct LLM translation (LLMT), Google Translate, and DeepL. The two AIPE configurations evaluated consistently outperform the generic translation baselines in terms of quality. LLMT does not match this quality, though it may suit less quality-sensitive domains. We observe how AIPE's gains vary according to pre-translation type, with fuzzy translation memory matches overrepresented among severe errors, and discuss deployment implications.

1 Introduction

The adoption of large language models (LLMs) in professional translation workflows has accelerated rapidly, raising practical questions for language service providers (LSPs) about where and how to deploy them. Two distinct paradigms have emerged: AI post-editing, where an LLM refines the output of an existing MT engine, and LLM-based direct translation, where the LLM translates from source without a pre-translation step. Both approaches promise quality improvements over generic neural MT, but independent, production-scale evidence comparing them is scarce. Most published evaluations focus on research-grade datasets or assess single language pairs and domains (Raunak et al., 2023), but industry deployments involve a far messier reality: heterogeneous content types and language pairs, mixed pre-translation input types (translation memory matches, generic MT, custom MT), and stringent business constraints on cost and turnaround. This paper reports on an evaluation designed to answer three questions: 1. Does AIPE deliver measurable quality improvements over generic MT baselines alone? 2. Is LLM-based direct translation a quality-equivalent and cheaper alternative to MT + AIPE? 3. How does pre-translation quality moderate AIPE's gains? A methodological note is needed. AIPE and direct LLM translation use additional resources, such as high-quality bilingual examples and style guides, while Google Translate and DeepL were not trained with glossaries or translation memories for this experiment. AIPE also has access to a pre-translation, which it refines rather than generating text from scratch. This difference is intentional and reflects real production conditions. As a result, the quality differences we observe come from the full process, including the initial translation, the model, the prompting strategy and the additional resources.

2 Background and Related Work

Post-editing MT output is well-established as a productivity measure in professional translation

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 59–65
Tilburg, The Netherlands, June 2026.

(Koponen, 2016). The advent of LLMs has prompted researchers to explore their use both as post-editors and as direct translators. Jiao et al. (2023) show that LLMs can achieve competitive translation quality on standard benchmarks. Moslem et al. (2023) demonstrate that LLMs can effectively adapt MT output when provided with in-context bilingual examples, closely analogous to the Translation Pairs mechanism used in our AIPE solution. Their findings on single language pairs motivated our investigation of whether this approach scales across ten target languages and diverse professional content types. Our study complements this prior work with production-scale evidence grounded in real client workflows with heterogeneous input types. We calculate automatic metrics on the 71,262-segment dataset, and human annotation of a stratified 6,618-segment subset (around 600 per language) by 60 professional translators using rankings and the DQF-MQM framework (Lommel et al., 2014). Lommel et al. (2014b) demonstrate its reliability across annotator profiles, supporting our use of both internal and external specialist linguists.

3 Systems

3.1 AIPE

AIPE is an LLM-based post-editing system that takes existing translations and improves them. The LLM receives the source and pre-translated text together with relevant context retrieved via a retrieval-augmented generation (RAG) strategy. This context includes human-reviewed bilingual segments (referred to as Translation Pairs throughout this paper) and language-specific style rules, which guide the LLM to correct errors, adjust terminology, and align the output with the expected tone and brand voice. AIPE targets MT outputs and fuzzy translation memory (TM) matches, skipping Exact and ICE (Internal Context Exact)[1] matches to preserve pre-approved content. The system respects inline tags and file structure, while also storing detailed metadata and post-editing results for tracking, analysis, and continuous improvement. Typically, this feature is used in conjunction with other technologies under the OPAL Enable product offering (Nunziatini et al., 2025), such as customized MT and AI Quality Estimation, to maximise its efficiency. Two AIPE configurations were evaluated:

- Generic MT + AIPE: using Google Translate output as the pre-translation, and
- Production AIPE: follows a typical real-life production setup in which we leverage TM(s) down to 75% fuzzy matches first, and then apply MT (generic or customized, depending on the use case) on the rest of the content. Then, AIPE is applied to all the segments, skipping exact and ICE matches. In more detail, pre-translation input types are distributed as follows:
  - MT segments (generic or customized): 45%
  - fuzzy TM matches equal or above 75%: 19%
  - ICE segments: 19%
  - exact TM matches: 17%

In both cases, Translation Pairs were retrieved where available.

3.2 LLMT

LLMT uses the same secure publicly available LLM, but translates directly from source, without any MT pre-translation. It receives the same style guides and Translation Pairs as AIPE. This configuration is cheaper and faster to operate since it eliminates the cost of an MT pre-translation step, and was hypothesized to approximate AIPE quality at lower cost.

3.3 Baseline MT Engines

Google Translate and DeepL served as generic MT baselines, translating directly from source with no content-specific resources such as glossaries, translation memories or style guides.

4 Experimental Setup

4.1 Automatic Scoring

The first part of the experiment is an automatic scoring exercise. In this exercise we compare, segment by segment, the translations generated by the different systems mentioned above against the gold standard (reference) human translation. The automatic evaluation dataset comprised 71,262 segments (553,518 words) spanning:

[1] An ICE match is a 100% match where the preceding and the following segments that are in the TM are the same as the previous and next segment in the translation. Since the segment matched as well as the segments before and after that match are identical to the earlier translation, the translation quality has already been verified.

Table 1: Number of segments per target language and domain

Target Language	Prod/Serv	Product	Marketing	Support	Med. Dev.	Total	%
fr-FR (French)	3,668	749	1,873	1,245	228	7,763	10.89%
es-ES (Spanish)	3,668	749	1,873	1,245	228	7,763	10.89%
de-DE (German)	3,668	749	1,873	1,245	228	7,763	10.89%
zh-CN (Chinese)	3,668	749	1,873	1,245	228	7,763	10.89%
ja-JP (Japanese)	3,668	749	1,873	1,245	228	7,763	10.89%
pt-BR (Portuguese)	3,668	749	1,873	1,245	228	7,763	10.89%
ko-KR (Korean)	3,668	749	1,873	1,245	228	7,763	10.89%
it-IT (Italian)	3,668	749	1,873	1,245	228	7,763	10.89%
ru-RU (Russian)	846	749	1,873	881	228	4,577	6.42%
tr-TR (Turkish)	850	749	1,873	881	228	4,581	6.43%
Total	31,040	7,490	18,730	11,722	2,280	71,262	100%
%	43.56%	10.51%	26.28%	16.45%	3.20%		

- 10 target target languages: en-US into fr-FR, es-ES, de-DE, zh-CN, ja-JP, pt-BR, ko-KR, it-IT, ru-RU, tr-TR
- 5 domains: Product/Service (44%), Product (11%), Marketing (26%), Support (17%), Medical Devices (3%)

Table 1 shows the distribution of segments across target languages and domains for the full automatic evaluation dataset.

4.2 Human Evaluation

A stratified sample of approximately 600 source segments per language (equivalent to 6,618 total source segments) was selected for human evaluation across the same accounts and target languages, with slightly smaller samples for ru-RU and tr-TR, reflecting their lower volume of translation requests across accounts. Five system outputs (Generic MT + AIPE, Production AIPE, LLMT, Google Translate, and DeepL) were assessed per segment by three professional translators per language, for a total of 60 annotators (30 specialist linguists for medical device content and 30 for the rest of the domain). The evaluation was conducted blindly and with system order randomized. Annotators performed two tasks:

- Error annotation using DQF-MQM, covering Accuracy, Fluency, Terminology, Style, Design, Verity and Locale Convention error types, each rated as Neutral, Minor, Major, or Critical.
- Preference ranking into three positions (Rank 1 = best, Rank 3 = worst), with ties permitted. Annotators could assign multiple systems to the same rank position, and no rank position was mandatory, meaning that if all outputs were deemed of average or insufficient quality, they could leave the Rank 1 position empty.

Linguists were provided with glossaries, style guides, and TMs as a reference if available.

5 Results

5.1 Automatic Metrics

We report edit distance (HTER-style, computed against human post-edits), ChrF (Popović, 2015), BLEU, and COMET (Rei et al., 2020).

System	HTER ↓	ChrF ↑	BLEU ↑	COMET ↑
Generic MT+AIPE	0.12	82.90	66.70	0.92
Production AIPE	0.11	82.96	70.02	0.92
Google Translate	0.21	66.57	42.10	0.89
LLMT	0.20	68.96	46.56	0.89
DeepL	0.22	64.96	41.29	0.88

Automatic metrics consistently rank the two AIPE-based configurations above all other systems, with Production AIPE achieving the lowest HTER (0.11) and highest chrF (82.96), BLEU (70.02), and COMET (0.92), followed closely by Generic MT+AIPE, while LLMT offers marginal gains over the Google Translate baseline and DeepL performs comparably to it. To validate whether these automatic metric differences reflect genuine translation quality gaps as perceived by expert translators, we conducted human evaluations.

5.2 System Ranking

Human rankings were consistent across all aggregation methods (total counts, fractional tie-aware scoring, majority vote, unanimous vote). The order from best to worst was: Generic MT + AIPE $>$ Production AIPE $>$ Google Translate $\approx$ LLMT $>$ DeepL.

Table 2: Row-normalized rank distribution by system

System	Rank 1	Rank 2	Rank 3
Generic MT + AIPE	57%	29%	14%
Production AIPE	47%	30%	23%
Google Translate	44%	34%	22%
LLMT	41%	32%	26%
DeepL	38%	33%	29%

As shown in Table 2, Generic MT + AIPE received Rank 1 in 57% of all its annotations, Production AIPE received 47% and DeepL 38%, which makes a 19-percentage-point gap between strongest and weakest. This ranking held across the individual target languages and domains. Minor variation occurred in the relative positions of LLMT and Google Translate depending on aggregation method, but neither system consistently broke from the middle tier.

5.3 Error Annotation

Error counts mirrored rankings. As we can see in Table 3, Generic MT + AIPE accounted for 15.8% of all annotated errors, Production AIPE for 19.1%, and DeepL for 23.6%, the highest of any system.

Table 3: Error distribution by systems (all errors)

System	% of Total Errors
DeepL	23.6%
LLMT	21.1%
Google Translate	20.4%
Production AIPE	19.1%
Generic MT + AIPE	15.8%

An important nuance concerns error severity. Production AIPE and Generic MT + AIPE share the same AIPE component; the only difference is the pre-translation it processes. Despite this, Production AIPE shows a higher share of major and critical errors (see Table 4). A possible explanation is the variability in pre-translation quality. In the Production AIPE setup, the inputs include fuzzy TM matches, which account for 19% of all segments but 33% of severe errors. This may suggest that when the input is less reliable or more heterogeneous, AIPE is more likely to preserve or amplify those issues.

Table 4: Error distribution by system (major and critical)

System	% Major & Critical
DeepL	25.0%
Production AIPE	21.7%
Google Translate	20.1%
LLMT	19.7%
Generic MT + AIPE	13.7%

Regarding the distribution of error types (rows normalized by system), Style, Terminology, and Accuracy errors are the most frequent categories across all systems (see Table 5). Within its own error distribution, Google Translate shows a higher share of Terminology errors (32%), as expected from a generic model, while, within its own distribution, Production AIPE shows a higher share of Accuracy errors (29%). This pattern may also be related to the characteristics of its pre-translation inputs (which includes fuzzy TM matches), although this remains a tentative interpretation that needs further investigation.

Table 5: Row-normalized error types by system (%)

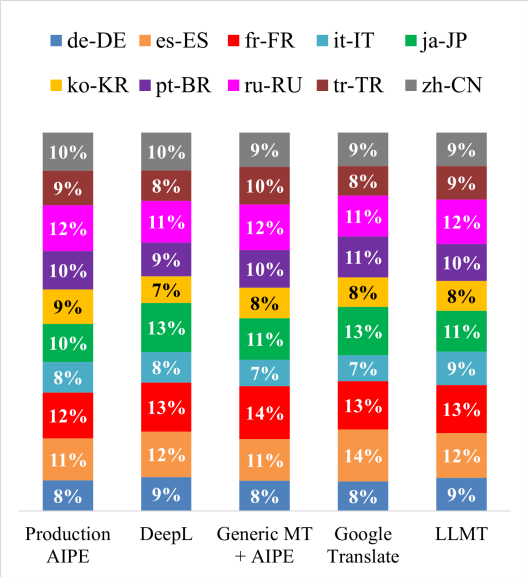
Error Type	Prod. AIPE	DeepL	Gen. MT + AIPE	Google Trans.	LLMT
Style	27.1	28.9	30.9	29.3	30.6
Terminology	22.2	29.0	24.3	32.3	19.9
Accuracy	29.1	24.3	24.4	22.9	25.1
Fluency	16.7	14.8	14.7	13.8	18.7
Design	3.7	1.9	4.5	0.8	4.6
Other	0.8	0.6	0.8	0.6	0.6
Locale conv.	0.4	0.4	0.3	0.2	0.3
Verity	0.1	0.1	0.1	0.1	0.1

5.4 Target Locale-Level Variation

The overall system ranking (Generic MT + AIPE lowest error share, followed by Production AIPE, with DeepL highest) holds across all ten target target languages, though the magnitude of performance differences varies by language. Total error

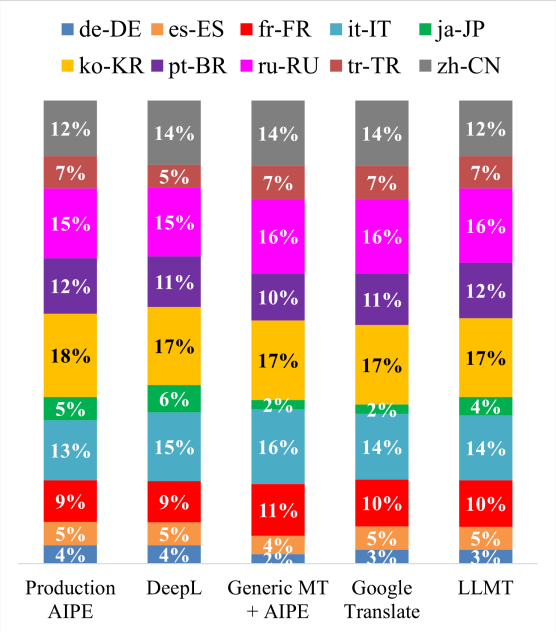
volume and major/critical error concentration do not always move together across target languages, which has practical implications for how deployment risk is assessed. es-ES, fr-FR, and ru-RU show relatively high total error shares (11–14%), while de-DE, it-IT, ko-KR, and tr-TR show lower total shares (7–10%) (see Table 6).

Table 6: Column-normalized error share per system by language



The severity picture (Table 7) differs: ko-KR, despite low total errors, shows a relatively high concentration of major and critical errors, while es-ES shows the reverse (high total errors but a lower share of severe ones). These patterns suggest that locale-level deployment decisions benefit from severity-weighted evaluation rather than total error counts alone. Results for ru-RU and tr-TR should be interpreted with additional caution given their smaller representation in the human evaluation dataset.

Table 7: Column-normalized major + critical error share per system by language



5.5 Content-Level Variation

Content type modulates AIPE's gains, though the dataset distribution means that findings for smaller specialties should be treated as directional rather than definitive. Medical Devices presents a distinct profile: across all systems, this specialty exhibits a higher share of major and critical errors relative to total errors than any other content type, consistent with the precision demands of regulated content (see Table 8).

Table 8: Mean major + critical errors per segment by system and content specialty

Content	Gen. MT + AIPE	Prod. AIPE	Google Trans.	LLMT	DeepL
Marketing	0.06	0.09	0.09	0.11	0.12
Med. Devices	0.22	0.33	0.33	0.30	0.35
Product	0.08	0.14	0.11	0.13	0.15
Prod./Service	0.09	0.15	0.15	0.14	0.20
Support	0.07	0.13	0.14	0.09	0.14

6 Conclusions

This paper presents production-scale evidence that AI post-editing (AIPE) consistently outperforms generic MT baselines and LLM direct translation (LLMT) across ten target languages and five domains. The quality gap between the two AIPE configurations further reveals that pre-translation quality is a key deployment variable: using generic MT alone as input produces fewer and less severe errors than combining MT with fuzzy TM matches,

suggesting that a single, consistent pre-translation source is preferable to mixed input types of varying quality. LLMT may suit domains where top-tier quality is not required, but it does not offer a quality-equivalent alternative. These findings provide concrete guidance for AIPE deployment decisions in professional localization workflows.

6.1 Deployment Implications

The findings have several concrete implications for LSPs considering similar deployments:

1. AIPE on generic MT is a strong default configuration. The marginal cost of a Google Translate pre-translation is low, and the quality gain over standalone LLM or MT translation is consistent and meaningful across target languages and domains.

2. LLMT has a role in cost-tiered workflows. It delivers quality comparable to standalone Google Translate while costing roughly 60% less (approximately $0.40 vs. $1.00 per language per 600 segments, based on actual production spend[2]). For content types where top-tier quality is not critical, LLMT represents a viable and more economical option that also benefits from style guides and Translation Pairs, which generic MT does not receive.

3. Fuzzy TM inputs require careful handling. They are over-represented among severe errors and may benefit from pre-filtering or a tailored prompting strategy before being passed to AIPE.

7 Limitations and Next Steps

The dataset, whilst large, was not perfectly balanced: Product/Service dominated the volume (44%). The inability to systematically distinguish custom from generic MT inputs in the AIPE condition limited the precision of comparisons across pre-translation types. Furthermore, both AIPE and LLMT are powered by a single LLM; other LLMs or generic MT engines may behave differently. Results reflect a specific LSP context and may not generalise directly to other deployment architectures or domains. Next steps include collecting a more balanced dataset, conducting match-type-controlled experiments to isolate the effect of pre-translation quality on AIPE performance, and evaluating alternative LLMs and MT engines to test the generalisability of these findings.

[2]Pricing as of November 2025

References

Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? Yes with GPT-4 as the engine. *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.

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

Lommel, Arle, Maja Popović, and Aljoscha Burchardt. 2014. Assessing inter-annotator agreement for translation error annotation. In *Proceedings of the Workshop on Automatic and Manual Metrics for Operational Translation Evaluation (MTE)*, pages 24–30, Reykjavik, Iceland.

Moslem, Yasmin, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland.

Nunziatini, Mara, Konstantinos Karageorgos, Aaron Schliem, and Mikaela Grace. 2025. OPAL Enable: Revolutionizing localization through advanced AI. In *Proceedings of Machine Translation Summit XX: Volume 2*, pages 83–85, Geneva, Switzerland. European Association for Machine Translation.

Popović, Maja. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal.

Raunak, Vikas, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. Leveraging GPT-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024, Singapore.

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.

Snover, Matthew, Bonnie Dorr, Richard 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 (AMTA)*, pages 223–231, Cambridge, MA.

Reasoning as Supportive Context for Machine Translation: A Case Study on Hindi to Bengali Language Pair

Kshetrimayum Boynao Singh[1,2]*, **Saksham Singh**[1]*, **Partha Pakray**[2], **Asif Ekbal**[1]
Department of Computer Science and Engineering
[1]Indian Institute of Technology Patna, India
[2]National Institute of Technology Silchar, India
{boynfrancis, meetsakshamsingh, pakraypartha, asif.ekbal}@gmail.com

Abstract

We investigate whether reasoning information can enhance machine translation when incorporated as supportive context during training and inference. Using Hindi-Bengali translation as a case study, we define five reasoning components: Key Terms, Syntactic, Semantic, Pragmatic, and Paraphrase. We conduct a complete ablation across all 31 possible combinations using Gemma-3-1B-Instruct and evaluate on multi-domain benchmark with BLEU, chrF, and TER. Evaluation results show that reasoning effectiveness depends on its type and composition rather than quantity. Combining multiple heterogeneous signals causes objective diffusion, degrading performance. The compact Semantic and Paraphrase combination proves optimal, and providing it during inference yields 23.86 BLEU compared to 22.12 from standard fine-tuning a +1.74 BLEU gain across eight domains. These findings demonstrate that targeted semantic guidance consistently and meaningfully improves the compact translation models.

1 Introduction

Neural machine translation (MT) has advanced substantially with multilingual pretraining and instruction-tuned language models. However, translation quality remains uneven across many language pairs, especially when model capacity is limited. This is true even for related Indic languages, such as Hindi and Bengali, where lexical overlap coexists with differences in script, morphology, register, and domain-specific usage.

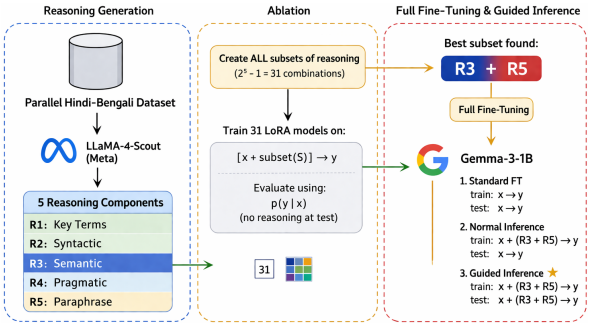


Figure 1: Overview of our reasoning-augmented MT methodology for Hindi-Bengali translation

Recent work has shown that reasoning signals can improve performance on tasks, such as arithmetic, logical inference, and commonsense reasoning (He et al., 2020). These findings have motivated the use of reasoning-oriented supervision in a broader range of natural language processing (NLP) problems. However, most prior work studies settings in which intermediate reasoning is either directly useful for the final prediction or is explicitly provided during inference (Lee et al., 2025) (Ghazaryan et al., 2025). Machine translation differs in an important way: the goal is not to generate reasoning traces, but to produce a fluent target sentence that faithfully preserves the meaning of the source. This raises a natural question: can reasoning information improve machine translation when it is used only as supportive context ? On the one hand, such information may clarify sentence meaning and foreground semantically important content. On the other hand, it may increase input complexity, introduce noisy supervision, or create a difference between training and inference conditions.

*Equal contribution.

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 67–76
Tilburg, The Netherlands, June 2026.

To study this, we investigate reasoning as supportive context for Hindi–Bengali translation. During fine-tuning, the model learns from structured reasoning annotations paired with parallel data, and at inference, it uses both source text and reasoning context to generate translations. We define five components: *Key Terms*, *Syntactic*, *Semantic*, *Pragmatic*, and *Paraphrase*, and perform ablations over all non-empty combinations, yielding 31 reasoning-based models alongside a standard baseline. All models are fine-tuned from Gemma-3-1B-Instruct and evaluated using BLEU, chrF, and TER (Papineni et al., 2002).

Results show that reasoning impact varies across setups; adding more components does not ensure better performance. Effectiveness depends on composition, not quantity. A compact *Semantic+Paraphrase* combination performs best and remains stable, indicating targeted reasoning is beneficial.

2 Related Work

2.1 Reasoning in Large Language Models

Recent advances in large language models (LLMs)(Zhu et al., 2024; He et al., 2024) have sparked growing interest in reasoning-oriented (Nguyen and Xu, 2025) supervision. Studies on chain-of-thought (Wang et al., 2025; Jiang et al., 2025) prompting, tree of thoughts (Yao et al., 2023), and related reasoning strategies (Fan et al., 2025), such as self-consistency and least-to-most prompting show that models can achieve significant gains on tasks, such as arithmetic reasoning, commonsense inference (Liu et al., 2023), and symbolic problem (Gaur and Saunshi, 2023) solving. In these settings, reasoning steps help models structure their internal representations (Dabre et al., 2023) and guide (Jiang et al., 2025) the generation of correct outputs. However, most of this work focuses on tasks where reasoning is directly part of the solution process. Machine translation differs from these scenarios because the final output is a fluent sentence rather than an explicit reasoning trace, making it unclear whether reasoning supervision transfers effectively to translation tasks.

2.2 Machine Translation for Indic Languages

Neural machine translation for the Indic languages has improved through multilingual pretrained models (Bala Das et al., 2023; Singh et al., 2025a; Singh et al., 2026a), larger parallel datasets and dedicated benchmarking efforts. These systems have enhanced cross-lingual transfer (Singh et al., 2023) and improved translation quality across several South Asian languages. Nevertheless, many approaches rely on large-scale models and multilingual training pipelines (Verma et al., 2025). Closely related language pairs, such as Hindi and Bengali, still pose meaningful challenges due to differences in script, morphological structure, and domain-specific language use. As a result, translation accuracy depends not only on lexical similarity but also on reliable semantic interpretation.

2.3 Translation Support Beyond Parallel Data

Several studies have investigated enriching translation models with additional forms of guidance beyond plain parallel data. These approaches include terminology constraints, syntactic annotations, alignment information (Chen et al., 2017) domain labels (Yan et al., 2018; Singh et al., 2026b), and paraphrastic variants. Such signals aim to help models preserve technical terminology (Gao et al., 2025), resolve structural ambiguity and capture alternative expressions of the same meaning. However, combining multiple support signals reflects nuanced interactions rather than cumulative improvements (Singh et al., 2025b). Auxiliary information may overlap in function and influence the model's attention, shaping how effectively the central translation objective is emphasized.

2.4 Research Gaps in Reasoning for MT

Despite growing interest in reasoning supervision, its role in MT remains underexplored. Existing work rarely examines whether reasoning-style information can improve translation when used during training time, as well as inference time, and how the translation empowers. Furthermore, examining which path of reasoning signals are beneficial for translation and how different forms of the other supportive information interact when combined. This work addresses these gaps by decomposing reasoning into multiple linguistic components and conducting a complete combinational ablation study. By evaluating all the reasoning subsets, we identify which forms of reasoning supports and improve the translation quality, and which combinations degrade it, as well as which reasoning section is required during the inference time in order to improve translation capabilities.

3 Dataset

3.1 Parallel Corpus

We conduct experiments on a Hindi–Bengali parallel corpus spanning eight domains, *viz.* Agriculture, Tourism, Governance, Climate, Healthcare, Science and Technology, Judiciary, and Education. After filtering for sentence quality and annotation completeness, the final training set contains 36,040 sentence pairs, and the test set contains 2,000 sentence pairs. The test set supports both overall and domain-wise evaluation, with 200 examples for each domain except Education, which contains 600 examples.

The training corpus contains 771,621 Hindi tokens and 603,131 Bengali tokens, with average sentence lengths of 21.4 and 16.7 tokens, respectively. Bengali also shows higher lexical diversity, with 61,284 unique word types compared to 45,170 in Hindi. The test set follows a similar pattern, containing 41,027 Hindi tokens and 32,061 Bengali tokens. Overall, the corpus provides a diverse benchmark for evaluating translation quality across domains with different linguistic characteristics, ranging from technical and descriptive text to formal institutional and instructional language.

3.2 Data Filtering and Quality Considerations

Before training, we filter the corpus to remove noisy or misaligned sentence pairs. This process enforces basic quality constraints, such as reasonable sentence length, alignment consistency, and valid reasoning annotations. Examples with incomplete, empty, or malformed reasoning fields are excluded. These steps reduce training noise and improve the reliability of the subsequent ablation analysis.

3.3 Reasoning Annotation Pipeline

We generate structured reasoning annotations for each training sentence using Llama-4-Scout-17B-16E-Instruct. To leverage its stronger reasoning ability in English, each Hindi–Bengali sentence pair is first mapped through an intermediate Hindi–English translation, from which reasoning annotations are produced and then adapted back into Hindi and Bengali. To ensure scalability and consistency, the reasoning context is systematically transferred using Google Translate. The resulting annotations are designed to provide supportive contextual signals for translation rather than exhaustive linguistic analysis. For each sentence pair, the pipeline generates five reasoning components: Key terms (R_1), Syntactic (R_2), Semantic (R_3), Pragmatics (R_4), and Paraphrase (R_5). When applicable, annotations are produced for both Hindi and Bengali to maintain cross-lingual alignment. Structurally inconsistent annotations are filtered before training, yielding stable and uniform inputs for all 31 reasoning-based ablation configurations.

3.4 Reasoning Components

The five reasoning components capture complementary types of translation-relevant information.

R_1: **Key terms** identify important content words, explain their role in context, and clarify ambiguous expressions.

R_2: **Syntactic** information describes grammatical structure, including dependencies, modifiers, and clause boundaries, to support preservation of compositional meaning.

R_3: **Semantic** reasoning captures the core meaning of a sentence by identifying entities, actions, and their relations.

R_4: **Pragmatics** models discourse-level cues, such as speaker intent, tone, and contextual dependence beyond literal sentence meaning.

R_5: **Paraphrase** provides semantically equivalent reformulations with different lexical or syntactic realizations while preserving the original meaning.

These components differ not only in function but also in length. As shown in Table 1, Syntactic and Key Terms are the most verbose annotations, averaging more than 80 tokens per instance, whereas Semantic and Paraphrase are substantially more compact, averaging 45 and 24 tokens, respectively. This variation allows us to study both the effect of reasoning type and the effect of reasoning scale on translation performance.

4 Methodology

Our goal is to examine whether structured reasoning can serve as supportive context for machine translation, and how its presence or absence at inference time affects performance. To this end, we adopt a three-phase methodology: reasoning augmentation using *Llama-4-Scout-17B-16E-Instruct*, an exhaustive ablation study on *google/gemma-3-1b-it*, and an evaluation of inference settings to determine which reasoning configuration best improves translation quality.

Training data	Sentences	Tokens	Vocabulary	# Characters	Avg Tokens/Sent
Hindi (S)	36,040	771,621	45,170	4,008,578	21.4
Bengali (T)	36,040	603,131	61,284	3,930,184	16.7
Hindi Key-terms (R_1)	36,040	2,946,267	48,260	15,414,987	81.7
Bengali Key-terms (R_1)	36,040	2,226,749	90,354	16,424,124	61.8
Hindi Syntactic (R_2)	36,040	3,077,551	66,757	14,149,631	85.4
Bengali Syntactic (R_2)	36,040	2,487,377	110,260	8,353,154	69.0
Hindi Semantic (R_3)	36,040	1,628,938	28,499	4,506,294	45.2
Bengali Semantic (R_3)	36,040	1,194,331	48,063	14,561,377	33.1
Hindi Pragmatics (R_4)	36,040	2,675,864	32,384	16,420,421	74.2
Bengali Pragmatics (R_4)	36,040	2,001,633	49,014	7,986,367	55.5
Hindi Paraphrase (R_5)	36,040	857,822	24,545	13,794,926	23.8
Bengali Paraphrase (R_5)	36,040	633,695	43,157	4,342,617	17.6

Table 1: Training corpus statistics for the Hindi-Bengali parallel data and the five reasoning components on both sides

Domain (Sentence)	Language	Tokens
Agriculture (200)	Hindi	3,956
	Bengali	3,249
Tourism (200)	Hindi	4,028
	Bengali	3,222
Governance (200)	Hindi	4,160
	Bengali	3,319
Climate (200)	Hindi	4,093
	Bengali	3,296
Healthcare (200)	Hindi	4,049
	Bengali	3,174
Science_Technology (200)	Hindi	4,117
	Bengali	3,218
Judiciary (200)	Hindi	4,191
	Bengali	3,112
Education (600)	Hindi	12,433
	Bengali	9,471
Overall (2000)	Hindi	41,027
	Bengali	32,061

Table 2: Domain-wise and overall statistics of the Hindi-Bengali test set, including sentence counts and token counts for Hindi and Bengali

4.1 Problem Formulation

Let x denote a Hindi source sentence and y its Bengali translation. In standard machine translation, the model learns the conditional mapping $p(y \mid x)$. We extend this setting by augmenting the input with structured reasoning. Let $\mathcal{R} = \{R_1, R_2, R_3, R_4, R_5\}$ denote the five reasoning components. For any subset $S \subseteq \mathcal{R}$, the model is trained on the inputs formed by concatenating x with the reasoning components in S, while the target remains the translation y.

We evaluate under two inference paradigms: **Zero-Reasoning Inference:** Evaluation under $p(y \mid x)$, where reasoning is not provided at test time. This isolates the effect of training-time exposure to reasoning.

Reasoning Guided Inference: Evaluation under $p(y \mid x, S_{\text{optimal}})$, where the best reasoning subset is identified during ablation study and is provided along with the source during testing.

4.2 Base Model and Reasoning Pipeline

We use *Gemma-3-1B-Instruct* as the translation backbone, as it is compact enough for auxiliary context to meaningfully affect translation behavior. All reasoning components (R_1 to R_5) for both training and test data are generated using *Llama-4-Scout-17B-16E-Instruct*. Reasoning is first produced in English and then translated into Hindi and Bengali to maintain structural consistency. To avoid target leakage, test-time reasoning in Guided Inference is generated strictly from the Hindi source sentence, without access to the gold Bengali translation.

4.3 Structured Training Format

Each training instance is formatted as a prompt containing the Hindi source sentence, the selected reasoning components in Hindi and Bengali, and an instruction to generate only the Bengali translation. The loss is applied only to the translation output, ensuring that the model is not trained to reproduce the reasoning itself. To identify which reasoning signals are useful, we perform a complete ablation over all non-empty combinations of the five components using LoRA. This results in $2^5 - 1 = 31$ reasoning-conditioned models, along with a standard source-target finetuning. All 31 models are trained under identical settings: 1 epoch, learning rate 1×10^{-4} with cosine scheduling, batch size 4, and gradient accumulation over 16 steps, yielding an effective batch size of 64. We use `paged_adamw_8bit` optimizer (Kingma and Ba, 2017) in `bfloat16` precision. During this phase, evaluation is conducted using Zero-Reasoning Inference to identify the optimal subset S_{optimal}.

4.4 Full Parameter Fine-Tuning and Inference Evaluation

After identifying the optimal combination ($R_3 + R_5$), we perform full parameter fine-tuning to evaluate the effect of training-inference alignment more rigorously. We consider three setups:

1. **Standard Finetune:** Full fine-tuning on

source-target pairs only, evaluated under zero-reasoning inference.

2. **Normal Inference:** Full fine-tuning with $R_3 + R_5$ during training, but evaluation without reasoning inference. This measures the penalty caused by reasoning when is not provided at during the inference time.

3. **Guided Inference:** Full fine-tuning with $R_3 + R_5$ during training, and evaluation with the same $R_3 + R_5$ context at test time. This removes the difference between training and testing and measures the full benefit of reasoning-augmented translation.

5 Results

Our evaluation proceeds in two stages. First, we conduct an exhaustive *LoRA* ablation study to assess the effect of reasoning complexity and identify the most effective supportive context. As a reference point, the original baseline *Gemma-3-1B-Instruct* model without task-specific fine-tuning achieves 4.99 BLEU, 28.87 chrF, and 137.41 TER on the test set, underscoring the need for task-specific adaptation. Second, we evaluate full-parameter fine-tuning to examine how performance changes when reasoning is not provided at test time versus when it is. The complete results for all 31 reasoning configurations are presented in Table 3.

Group	Model	BLEU	chrF	TER↓
Baseline	Gemma-3-1B-Instruct	4.99	28.87	137.41
${}^5C_1$ *LoRa*	R_1 (Key terms)	16.13	47.37	71.57
	R_2 (Syntactic)	14.77	45.31	74.13
	R_3 (Semantic)	**16.58**	**48.03**	**70.23**
	R_4 (Pragmatics)	14.45	43.36	75.36
	R_5 (Paraphrase)	15.27	44.64	74.08
${}^5C_2$ *LoRa*	R_1+R_2	15.03	45.88	74.15
	R_1+R_3	14.91	44.41	75.40
	R_1+R_4	13.30	40.25	79.89
	R_1+R_5	14.94	44.62	74.50
	R_2+R_3	15.54	46.64	72.30
	R_2+R_4	14.86	45.35	73.46
	R_2+R_5	15.89	47.07	71.39
	R_3+R_4	15.88	47.32	71.70
	R_3+R_5	**16.91**	**48.92**	**69.50**
	R_4+R_5	15.72	46.92	70.73
${}^5C_3$ *LoRa*	R_1+R_2+R_3	15.70	46.82	71.36
	R_1+R_2+R_4	14.42	44.57	74.76
	R_1+R_2+R_5	15.26	45.66	72.70
	R_1+R_3+R_4	13.72	44.97	75.57
	R_1+R_3+R_5	15.78	46.96	72.03
	R_1+R_4+R_5	14.42	44.27	75.76
	R_2+R_3+R_4	15.42	46.31	72.24
	R_2+R_3+R_5	15.52	46.52	71.62
	R_2+R_4+R_5	15.09	45.64	72.95
	R_3+R_4+R_5	**15.85**	**48.09**	**71.72**
${}^5C_4$ *LoRa*	R_1+R_2+R_3+R_4	14.45	44.75	75.57
	R_1+R_2+R_3+R_5	13.91	42.61	75.57
	R_1+R_2+R_4+R_5	13.98	44.17	76.69
	R_1+R_3+R_4+R_5	13.95	43.48	76.65
	R_2+R_3+R_4+R_5	**14.95**	**45.65**	**73.58**
${}^5C_5$ *LoRa*	Gemma all R_1 to R_5	11.64	38.35	84.13
SFT	Finetune with *LoRa*	17.55	50.00	67.41
	Full Finetune	**22.12**	**56.06**	**60.51**
Full FT R_3+R_5	Standard Inference	22.26	55.54	61.61
	Reasoning Inference	**23.86**	**57.57**	**58.55**

Table 3: Translation performance metrics for all 31 *LoRA* reasoning ablation configurations and the full-parameter fine-tuning setups. The results demonstrate that the Guided Inference approach using the $R_3 + R_5$ context yields the highest overall performance.

5.1 Phase 1: Ablation Study and the Role of Semantic Guidance

The Phase 1 ablation results show that the translation quality depends more on the type of reasoning than on the amount of reasoning added. Standard *LoRA* fine-tuning on source-target pairs achieves the best score of 17.55 BLEU, but performance does not improve monotonically as more reasoning components are introduced across ${}^5C_1$ to ${}^5C_5$. Instead, larger and more heterogeneous combinations often reduce performance, with the full five-component setting (${}^5C_5$) dropping to 11.64 BLEU and most four-component settings (${}^5C_4$) remaining below 15 BLEU points. This suggests that excessive reasoning can dilute the core translation objective and burden a compact model with unnecessary context. In contrast, compact semantic guidance proves more effective. Among single-component settings, Semantic reasoning (R_3) performs the best with 16.58 BLEU, while among pairwise settings, $R_3 + R_5$ achieves the highest score of 16.91 BLEU, showing that meaning-focused support and paraphrastic reformulation help translation more than heavier reasoning combinations.

5.2 Phase 2: Training-Inference Alignment

We next evaluate the best subset, $R_3 + R_5$, under full-parameter fine-tuning. As shown in Table 3, the standard full fine-tuning reaches 22.12 BLEU. When the model is trained with $R_3 + R_5$ but evaluated without reasoning, performance increases only slightly to 22.26 BLEU points. This small gain suggests a difference in training and testing: the model learns under reasoning-augmented inputs but cannot fully exploit that information when it is removed at test time. When the same $R_3 + R_5$ context is provided during inference, performance rises to 23.86 BLEU, with corresponding gains in chrF and TER. This result shows that reasoning is most effective when training and inference conditions remain aligned.

Model Configuration	Agri.	Tour.	Gov.	Clim.	Health	Sci&T	Judic.	Educ.	Overall
LoRA: R_3 Only	15.06	7.27	12.06	14.70	15.90	22.00	13.66	20.72	16.58
LoRA: $R_3 + R_5$	15.20	7.82	11.87	14.92	16.32	23.03	13.30	21.18	16.91
LoRA: $R_3 + R_4 + R_5$	14.89	7.50	12.29	13.99	14.84	20.43	12.58	20.16	15.85
LoRA: $R_2 + R_3 + R_4 + R_5$	12.59	7.07	10.67	14.03	14.37	19.20	11.52	19.18	14.95
LoRA: All 5	10.12	5.12	8.63	11.20	11.08	19.05	7.08	14.72	11.64
Full SFT: Standard Finetune	19.92	12.87	17.44	**21.95**	19.18	24.85	18.19	28.27	22.12
Full SFT: $R_3 + R_5$ (Normal Test)	20.20	12.26	17.73	19.90	19.68	26.31	**19.60**	28.13	22.26
Full SFT: $R_3 + R_5$ (Guided Inference)	**21.86**	**15.32**	**20.32**	21.16	**20.73**	**30.50**	19.12	**29.12**	**23.86**

Table 4: Domain-wise BLEU scores across 8 distinct evaluation domains. Guided Inference demonstrates broad robustness, yielding significant improvements in highly technical domains like Science & Technology and Governance

5.3 Domain-Wise Analysis

We further evaluate performance across eight domains, as shown in Table 4. The guided inference setup with $R_3 + R_5$ improves over the standard full fine-tuning in nearly all the domains. The largest gains appear in Science & Technology (24.85 to 30.50 BLEU), Governance (17.44 to 20.32 BLEU), and Tourism (12.87 to 15.32 BLEU). Although, Climate domain shows a small drop relative to the finetune, the overall trend indicates that compact semantic support improves robustness across diverse domains, especially in technically demanding settings.

6 Discussion

6.1 Aligning Context with Translation

The contrast across reasoning configurations reveals an important trade-off in context-augmented translation. For compact models, adding large amount of heterogeneous reasoning can weaken performance instead of improving it. We refer to this effect as *objective diffusion*: the model must allocate limited capacity to processing multiple forms of auxiliary context, which can distract from the core translation objective. In contrast, the strong performance of $R_3 + R_5$ highlights the value of *semantic anchoring*. Semantic reasoning (R_3) helps the model focus on the core meaning of the source sentence, while Paraphrase (R_5) provides a meaning-preserving reformulation that supports flexible generation. Because both the components operate at the level of semantic preservation, they align more directly with the goal of translation than broader and more heterogeneous reasoning combinations.

6.2 The Role of Guided Inference

Our results also show that training-inference alignment is critical in reasoning-augmented MT. When a model is trained with reasoning-rich inputs but evaluated without reasoning, the benefit is limited. This suggests that the model learns to rely on the additional semantic structure during training, and removing it at test time creates a difference. Reasoning guided inference addresses this issue by supplying source-derived reasoning during testing, thereby preserving the input structure seen in training. Under this setup, performance improves from 22.12 BLEU to 23.86 BLEU. This gain indicates that reasoning is most effective not as an isolated training signal, but as a consistent form of supportive context available throughout the translation process.

7 Conclusion

In this paper, we study whether structured reasoning can improve machine translation when used as supportive context for Hindi–Bengali translation. Through a complete 31-model ablation studies, we found that reasoning is not uniformly beneficial: large heterogeneous combination often degrade performance, whereas compact semantic guidance is more effective. In particular, the combination of Semantic and Paraphrase ($R_3 + R_5$) emerged as the strongest configuration. Our full-parameter experiments further showed that the benefit of reasoning depends on training-inference alignment. When reasoning is used during training but removed at test time, gains remain small. When the same $R_3 + R_5$ context is also provided during inference, performance rises from 22.12 BLEU to 23.86 BLEU points, with improvements across multiple domains. Overall, our findings suggest that reasoning can benefit compact translation models when it is semantically focused, source-grounded, and used consistently across training and inference. More broadly, the results highlight that the effectiveness of supportive context depends not on its quantity, but on how closely it aligns with the translation objective.

Acknowledgment

The authors gratefully acknowledge the COIL-D (Centre of Indian Language Data) Project under Bhashini, funded by the Ministry of Electronics and Information Technology (MeitY), Government of India, for providing the resources and computational infrastructure that supported this research.

References

Bala Das, Sudhansu, Atharv Biradar, Tapas Kumar Mishra, and Bidyut Kr. Patra. 2023. Improving multilingual neural machine translation system for indic languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(6), June.

Chen, Huadong, Shujian Huang, David Chiang, and Jiajun Chen. 2017. Improved neural machine translation with a syntax-aware encoder and decoder. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1936–1945, Vancouver, Canada, July. Association for Computational Linguistics.

Dabre, Raj, Bianka Buschbeck, Miriam Exel, and Hideki Tanaka. 2023. A study on the effectiveness of large language models for translation with markup. In Utiyama, Masao and Rui Wang, editors, *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 148–159, Macau SAR, China, September. Asia-Pacific Association for Machine Translation.

Fan, Yuchun, Yongyu Mu, YiLin Wang, Lei Huang, Junhao Ruan, Bei Li, Tong Xiao, Shujian Huang, Xiaocheng Feng, and Jingbo Zhu. 2025. SLAM: Towards efficient multilingual reasoning via selective language alignment. In Rambow, Owen, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9499–9515, Abu Dhabi, UAE, January. Association for Computational Linguistics.

Gao, Hui, Jing Zhang, Peng Zhang, and Chang Yang. 2025. Consistency rating of semantic transparency: an evaluation method for metaphor competence in idiom understanding tasks. In Rambow, Owen, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10460–10471, Abu Dhabi, UAE, January. Association for Computational Linguistics.

Gaur, Vedant and Nikunj Saunshi. 2023. Reasoning in large language models through symbolic math word problems.

Ghazaryan, Gayane, Erik Arakelyan, Isabelle Augenstein, and Pasquale Minervini. 2025. SynDARin: Synthesising datasets for automated reasoning in low-resource languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6459–6466, Abu Dhabi, UAE, January. Association for Computational Linguistics.

He, Jie, Tao Wang, Deyi Xiong, and Qun Liu. 2020. The box is in the pen: Evaluating commonsense reasoning in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3662–3672, Online, November. Association for Computational Linguistics.

He, Zhiwei, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Jiang, Gangwei, Yahui Liu, Zhaoyi Li, Wei Bi, Fuzheng Zhang, Linqi Song, Ying Wei, and Defu Lian. 2025. What makes a good reasoning chain? uncovering structural patterns in long chain-of-thought reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6490–6514, Suzhou, China, November. Association for Computational Linguistics.

Kingma, Diederik P. and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Lee, Minjae, Youngbin Noh, and Seung Jin Lee. 2025. A testset for context-aware LLM translation in Korean-to-English discourse level translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1632–1646, Abu Dhabi, UAE, January. Association for Computational Linguistics.

Liu, Xuebo, Yutong Wang, Derek F. Wong, Runzhe Zhan, Liangxuan Yu, and Min Zhang. 2023. Revisiting commonsense reasoning in machine translation: Training, evaluation and challenge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15536–15550, Toronto, Canada, July. Association for Computational Linguistics.

Nguyen, Lam and Yang Xu. 2025. Reasoning for translation: Comparative analysis of chain-of-thought and tree-of-thought prompting for LLM translation. In Zhao, Jin, Mingyang Wang, and Zhu Liu, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 259–275, Vienna, Austria, July. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Singh, Kshetrimayum Boynao, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Ningthoujam Justwant Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2023. A comparative study of transformer and transfer learning MT models for English-Manipuri. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 791–796, Goa University, Goa, India, December. NLP Association of India (NLPAI).

Singh, Kshetrimayum Boynao, Asif Ekbal, and Partha Pakray. 2025a. Evaluating IndicTrans2 and ByT5 for English–Santali machine translation using the ol chiki script. In *Proceedings of the 1st Workshop on Multimodal Models for Low-Resource Contexts and Social Impact (MMLoSo 2025)*, pages 95–100, Mumbai, India, December. Association for Computational Linguistics.

Singh, Ningthoujam Avichandra, Boynao Kshetrimayum, Ningthoujam Justwant Singh, Thoudam Doren Singh, and Shilpa Sharma. 2025b. Quantifying the impact of data scale and quality on neural machine translation for low-resource language pair. In *2025 OITS International Conference on Information Technology (OCIT)*, pages 1–6.

Singh, Kshetrimayum Boynao, Soham Bhattacharjee, Baban Gain, Asif Ekbal, and Partha Pakray. 2026a. A comparative study of fine-tuned mbart and indictrans2 for english-hindi legal machine translation. *TechRxiv*, 2026(0102).

Singh, Kshetrimayum Boynao, Soham Bhattacharjee, Baban Gain, Asif Ekbal, and Partha Pakray. 2026b. A comparative study of fine-tuned mbart and indictrans2 for english-hindi legal machine translation. *TechRxiv*, 2026(0102).

Verma, Sshubam, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. MILU: A multi-task Indic language understanding benchmark. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10076–10132, Albuquerque, New Mexico, April. Association for Computational Linguistics.

Wang, Jiaan, Fandong Meng, Yunlong Liang, and Jie Zhou. 2025. DRT: Deep reasoning translation via long chain-of-thought. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6770–6782, Vienna, Austria, July. Association for Computational Linguistics.

Yan, Shen, Leonard Dahlmann, Pavel Petrushkov, Sanjika Hewavitharana, and Shahram Khadivi. 2018. Word-based domain adaptation for neural machine translation. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 31–38, Brussels, October 29-30. International Conference on Spoken Language Translation.

Yao, Shunyu, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Zhu, Wenhao, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In Duh, Kevin, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico, June. Association for Computational Linguistics.

A Appendix

B Hardware, Software, and Reproducibility

All experiments were conducted on a high-performance computing server equipped with two NVIDIA A100 GPUs, each with 80GB of memory, enabling efficient large-scale training and inference. The computational environment was configured using Python 3.13.2, along with PyTorch 2.7.1 and Transformers 4.57.3 for model implementation and experimentation. These libraries ensured compatibility with modern large language model architectures and optimized GPU acceleration.

To maintain consistency across experiments, we used fixed configurations for model loading, batching, and decoding strategies. We also experimented with a validation set during training; however, the scores remained nearly identical to training without it. Therefore, the final model was trained without a validation set to fully utilize the available training data. The inference process was optimized through batch processing and parallel execution for efficient GPU utilization. Additionally, external API-based components used for reasoning generation were managed through multiple API keys and a load-balanced request mechanism to ensure stability and scalability.

For reproducibility, we maintained detailed configurations, cached intermediate outputs such as reasoning annotations, and fixed key parameters wherever applicable. We will publicly release the complete codebase, including pre-processing scripts, inference pipelines, and configuration files, to support replication and further research. To support reproducibility, all code and datasets are publicly available on GitHub[1] and our research group webpage.[2]

[1] `https://github.com/helloboyn/RSC4MT`

[2] `https://ai-nlp-ml.github.io/resources.html`

Reasoning Generation Prompt

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are an expert linguistic analyst. Analyze the given Hindi sentence and generate
5 types of reasoning.

Return ONLY one valid JSON object with exactly these keys:
"key_terms", "syntactic", "semantic", "pragmatic", "paraphrase"

Strict output rules for every value:
- Do NOT start with phrases like "The sentence", "This sentence means", or similar.
- Each value must be a single clean sentence or phrase.
- Do not add explanations outside JSON.
- Do not mention that you are analyzing.
- Avoid redundancy, filler, and obvious restatements.
- If a field is not applicable, return an empty string "".

{
  "key_terms": "Identify the most important content words (nouns, verbs, key entities,
  domain-specific terms). Focus only on meaningful lexical units essential for translation.
  Avoid full sentence repetition.",

  "syntactic": "Analyze the syntactic structure using linguistic terms: grammatical
  ↪  relations,
  subject-object-verb roles, modifiers, clause structure, and word order.
  Do not start with words like 'sentence' or 'phrase'.",

  "semantic": "Provide a precise one-sentence semantic interpretation. Capture the exact
  ↪  meaning,
  resolve ambiguity, and ensure the intended sense of words is clear.",

  "pragmatic": "Analyze contextual and pragmatic aspects in one sentence: tone, intent,
  ↪  formality,
  audience, domain, and any implied meaning or communicative purpose.",

  "paraphrase": "Generate a semantically equivalent paraphrase in one sentence, improving
  ↪  clarity
  and reducing ambiguity while preserving exact meaning."
}
<|eot_id|><|start_header_id|>user<|end_header_id|>
Sentence: "{INPUT_SENTENCE}"
Return valid JSON only.
<|eot_id|>
```

Table 5: Prompt used for generating structured linguistic reasoning from Hindi sentences. The prompt instructs the model to produce five components; key terms, syntactic structure, semantic interpretation, pragmatic context, and paraphrase used to enhance translation quality

Output Sample 1: Qualitative Comparison of Reasoning-Augmented Translation Models

Source: दोपहर के भोजन के उपरांत, शिष्टमंडल को ऐतिहासिक मेलबोर्न क्रिकेट ग्राउंड (एमसीजी) ले जाया गया।

Target: দুপুরের খাওয়ার পর, প্রতিনিধিদলকে ঐতিহাসিক মেলবোর্ন ক্রিকেট গ্রাউন্ড (এম সি জি) নিয়ে যাওয়া হয় ।

R3 LORA: দোপহর ের খাবার পর, বডিকে ঐতিহাসিক মেলবর্ন ক্রিকিং গ্রাউন্ড (এমসিজি) নিয়ে যাওয়া হয়েছিল।

R3R5 LORA: দোপহর ের খাবার পর, মাননীয়দের ঐতিহাসিক মেলবর্ন কিকিং গ্রাউন্ড (এমসিজি) নিয়ে যাওয়া হয়েছিল।

R3R4R5 LORA: দুপুর বেলার পরে, বডিমন্ডন ক্যানিংক্রিম গ্রাউন্ড (এমসিজি) - যাত্মকে নিয়ে যাওয়া হয়েছিল।

R2R3R4R5 LORA: দুপুর খাবারের পর, বডিকে ঐতিহাসিক মকম্বন ক্রিকেট গ্রাউন্ড (এমসিজি) নিয়ে যাওয়া হয়।

ALL 5 LORA: বিকেল ব্রেক করার পর, চিঠিটি ঐতিহাসিক ম chlorides ্রি ন গ্রাউন্ডে (এমসিজি) নিয়ে যাওয়া হয়।

FULL FT SFT: দুপুরের খাবারের পর, দলীয়কে ঐতিহাসিক মেলবোন ক্রিকেট মাঠ (এমসিজি) নিয়ে যাওয়া হয়েছিল।

FULL FT R3R5:

NORMAL TEST: দুপুরের খাবারের পর, প্রতিনিধিদল ঐতিহাসিক মেলবরিন ক্রিকেট গ্রাউন্ডে (এমসিজি) পৌঁছায়।

GUIDED TEST: দুপুরের খাবারের পর, প্রতিনিধিদলকে ঐতিহাসিক মেলবোর্ন ক্রিকেট গ্রাউন্ড (এমসিজি) নিয়ে যাওয়া হয়।

Table 6: The source sentence states that "After lunch, the delegation was taken to the historic Melbourne Cricket Ground (MCG)" **R3R4R5 LoRA** shows clear linguistic errors, including incorrect formations such as "বডিমন্ডন ক্যানিংক্রিম" and the non-Bengali word "যাত্মকে". The **Full FT SFT** output contains inappropriate lexical choice ("দলীয়েক"), reducing naturalness. In Normal Test, "মেলবোর্ন" is mistransliterated as "মেলবরিন", while the Guided Test correctly produces "মেলবোর্ন", improving accuracy. Additionally, the Guided output uses proper case marking ("প্রতিনিধিদলকে") and construction ("নিয়ে যাওয়া হয়"), though it slightly alters voice. Overall, the **Guided Test** yields the most linguistically accurate and natural Bengali translation.

Output Sample 2: Qualitative Comparison of Reasoning-Augmented Translation Models

Source: वहीं, पंजाब स्टेट पावर कॉर्पोरेशन लिमिटेड रावी नदी पर रंजीत सागर बांध का ख्याल रखता है।

Target: অন্য দিকে, পাঞ্জাব স্টেট পাওয়ার কর্পোরেশন লিমিটেড রাভি নদীর উপর রঞ্জিত সাগর বাঁধের তত্ত্বাবধান করে।

R3 LORA: অন্যদিকে, পংজাব স্টেট পাওয়ার কর্পোরেশন লিমিটেড রাভি নদীকে রিংজাত সাগর বাঁধের দায়িত্বে রাখে।

R3R5 LORA: অন্যদিকে, পংজাব স্টেট পাওয়ার কর্পোরেশন লিমিটেড রাভি নদীকে রিংজাত সাগর বাঁধের দায়িত্বে রাখে।

R3R4R5 LORA: অন্যদিকে, পংজাব স্টেট পাওয়ার কর্পোরেশন লিমিটেড রাভি নদীকে নিয়ন্ত্রণ করে।

R2R3R4R5 LORA: অন্যদিকে, পংজাব স্টেট পাওয়ার কর্পোরেশন লিমিটেড রায়গড়□ের উপর রঞ্জিত সাগর বাঁধের যত্ন নেয়।

ALL 5 LORA: অন্যদিকে, পংজাব স্টেট পাবর কর্পোরেশন লিমিটেড রাবী নদীর উপর রিংজি সার্কিটের খেয়াল রাখে।

FULL FT SFT: একইভাবে, পাঞ্জাব স্টেট পাওয়ার কর্পোরেশন লিমিটেড রাস্তার নদীতে রঞ্জিত সাগর বাঁধের যত্ন নেয়।

FULL FT R3R5:

NORMAL TEST: একই সঙ্গে, পাঞ্জাব স্টেট পাওয়ার কর্পোরেশন লিমিটেড রাবী নদীর উপর রঞ্জিত সাগর বাঁধের যত্ন নেয়।

GUIDED TEST: একই সময়ে, পাঞ্জাব স্টেট পাওয়ার কর্পোরেশন লিমিটেড রাভি নদীর উপর রঞ্জিত সাগর বাঁধের তত্ত্বাবধান করে।

Table 7: The source sentence states that "Meanwhile, Punjab State Power Corporation Limited takes care of the Ranjit Sagar Dam on the Ravi River." The reference translation expresses this meaning using a formal and contextually appropriate predicate, "তত্ত্বাবধান করে". Among the reasoning-based variants, **R3 LoRA** and **R3R5 LoRA** preserve the main entities and the overall meaning reasonably well, but their use of "দায়িত্বে রয়েছে" is slightly less precise than the reference and weakens the locative relation. In contrast, **R3R4R5 LoRA**, **R2R3R4R5 LoRA**, and **All 5 LoRA** show noticeable semantic degradation, including omission or distortion of key content words and reduced adequacy. The **Full FT SFT** and **Full FT R3R5** normal outputs recover most of the source content, but the phrase "যত্ন নেয়" is less suitable for describing institutional oversight. Overall, the **Full FT R3R5 Guided Test** output is closest to the reference "রাভি" over "রাবী", as it best preserves the entities, relation, and formal administrative tone of the source sentence.

Embedding Similarity Is Not Quality Estimation: Lessons from Replacing a Dedicated QE Model

Dimitrios Zaikis
Welocalize
dimitrios.zaikis@welocalize.com

Andrea Biondo
Welocalize
andrea.biondo@welocalize.com

Matthew Dixon
Welocalize
matthew.dixon@welocalize.com

Konstantinos Karageorgos
Welocalize
konstantinos.karageo@welocalize.com

Aaron Schliem
Welocalize
aaron.schliem@welocalize.com

Abstract

Machine translation quality estimation (QE) typically relies on dedicated neural models trained on human judgments. We evaluate whether cosine similarity over general-purpose embeddings can serve as a lightweight alternative, using Gemini embeddings as the scoring backbone. Through three experiments (rogue dimension analysis, score calibration, and a learned calibration head) and a root cause analysis, we find that cosine similarity between source and translation saturates in the 0.94–0.99 range because even poor translations preserve most of the source semantics, leaving an Area Under the ROC Curve (AUC) ceiling of approximately 0.63. However, a LightGBM classifier trained on normalized cosine and surface-level text features breaks through this ceiling (AUC 0.751), with the improvement driven primarily by features orthogonal to embedding similarity. These findings demonstrate that raw embedding similarity cannot serve as a drop-in QE replacement and identify learned calibration as a viable lightweight path forward.

1 Introduction

Segment-level quality estimation (QE) for machine translation aims to predict whether a translated segment is suitable for release without human review (Zhao et al., 2024). In practice, QE scoring determines how many segments can be approved automatically, directly affecting the cost and speed of the review process (Gladkoff et al., 2024; Cabeça et al., 2023).

The system under study previously used a dedicated neural QE model from the COMET family (Rei et al., 2020), built on XLM-RoBERTa-large and trained on human quality annotations. The replacement was motivated by operational considerations: general-purpose embedding models reduce infrastructure complexity, provide broad multilingual coverage, and were already deployed in the pipeline for other tasks. This model was replaced by a scorer based on cosine similarity over Gemini embeddings (Lee et al., 2025), a general-purpose embedding model that supports a `SEMANTIC_SIMILARITY` task type designed for pairwise text comparison. However, translations inherently preserve most of the source semantics, which may make cosine similarity a weak discriminator of quality differences. The question motivating this work is whether such a general-purpose approach can match the performance of a dedicated QE model, or whether the semantic overlap between source and translation creates a fundamental ceiling.

Initial evaluation revealed a performance difference: AUC 0.631 for the dedicated model versus 0.627 for the embedding scorer. We investigate the causes of this difference and whether calibration or feature engineering can reduce it.

We report on three experiments and a root cause analysis:

1. **Rogue dimension analysis**: Outlier embedding dimensions exist but are harmless; removing them does not change quality rankings (§5.1).

2. **Score calibration**: Redistributing or recalibrating scores cannot improve discrimination

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 77–82
Tilburg, The Netherlands, June 2026.

because the ranking is preserved (§5.2).

3. **Learned calibration head**: A LightGBM classifier combining normalized cosine with surface-level features substantially exceeds both baselines (§5.3).

4. **Root cause identification**: Cosine similarity saturates for source-translation pairs because even poor translations preserve semantic content, creating a hard ceiling on discrimination (§5.4).

These experiments explain *why* embedding similarity falls short as a QE proxy and demonstrate that the saturation ceiling can be overcome with a learned model that exploits signals beyond cosine similarity. These findings are directly relevant to localization teams evaluating whether general-purpose embeddings can replace dedicated QE models in production pipelines (Sun and Wang, 2024; Zhao et al., 2024).

2 Background

Quality Estimation. QE aims to predict translation quality without reference translations (Zhao et al., 2024). State-of-the-art QE models such as COMET (Rei et al., 2020), CometKiwi (Rei et al., 2022), and xCOMET (Guerreiro et al., 2024) use cross-lingual encoders (e.g., XLM-RoBERTa) to produce source and translation embeddings, then construct a feature vector $[\mathbf{e}_s; \mathbf{e}_t; |\mathbf{e}_s - \mathbf{e}_t|; \mathbf{e}_s \odot \mathbf{e}_t]$, where $\mathbf{e}_s$ and $\mathbf{e}_t$ are the source and translation embeddings, ; denotes concatenation, and $\odot$ the element-wise product. This vector feeds into a trained regression head. The design explicitly preserves element-wise differences, which allows the model to detect subtle quality signals even when overall semantic similarity is high.

Embedding Similarity for Evaluation. Recent work has explored embedding similarity as a lightweight alternative to trained QE models. The underlying premise is that a faithful translation preserves the semantic content of the source, so cosine similarity between their embeddings should correlate with translation quality: higher similarity implies better preservation of meaning. This reduces QE to a semantic similarity measurement, avoiding the need for quality-annotated training data. Sun et al. (2024) investigate textual similarity as a metric for MT quality estimation with mixed results depending on task formulation. Dinh et al. (2024) propose a k-nearest neighbors approach over embedding representations for QE. Steck et al. (2024) show that cosine similarity over embeddings does not always capture the properties users expect.

Rogue Dimensions. Timkey and van Schijndel (2021) show that transformer language models contain a small number of "rogue dimensions" with abnormally large means that can dominate cosine similarity. They propose z-score standardization as a fix. Whether this affects modern embedding models used for QE has not, to our knowledge, been investigated.

3 System Description

The QE system under study operates as a component within a translation pipeline. For each segment, it produces a quality score used to make a binary accept/reject decision via a per-group threshold.[1]

Dedicated QE Model. The dedicated model uses XLM-RoBERTa-large (1024-dimensional embeddings) to encode source and translation independently, constructs a 4096-dimensional feature vector via concatenation and element-wise operations, and passes this through a trained regression head. The raw model output is rescaled to $[0, 1]$ through post-processing normalizations.

Embedding-Based Scorer. The replacement scorer uses Gemini embeddings (`gemini-embedding-001`) with the `SEMANTIC_SIMILARITY` task type (Lee et al., 2025). Source and translation are embedded independently into 1536-dimensional vectors.[2] The scoring steps are:

1. Compute cosine similarity: $\cos(\mathbf{e}_s, \mathbf{e}_t)$
2. Rescale to $[0, 1]$: $(\cos + 1)/2$
3. Apply post-processing normalizations
4. Compare against the per-group threshold for the accept/reject decision

[1] A group is defined by account, language pair, and content domain.

[2] Truncated from the model's maximum 3072 dimensions via its built-in variable-dimensionality support.

Account	N	% Perf.	Src	Top targets
A	4,318	38.1	en, sv	zh, pt, ru, fr
B	2,666	36.2	en	de, ja, fr, es
C	1,032	56.0	en	fr, es, de, it
D	700	27.4	en	ja, es, fr, zh
E	687	39.0	en	pt, it, ja, zh
F	597	78.2	en	fr, es, pt, it
Total	**10,000**	**41.1**		**38 target langs**

Table 1: Dataset summary. Accounts are anonymized. Source languages are predominantly English with minor Swedish source. "% Perf." is the proportion labeled as acceptable.

4 Experimental Setup

Dataset. We sampled 10,000 translation segments from a pool of 2.1 million segments across six accounts (Table 1). The data are proprietary production content from a large language service provider; no raw translation text is disclosed. The segments are predominantly English source text translated into 38 target languages, covering domains such as technical documentation, user interfaces, and support content. Each segment carries an independent human quality annotation indicating whether the translation meets quality standards: 41.1% are labeled as acceptable, 58.9% require review.

Evaluation Metrics. We evaluate using three metrics that reflect both statistical discrimination and practical applicability:

- **AUC-ROC**: threshold-independent measure of the scorer's ability to separate acceptable from non-acceptable segments.
- **Qualifying accounts**: number of accounts (out of 6) where a threshold can be found satisfying minimum precision (≥ 0.8), minimum support (≥ 50 segments), and minimum volume constraints.
- **Acceptance rate**: percentage of words accepted at the optimal threshold, a key throughput indicator.

Protocol. We use a 70/30 train/test split stratified by the quality label (7,000/3,000 segments). All calibration models are fit on the training set and evaluated on the held-out test set. All data were used in accordance with applicable data processing agreements.

Finding	Value	Interpretation
Top rogue dim (115)	$\mu{=}-0.242$	$11\times$ std
Top-5 contribution	24.9%	$\approx$1/4 of total score
r^2 after removal	0.9991	Near-perfect rank preservation
Max per-account Δ	0.14%	No group $> 5\%$

Table 2: Rogue dimension analysis. Despite accounting for nearly a quarter of total cosine similarity, rogue dimensions act as constant offsets: removing them changes rankings by only 0.09%.

5 Experiments and Results

5.1 Experiment 1: Rogue Dimension Analysis

Hypothesis. Following Timkey and van Schijndel (2021), Gemini embeddings may contain rogue dimensions whose abnormally large means dominate cosine similarity and distort quality rankings. Removing them might improve discrimination.

Method. We compute per-dimension mean and standard deviation across the full corpus of 20,000 embedding vectors (10K source + 10K target). We flag dimensions where $|\mu_d|$ exceeds 10 standard deviations of the distribution of dimension means. We then measure: (a) how much the top-k rogue dimensions contribute to total cosine similarity, and (b) the r^2 between original and post-removal scores to see whether rankings change.

Results. Rogue dimensions are present: dimension 115 has a mean of -0.242, roughly $11\times$ the standard deviation of all dimension means, and the top 5 dimensions account for 24.9% of total cosine similarity. However, as Table 2 shows, removing them has negligible impact on rankings.

The explanation is that rogue dimensions add a near-constant term to all similarity scores. Since the accept/reject decision depends on *rankings* (threshold-based classification), a constant offset is harmless. This extends the findings of Timkey and van Schijndel (2021) to Gemini embeddings: rogue dimensions exist but do not distort quality rankings.

5.2 Experiment 2: Score Calibration

Hypothesis. The raw cosine similarity scores may be poorly calibrated for the accept/reject decision. Post-processing could improve discrimination.

Methods. Three calibration approaches of increasing complexity were tested:

Method	AUC	Qual.	Acc. %
Dedicated model (COMET)	*0.631*	—	—
Baseline (cosine)	0.627	1/6	14.4
Monotonic norm.	0.627	1/6	14.4
Isotonic regression	0.627	1/6	17.1
Feature combination	0.561	0/6	0.0
LightGBM cal. head	0.751	2/6	34.8

Table 3: Score calibration results on the test set (N=3,000). The dedicated COMET model is shown for reference (threshold-based metrics not directly comparable due to different score distributions). "Qual." = accounts meeting minimum quality criteria. "Acc. %" = best per-account word acceptance rate at optimal threshold. The LightGBM calibration head (§5.3) is trained on normalized cosine and surface-level features.

1. **Monotonic normalization**: an unsupervised monotonic transform that maps scores to a uniform distribution on $[0, 1]$.

2. **Isotonic Regression** (IR): a supervised, non-parametric method that learns a monotone mapping from raw scores to empirical probabilities of being acceptable (Niculescu-Mizil and Caruana, 2005).

3. **Feature Combination** (FC): logistic regression over multiple features extracted from the embedding pairs, including cosine similarity and various distance measures.

Results. Table 3 presents the results. Both the monotonic normalization and isotonic regression produce AUC identical to the baseline (0.627). This confirms a mathematical property of AUC: monotonic transforms preserve rankings by definition and therefore cannot change it.

The feature combination approach performs *worse* than the baseline (AUC 0.561 vs. 0.627), with zero qualifying accounts. Gemini embeddings are perfectly L2-normalized by design, so the norm ratio across all 10,000 pairs has zero variance, making norm-based features constant and uninformative. The remaining distance-based features are monotonic functions of cosine similarity for L2-normalized vectors, so they carry no additional signal.

5.3 Experiment 3: Learned Calibration Head

Hypothesis. A learned model that combines embedding-derived features with surface-level text quality indicators may overcome the cosine saturation ceiling by exploiting signals orthogonal to semantic similarity.

Method. We train a LightGBM (Ke et al., 2017) gradient-boosted tree classifier to predict whether a segment is acceptable, framing QE as binary classification. The feature set combines embedding-derived similarity measures with surface-level text quality indicators. Training uses binary cross-entropy loss with word-count-weighted samples, aligning the objective with the downstream throughput metric (accepted words rather than accepted segments). Hyperparameters are tuned via cross-validation on the training set.

Crucially, the model does *not* use the source and target embeddings directly (no concatenation, no element-wise operations). All embedding information is mediated through cosine similarity and its normalized variants. This design reflects the finding that the 1536-dimensional embeddings contain mostly linguistic content rather than quality signal, making high-dimensional features noisy for a 7,000-sample dataset.

Results. The LightGBM calibration head achieves AUC 0.751 (Table 3), substantially above both the cosine baseline (0.627) and the dedicated COMET model (0.631). Two of six accounts qualify for deployment (vs. one for the cosine baseline), and the best per-account acceptance rate reaches 34.8% (vs. 14.4%). In operational terms, this means substantially more translation volume can be auto-approved, reducing human review load.

Feature importance analysis reveals that surface-level features collectively outweigh the embedding-derived features. This confirms that the improvement comes primarily from signals orthogonal to embedding similarity, capturing error patterns that cosine similarity cannot detect.

5.4 Root Cause: Score Saturation

The experiments above show that the AUC ceiling of around 0.63 is not caused by rogue dimensions, poor score calibration, or unexploited embedding features. The root cause is simpler: **cosine similarity measures semantic overlap, and translations inherently have near-perfect semantic overlap with their source.**

Even a poor translation typically covers the same topic, mentions the same entities, and uses related vocabulary, all of which embedding similarity captures. The quality-relevant differences

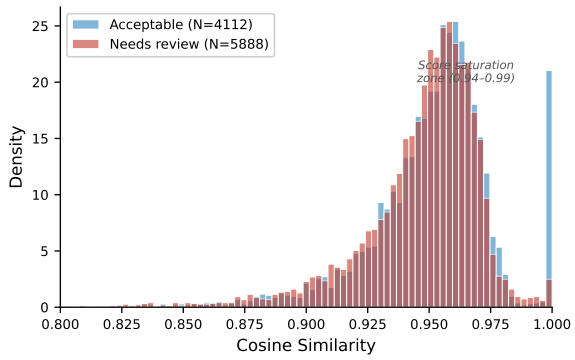


Figure 1: Distribution of cosine similarity scores for acceptable vs. non-acceptable segments. Both distributions overlap heavily in the 0.94–0.99 range, leaving minimal room for threshold-based separation.

(a dropped negation, an incorrect number, wrong terminology) are subtle linguistic details that produce negligible movement in a 1536-dimensional embedding space.

Empirically, cosine similarity scores cluster in the 0.94–0.99 range for both acceptable and non-acceptable segments (Figure 1). The mean score difference between the two classes is only 0.004 (acceptable: 0.951, needs review: 0.947), compressed into a range of roughly 0.05 units, which severely limits threshold-based discrimination.

This contrasts with how dedicated QE models work. The $[\mathbf{e}_s; \mathbf{e}_t; |\mathbf{e}_s - \mathbf{e}_t|; \mathbf{e}_s \odot \mathbf{e}_t]$ feature vector explicitly preserves element-wise differences, and the regression head learns to amplify divergences that signal quality problems. Raw cosine similarity collapses all pairwise information into a single scalar, discarding the fine-grained signal that QE requires. Even the dedicated COMET model only reaches AUC 0.631 on this data (Table 3), which suggests that segment-level QE on heterogeneous multilingual data is inherently challenging.

6 Discussion

Generalizability. We argue that the saturation effect is structural: it follows from what cosine similarity measures, not from the specific embedding model. Any model optimized for semantic similarity will produce high scores for source-translation pairs, because adequate translations *are* semantically similar by definition. The LightGBM result reinforces this: the improvement comes primarily from non-embedding features, confirming that cosine similarity itself provides limited quality signal. Embedding similarity may still be useful for *system-level* comparisons (ranking different MT engines), where outputs genuinely differ in semantic fidelity. The limitation is specific to *segment-level* QE within a single system.

Calibration vs. Discrimination. Monotonic transforms preserve rankings by construction and therefore cannot improve AUC (Niculescu-Mizil and Caruana, 2005). It is important to distinguish calibration problems (fixable via score remapping) from discrimination problems (which require a different model).

Practical Recommendations. Embedding cosine similarity should not be treated as a drop-in replacement for trained QE models. The LightGBM calibration head demonstrates a viable middle ground: a lightweight learned model that combines normalized cosine with surface-level text features can substantially outperform both raw cosine and the dedicated COMET model, without requiring the large-scale human annotation data that dedicated QE models depend on. When a dedicated model is unavailable, such learned calibration approaches are preferable to attempting to post-process similarity scores.

Limitations. This study uses a single embedding model (Gemini), a single pipeline, binary quality labels (not continuous MQM scores), and data from a single domain. Nevertheless, we expect other general-purpose embedding models to exhibit similar saturation, since the effect follows from what cosine similarity measures rather than from the specific model. The AUC difference between models (0.004) is small and we do not report bootstrap confidence intervals; however, the core finding rests not on the magnitude of this difference but on the saturation mechanism that bounds it. Evaluating this hypothesis across other embedding models, and testing the learned calibration approach on larger datasets with continuous quality labels, are immediate next steps.

7 Conclusion

This paper presented a systematic investigation into using Gemini embedding cosine similarity as a replacement for dedicated neural QE. Rogue dimension analysis and score calibration experiments confirm that cosine similarity saturates for source-translation pairs because even poor translations preserve most of the source semantics, resulting in an AUC ceiling of approximately 0.63.

This ceiling cannot be addressed through calibration or feature engineering; it reflects a fundamental limitation of reducing a high-dimensional pairwise comparison to a single similarity scalar.

However, the saturation ceiling is not a hard limit on the data itself. A LightGBM calibration head trained on normalized cosine and surface-level text features achieves AUC 0.751, substantially exceeding both the cosine baseline and the dedicated COMET model. The improvement is driven primarily by features orthogonal to embedding similarity, confirming that the quality signal cosine collapses can be recovered through complementary indicators. This suggests that the path forward for lightweight QE lies in learned models that combine embedding-derived signals with surface-level text features, rather than in better embeddings or more sophisticated similarity measures.

References

Cabeça, Mariana, Marianna Buchicchio, Madalena Gonçalves, Christine Maroti, João Godinho, Pedro Coelho, Helena Moniz, and Alon Lavie. 2023. Quality fit for purpose: Building business critical errors test suites. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 451–460, Tampere, Finland, June. European Association for Machine Translation.

Dinh, Tu Anh, Tobias Palzer, and Jan Niehues. 2024. Quality estimation with k-nearest neighbors and automatic evaluation for model-specific quality estimation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 133–146, Sheffield, UK, June. European Association for Machine Translation (EAMT).

Gladkoff, Serge, Lifeng Han, Gleb Erofeev, Irina Sorokina, and Goran Nenadic. 2024. MTUncertainty: Assessing the need for post-editing of machine translation outputs by fine-tuning OpenAI LLMs. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 337–346, Sheffield, UK, June. European Association for Machine Translation (EAMT).

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

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 3149–3157.

Lee, Jinhyuk, Feiyang Chen, Sahil Dua, Daniel Cer, and Madhuri Shanbhogue. 2025. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*.

Niculescu-Mizil, Alexandru and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pages 625–632. ACM.

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. Association for Computational Linguistics.

Rei, Ricardo, Marcos Treviso, Nuno M. Yu, António C. Guerreiro, José G. C. de Souza, and Alon Lavie. 2022. CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645. Association for Computational Linguistics.

Steck, Harald, Chaitanya Ekanadham, and Nathan Kallus. 2024. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM Web Conference 2024*. ACM.

Sun, Kun and Rong Wang. 2024. Textual similarity as a key metric in machine translation quality estimation. *arXiv preprint arXiv:2406.07440*.

Timkey, William and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4527–4546. Association for Computational Linguistics.

Zhao, Haofei, Yilun Liu, Shimin Tao, Weibin Meng, Yimeng Chen, Xiang Geng, Chang Su, Min Zhang, and Hao Yang. 2024. From handcrafted features to LLMs: A brief survey for machine translation quality estimation. *arXiv preprint arXiv:2403.14118*.

Enhancing LLM Translation Performance for Spanish–Valencian through Supervised Fine-Tuning and Reinforcement Learning

Paula Guerrero Castelló
University of the Basque Country (UPV/EHU)
pguerrero005@ikasle.ehu.eus

Abstract

Valencian, the Western Catalan variety used in the Valencian Community of Spain, lacks a dedicated language code in most multilingual machine translation (MT) systems, and is systematically rendered closer to the standard written Eastern Catalan used in Catalonia. We address this gap by adapting TranslateGemma-4B-IT, a 4-billion-parameter instruction-tuned (IT) large language model (LLM) specialized for translation, via three post-training strategies using only public corpora and Quantized Low-Rank Adaptation (QLoRA): (i) supervised fine-tuning (SFT); (ii) Group Relative Policy Optimization (GRPO), a reinforcement learning (RL) technique, with chrF plus a naturalness reward (GRPOv1); and (iii) GRPO with a composite automatic-metric reward (GRPOv2). Our results suggest that reward-function alignment with the target dialect is a key determinant of RL success in low-resource dialectal MT.

1 Introduction

Valencian belongs to the Western branch of the Catalan language family and holds official status alongside Spanish (ES) in the Valencian Community.

Its morphological and lexical properties diverge in important ways from the standard written Eastern Catalan used in Catalonia. Among the most frequent contrasts are verbal morphology (nearly all subjunctive forms differ, e.g. *diga/digui*) and third-person possessives (*seua/seu* vs. *seva/seu*). Some temporal and lexical items also differ: *hui/avui*, *xicotet/petit*, *vore/veure*.

In current NLP practice, however, these two varieties are conflated under the ISO 639-1 code "ca": most parallel corpora assign this label to the standard written Eastern Catalan used in Catalonia, and translation models include no dialect-specific code. Although Valencian does hold a dedicated code in the finer-grained ISO 639-6 standard (vlca), and several open-source software packages already employ this variety, these dialectal identifiers remain unsupported by major MT frameworks.

TranslateGemma-4B-IT (Finkelstein et al., 2026) is a representative example of this limitation. The model conditions on a target_lang_code prompt field, but only "ca" is available for the entire Catalan continuum. Zero-shot prompting with this code predominantly produces Eastern Catalan morphological forms regardless of the user's intended regional target.

Adapting such a model without full fine-tuning, without proprietary data, and without sufficient computational resources poses a realistic challenge for practitioners working on regional varieties, and thus constitutes the central motivation of this paper[1].

We make three specific contributions. First, we establish a reproducible QLoRA-based SFT approach for the ES–VLCA direction, where QLoRA, an efficient fine-tuning method that applies low-rank adapters to quantized models, substantially closes the zero-shot quality gap. Second, we conduct a systematic comparison of two GRPO reward designs for dialectal MT: one that incorporates a learned naturalness classifier, and one relying on automatic reference-based metrics and type-token

[1]Code available at https://github.com/guerreropaula/spanish-valencian-mt-rl.

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 83–90
Tilburg, The Netherlands, June 2026.

ratio (TTR). Third, we introduce a *dialectal Valencian score* (DVS) as a complementary evaluation axis alongside standard MT metrics, and use it to demonstrate that the reward design choice has opposing consequences for translation quality and dialect production.

2 Background

Dialectal and low-resource MT. Neural approaches to dialectal variation have progressed from rule-based dialectal preprocessing pipelines (Salloum and Habash, 2012) to unified neural models that utilize multitask learning (Baniata et al., 2018). Nevertheless, much of the research in dialectal MT has primarily focused on Arabic varieties. Regarding the Iberian languages, there is comparatively little prior work, and systems addressing the Catalan–Valencian divergence have received limited attention despite the growing availability of resources from the AINA initiative (Villegas, 2023) and the PlanTL-GOB-ES corpora (Secretaría de Estado de Telecomunicaciones y Sociedad de la Información, 2015). In this context, Apertium (Forcada et al., 2011), an open-source rule-based MT platform currently used by several Valencian institutions, represents one of the main references for Valencian MT.

The most closely related prior work is Galiano Jiménez et al. (2024), who developed systems for the WMT 2024 shared task on translation into low-resource languages of Spain, exploiting Valencian and Catalan as auxiliary transfer-learning languages. Our approach differs in that we target ES–VLCA directly, adapt a translation-specialized LLM via QLoRA-based post-training rather than building dedicated MT pipelines, and introduce a comparison of RL reward designs for dialectal adaptation. The challenge is compounded by the fact that the two varieties share a large common vocabulary, so models trained on standard Eastern Catalan can achieve acceptable metric scores on Valencian test sets while still producing incorrect morphological forms.

RL for MT refinement. Sequence-level training with automatic metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin and Hovy, 2003) as a reward was proposed by Ranzato et al. (2016) and extended to human preference signals by Kreutzer et al. (2018). Wu et al. (2018) provide a systematic study of RL training strategies for neural MT (NMT), showing that reward shaping—the practice of designing intermediate rewards to guide policy updates—and baseline design are critical for stable training. Work using quality-aware decoding with learning-based metrics such as COMET and BLEURT (Rei et al., 2020; Sellam et al., 2020) has further shown that these metrics capture translation quality better than n-gram overlap alone (Fernandes et al., 2022). In this context, GRPO (Shao et al., 2024), originally developed for mathematical reasoning, eliminates the value network (critic) by normalizing rewards within sampled output groups, therefore reducing memory requirements and making it tractable for billion-parameter models under constrained hardware. Recent work has applied similar RL objectives to GRPO directly to MT, including Feng et al. (2025) and the WMT 2025 submission by Wang et al. (2025), further highlighting the growing adoption of RL-based refinement methods for translation LLMs.

Naturalness in MT. Improving the naturalness and fluency of MT output is a distinct and understudied goal. MT outputs exhibit reduced lexical diversity and increased source-language interference compared to human translation: what is referred to as machine *translationese* (Vanmassenhove et al., 2021). Metrics such as TTR and MTLD are designed to capture such effects. In parallel, Lai et al. (2025) propose an RL-based alignment framework for neural MT that uses COMET and binary classifiers—trained on human translation (HT), machine translation (MT), and original target-language data (OR)—as reward models to approximate human preference. Their findings directly motivate both the TTR component in our GRPOv2 reward and our decision to test a learned HT/MT naturalness classifier as part of the GRPOv1 reward.

Parameter-efficient fine-tuning. Low-Rank Adaptation (LoRA) (Hu et al., 2022) constrains weight updates to low-rank decompositions; combined with 4-bit NF4 quantization as in Quantized LoRA (QLoRA) (Dettmers et al., 2023), it enables adaptation of multi-billion-parameter models on a single GPU, which is the hardware regime assumed throughout this work. We therefore adopt this approach in our experiments.

3 System description

3.1 Training pipeline overview

Figure 1 illustrates the two training pipelines investigated. Both share an SFT stage that adapts the

base model to the ES–VLCA direction; they diverge only in the subsequent GRPO reward design.

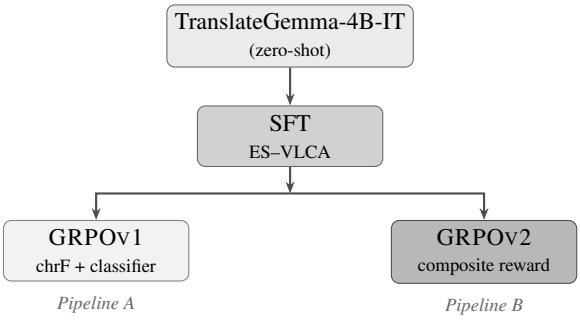


Figure 1: Training pipelines. Both share the SFT stage and differ only in the GRPO reward signal applied thereafter; Pipeline B performs best.

3.2 Base model

TranslateGemma-4B-IT (Finkelstein et al., 2026) is a 4-billion-parameter instruction-tuned model pre-trained to optimize translation quality across 55 languages. Following the authors' specification, we use the structured JSON template with "`ca`" as target language code throughout all experimental conditions; no Valencian-specific code is available.

3.3 Data

SFT corpus. We use the GPLSI AMIC parallel dataset,[2] which provides 50,000 ES–VLCA sentence pairs sourced from news articles published by the *Associació de Mitjans d'Informació i Comunicació* (AMIC). The target Valencian side exposes the model to the morpho-lexical features that distinguish the desired variety from other Catalan varieties.

GRPO corpus. The Spanish source side of the same corpus is reused for RL training: 5,000 sentences for GRPOv1 and 10,000 for GRPOv2. These subsets are randomly sampled from the 50,000 pairs. The smaller corpus for GRPOv1 reflects the higher per-step computational cost of the classifier reward, which made training on more sentences infeasible given the available hardware. Note that this difference in corpus size means the two GRPO variants are not fully matched in compute; readers should bear this in mind when comparing their results directly. Valencian references are used solely to compute reward signals and are never provided to the policy as decoding context.

HT/MT classifier training data. The naturalness classifier, further explained in Section 3.6, is trained on professional human translations drawn from TildeMODEL (general), DOGC (official/legal), and Europarl (parliamentary), paired with MT outputs from OPUS-MT (Tiedemann and Thottingal, 2020) and NLLB-200 (Costa-jussà et al., 2022), yielding 108,000 training and 12,000 validation instances. Each system produces 50% of the translations to prevent system-specific bias. We note that all classifier training data derive from standard Eastern Catalan corpora. This distribution gap between standard Eastern Catalan and Valencian is the central limitation we identify in Section 6. The use of a standard-Catalan classifier was motivated by the absence of sufficiently large Valencian human-translation corpora suitable for classifier training.

Evaluation. All systems are evaluated on the GPLSI ES–VLCA translation test set,[3] a 1,000-sentence held-out set of professional news translations, which ensures adherence to dialectal particularities.

3.4 Supervised fine-tuning

QLoRA targets all attention and feed-forward projection matrices with rank r=16, scaling α=32, and 4-bit NF4 double quantization, yielding 32.8M trainable parameters (0.76% of 4B). The model is trained for three epochs over the 50,000 parallel sentence pairs.

3.5 HT/MT naturalness classifier

We fine-tune RoBERTa-ca (Armengol-Estapé et al., 2021) as a binary classifier estimating $P(\mathrm{HT} \mid \mathrm{hypothesis})$. On the standard-Catalan validation set it attains accuracy 74.7%, F1 74.5%, recall 74.7%, and precision 75.2%. We acknowledge that an F1 of approximately 75% represents a moderate level of reliability; its limitations in the context of reward modeling are discussed in Section 6.

3.6 GRPOv1: hybrid reward

Following SFT, we apply GRPO-based RL to further refine the model. GRPO is a policy-gradient algorithm in which rewards are normalized within a group of sampled outputs for the same input, eliminating the need for a separate value network (Shao

[2] `https://huggingface.co/datasets/gplsi/amic_parallel`

[3] `https://huggingface.co/datasets/gplsi/ES-VA_translation_test`

et al., 2024). For each source sentence, the policy samples $G{=}4$ independent hypotheses. Within-group reward normalization produces advantage estimates used by the GRPO objective with Kullback–Leibler (KL) divergence penalty $\beta{=}0.04$ and learning rate $5{\times}10^{-6}$. In this first variant, the reward combines reference fidelity and estimated naturalness:

$$r = (1-\alpha)\, r_c + \alpha\, r_t \tag{1}$$

where $r_c = \mathrm{chrF}/100$ and $r_t = P(\mathrm{HT}\,|\,\mathrm{hyp})$. To prevent the naturalness term from destabilizing early training, α is held at zero for the first 50 warm-up steps and then increases linearly to 0.3. Training proceeds for 100 steps; the checkpoint at step 80 is retained based on held-out reward.

3.7 GRPOv2: composite automatic-metric reward

GRPOv2 eliminates the classifier dependency, replacing it with a combination of three automatic signals:

$$r = 0.5\,\hat{r}_{\mathrm{chrF}} + 0.3\,\hat{r}_{\mathrm{COMET}} + 0.2\,\mathrm{TTR} - \mathbf{1}[\mathrm{copy}] \tag{2}$$

where $\hat{r}(\cdot)$ denotes min-max normalization to $[0, 1]$. The chrF and COMET components capture surface-level similarity and semantic adequacy, respectively, while the type-token ratio (TTR) promotes lexical diversity and discourages repetitive outputs.

The $\mathbf{1}[\mathrm{copy}]$ term equals 1 when the hypothesis reproduces the source sentence verbatim, enforcing a penalty for copying. This is motivated by the high lexical overlap between Spanish and Valencian as a closely related language pair. Training runs for 200 steps over 10,000 sentences (PPO clip[4] $\varepsilon{=}0.2$, $\beta{=}0.04$). The final model is selected based on the highest mean reward.

4 Evaluation

Standard MT metrics. We report chrF, BLEU, TER, BLEURT, and COMET (Popović, 2015; Papineni et al., 2002; Snover et al., 2006; Sellam et al., 2020; Rei et al., 2020). TER is expressed as a percentage throughout.

[4]PPO (Proximal Policy Optimization) clip is a hyperparameter ε that constrains how far the updated policy may deviate from the previous one in a single gradient step.

System	chrF↑	BLEU↑	TER%↓	BLEURT↑	COMET↑
Baseline	69.02	39.22	40.30	0.258	0.906
SFT	83.16	60.16	22.80	0.524	0.934
GRPOv1	81.65	56.94	23.96	0.481	0.926
GRPOv2	**84.68**	**62.16**	**20.63**	**0.544**	**0.936**

Table 1: Translation quality on the 1,000 ES–VLCA test set. Best in bold. TER: lower is better; reported as percentage.

Dialectal Valencian score (DVS). We introduce the *dialectal Valencian score* (DVS), which measures how frequently a system produces Valencian-specific morpho-lexical surface forms rather than their Eastern Catalan equivalents. We compile a vocabulary of 35 contrastive pairs selected on two criteria: (i) minimum frequency of five occurrences in the 1,000-sentence test set for at least one variant, and (ii) representation across morphological categories (possessives, temporal adverbials, verbal stems, and common lexical alternates), with a roughly equal split across categories. Examples include *seua/seva* (possessive), *hui/avui* (temporal adverb), and *xicotet/petit*, *vore/veure* (lexical alternates). The 35-pair list is provided in Appendix A. DVS is independent of the reference-based metrics above: a system can score high on chrF while still not adhering to Valencian morphology and lexicon.

5 Results

5.1 Automatic translation quality

Table 1 reports metric scores on the test set. SFT delivers the largest single-step absolute improvement: chrF rises 14.1 points, BLEU 20.9 points, and TER falls 17.5 percentage points relative to the zero-shot baseline. GRPOv2 consistently outperforms SFT across all five metrics, with further gains of 1.5 chrF, 2.0 BLEU, and 2.2 TER. By contrast, GRPOv1 regresses below SFT on every metric, a pattern we analyze in Section 6.

5.2 Dialectal Valencian score

Figure 2 and Table 2 present DVS results. SFT achieves the highest overall score (41.0%), confirming that direct supervised exposure to Valencian forms is the most effective lever for dialect adaptation. GRPOv2 retains most of this gain (36.2%), indicating that character-level reward partially reinforces Valencian-specific tokens. GRPOv1, by contrast, collapses to 15.9%, less than half the value of the SFT checkpoint it is initialized from.

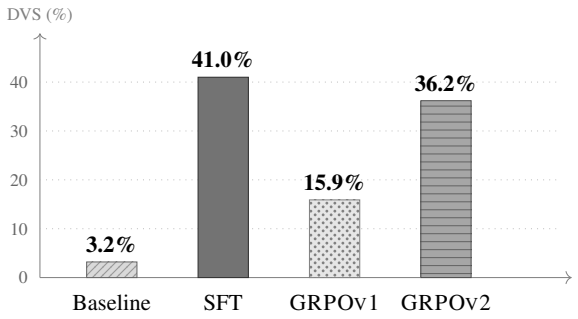


Figure 2: Dialectal Valencian score (DVS): percentage of Valencian forms produced over the 1,000-sentence test set. DVS is computed where at least one contrastive variant is contextually present.

CA	VLCA	Base	SFT	v1	v2
petit	*xicotet*	0%	100%	0%	100%
veure	*vore*	0%	83%	0%	17%
avui	*hui*	0%	71%	0%	86%
seva	*seua*	0%	100%	45%	100%

Table 2: Feature-level DVS for four representative pairs (see Appendix A for the full 35-pair list). GRPOv1 (v1) scores 0% on most Valencian features despite initializing from the SFT checkpoint.

5.3 Qualitative analysis

Table 3 presents a representative example from the test set. All four systems preserve the propositional content; the difference lies in the third-person feminine possessive *seva/seua* ("su"). The baseline and GRPOv1 produce the Eastern Catalan *seva*; SFT and GRPOv2 produce the Valencian form *seua*.

6 Analysis

Classifier-induced dialect suppression in GRPOv1. The performance collapse of GRPOv1 can be traced directly to the distribution mismatch of its classifier. The reward in GRPOv1 penalizes Valencian-specific forms, because the classifier was trained exclusively on standard Eastern Catalan data and consequently assigns lower naturalness scores to Valencian morphology. This behavior is consistent with the finding of Lai et al. (2025) that naturalness classifiers underperform as reward signals when there is a mismatch between the training data and the target-side data used during alignment; we extend this observation to the cross-variety setting. As shown in Table 2, feature-level DVS data support this interpretation: GRPOv1 produces the Valencian possessive *seua* in only 45% of instances despite the SFT model using it in 100%.

System	Output
ES source	*Su participación en este concurso le dio la oportunidad de entrar al mundo artístico.*
Baseline	La **seva** participació … l'oportunitat d'entrar al món artístic.
SFT	La *seua* participació … l'oportunitat d'ingressar al món artístic.
GRPOv1	La **seva** participació … l'oportunitat d'ingressar al món artístic.
GRPOv2	La *seua* participació … l'oportunitat d'ingressar al món artístic.

Table 3: Qualitative comparison. **Bold**: Eastern Catalan form; *italic*: Valencian form.

Composite reward and metric complementarity. The three components of the GRPOv2 reward serve distinct roles. ChrF provides character-level supervision that rewards partial morphological matches, which is particularly important for Valencian-specific suffixes, verb inflections, and possessive forms that differ minimally from their standard Catalan counterparts; this may account for the near-complete preservation of dialectal accuracy relative to the SFT checkpoint. The COMET term introduces semantic evaluation, ensuring that dialectal variation does not come at the cost of meaning preservation. TTR penalizes repetition without imposing any preference over the lexical inventory of the target variety, thus avoiding the classifier-induced dialect suppression documented in GRPOv1. Finally, the copy penalty discourages verbatim source copying, motivated by the high similarity between the language pair. The effectiveness of this configuration suggests that multi-objective RL rewards combining fidelity and semantic quality signals are more robust than single-metric or classifier-based approaches for dialectal low-resource translation.

SFT as the primary dialect-adaptation signal. The highest DVS is achieved by SFT rather than by any RL variant. This reveals a tension in the reward: scalar reward signals provide no explicit incentive for producing Valencian-specific lexical choices. Direct SFT, by contrast, supplies explicit examples of the desired dialectal distribution, thus providing an unambiguous policy gradient toward the desired Valencian variety. Accordingly, future work should prioritize the curation of high-quality, dialect-annotated parallel corpora; enlarging the SFT training data is likely to result in higher returns for dialect adaptation than further RL tuning over resource-constrained settings.

7 Limitations

The training and evaluation data derive from a single domain (news journalism), and performance on legal, administrative, or literary Valencian text remains untested. DVS, while interpretable and useful in this work, captures only lexical surface contrasts and does not account for morphosyntactic divergences (verbal periphrases, clitic ordering, verbal paradigm differences such as subjunctive forms that are equally characteristic of Valencian). Furthermore, the current DVS is primarily based on the written standard established by the Valencian normative authority (Acadèmia Valenciana de la Llengua, 2006); a DVS inclusive of all competing standards would likely yield different results.

The automatic naturalness signals in GRPOv2 (type-token ratio, chrF, COMET) remain shallow proxies for fluency that would ideally be validated against native-speaker judgments. Training hardware constraints bound the total compute invested and may affect reproducibility across configurations. The non-comparability of compute between GRPOv1 and GRPOv2 (5,000 vs. 10,000 training sentences) is an additional caveat when interpreting their relative performance. Finally, our experiments cover a single language pair and domain; the conclusions drawn about reward-function alignment and dialect suppression should therefore be interpreted as initial findings that warrant replication across other dialectal pairs before being further generalized.

8 Conclusion

We have shown that a pre-trained translation-specialized LLM can be adapted to the under-resourced Valencian dialect through lightweight post-training on publicly available corpora and commodity hardware. SFT on 50,000 in-domain pairs constitutes the single largest improvement (+14.1 chrF, +20.9 BLEU) and achieves the highest dialectal form production rate (41.0% DVS). GRPOv2 improves on SFT across all five standard MT metrics and retains 36.2% dialectal adequacy, establishing a strong baseline for the ES–VLCA direction. The regression of GRPOv1 has implications beyond this language pair: a naturalness classifier trained on a related standard variety is actively harmful when applied as an RL reward signal for dialectal adaptation, since, as demonstrated, it systematically penalizes the target dialect's characteristic forms. This outcome underscores the importance of aligning reward models with the actual target variety. Future work should address this by (i) training the naturalness classifier exclusively on high-quality professional Valencian translations; (ii) scaling the SFT corpus via controlled data augmentation; (iii) extending DVS to morphosyntactic features beyond surface contrasts, while accounting for competing Valencian written norms; and (iv) conducting human evaluation with native Valencian speakers to disentangle automatic-metric artifacts from real translation quality improvements.

Acknowledgments

We sincerely thank Antonio Toral, Nora Aranberri, Mikel Forcada, Gorka Azkune, and Eneko Agirre for their valuable support and guidance throughout the development of this work.

References

Acadèmia Valenciana de la Llengua. 2006. *Gramàtica normativa valenciana*. Acadèmia Valenciana de la Llengua, València.

Armengol-Estapé, Jordi, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Adriana Gonzalez-Agirre, Maite Melero, and Marta Villegas. 2021. Are Multilingual Models the Best Choice for Moderately Underresourced Languages? A Comprehensive Assessment for Catalan. In *Findings of ACL-IJCNLP 2021*, pages 4933–4946.

Baniata, Laith H., Seyoung Park, and Seong-Bae Park. 2018. A Neural Machine Translation Model for Arabic Dialects that Utilises Multitask Learning (MTL). *Computational Intelligence and Neuroscience*, 2018.

Costa-jussà, Marta R., Angela Fan, Shruti Bhosale, et al. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*.

Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36.

Feng, Zhaopeng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Zhe Xu, Yao Hu, Jian Wu, and Zuozhu Liu. 2025. MT-R1-Zero: Advancing LLM-based Machine Translation via R1-Zero-like Reinforcement Learning. In *Findings of ACL: EMNLP 2025*, pages 18685–18702, Suzhou, China. Association for Computational Linguistics.

Fernandes, Patrick, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and André F. T. Martins. 2022. Quality-Aware Decoding for Neural Machine Translation. In *Proceedings of*

NAACL 2022, pages 1396–1412, Seattle. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2601.09012*.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Galiano Jiménez, Aaron, Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2024. Universitat d'Alacant's Submission to the WMT 2024 Shared Task on Translation into Low-Resource Languages of Spain. In *Proceedings of the Ninth Conference on Machine Translation*, pages 885–891, Miami, Florida, USA. Association for Computational Linguistics.

Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations*.

Kreutzer, Julia, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018. Can Neural Machine Translation be Improved with User Feedback? In Bangalore, Srinivas, Jennifer Chu-Carroll, and Yunyao Li, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 92–105, New Orleans - Louisiana, June. Association for Computational Linguistics.

Lai, Huiyuan, Esther Ploeger, Rik Van Noord, and Antonio Toral. 2025. Multi-perspective Alignment for Increasing Naturalness in Neural Machine Translation. In Che, Wanxiang, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28071–28084, Vienna, Austria, July. Association for Computational Linguistics.

Lin, Chin-Yew and Eduard Hovy. 2003. Automatic evaluation of summaries using N-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, NAACL '03, pages 71–78, Edmonton, Canada. Association for Computational Linguistics.

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, Pennsylvania, USA, July. 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, Portugal, September. Association for Computational Linguistics.

Ranzato, Marc'Aurelio, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence Level Training with Recurrent Neural Networks. In *Proceedings of the International Conference on Learning Representations*.

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, November. Association for Computational Linguistics.

Salloum, Wael and Nizar Habash. 2012. Elissa: A Dialectal to Standard Arabic Machine Translation System. In Kay, Martin and Christian Boitet, editors, *Proceedings of COLING 2012: Demonstration Papers*, pages 385–392, Mumbai, India, December. The COLING 2012 Organizing Committee.

Secretaría de Estado de Telecomunicaciones y Sociedad de la Información. 2015. Plan de Impulso de las Tecnologías del Lenguaje. Technical report, Ministerio de Industria, Energía y Turismo.

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, July. Association for Computational Linguistics.

Shao, Zhihong, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA 2006*, pages 223–231.

Tiedemann, Jörg and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In Martins, André, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal, November. European Association for Machine Translation.

Vanmassenhove, Eva, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In Merlo, Paola, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online, April. Association for Computational Linguistics.

Villegas, Marta. 2023. El projecte AINA, la IA i les tecnologies del llenguatge. *Terminàlia*, 27:80–84.

Wang, Hao, Linlong Xu, Heng Liu, Yangyang Liu, Xiaohu Zhao, Bo Zeng, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. Marco Large Translation Model at WMT2025: Transforming Translation Capability in LLMs via Quality-Aware Training and Decoding. In *Proceedings of WMT 2025*.

Wu, Lijun, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. A Study of Reinforcement Learning for Neural Machine Translation. In *Proceedings of EMNLP 2018*, pages 3612–3621, Brussels, Belgium. Association for Computational Linguistics.

A DVS Contrastive Vocabulary

Table 4 lists all 35 contrastive pairs used to compute the *dialectal Valencian score*. The pairs are grouped by morphological category. All pairs meet the minimum frequency threshold of five occurrences in the 1,000-sentence test set. The list reflects primarily the AVL written standard.

Category	CA form	VLCA form
Determiners & possessives		
	aquesta	*esta*
	aquest	*este*
	aquestes	*estes*
	aquests	*estos*
	seva	*seua*
	seves	*seues*
	darrer	*últim*
	darrers	*últims*
	darrera	*última*
Verbal stems & inchoative endings		
	tenir	*tindre*
	obtenir	*obtindre*
	veure	*vore*
	segueix	*seguix*
	segueixen	*seguixen*
	requereix	*requerix*
	divideix	*dividix*
	constitueixen	*constituïxen*
	absorbeixen	*absorbixen*
Lexical alternates		
	nens	*xiquets*
	nen	*xiquet*
	nena	*xiqueta*
	nenes	*xiquetes*
	petit	*xicotet*
	petits	*xicotets*
	petita	*xicoteta*
	feina	*faena*
	feines	*faenes*
	cop	*colp*
	cops	*colps*
	avui	*hui*
	servei	*servici*
	serveis	*servicis*
	mirall	*espill*
	tomàquet	*tomaca*
	tomàquets	*tomaques*

Table 4: Full DVS contrastive vocabulary grouped by morphological category. Pairs reflect the AVL written standard. Each entry represents a CA–VLCA surface-form pair; DVS is computed over sentences where at least one variant of each pair is contextually present.

Towards Visually-Guided Movie Subtitle Translation for Indic Languages

Tarun Chintada[*], **Kshetrimayum Boynao Singh**[*], **Asif Ekbal**
Department of Computer Science and Engineering
Indian Institute of Technology Patna, India
{tarunchintada1, boynfrancis, asif.ekbal}@gmail.com

Abstract

Movie subtitle translation is inherently multimodal, yet text-only systems often miss visual cues needed to convey emotion, action, and social nuance, especially for low-resource Indic languages (English to Hindi, Bengali, Telugu, Tamil and Kannada). We present a case study on five full-length films and compare two lightweight visual grounding strategies: structured attribute summaries from a 5-minute sliding window and free-text summaries of inter-subtitle visual gaps. Our analysis shows that temporal misalignment between subtitles and frames is a major obstacle in long-form video, often rendering indiscriminate visual grounding ineffective. However, oracle selective[1] grounding, which replaces only the lowest-quality 20-30% of baseline segments with visual-enhanced outputs, consistently improves COMET over the text-only baseline while requiring far less visual processing. Among the two approaches, coarse attribute-based visual context summarization is more robust, capturing scene-level emotion and contextual subtle cues that text alone often misses.

1 Introduction

The global demand for movie subtitle translation has grown exponentially with the rise of streaming platforms and international releases. Subtitles must convey meaning under strict spatial and temporal constraints, often compressing conversational speech, idiomatic expressions, and cultural references into short, timed segments. For low-resource Indian languages characterised by rich morphology (Singh et al., 2025a), diglossia, and sparse parallel corpora, these challenges are magnified, and text-only machine translation (MT) systems frequently produce translations that are either literal or contextually inadequate (Singh et al., 2025b; Artetxe et al., 2020).

Films are inherently multimodal: meaning is distributed across dialogue, visual scenes, character actions, and emotional cues. In principle, incorporating visual context could disambiguate references, resolve honorifics, and ground translations in the on-screen situation (Elliott et al., 2016). Yet, unlike traditional multimodal machine translation (MT) tasks (e.g., translating image captions), movie subtitle translation presents two distinctive difficulties.

First, most subtitle segments do not rely on visual information. The majority of dialogues are conversational and can be translated accurately from text alone. Visual grounding is beneficial only in a minority of cases-action cues (Gain et al., 2025; Kumar et al., 2025), emotion-driven (Shen et al., 2025) exchanges, or references to visible objects. For instance, translating the line *"He's coming!"* requires knowing whether the threat is a person, animal, or vehicle; such information is visually available. In contrast, a typical exchange like *"How was your day?"* gains little from the accompanying visuals. This asymmetry makes indiscriminate application of computationally expensive visual processing inefficient and often unnecessary.

Second, visual and textual streams in long-form movies are often misaligned. Subtitles are generated independently of video frames, and cumulative

[*]Equal contribution.

[1]Oracle selective refers to a specialized, often theoretical, method used to achieve the absolute best outcome

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 91–101

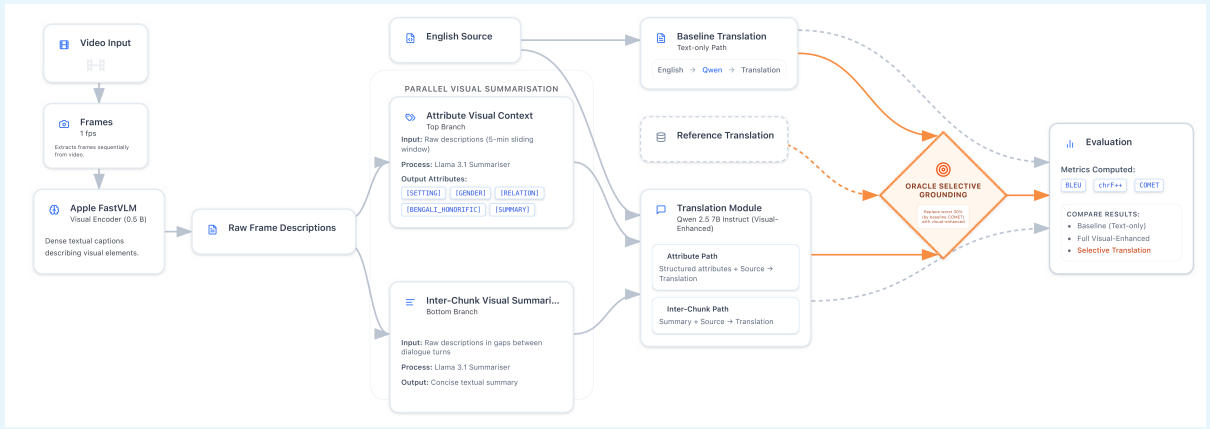


Figure 1: Architecture of the multimodal subtitle translation pipeline

temporal drift can cause a substantial fraction of subtitles to be paired with irrelevant or misleading visuals (Wang and Zhao, 2024). Over a 180-minute film, drift as small as one second per hour accumulates to a three-minute mismatch, affecting a notable portion of subtitle segments. When visual context is misaligned, it ceases to be helpful and can actively degrade translation quality (Appicharla et al., 2024), a phenomenon rarely discussed in the multimodal MT literature (Radinger, 2025).

Motivated by these practical realities, we conduct a case study that systematically compares two summarization-based strategies for integrating visual context into subtitle translation for five Indian languages: Hindi, Bengali, Telugu, Kannada, and Tamil. These languages represent a range of morphological complexity and cultural nuances, making them ideal for studying multimodal subtitle translation in low-resource settings (Eberhard et al., 2025). The two strategies are:

1. **Attribute Visual Context (Attr-VC)**: Aggregates a 5-minute sliding window of raw visual descriptions and summarizes them using Llama 3.1 into structured attributes (e.g., setting, gender, honorifics, emotional intent).

2. **Inter-Chunk Visual Summarization (Inter-VS)**: Summarizes the visual content occurring between dialogue turns (the gaps) into a free-text description using Llama 3.1.

Using subtitles from five full-length movies spanning diverse genres, we evaluate these methods under realistic conditions. A central finding is that applying visual context indiscriminately often degrades performance due to temporal misalignment. However, an *oracle selective grounding* that replaces the worst 20-30% of baseline segments (by baseline COMET) with visual-enhanced translations consistently improves semantic adequacy (COMET) over the text-only baseline, recovering most of the gain while using only a fraction of the visual processing. Coarse attribute-based summarization proves particularly robust, capturing emotional tone and scene-level cues that text alone cannot convey. The key attribute of this work are:

1. **A comparative case study** of two visual summarization strategies for low-resource subtitle translation.
2. **Identification and quantification** of temporal misalignment as a major practical obstacle in long-form multimodal MT.
3. **Empirical evidence** that coarse attribute summarization is resilient to drift and that selective grounding can recover most of the gain.

2 Dataset Preparation and Resources

To evaluate multimodal subtitle translation in realistic settings, we curate a dataset derived from five commercially released movies selected to ensure diversity in genre, narrative style, and visual complexity: *Titanic* (1997), *Skyfall* (2012), *Oppenheimer* (2023), *Spider-Man 2* (2004), and *Avatar 2* (2022). These films span romance, action, historical drama, science fiction, and superhero genres, providing a wide range of dialogue types and visually grounded scenes.

2.1 Movies and Visual Data

For each movie, we extract video frames at 24-fps and align them with subtitle timestamps. Table 1 summarizes the visual data, including total dura-

Movie	Genre	Duration	Total Frames	Extracted Frames
Titanic (1997)	Romance, Drama, Epic	3:14:54	280,305	11,694
Skyfall (2012)	Action, Adventure, Spy Thriller	2:23:10	205,952	8,589
Oppenheimer (2023)	Biographical, History, Drama	3:00:22	259,472	10,822
Spider-Man 2 (2004)	Superhero, Action, Sci-Fi	2:15:48	195,360	8,148
Avatar 2 (2022)	Sci-Fi, Action, Adventure	3:12:38	277,117	11,558

Table 1: Visual data statistics for the five movies. Extracted frames are sampled within the time span of each subtitle segment.

Movie	Total Pairs	Avg Words	Avg Chars
Titanic	1,991	5.64	29.57
Skyfall	1,140	5.70	29.74
Oppenheimer	3,519	5.68	31.77
Spider-Man 2	1,049	5.83	30.64
Avatar 2	1,444	6.44	32.86
Total	**9,143**	**5.86**	**30.92**

Table 2: Subtitle statistics per movie. Total pairs represent the number of English subtitle segments.

tion, total frames, and the number of frames extracted within subtitle time spans (i.e., frames that fall within the time window of a subtitle segment). The extracted frames serve as the visual input for our multimodal methods.

2.2 Subtitle Corpora

We extract subtitles from publicly available sources[2] and temporally align them with the corresponding video segments. All subtitle pairs are preprocessed to remove noise, normalize punctuation, and filter excessively long or short segments. Table 2 provides detailed statistics for each movie, including the number of subtitle pairs (English source and target language), average English source length in words, and average English source length in characters. Parallel subtitles are available for Hindi (all movies except Avatar 2), Bengali and Telugu (all movies), Kannada (all movies except Spider-Man 2), and Tamil (all movies except Titanic). This selection ensures coverage of linguistically diverse Indian languages while respecting the availability of high-quality parallel subtitles. All movies were legally purchased as DVDs. Frame extraction for research constitutes fair use and follows standard practice in video-language benchmarks.

2.3 Data Release

To foster reproducibility and further research, we will release the curated movie-subtitle-visual alignment data for all five languages under a fair-use educational/research license. The release includes the English source, reference translations, and extracted visual descriptions. The code and instructions for reproducing the experiments will also be made publicly available.

[2]http://subtitlecat.com

3 Methodology

Our methodology is designed for real-world applicability: all models are used off-the-shelf in a zero-shot setting, no fine-tuning or training is performed, and the pipeline is fully reproducible. We use Qwen-2.5-7B-Instruct (Qwen et al., 2025) as the translation model, Llama-3.1-8B-Instruct (Meta, 2024) for summarization (Datta et al., 2025), and Apple FastVLM-0.5B (Vasu et al., 2025) for visual description extraction. The pipeline is depicted in Figure 1. For the text-only baseline, Qwen is prompted with the English source only. This yields the text-only translation against which visual-enhanced methods are compared.

3.1 Visual Context Generation

From each movie, we sample frames at 1-fps and obtain raw textual descriptions using FastVLM-0.5B. These descriptions are then summarized by Llama-3.1 into two distinct forms:

Attribute Visual Context (Attr-VC) A 5-minute sliding window (centered on the subtitle start) is aggregated and summarized into structured attributes: `[SETTING]`, `[GENDER]`, `[RELATION]`, `[HONORIFIC]`, and `[SUMMARY]`. This yields a coarse, high-level scene description.

Inter-Chunk Visual Summarization (Inter-VS) The raw descriptions that fall between the end of the previous subtitle and the start of the current subtitle (the visual gap) are summarized into a free-text description. This captures visual events that occur between dialogue turns. The full prompts used for these summarization tasks are provided in Table 6. Both the summaries are concatenated with the English source using the same prompt template, which instructs the model to ground its translation in the visual context.

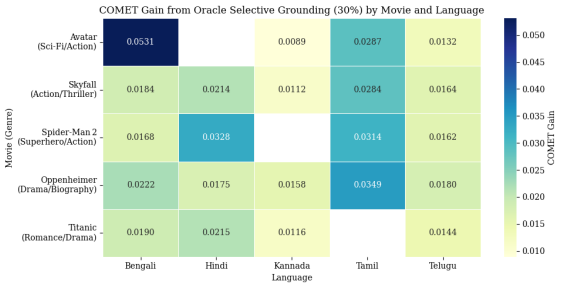


Figure 2: COMET gain from Oracle Selective Grounding (30%) by movie and language. Gain is computed as the difference between the oracle selective translation (replacing the worst 30% of baseline segments by baseline COMET) and the text-only baseline, using the better of the two visual summarisation methods for each pair. Movie names are followed by their genre in parentheses. Darker shades indicate larger gains.

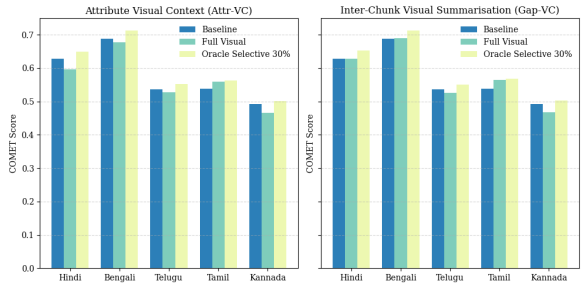


Figure 3: Language-wise COMET scores for the two visual summarization methods. For each method, bars show the baseline (text-only), full visual-enhanced translation, and oracle selective grounding (replacing the worst 30% of baseline segments by baseline COMET). The oracle selective consistently improves COMET over the baseline across all languages, with relative gains of 2-5%.

3.2 Oracle Selective Grounding

To estimate the upper bound of selective visual grounding, we compute per-segment COMET scores for the baseline translations against the reference. We then replace the worst $k\%$ of segments (by baseline COMET) with the corresponding visual-enhanced translation (from either Attr-VC or Inter-VS). We experiment with $k = 20\%$ and 30%. This oracle analysis shows the potential improvement if one could perfectly identify low-quality baseline segments; it does not require any training and represents an upper bound for practical quality-estimation systems.

4 Evaluation Results

4.1 Evaluation Setup

We evaluate on the full curated test set (all aligned subtitle segments) using corpus-level BLEU (Papineni et al., 2002), chrF++ (Popović, 2015), and COMET (Rei et al., 2020).

4.2 Results and Analysis

We compare the two visual summarisation strategies *Attribute Visual Context (Attr-VC)* and *Inter-Chunk Visual Summarization (Inter-VS)* against a text-only baseline. We perform experiments with five movies in five Indian languages (Hindi, Bengali, Telugu, Tamil, Kannada) using Qwen-2.5-7B.

Results for full per-movie, per-language are shown in Table 3. Table 4 summarizes the language-wise COMET improvements.

4.3 Overall Observations

Both summarisation methods show that applying visual context indiscriminately often degrades performance compared to the text-only baseline. For example, in several movie-language pairs (e.g., Oppenheimer Hindi, Skyfall Bengali), full VT COMET is lower than the baseline. This is directly attributable to temporal misalignment: when visual frames do not match the spoken dialogue, the model is misled. However, oracle selective grounding consistently improves COMET over the baseline in almost all cases, recovering most of the potential gain while using only 30% of the visual processing. Figure 2 visualises the per-movie, per-language COMET gain from oracle selective grounding, highlighting that action-rich movies (e.g., Skyfall) show larger improvements. Human evaluation with a small set (30 examples, each for Telugu and Hindi) confirmed that selective grounding with oracle significantly improves adequacy: average score increased from 2.9 (baseline) to 4.1 (selective) on a 1-5 scale.

4.4 Comparison of Summarization Methods

Attr-VC, which aggregates a 5-minute window into coarse attributes, proves more robust to drift than Inter-VS. In many cases (e.g., Avatar Bengali, Oppenheimer Telugu, Titanic Hindi), its selective 30% COMET surpasses the full VT of Inter-VS. The attribute-based representation, by ignoring precise timing, effectively filters out irrelevant frames. Inter-VS, while capturing finer-grained visual events, is more sensitive to misalignment; its full VT often underperforms the baseline, but selective grounding still yields gains.

Movie	Lang	Baseline			5-Minute Slide Visual Attribute						Inter-Chunk Visual Summarisation					
		BLEU	chrF++	COMET	Visual-Enhanced			Oracle Selective			Visual-Enhanced			Oracle Selective		
					BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET
Avatar	Ben	5.68	28.71	0.6298	6.95	27.95	**0.7014**	6.92	**29.98**	**0.6829**	**8.10**	28.24	0.7137	6.91	29.80	0.6865
Avatar	Tel	4.30	19.66	0.5257	3.28	18.23	0.5154	4.38	**19.79**	0.5390	**3.67**	18.32	0.5153	**4.67**	19.85	0.5390
Avatar	Tam	3.85	23.49	0.5352	4.08	22.94	0.5545	4.24	24.36	0.5580	**4.57**	**23.62**	**0.5613**	**4.33**	**24.50**	**0.5639**
Avatar	Kan	3.50	18.94	0.4857	2.34	15.20	0.4582	3.39	18.65	0.4933	**2.23**	**15.28**	**0.4612**	**3.33**	18.37	**0.4946**
Oppenh.	Ben	8.05	29.38	0.7026	5.47	25.09	0.6735	8.03	**29.41**	0.7248	**6.37**	26.26	0.6858	**8.08**	29.50	0.7237
Oppenh.	Hin	11.76	31.64	0.6467	8.62	27.28	0.5972	11.83	31.74	0.6642	**9.62**	**28.72**	**0.6297**	**11.86**	**32.06**	**0.6690**
Oppenh.	Tel	4.04	18.86	0.5475	3.29	17.77	0.5387	3.89	**19.13**	**0.5654**	**3.53**	18.05	0.5379	**3.96**	19.21	0.5647
Oppenh.	Tam	3.15	20.60	0.5366	3.15	21.37	0.5654	3.36	21.85	0.5630	**3.47**	**21.63**	**0.5690**	**3.66**	**22.18**	**0.5715**
Oppenh.	Kan	2.95	16.81	0.4938	2.23	14.63	0.4740	2.96	17.20	0.5066	**2.30**	14.25	0.4735	2.93	17.05	**0.5090**
Skyfall	Ben	6.31	27.30	0.6914	4.10	23.51	0.6588	**6.04**	27.39	**0.7098**	**4.55**	**23.68**	0.6612	5.93	27.14	0.7056
Skyfall	Hin	6.31	25.65	0.6026	5.74	25.28	0.5882	**6.53**	**26.68**	**0.6258**	**6.41**	25.86	0.6098	6.65	26.47	0.6245
Skyfall	Tel	2.47	17.68	0.5288	2.13	16.66	0.5248	**2.24**	**18.00**	**0.5478**	1.41	16.84	0.5157	2.16	17.86	0.5454
Skyfall	Tam	2.33	21.09	0.5350	2.22	21.11	0.5581	2.59	21.78	0.5581	**1.83**	20.89	**0.5595**	**2.66**	**22.02**	**0.5639**
Skyfall	Kan	1.59	16.88	0.4920	1.76	14.38	0.4668	1.59	16.90	0.5013	**1.54**	**14.36**	**0.4682**	1.58	16.80	**0.5038**
Spider2	Ben	9.58	26.81	0.7190	6.69	24.44	0.6902	9.10	26.97	0.7359	**8.55**	**25.77**	**0.7021**	**9.47**	**27.18**	**0.7350**
Spider2	Hin	12.33	29.15	0.6459	10.13	27.71	0.6286	12.61	**30.20**	0.6746	**12.19**	29.17	**0.6532**	**12.93**	30.32	**0.6786**
Spider2	Tel	5.22	18.47	0.5407	3.57	17.69	0.5349	5.04	18.84	0.5567	3.40	17.73	0.5342	**5.04**	**18.74**	**0.5569**
Spider2	Tam	4.12	21.65	0.5448	3.27	21.39	0.5601	4.29	22.73	0.5684	**4.01**	**22.09**	**0.5694**	**4.32**	**22.83**	**0.5761**
Titanic	Ben	9.59	25.87	0.6960	7.04	22.97	0.6616	9.55	26.11	0.7130	**8.65**	**24.97**	**0.6849**	**9.82**	**26.41**	**0.7150**
Titanic	Hin	11.98	26.59	0.6152	9.12	24.45	0.5711	11.92	27.12	0.6321	**12.29**	**26.90**	**0.6176**	**12.66**	**27.62**	**0.6367**
Titanic	Tel	5.03	17.82	0.5350	3.85	16.99	0.5211	4.92	**18.01**	0.5481	**4.32**	17.52	0.5296	**5.11**	18.21	0.5494
Titanic	Kan	4.95	17.16	0.4950	3.11	14.04	0.4670	4.73	16.97	0.5037	**3.14**	**14.45**	**0.4665**	**4.86**	**17.03**	**0.5049**

Table 3: Comparison of two visual summarization strategies. *5-Minute Slide Visual Attribute* aggregates a 5-minute sliding window into structured attributes (setting, gender, honorifics, emotion); *Inter-Chunk Visual Summarization* summarizes the visual content between dialogue turns. For each method, we report *Visual-Enhanced* (using the full visual context for all segments) and *Oracle Selective* (replacing the worst 30% of baseline segments by baseline COMET with the visual-enhanced translation). This oracle shows the upper bound of selective grounding. Metrics are corpus-level BLEU, chrF++, and COMET. **Bold** indicates the higher score between the two methods for the same condition (Visual-Enhanced or Oracle Selective).

Language	Attr-VC		Inter-VS	
	Full	Sel30	Full	Sel30
Hindi	-5.0%	+3.4%	0.0%	+3.9%
Bengali	-1.6%	+3.7%	+0.3%	+3.7%
Telugu	-1.6%	+3.0%	-1.7%	+2.9%
Tamil	+4.0%	+5.9%	+5.0%	+5.8%
Kannada	-5.1%	+2.3%	-4.9%	+2.4%

Table 4: Language-wise average COMET improvement (Δ) over baseline for each method and condition. Positive values indicate improvement. Oracle Selective replaces the worst 30% of baseline segments by baseline COMET.

4.5 Language-Wise Trends

Table 4 aggregates COMET improvement over the baseline for each language. For Attr-VC selective, all languages show positive gains (range +2.3% to +5.9%). The gains are larger for morphologically rich languages (Bengali, Tamil, Kannada) where visual cues (e.g., honorifics, emotional tone) help resolve pragmatic ambiguities. Inter-VS selective also yields consistent improvements, though slightly lower for some languages. The COMET improvements across languages are further illustrated in Figure 3.

4.6 Summary of Key Findings

Our case study yields three actionable insights:

1. **Coarse attribute-based summarization is robust to temporal drift** By aggregating visual information over a 5-minute window into structured attributes, Attr-VC achieves pragmatic gains without being misled by misaligned frames.

2. **Selective visual grounding can recover most of the gain** An oracle that replaces only the worst 20-30% of baseline segments with visual-enhanced translations consistently improves COMET over baseline, using a fraction of the visual processing.

3. **Alignment quality often outweighs architectural complexity** Fine-grained summarization methods that rely on precise temporal alignment are fragile. For real-world deployment, drift-tolerant architectures are preferred.

5 Discussion

Our case study reveals a central tension: while visual context can provide critical grounding for a small subset of segments, it is irrelevant or even harmful for the majority. This asymmetry, combined with temporal misalignment, shapes the relative performance of the two summarization strategies and the effectiveness of selective grounding.

5.1 When Visual Context Helps and When It Does Not

Only a minority of subtitle segments truly depend on visual information. Action cues ("He's coming!"), emotion-driven exchanges ("I'm so sorry"),

and references to on-screen objects ("That one") require visual grounding to resolve ambiguity. For the vast majority of conversational dialogue, text alone is sufficient, and adding visual context adds no benefit. In our dataset, we estimate that fewer than 15% of subtitles are visually grounded in this sense. This explains why even the best-performing method oracle selective grounding-achieves only a modest overall COMET gain (2-5%) while replacing only 20-30% of segments.

5.2 Why Attribute Summarization Outperforms Gap Summarization?

Attr-VC aggregates a 5-minute sliding window into high-level attributes. This coarse representation is rarely misleading for neutral segments and provides valuable pragmatic context for the minority that need it. Moreover, it is inherently robust to misalignment because it aggregates over longer time windows, effectively ignoring irrelevant frames. Inter-VS summarizes the visual content between dialogue turns. This finer-grained representation captures visual events that may be directly relevant, but it is more sensitive to drift. When visual frames are misaligned, Inter-VS can introduce misleading information, causing its full visual-enhanced translations to sometimes underperform the baseline. However, when applied selectively (only to the worst baseline segments), Inter-VS still yields substantial gains on the replaced set.

5.3 The Role of Oracle Selective Grounding

Our oracle analysis replaces the worst 20-30% of baseline segments (by baseline COMET) with the corresponding visual-enhanced translation. This shows the upper bound of what could be achieved with a perfect quality-estimation model. For both methods, selective grounding consistently lifts COMET above the baseline, often matching or even exceeding the full visual-enhanced performance of the other method.

5.4 Insights for Low-Resource Indian Languages

For morphologically rich languages such as Bengali and Kannada, the small subset of visually grounded segments often includes honorifics, implicit referents, or emotional tone areas, where text-only models are weakest. Coarse attribute summarisation helps in these cases without harming the rest, yielding clear COMET gains in the selective setting (e.g., +3.7% for Bengali, +2.3% for Kannada). This suggests that for low-resource settings, investing in reliable, drift-tolerant visual abstractions is more practical than pursuing fine-grained fusion or high-frequency visual processing.

5.5 Practical Implications for Movie Localisation

Our findings lead to three actionable recommendations for deploying multimodal subtitle translation:

1. **Be selective:** Not every subtitle needs visual context. An oracle study shows that replacing only the worst 20-30% of baseline segments can recover most of the gain. A practical system would use a quality-estimation model (e.g., based on sentence length, emotion words, or visual confidence) to trigger visual enhancement only when needed.

2. **Prefer robust architectures:** Attribute-based summarization (5-minute sliding window condensed into structured cues) offers a lightweight, drift-tolerant solution suitable for production pipelines.

3. **Fix alignment before using fine-grained context:** If a more detailed visual context is desired (e.g., gap-based summarization), pre-processing steps such as dynamic time warping (DTW) or audio-visual synchronisation are essential to avoid performance degradation.

6 Conclusion

In this paper, we compare two summarization strategies for integrating visual context into subtitle translation for five low-resource Indian languages. We find that temporal misalignment- a common real-world issue- causes full visual-enhanced translation to often underperform the text-only baseline. However, an oracle selective grounding that replaces only the worst 20-30% of baseline segments with visual-enhanced translations consistently improves semantic adequacy (COMET) across all the languages, recovering most of the potential gain while using a fraction of the visual processing. Coarse attribute-based summarization proves particularly robust to drift, capturing emotional tone and scene-level cues that text alone cannot convey. Our results underscore that alignment quality often outweighs architectural complexity and that selective visual grounding offers a practical path to efficient, deployable multimodal subtitle translation.

Limitations

Our study is limited to five movies and five languages; the proportion of visually grounded segments may vary by genre and across different types of audiovisual content. The oracle selective grounding demonstrates an upper bound; future work should develop automatic quality-estimation models that can identify low-quality baseline segments without reference translations, enabling practical selective grounding. Human evaluation of pragmatic adequacy (e.g., honorifics, emotional tone) would complement automatic metrics to better capture the subtle benefits of selective visual grounding. Additionally, exploring more sophisticated alignment techniques (e.g., dynamic time warping) could further reduce temporal misalignment and improve the robustness of fine-grained visual context.

Acknowledgements

The authors would like to express their sincere gratitude to the project, Centre of Indian Language Data (COIL-D) under Bhashini, funded by the Ministry of Electronics and Information Technology (MeitY), Government of India for its generous support.

References

Appicharla, Ramakrishna, Baban Gain, Santanu Pal, Asif Ekbal, and Pushpak Bhattacharyya. 2024. A case study on context-aware neural machine translation with multi-task learning. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 246–257, Sheffield, UK, June. European Association for Machine Translation (EAMT).

Artetxe, Mikel, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online, July. Association for Computational Linguistics.

Datta, Debtanu, Shounak Paul, Kshetrimayum Boynao Singh, Sandeep Kumar, Abhinav Joshi, Shivani Mishra, Sarika Jain, Asif Ekbal, Pawan Goyal, Ashutosh Modi, and Saptarshi Ghosh. 2025. Findings of the JUST-NLP 2025 shared task on summarization of Indian court judgments. In *Proceedings of the 1st Workshop on NLP for Empowering Justice (JUST-NLP 2025)*, pages 5–11, Mumbai, India, December. Association for Computational Linguistics.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig. 2025. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 28th edition.

Elliott, Desmond, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Gain, Baban, Dibyanayan Bandyopadhyay, Samrat Mukherjee, Chandranath Adak, and Asif Ekbal. 2025. Impact of visual context on noisy multimodal nmt: An empirical study for english to indian languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 24(8), August.

Kumar, Deepak, Baban Gain, Kshetrimayum Boynao Singh, and Asif Ekbal. 2025. Does vision still help? multimodal translation with CLIP-based image selection. In Nakazawa, Toshiaki and Isao Goto, editors, *Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025)*, pages 115–123, Mumbai, India, December. Association for Computational Linguistics.

Meta. 2024. Llama 3.1: The llama 3.1 collection of multilingual large language models. `https://huggingface.co/meta-llama/Llama-3.1-8B`, July. Model release date: July 23, 2024. Accessed: 2026-03-27.

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, Pennsylvania, USA, July. 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, Portugal, September. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Radinger, Anke. 2025. Subtitling in audiovisual translation studies. In *Researching Subtitling Processes*, pages 29–53. Springer.

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, November. Association for Computational Linguistics.

Shen, Zhiyu, Yunhe Pang, Yanghui Rao, and Jianxing Yu. 2025. CoE: A clue of emotion framework for emotion recognition in conversations. In Che, Wanxiang, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23548–23563, Vienna, Austria, July. Association for Computational Linguistics.

Singh, Kshetrimayum Boynao, Asif Ekbal, and Partha Pakray. 2025a. Evaluating IndicTrans2 and ByT5 for English–Santali machine translation using the ol chiki script. In Shukla, Ankita, Sandeep Kumar, Amrit Singh Bedi, and Tanmoy Chakraborty, editors, *Proceedings of the 1st Workshop on Multimodal Models for Low-Resource Contexts and Social Impact (MMLoSo 2025)*, pages 95–100, Mumbai, India, December. Association for Computational Linguistics.

Singh, Kshetrimayum Boynao, Deepak Kumar, and Asif Ekbal. 2025b. Evaluation of LLM for English to Hindi legal domain machine translation systems. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 823–833, Suzhou, China, November. Association for Computational Linguistics.

Vasu, Pavan Kumar Anasosalu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokul Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, and Hadi Pouransari. 2025. Fastvlm: Efficient vision encoding for vision language models.

Wang, Yuqing and Yun Zhao. 2024. TRAM: Benchmarking temporal reasoning for large language models. In Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6389–6415, Bangkok, Thailand, August. Association for Computational Linguistics.

A Appendix

A.1 Ethical Considerations

Multimodal systems may inherit and amplify biases present in movies, where visual cues can reinforce stereotypes. Temporal misalignment between audio and visuals may also lead to inappropriate translations of culturally sensitive content. For Indian languages, honorifics and regional variations require careful handling to ensure accuracy. We advocate for human oversight, transparent reporting, and evaluation frameworks that assess cultural appropriateness. As selective grounding shows that only some segments need visual context, applying visual enhancement through a bias-aware filter can help reduce potential harms.

A.2 Visual Feature Extraction

1. The original `.mkv` video is compressed to $\approx$ 1-GB and converted to `.mp4`.

2. Frames are extracted at 1-fps and passed to Apple FastVLM-0.5B to obtain a raw textual description per frame.

3. Raw descriptions are cleaned (removing repeated phrases, normalising punctuation) to produce a final description per frame.

For each subtitle, the visual context is formed by concatenating the cleaned descriptions of all frames whose timestamps fall within the subtitle's time window. When multiple frames belong to the same segment, the descriptions are aggregated into a single summary (for the gap-based method, the window is the interval between the previous subtitle's end and the current subtitle's start).

A.3 Context-Aware Translation Pipeline

The translation pipeline uses the same Qwen-2.5-7B-Instruct model for both baseline and visual-enhanced translations. The only difference is the input prompt. For the baseline, the model receives only the English source. For visual-enhanced, we prepend the summarised visual context (Attr-VC or Inter-VS) to the source using the visual-enhanced prompt template shown in Table 5. The prompt instructs the model to ground its translation in the visual scene, paying attention to gender, honorifics, and emotional tone. The same greedy decoding parameters described in hyperparameterss are used for all runs.

A.4 Hyperparameters

All inference runs use greedy decoding to ensure reproducibility.

For the translation model (Qwen-2.5-7B-Instruct):

- `max_new_tokens = 100`
- `do_sample = False`
- `repetition_penalty = 1.1`
- temperature and top-p left at default (1.0 and 1.0, effectively greedy).

For the summarization model (Llama-3.1-8B-Instruct):

- `max_new_tokens = 256`
- `do_sample = False`
- temperature and top-p at default (1.0 and 1.0).

A.5 Oracle Selective Grounding

Per-segment COMET scores for the baseline translation (against the reference) are computed using the `Unbabel/wmt22-comet-da` model. The worst $k\%$ of segments (by baseline COMET) are replaced with the corresponding visual-enhanced translation (from either Attr-VC or Inter-VS). We report results for $k = 20\%$ and 30%. Appendix 8

B Additional Analysis

B.1 Window Size Considerations

The choice of a 5-minute sliding window for Attribute Visual Context was guided by the need to capture enough local scene context while remaining robust to temporal drift. A larger window (e.g., 10 minutes) would aggregate more visual information, but it could also include more irrelevant frames, potentially increasing the risk of hallucination when the visual context does not align with the dialogue. A smaller window (e.g., 2 minutes) would be more sensitive to misalignment. The 5-minute window offers a reasonable trade-off, as evidenced by the improvement in COMET over the baseline for most language-movie pairs.

B.2 Oracle Selective Grounding

The full oracle selective results for both summarization methods are presented in Table 8. Replacing only the worst 20-30% of baseline segments consistently lifts COMET above the baseline, demonstrating that most of the gain can be achieved with a fraction of the visual processing.

C Additional Analysis

The 5-minute sliding window offers a reasonable trade-off between context and robustness; a larger window would risk including irrelevant frames, a smaller window would be more sensitive to misalignment. Oracle selective grounding (Table 8) shows that replacing only the worst 20-30% of baseline segments recovers most of the gain with minimal visual processing. Fine-grained fusion methods (visual prefixing and cross-attention fusion) failed under misalignment (e.g., "He is very kind" $\rightarrow$ "He drives fast") because they assume perfect alignment, whereas coarse attribute summarization ignores irrelevant frames. This reinforces the idea that alignment quality often outweighs architectural complexity, and that selective grounding offers a practical path to efficient visual-guided translation.

C.1 Clarifications

Movie subtitle translation is inherently difficult (fragmented speech, cultural references, zero-shot domain adaptation); our baseline scores reflect this challenge, and our contribution lies in the *relative* COMET gain, not absolute SOTA. We use Qwen-2.5-7B-Instruct because it supports zero-shot instruction following without fine-tuning, unlike dedicated Indic models (e.g., IndicTrans2) that are optimised for sentence-level translation and do not accept multi-field visual context prompts. FastVLM-0.5B prioritises efficiency (85× faster than LLaVA); using a smaller VLM makes gains harder to achieve, so our positive results demonstrate robustness. FastVLM outputs raw descriptions; we summarise them with Llama 3.1 to obtain structured attributes or free-text gap summaries, because direct prompting of FastVLM for structured output is not feasible. Our experiments compare visual-enhanced translation against an identical text-only baseline, isolating the effect of visual context; comparing across different MT systems would confound architectural differences.

C.2 Evaluation Metrics

We report corpus-level BLEU (Papineni et al., 2002) using SacreBLEU and chrF++ (Popović, 2015) with `word_order=2`. COMET (Rei et al., 2020) is computed with the `wmt22-comet-da` model using default settings.

C.3 Data and Code Availability

To foster reproducibility, the curated movie-subtitle-visual alignment data for all five languages will be released under a fair-use educational/research license. The release includes English sources, reference translations, and extracted visual descriptions. All the codes and datasets used for extraction, summarization, translation, and evaluation are available at GitHub [3] and our group page.[4]

[3] `https://github.com/Tarunc224/visually-guided-subtitle-translation`

[4] https://ai-nlp-ml.github.io/resources.html

Baseline Prompt (text-only)

```
<|im_start|>system
You are a translation expert. Translate dialogue from English to {target_language}.
RULES:
- Provide ONLY the translated {target_language} dialogue.
- DO NOT include explanations, or English text.
<|im_end|>
<|im_start|>user
[SOURCE]: "{row['english_dialogue']}"
[TASK]: Translate to {target_language} dialogue.
<|im_end|>
<|im_start|>assistant
```

Visual-Enhanced Prompt

```
<|im_start|>system
You are a cinematic multimodal translator specializing in English-to-{target_language}.
Your goal is to provide a "grounded translation" where the choice of words depends on the visual scene.

RULES:
1. GENDER: Use the Visual Context to identify speaker/listener gender.
2. HONORIFICS: Determine social hierarchy from the scene (Formal vs. Informal).
3. LOOSE MEANING: Prioritize emotional intent and natural {target_language} flow.
4. Output ONLY the translated {target_language} dialogue text. No names, no English.
<|im_end|>
<|im_start|>user
[VISUAL CONTEXT]: {row['visual_context']}
[ENGLISH SOURCE]: "{row['english_source']}"
[TASK]: Based on the visual scene, provide the most natural {target_language} translation.
<|im_end|>
<|im_start|>assistant
```

Table 5: Comparison of baseline and visual-enhanced prompts.

Summarisation Method	Prompt Template

Attribute Visual Context (Attr-VC)	`<|begin_of_text|><|start_header_id|>system<|end_header_id|>`
	`Identify these cinematic attributes to guide {target_language} translation:`
	`[SETTING]: (e.g., Formal, Public, Intimate)`
	`[GENDER]: (Speaker/Listener gender)`
	`[RELATION]: (e.g., Stranger, Family, Hostile)`
	`[HONORIFIC]: (language-specific, e.g., APNI/TUMI for Bengali)`
	`[SUMMARY]: (One sentence factual summary with emotional intent)`
	`Output ONLY these tags.<|eot_id|><|start_header_id|>user<|end_header_id|>`
	`Visual Data: {sample[:3000]}<|eot_id|>`
	`<|start_header_id|>assistant<|end_header_id|>`

Inter-Chunk Visual Summarisation (Inter-VS)	`<|begin_of_text|><|start_header_id|>system<|end_header_id|>`
	`You are a movie analyzer. Summarize the following visual descriptions`
	`from {start_sec}s to {end_sec}s of the movie into 2-3 sentences.`
	`Focus ONLY on the current location and character actions.`
	`Do not use introductory filler.`
	`<|eot_id|><|start_header_id|>user<|end_header_id|>`
	`Visual Data: {text_blob[:2500]}`
	`<|eot_id|><|start_header_id|>assistant<|end_header_id|>`

Table 6: Prompt templates used for visual summarization. Both the prompts are given to Llama-3.1-8B-Instruct. The attribute prompt outputs structured tags; the inter-chunk prompt outputs a free-text sentence. Placeholders in braces are replaced with actual data at runtime.

Scenario	Source	Baseline (literal)	Visual-Enhanced	Reference	Explanation
			Visual context helps		
Emotion	"I'm so sorry."	"I am sorry."	"I am truly sorry."	"I am truly sorry."	Facial expression conveys deeper remorse, captured by visual summary.
Action	"He's coming!"	"He is coming."	"The man is coming!"	"The man is coming!"	Visual identifies gender and urgency.
Honorific	"Please sit."	"Sit."	"Please sit (formal)."	"Please sit (formal)."	Formal setting triggers correct honorific.
			Visual context backfires (temporal misalignment)		
Misalign	"He is very kind."	"He is very kind."	"He drives fast."	"He is very kind."	Car chase frame (drift) causes hallucination.
Misalign	"I'm flying."	"I'm flying."	"I'm swimming."	"I'm flying."	Calm ocean scene (drift) misleads.
			Visual context backfires (irrelevant visuals)		
Neutral	"How was your day?"	"How was your day?"	"How was your day?"	"How was your day?"	No improvement; adds processing overhead.

Table 7: When visual context helps (emotion, action, honorifics) and when it backfires (temporal misalignment, irrelevant visuals). Visual-enhanced outputs are from the better of the two summarization methods for each case. The misalignment examples are drawn from actual data where cumulative drift paired a kind remark with a car chase, and a flying statement with a calm ocean scene.

Movie	Language	Baseline COMET	5-Minute Slide Visual Attribute		Inter-Chunk Visual Summarisation	
			Oracle selective 20%	Oracle selective 30%	Oracle selective 20%	Oracle selective 30%
Avatar	Bengali	0.6298	0.6682	0.6829	0.6709	0.6865
Avatar	Telugu	0.5257	0.5354	0.5390	0.5361	0.5390
Avatar	Tamil	0.5352	0.5580	0.5650	0.5569	0.5639
Avatar	Kannada	0.4857	0.4933	0.4943	0.4928	0.4946
Oppenhimer	Bengali	0.7026	0.7224	0.7248	0.7210	0.7237
Oppenhimer	Hindi	0.6467	0.6617	0.6642	0.6643	0.6690
Oppenhimer	Telugu	0.5475	0.5604	0.5654	0.5600	0.5647
Oppenhimer	Tamil	0.5366	0.5630	0.5715	0.5639	0.5715
Oppenhimer	Kannada	0.4938	0.5066	0.5096	0.5065	0.5090
Skyfall	Bengali	0.6914	0.7064	0.7098	0.7044	0.7056
Skyfall	Hindi	0.6026	0.6223	0.6258	0.6216	0.6245
Skyfall	Telugu	0.5288	0.5430	0.5478	0.5415	0.5454
Skyfall	Tamil	0.5350	0.5581	0.5659	0.5561	0.5639
Skyfall	Kannada	0.4920	0.5013	0.5038	0.5014	0.5038
Spider 2	Bengali	0.7190	0.7337	0.7359	0.7305	0.7350
Spider 2	Hindi	0.6459	0.6684	0.6746	0.6717	0.6786
Spider 2	Telugu	0.5407	0.5527	0.5567	0.5527	0.5569
Spider 2	Tamil	0.5448	0.5684	0.5753	0.5700	0.5761
Titanic	Bengali	0.6960	0.7114	0.7130	0.7120	0.7150
Titanic	Hindi	0.6152	0.6293	0.6321	0.6320	0.6367
Titanic	Telugu	0.5350	0.5456	0.5481	0.5465	0.5494
Titanic	Kannada	0.4950	0.5037	0.5065	0.5049	0.5063

Table 8: Oracle selective COMET scores for the two summarization methods. "Oracle selective 20%" and "Oracle selective 30%" replace the worst 20% and 30% of baseline segments (by baseline COMET) with the corresponding visual-enhanced translation. The full visual-enhanced results are reported in Table 3.

Is a Picture Worth a Thousand Words? Exploration and Implementation Considerations for Visual Context in Translation Workflows

Vera Senderowicz Guerra
Welocalize
vera.senderowicz@welocalize.com

Olesia Khrapunova
Welocalize
olesia.khrapunova@welocalize.com

Abstract

Vision-language models (VLMs) have the potential to enhance machine translation (MT) by leveraging visual context alongside text, yet their real utility for production workflows remains unclear. We conduct a unified, multi-condition evaluation of six leading VLMs–both open and proprietary–on two benchmarks (CoMMuTE and CaMMT), targeting lexical and cultural disambiguation respectively. We complement this with a domain-style case study simulating technical documentation localization. Results show that model performance varies widely, and the benefit of relevant images does not transfer uniformly across use cases. Proprietary models are notably sensitive to irrelevant images while open-source models are generally more stable. Contradicting visuals, by contrast, degrade translation across all models. Taken together, our findings show that rigorous evaluation is a necessary precondition for production deployment: metric gains can mask real accuracy losses, model sensitivity to irrelevant images should inform model selection, and avoiding contradicting visuals is a hard requirement for any pipeline.

1 Introduction

The extent to which translation choices depend on the various conceptions of *context* has been widely studied (Dimitriu, 2015; House, 2006; Damaskinidis, 2016). In Machine Translation (MT), context retrieval is especially challenging, as it is inherently constrained by the model's input, with the available context limited to the information the model is exposed to. This has motivated a growing body of work on context-aware MT (Voita et al., 2019; Fernandes et al., 2021; Castilho and Knowles, 2025).

Organizations deploying MT in production workflows increasingly work with content that includes visual context: technical documentation, UI strings, marketing assets. In theory, visual context can inform translation decisions –from guiding lexical choice to signaling register or style– but a natural question arises: should available images be fed to the model? And if so, what happens when the image attached to a segment is not the most relevant to translate its content? Before visual context can be reliably deployed in production MT, we need to understand precisely when and why images help or hurt.

Our results show that relevant images consistently improve lexical disambiguation, but this benefit does not necessarily transfer to other use cases, which is precisely why rigorous evaluation across conditions and scenarios is necessary prior to production deployment. We focus mainly on disambiguation, a task that provides a straightforward-to-measure signal, and evaluate six VLMs from both open and proprietary providers. We make the following contributions:

- A **multi-condition experimental design** (no image, correct image, random image, and contradicting image when available) that disentangles the impact of *image content* versus *image presence*;
- A **unified evaluation protocol** across both *lexical* and *cultural disambiguation*, enabling direct comparison under consistent conditions

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 103–115

and overlapping languages to assess robustness transference across tasks;

- An illustrative **practitioner-oriented case study** testing whether benchmark findings hold in a realistic localization pipeline, illustrating how to measure visual context impact in production-like conditions.

2 Related Work

Recent work has benchmarked visual context for machine translation, including lexical and cultural disambiguation tasks. The CoMMuTE benchmark (Futeral et al., 2023; Futeral et al., 2025) evaluates multimodal translation on sentences with ambiguous words, but compares only text-only and text+image conditions, without establishing whether models exploit image *content* or merely react to image *presence*.

The CaMMT benchmark (Villa-Cueva et al., 2025) performs evaluation on cultural disambiguation and includes controls with unrelated images, finding that they can degrade quality, supporting that gains under the correct image reflect content rather than mere visual input. However, no previous study has jointly assessed both benchmarks, so transfer of visual grounding across these task types remains an open question.

Concurrently, Liu et al. (2025) propose a reasoning-based framework that detects ambiguity before invoking visual context, but do not evaluate model behavior under irrelevant or contradicting images. Multimodal MT has a longer lineage in the WMT shared tasks (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018) and the Multi30K dataset (Elliott et al., 2016). Shen et al. (2024) survey the resulting body of work across architectures, training strategies, and evaluation paradigms. We depart from this line by evaluating modern instruction-tuned VLMs with native multimodal interfaces, and by systematically controlling for image content versus mere image presence, a confound noted in prior work (Elliott, 2018) but not part of the shared-task evaluation protocol itself. We additionally bridge benchmark evaluation to production deployment through a domain-style case study; a step that, to our knowledge, remains unaddressed in prior multimodal MT work.

	CoMMuTE	**CaMMT**
Disambiguation type	Lexical	Cultural
Translation direction	EN→X	EN→X
Languages used	FR, DE, AR, ZH, RU	RU, ZH
Items per language	150–162 pairs	44 (ZH); 57 (RU)
Images per item	2 (disambiguating)	1
Reference translations	Correct + incorrect	Regional only

Table 1: Overview of the two evaluation benchmarks.

3 Experimental Design

3.1 Datasets

We perform our evaluations on the two benchmarks that target complementary forms of visual disambiguation in translation, summarized in Table 1.

Each item in **CoMMuTE** consists of an ambiguous English sentence paired with two images that trigger different (unambiguous) target-language translations. For each image, a correct reference translation is provided; each image's correct reference translation constitutes an incorrect translation for the other image in that specific item. See Appendix A for a sample item.

In **CaMMT**, each item pairs a source sentence in the regional language with English reference translations and an image depicting a culturally-specific context. For items containing Culturally-Specific Items (CSIs), the dataset provides two English translation variants: a *conserved* version that preserves the original cultural term, and a *substituted* version that replaces it with a target-language generic equivalent. Since our goal is to perform a direct comparison with CoMMuTE and be able to assess the transfer of capabilities across tasks, we adapt CaMMT by reversing the translation direction and using only the substituted source condition with the generic term, analogous to the CoMMuTE's lexical ambiguity case (see Appendix B for a sample item).

3.2 Conditions and Research Questions

Each sentence is translated under multiple visual conditions, framed around specific research questions:

RQ1 (CoMMuTE + CaMMT): Does a relevant image improve translation? We compare a *correct image* (matching the intended meaning)

Model	Provider	Access
Qwen3-VL-8B	Alibaba	Open
Aya Vision 8B	Cohere	Open
InternVL3-8B	OpenGVLab	Open
GPT-4o	OpenAI	API
Claude Sonnet 4.6	Anthropic	API
Gemini 2.5 Flash	Google	API

Table 2: Models evaluated. Open-source models are all 8B parameters; proprietary model sizes are undisclosed.

against a text-only *no image* baseline.

RQ2 (CoMMuTE + CaMMT): Does any image affect translation? We compare a *random image* (an unrelated image from the same dataset) against the *no image* baseline.

RQ3 (CoMMuTE): Does a misleading image actively degrade translation? We compare a *contradicting image* (encoding the opposite meaning) against the *no image* baseline.

3.3 Models

We evaluate six VLMs (Table 2), spanning multiple providers and model families, both open-source and proprietary. Open-source models were fixed at 8B parameters due to local compute constraints and use greedy decoding (`do_sample=False`), while proprietary models are queried with each provider's default sampling configuration. We deliberately retained these defaults to reflect out-of-the-box practitioner behavior. We acknowledge that this introduces a determinism asymmetry between the two groups, and that proprietary results are subject to additional run-to-run variance not captured in our single-run point estimates. We therefore frame open-vs-proprietary differences as descriptive rather than causal.

All models receive the same prompt, and the image (when applicable) is provided as visual input alongside the text in a single request, using each model's native multimodal interface. The full prompt template is given in Appendix C. Beyond the structured output instruction, we deliberately avoid further prompt engineering, as our goal is to isolate the effect of the image as much as possible.

3.4 Evaluation Metrics

Reference-based translation quality. We compute chrF (Popović, 2015) against both the correct-meaning and incorrect-meaning references (for CoMMuTE) and against the regional reference (for CaMMT), across all conditions. We use it as our primary metric because character n-gram overlap is robust to segmentation and morphological variation across scripts. We additionally report BLEU (Papineni et al., 2002) in the Appendix for comparability with prior MT literature and industrial practice, where it remains a familiar benchmark metric. Given per-language sample sizes (150–162 for CoMMuTE, 44–57 for CaMMT), small absolute differences (within ± 2 chrF) should be interpreted with caution.

Disambiguation gap (CoMMuTE only). Futeral et al. (2023) measure disambiguation via *pair accuracy*. We instead report the *disambiguation gap*, defined as chrF(correct reference) $-$ chrF(incorrect reference), which preserves the magnitude and direction of the effect, enabling finer-grained comparison across models, languages, and conditions, while also disentangling sense selection from overall translation quality, a distinction that raw chrF against a single reference cannot make.

4 Results

We organise results by benchmark: CoMMuTE (lexical; RQ1-RQ3) in Section 4.1, then CaMMT (cultural; RQ1, RQ2) alongside a cross-task comparison in Section 4.2. Runtime and API cost relative to chrF performance are reported in Appendix D, providing guidance for model selection based on both translation quality and time/cost efficiency.

4.1 Lexical Disambiguation (CoMMuTE)

Figure 1 reports chrF deltas averaged across all five languages. Relevant images do improve translation (RQ1), consistent with Futeral et al. (2023): all six models gain between $+2.1$ and $+5.8$ chrF, with GPT-4o and Claude 4.6 benefiting the most. However, not just any image helps (RQ2). Random images leave most models within ± 1 chrF of the baseline, confirming that the RQ1 gains stem from image *content*, not mere image *presence*. Only two models are negatively affected by irrelevant visual input: Gemini 2.5 Flash loses -9.2 chrF on average, and Claude 4.6 drops moderately (-2.4). Misleading images do degrade translation (RQ3): all models lose chrF, with losses generally proportional to their correct-image gains, though not perfectly symmetric. Notably, Gemini 2.5 loses more from contradicting images (-6.4) than it gains from correct ones ($+2.1$), consistent with chrF cap-

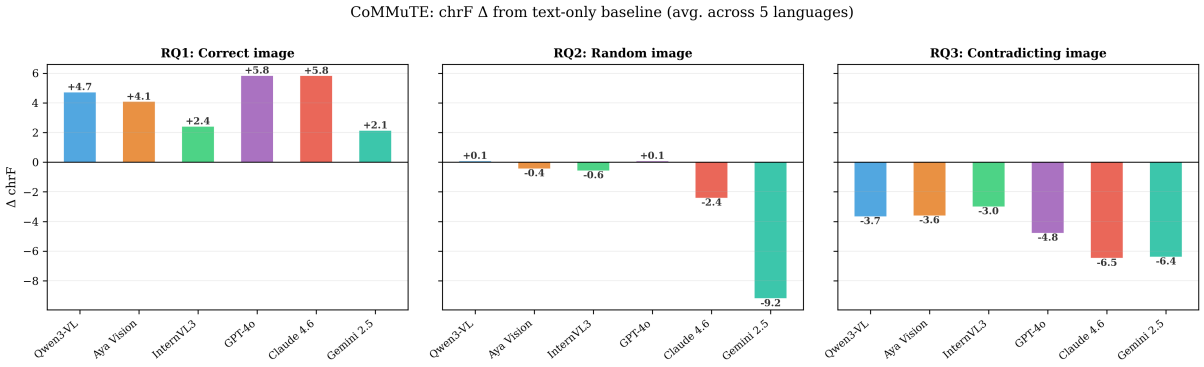


Figure 1: CoMMuTE: mean chrF change from the text-only baseline, averaged across five languages (FR, DE, AR, RU, ZH). Each panel corresponds to one visual condition (RQ1–RQ3). BLEU deltas follow the same pattern (Appendix E); per-language breakdowns and absolute values can be found in Appendix F.

turing translation-wide effects beyond the disambiguated term alone.

Per-language breakdowns (Appendix F) reveal that effect sizes are language- and model-dependent. Correct-image gains are largest on French for all three proprietary models, but open-source models show a different pattern: Qwen3-VL and InternVL3 gain most on Chinese, and Aya Vision peaks on French and German equally. In relative terms, gains tend to be proportionally larger for lower-baseline languages, though the pattern is not uniform across models; the clear exception is Gemini 2.5 Flash, which gains almost exclusively on French. Open-source models are largely robust to irrelevant images, whereas proprietary models diverge: GPT stays mostly robust across languages, while Claude shows random-image penalty and Gemini degrades sharply across all languages. The disambiguation gap heatmap in Appendix G corroborates these. Notably, Gemini's near-zero random-image gaps suggest that its sharp raw chrF drops are not accompanied by a systematic shift in sense selection.

4.2 Cultural Disambiguation (CaMMT) and Cross-Task Comparison

Figure 2 places CoMMuTE and CaMMT chrF deltas side by side on Russian and Chinese, the two languages shared between benchmarks. The correct-image benefit does *not* generalise from lexical to cultural disambiguation (RQ1): CoMMuTE gains are positive on both languages, whereas CaMMT deltas are slightly negative on Russian for all models except GPT-4o, and mixed on Chinese: marginally positive for most (Qwen $+2.5$ being the largest) but negative for Gemini (-2.7). Random-image sensitivity amplifies on CaMMT (RQ2): all three proprietary models degrade on both languages –Claude 4.6 up to -17.1 chrF (Russian) and Gemini 2.5 up to -27.1 (Chinese)– while open-source models remain largely stable on both tasks.

5 Domain-Style Case Study

Benchmark segments are cleaner and shorter than production strings (Vijayan et al., 2024), so understanding where benchmark behavior transfers and where it differs is precisely what makes systematic domain-grounded evaluation a prerequisite for production deployment. To probe feasibility in more realistic settings without requiring proprietary data, we add a small illustrative case study built around a fictitious company, *Nordic-Trans Motors* (heavy-vehicle service documentation). We use the synthetic figure in Appendix I and attach it to two English fragments of fairly comparable length: one **matched** text, whose translation can benefit from the figure and table, and one **mispaired** text that belongs to the same domain but is not related to the image. This setup mirrors a likely production trade-off: without a robust method for automatically verifying image–string relevance, available visuals would need to be linked to multiple segments, resulting in potentially non-helpful –or even harmful– pairings.

Each text is translated as a single unit, without an image and with the synthetic image, under the same prompt as in Section 3. We limit this case study to two languages and a small number of samples, both for practical constraints and to keep the qualitative analysis transparent and easy to follow. Source texts, hypotheses, and human-reviewed reference translations appear in Appendix J. We contrast Qwen and GPT only, as they rank among the

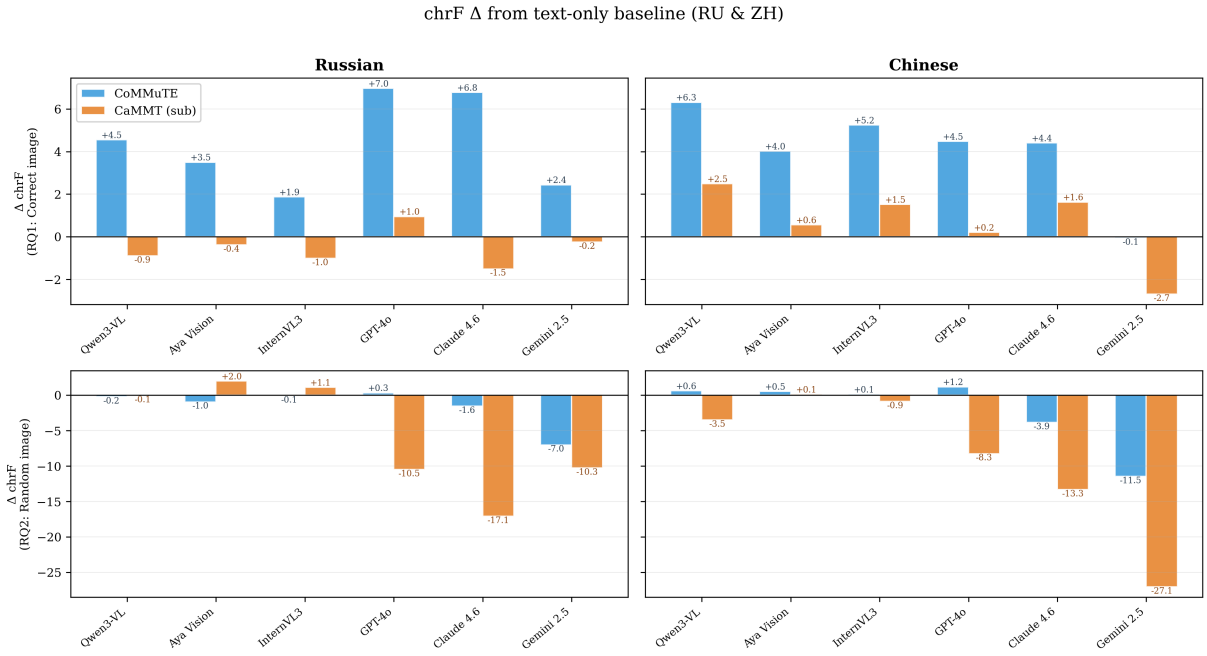


Figure 2: CaMMT and cross-task comparison: mean chrF change from the text-only baseline on Russian and Chinese for correct and random images (RQ1–RQ2). CoMMuTE (blue) vs. CaMMT substituted-source (orange). BLEU deltas (in Appendix H) follow a similar, although not identical, pattern.

strongest correct-image lift for open vs. proprietary families on CoMMuTE (Section 4.1) while remaining comparatively stable under misleading vision inputs. Tables 3 and 4 summarise chrF and BLEU computed over each text as a whole against our synthetic references, for the matched and mispaired conditions, respectively. Each table also reports reference-free COMET-QE (Rei et al., 2020) as a sanity check on fluency and adequacy relative to the English source.

COMET-QE mostly agrees with chrF/BLEU across matched conditions. However, across mispaired conditions, COMET-QE frequently diverges from chrF and BLEU. A human assessment provides the following contextualizing insights:

French. The matched image boosts translation adequacy in aspects beyond disambiguation: GPT moves from an incorrect/non-standard translation for *threadlocker* and an awkward *callout* calque to workshop-credible wording (*frein-filet*, *repère*, *routage*); Qwen similarly restructures phrasing and avoids the *callout* calque. Under the mispaired condition, outputs stay on topic –no accuracy issues stem from the unrelated drawing–, and relevantly, the image's presence still nudges surface wording for French: Qwen corrects a garbled term (*télématicie* → *télémétrie*) and aligns the English *forced*, from *obligatoire* to *forcé*, while GPT shows minor lexical swaps in French and significant metrics' boosts. These shifts suggest that even domain-grounding from an irrelevant image can improve output quality, a hypothesis that only surfaced when moving beyond benchmark conditions. However, COMET-QE edges downward despite chrF/BLEU gains, suggesting these surface improvements do not register as adequacy gains from a source-fidelity perspective.

Russian. Both conditions serve as a cautionary case: with the matched image, GPT shows significant drops in MT metrics, and while Qwen's metrics rise with the figure, the model shifts from one incorrect term to another (*cuff* → *cylinder head* instead of *manifold* in the output). With the mispaired image, neither model's translation meaningfully improves or degrades on human inspection, yet metrics diverge sharply: chrF and BLEU drop for Qwen but rise substantially for GPT, while COMET-QE moves in the opposite direction for both. These disagreements illustrate how metric gains under visual context do not reliably indicate actual translation improvement.

6 Conclusion

Our study finds potential for VLMs in MT workflows, not only for disambiguation but also for broader quality improvements in production-like contexts. Our benchmark evaluation was designed not as an end in itself, but to establish conditions

Lang	Model	chrF		BLEU		COMET–QE	
		text	+fig.	text	+fig.	text	+fig.
FR	Qwen3-VL-8B	58.66	63.37	32.83	36.00	−0.403	−0.266
	GPT-4o	55.96	64.66	14.91	34.66	−0.618	−0.424
RU	Qwen3-VL-8B	38.70	41.84	7.56	11.19	−0.243	−0.083
	GPT-4o	57.66	52.68	23.11	13.72	−0.114	−0.153

Table 3: Matched (on-topic) practical segment. **text** = translation without the image; **+fig.** = translation with that figure. COMET-QE (`Unbabel/wmt20-comet-qe-da`) is a reference-free quality estimator with unbounded scores; only relative comparisons within the same source text are meaningful. Higher values indicate higher estimated quality.

Lang	Model	chrF		BLEU		COMET–QE	
		text	+fig.	text	+fig.	text	+fig.
FR	Qwen3-VL-8B	76.26	77.87	62.07	62.49	−0.227	−0.265
	GPT-4o	83.03	88.73	66.87	76.97	−0.238	−0.258
RU	Qwen3-VL-8B	68.24	63.14	31.82	25.50	−0.147	−0.139
	GPT-4o	62.60	69.55	17.70	41.17	−0.236	−0.137

Table 4: Mispaired (off-topic) practical segment. **text** = translation without the image; **+fig.** = translation with unrelated figure. COMET-QE (`Unbabel/wmt20-comet-qe-da`) is a reference-free quality estimator with unbounded scores; only relative comparisons within the same source text are meaningful. Higher values indicate higher estimated quality.

under which visual context can be trusted: a necessary step before any deployment decision.

Our cross-task comparison shows that a model's ability on one task does not predict its performance on another: models that benefit from images in lexical disambiguation may see limited gains, or even losses, on cultural or domain-specific content. Model robustness also varies: open-source models and GPT remain largely stable when images are irrelevant, whereas Claude and Gemini show significant sensitivity to loosely matched inputs. On the benchmarks, this sensitivity causes quality drops, yet in the French case study, even a mispaired image improved surface quality for both models tested, suggesting that domain-grounding effects can emerge even without direct image–text relevance. Contradicting images were not tested in the case study, but on CoMMuTE they consistently harm outputs across all models. This suggests that reliable image–text matching is a strict requirement for any production pipeline.

Future directions include dedicated, production-scale benchmarks that span domain complexity, content types, and segment lengths across multiple conditions, as well as for advanced, scalable methods to reliably match images with the right source texts (e.g., via multimodal embeddings or hybrid retrieval). Until such tooling matures, deploying visual-context MT at scale will require care, especially when irrelevant or misleading images could negatively impact translation quality.

Overall, our results suggest that the benefits of vision-augmented MT are real but contingent. For practical adoption, organizations should not treat visual context as a universal improvement: they should carefully analyze the content types in their workflows, understand the specific challenges they expect visually-encoded information to address, verify image–text relevance, select models that provide robustness to irrelevant inputs, and design pipelines and evaluation methods targeting quality degradation signals that automatic metrics alone may miss.

7 Limitations

All results come from a single inference run; reported deltas are point estimates without confidence intervals. The domain-style case study uses a single synthetic image, two source texts, two languages, and two models; its findings are intended as illustrative hypotheses rather than generalisable conclusions.

8 Sustainability Statement

All local inference ran on a single Apple M4 Max (14-core, 36 GB unified memory) over approximately 20 h. Proprietary models were queried via API for 9.6 h. Using the Green Algorithms calculator (Lannelongue et al., 2021), we estimate a carbon footprint of 387.93 gCO_2e (2.27 kWh) for the local runs, equivalent to 0.42 tree-months, assuming European energy mix.

References

Barrault, Loïc, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels, October. Association for Computational Linguistics.

Castilho, Sheila and Rebecca Knowles. 2025. A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, 31(4):986–1016.

Damaskinidis, George. 2016. The visual aspect of translation training in multimodal texts. *Meta*, 61(2):299–319.

Dimitriu, Rodica. 2015. The many contexts of translation (studies). *Linguaculture*, 6(1):5–23, Jun.

Elliott, Desmond, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics.

Elliott, Desmond, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In Bojar, Ondřej, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, editors, *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark, September. Association for Computational Linguistics.

Elliott, Desmond. 2018. Adversarial evaluation of multimodal machine translation. In Riloff, Ellen, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium, October-November. Association for Computational Linguistics.

Fernandes, Patrick, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In Zong, Chengqing, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online, August. Association for Computational Linguistics.

Futeral, Matthieu, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. 2023. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5394–5413, Toronto, Canada, July. Association for Computational Linguistics.

Futeral, Matthieu, Cordelia Schmid, Benoît Sagot, and Rachel Bawden. 2025. Towards zero-shot multimodal machine translation. In Chiruzzo, Luis, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 761–778, Albuquerque, New Mexico, April. Association for Computational Linguistics.

House, Juliane. 2006. Text and context in translation. *Journal of Pragmatics*, 38(3):338–358. Special Issue: Translation and Context.

Lannelongue, Loïc, Jason Grealey, and Michael Inouye. 2021. Green algorithms: Quantifying the carbon footprint of computation. *Advanced Science*, 8(12):2100707.

Liu, Danyang, Fanjie Kong, Xiaohang Sun, Dhruva Patil, Avijit Vajpayee, Zhu Liu, Vimal Bhat, and Najmeh Sadoughi. 2025. Detect, disambiguate, and translate: On-demand visual reasoning for multimodal machine translation with large vision-language models. In Chiruzzo, Luis, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1559–1570, Albuquerque, New Mexico, April. Association for Computational Linguistics.

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, Pennsylvania, USA, July. 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, Portugal, September. Association for Computational Linguistics.

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, November. Association for Computational Linguistics.

Shen, Huangjun, Liangying Shao, Wenbo Li, Zhibin Lan, Zhanyu Liu, and Jinsong Su. 2024. A survey on multi-modal machine translation: Tasks, methods and challenges.

Specia, Lucia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In Bojar, Ondřej, Christian Buck, Rajen Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Jörg Tiedemann, and Marco Turchi, editors, *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany, August. Association for Computational Linguistics.

Vijayan, Vipin, Braeden Bowen, Scott Grigsby, Timothy Anderson, and Jeremy Gwinnup. 2024. The case for evaluating multimodal translation models on text datasets.

Villa-Cueva, Emilio, Sholpan Bolatzhanova, Diana Turmakhan, Kareem Elzeky, Henok Biadglign Ademtew, Alham Fikri Aji, Israel Abebe Azime, Jinheon Baek, Frederico Belcavello, Fermin Cristobal, Jan Christian Blaise Cruz, Mary Dabre, Raj Dabre, Toqeer Ehsan, Naome A Etori, Fauzan Farooqui, Jiahui Geng, Guido Ivetta, Thanmay Jayakumar, Soyeong Jeong, Zheng Wei Lim, Aishik Mandal, Sofia Martinelli, Mihail Minkov Mihaylov, Daniil Orel, Aniket Pramanick, Sukannya Purkayastha, Israfel Salazar, Haiyue Song, Tiago Timponi Torrent, Debela Desalegn Yadeta, Injy Hamed, Atnafu Lambebo Tonja, and Thamar Solorio. 2025. Cammt: Benchmarking culturally aware multimodal machine translation.

Voita, Elena, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In Korhonen, Anna, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy, July. Association for Computational Linguistics.

A CoMMuTE Sample Item

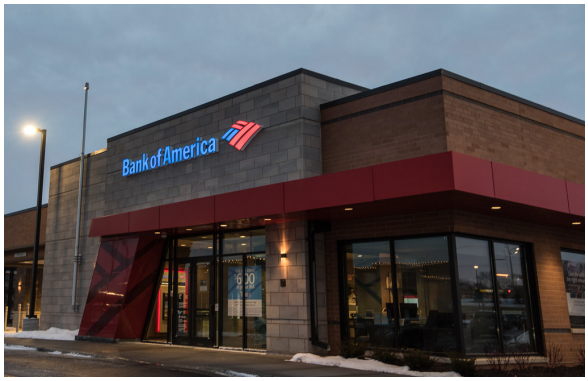


Figure 3: CoMMuTE sample item (Futeral et al., 2023). Source: *"He finally made it to the bank."* Image A (left): *rive* 'river bank'; Image B (right): *banque* 'financial bank'.

B CaMMT Sample Item

Figure 4: CaMMT sample item (Villa-Cueva et al., 2025). Substituted source: *"This dish is called a beet soup."* Conserved source: *"This dish is called borsch."* Russian reference: "Это блюдо называется борщ." The image provides the cultural context needed to recover the specific term борщ from the substituted source. Under the conserved source, the term is already present in the input; the image may help the model preserve it rather than paraphrase.

C Prompt Template

All models receive the following prompt. When an image condition applies, the image is provided as visual input alongside the text in a single request, using each model's native multimodal interface (chat-template processing for local models; base64 encoding for API models).

```
Translate the following English sentence into [LANGUAGE]: `[SENTENCE]'. Return your
translation as a JSON string as follows: {"translation": "<your translation>"}.
```

Output is parsed by extracting the `translation` field from the returned JSON. When a model does not produce valid JSON, the raw output is used as a fallback.

D Efficiency

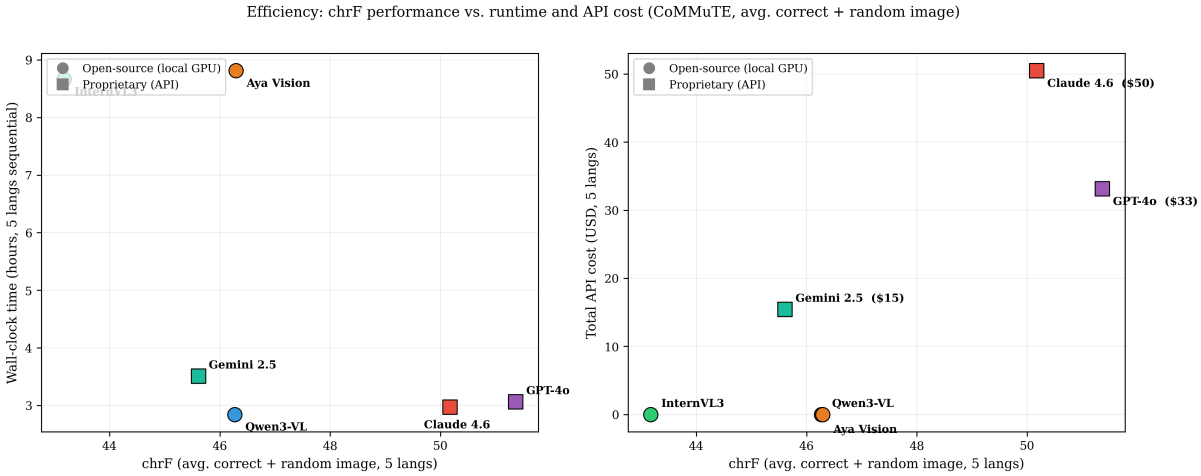


Figure 5: Efficiency trade-offs on CoMMuTE. The x-axis reports chrF averaged over the *correct-image* and *random-image* conditions (5 languages), reflecting realistic mixed-relevance usage. Left: wall-clock runtime. Right: total API cost (open-source models sit at $0). Squares denote proprietary (API) models; circles denote open-source (local GPU) models.

E CoMMuTE: BLEU Deltas

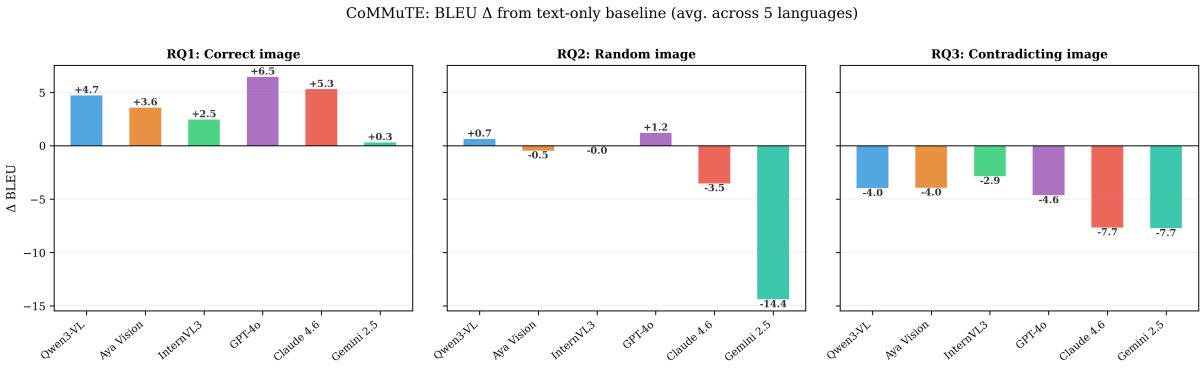


Figure 6: CoMMuTE: mean BLEU change from the text-only baseline, averaged across five languages (FR, DE, AR, RU, ZH). Layout mirrors Figure 1.

F CoMMuTE: Per-Language Breakdowns

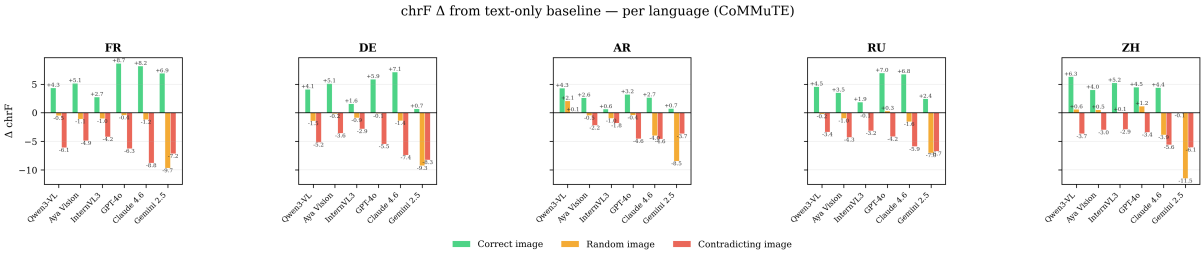


Figure 7: CoMMuTE: chrF deltas from text-only baseline, per language.

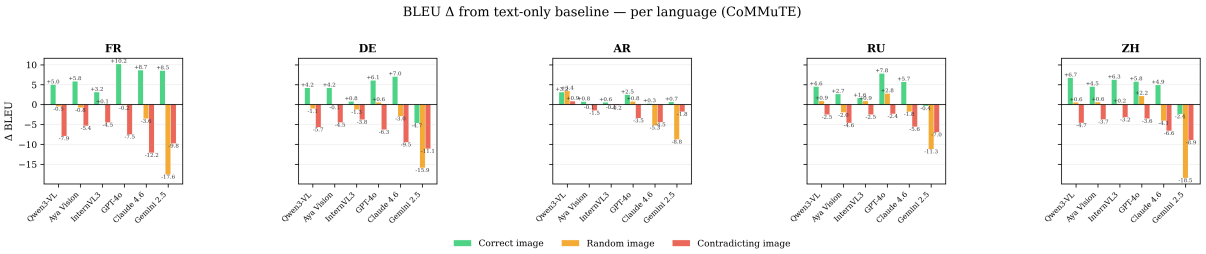


Figure 8: CoMMuTE: BLEU deltas from text-only baseline, per language.

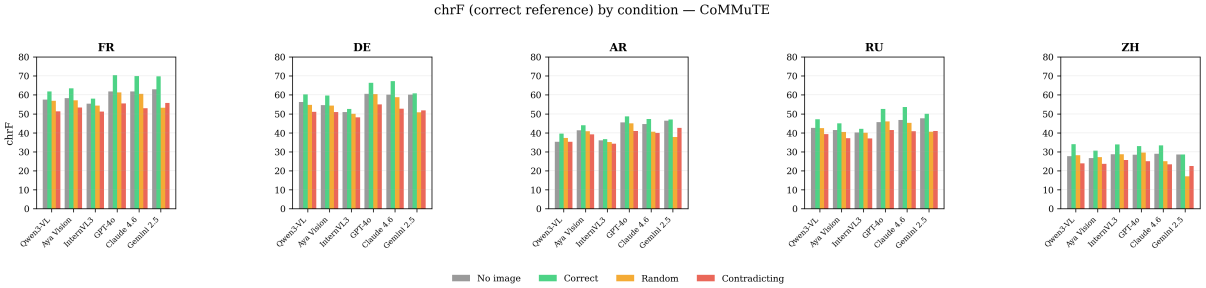


Figure 9: CoMMuTE: absolute chrF scores (correct reference) by condition and language.

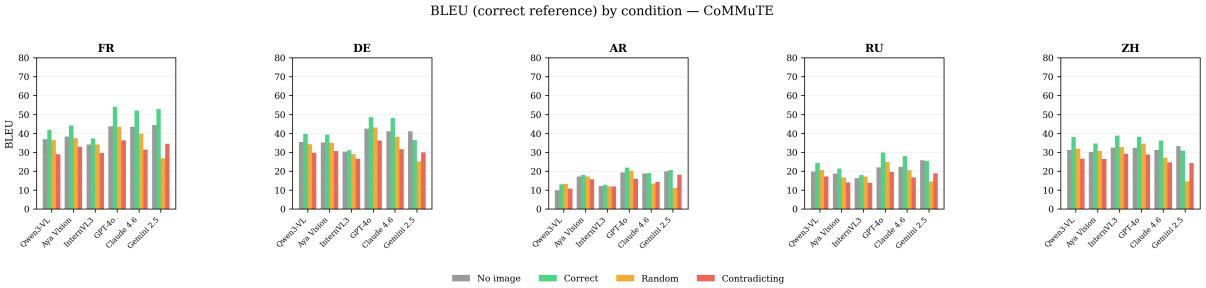


Figure 10: CoMMuTE: absolute BLEU scores (correct reference) by condition and language.

G CoMMuTE: Disambiguation Maps

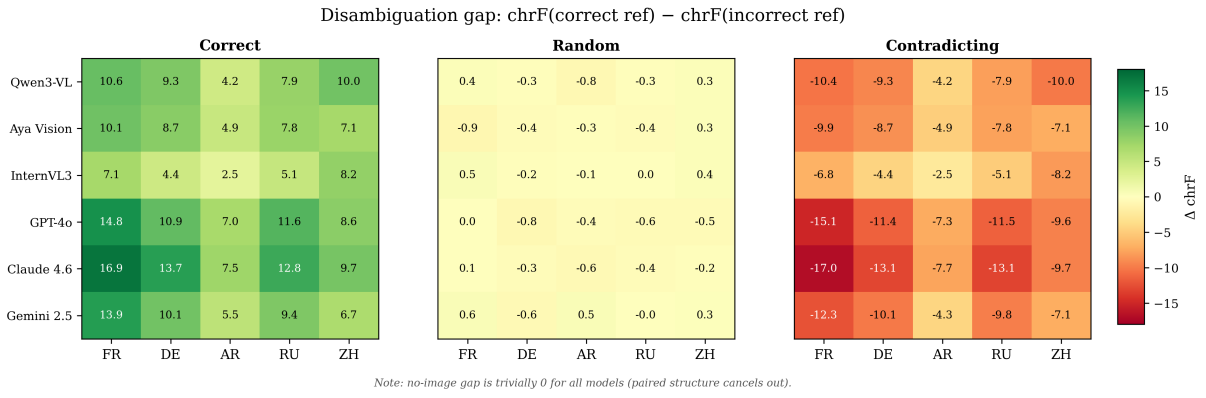


Figure 11: CoMMuTE disambiguation gap by visual condition and language. Positive values indicate the translation is closer to the intended meaning; negative values indicate it is closer to the wrong one.

H CaMMT and cross-task comparison: BLEU Deltas

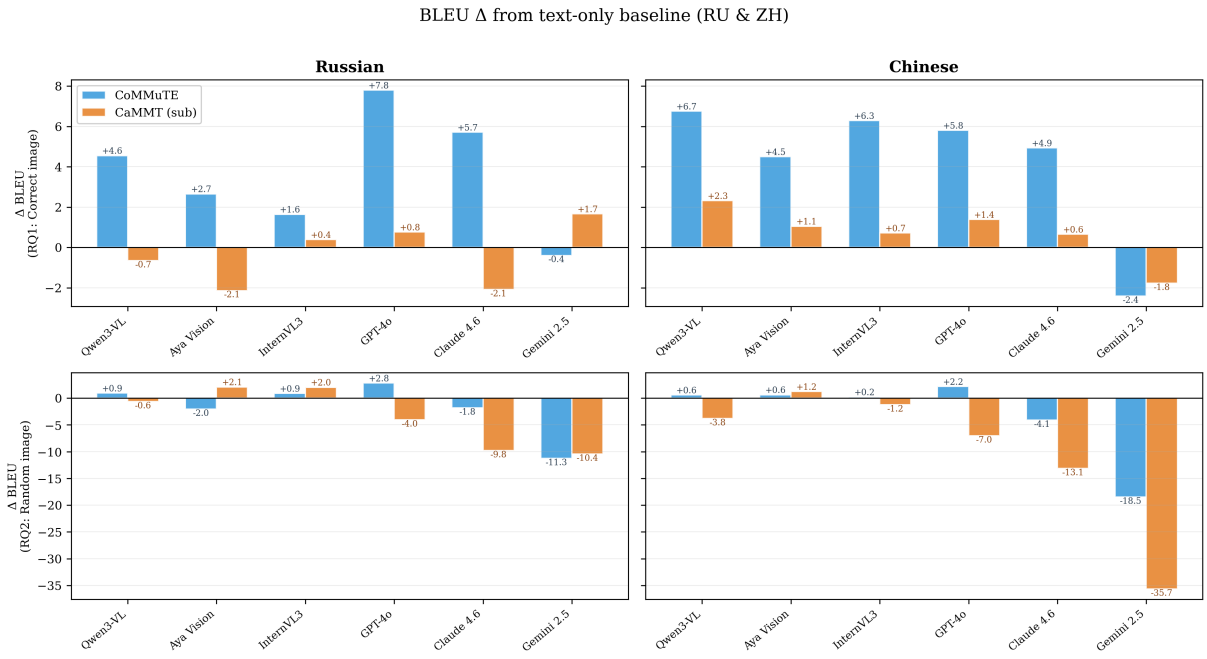
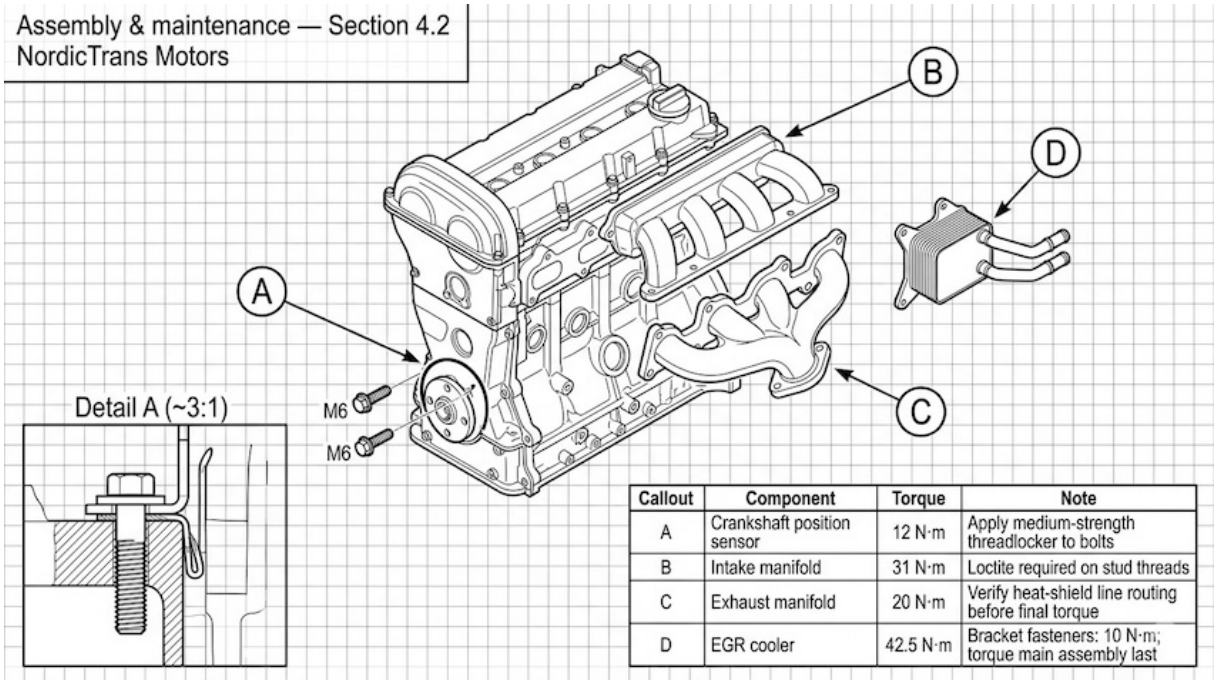


Figure 12: CaMMT and cross-task comparison: BLEU delta from the text-only baseline on Russian and Chinese.

I Synthetic *NordicTrans Motors* service illustration



Callout	Component	Torque	Note
A	Crankshaft position sensor	12 N·m	Apply medium-strength threadlocker to bolts
B	Intake manifold	31 N·m	Loctite required on stud threads
C	Exhaust manifold	20 N·m	Verify heat-shield line routing before final torque
D	EGR cooler	42.5 N·m	Bracket fasteners: 10 N·m; torque main assembly last

Figure 13: Synthetic image used as the shared visual for both the on-figure and off-figure English procedures in the practical assessment.

J Practical Assessment: Source, Hypotheses and Reference Translations

Condition	Source (EN)	Reference	Qwen3-VL-8B, no img	Qwen3-VL-8B, w/ img	GPT-4o, no img	GPT-4o, w/ img
Matched (FR)	Refer to Figure 4.2 and the Section 4.2 assembly table. Apply threadlocking compound to manifold studs only for the manifold callout with the higher torque value, as specified in the note for that row, and do not apply final tightening to the other manifold callout until the routing-related note for that row has been met.	Se reporter à la figure 4.2 et au tableau d'assemblage de la section 4.2. Appliquer un composé frein-filet sur les goujons du collecteur uniquement pour le repère de collecteur présentant la valeur de couple la plus élevée, conformément à la note figurant sur la ligne correspondante, et ne pas effectuer le serrage final sur l'autre repère de collecteur tant que la note relative au routage figurant sur la ligne de ce repère n'a pas été respectée.	Veuillez vous référer à la figure 4.2 et au tableau d'assemblage de la section 4.2. Appliquez uniquement un composé anti-décollement des filets aux boulons du collecteur correspondant à l'appel de collecteur avec la valeur de couple la plus élevée, tel que spécifié dans la note de cette ligne, et n'appliquez pas le serrage final aux autres appels de collecteur jusqu'à ce que la note liée au routage de cette ligne soit respectée.	Veuillez consulter la figure 4.2 et le tableau d'assemblage de la section 4.2. Appliquez uniquement un composé anti-décollement des filets aux boulons du collecteur pour le collecteur dont la valeur de couple est la plus élevée, conformément à la note spécifiée dans la ligne correspondante, et n'appliquez pas le serrage final au collecteur correspondant jusqu'à ce que la note relative au routage pour cette ligne soit respectée.	Consultez la Figure 4.2 et la table d'assemblage de la Section 4.2. Appliquez un composé de freinage de filetage uniquement aux goujons du collecteur pour l'appel de collecteur avec la valeur de couple plus élevée, comme spécifié dans la note pour cette rangée, et n'effectuez pas le serrage final de l'autre appel de collecteur jusqu'à ce que la note relative à l'acheminement pour cette rangée ait été respectée.	Consultez la figure 4.2 et le tableau d'assemblage de la section 4.2. Appliquez un **frein-filet** uniquement aux goujons du collecteur pour le **repère** du collecteur avec la valeur de couple la plus élevée, comme spécifié dans la note pour cette ligne, et n'effectuez pas le serrage final pour l'autre **repère** du collecteur tant que la note relative au **routage** pour cette ligne n'est pas respectée.
Matched (RU)	Refer to Figure 4.2 and the Section 4.2 assembly table. Apply threadlocking compound to manifold studs only for the manifold callout with the higher torque value, as specified in the note for that row, and do not apply final tightening to the other manifold callout until the routing-related note for that row has been met.	Обратитесь к рис. 4.2 и таблице сборки раздела 4.2. Нанесите фиксатор резьбы на шпильки крепления коллектора только для той позиции коллектора, которой в таблице соответствует большее значение момента затяжки, в соответствии с примечанием к строке этой позиции; не выполняйте окончательную затяжку для другой позиции коллектора, пока не будут выполнены требования примечания о прокладке линий, указанного для строки этой позиции.	См. Рисунок 4.2 и таблицу сборки раздела 4.2. Применяйте клей для **нитей** только к болтам **манжеты** с более высоким значением крутящего момента, как указано в примечании к этой строке, и не применяйте окончательную затяжку к другому **обозначению манжеты** до тех пор, пока не будет выполнено примечание, связанное с маршрутизацией, для этой строки.	См. рис. 4.2 и таблицу сборки раздела 4.2. Применяйте клей для **резьбы** только к болтам **головки блока цилиндров**, указанным в строке с наибольшим значением крутящего момента, как указано в примечании к этой строке, и не применяйте окончательную затяжку к другому **болту головки блока цилиндров**, пока не будет выполнено примечание, связанное с маршрутизацией, для этой строки.	Обратитесь к Рисунку 4.2 и таблице сборки Раздела 4.2. Нанесите фиксатор резьбы только на **шпильки** коллектора для той позиции коллектора, где указано более высокое значение крутящего момента, как указано в примечании для этой строки, и не выполняйте окончательное затягивание другой позиции коллектора, пока примечание относительно маршрута для этой строки не будет выполнено.	Обратитесь к Рисунку 4.2 и таблице сборки в Разделе 4.2. Наносите фиксатор резьбы на **штифты** коллектора только для обозначения коллектора с более высоким крутящим моментом, как указано в примечании для этой строки, и не производите окончательную затяжку для другого обозначения коллектора, пока не будет выполнено примечание, связанное с маршрутизацией, для этой строки.
Mispaired (FR)	If the telematics gateway logs J1939 DM1 frames indicating SCR system fault SPN 3364 while DEF tank level reads above 10%, perform a forced Dosing Air Pressure Test via the diagnostics menu before replacing the DEF pump assembly.	Si la passerelle télémétrique enregistre des trames DM1 J1939 signalant un défaut du système SCR SPN 3364 alors que le niveau du réservoir DEF reste supérieur à 10%, effectuez un test forcé de pression d'air de dosage via le menu de diagnostic avant de remplacer l'ensemble de la pompe DEF.	Si la passerelle de télématicie enregistre des cadres J1939 DM1 indiquant une défaillance du système SCR SPN 3364 pendant que le niveau du réservoir de DEF est supérieur à 10 %, effectuez un test obligatoire de pression d'air de dosage via le menu de diagnostic avant de remplacer l'ensemble de la pompe DEF.	Si la passerelle de **télémétrie** enregistre des cadres J1939 DM1 indiquant une défaillance du système SCR SPN 3364 pendant que le niveau du réservoir DEF est supérieur à 10 %, effectuez un test forcé de pression d'air de dosage via le menu de diagnostic avant de remplacer l'assemblage de pompe DEF.	Si la passerelle télématique enregistre des trames J1939 DM1 indiquant une défaillance du système SCR SPN 3364 alors que le niveau du réservoir de DEF est supérieur à 10 %, effectuez un test de pression d'air de dosage forcé via le menu de diagnostic avant de remplacer l'ensemble de la pompe DEF.	Si la passerelle télématique enregistre des trames DM1 J1939 indiquant un défaut du système SCR SPN 3364 alors que le niveau du réservoir DEF est supérieur à 10 %, effectuez un test de pression d'air de dosage forcé via le menu de diagnostic avant de remplacer l'ensemble de la pompe DEF.
Mispaired (RU)	If the telematics gateway logs J1939 DM1 frames indicating SCR system fault SPN 3364 while DEF tank level reads above 10%, perform a forced Dosing Air Pressure Test via the diagnostics menu before replacing the DEF pump assembly.	Если телематический шлюз фиксирует кадры DM1 J1939 с индикацией неисправности системы SCR SPN 3364 при уровне мочевины в баке выше 10%, выполните принудительную проверку давления воздуха дозирования через меню диагностики перед заменой узла насоса подачи мочевины.	Если телематический шлюз фиксирует кадры J1939 DM1, указывающие на неисправность системы SCR SPN 3364, при условии, что уровень топлива в баке DEF превышает 10%, выполните принудительный тест давления воздуха для дозирования через меню диагностики перед заменой агрегата насоса DEF.	Если шлюз телематики фиксирует кадры J1939 DM1, указывающие на неисправность системы SCR SPN 3364, при условии, что уровень топлива в баке DEF превышает 10%, выполните принудительный тест давления воздуха для дозирования через меню диагностики перед заменой агрегата насоса DEF.	Если телематический шлюз регистрирует кадры J1939 DM1, указывающие на неисправность системы SCR с SPN 3364, в то время как уровень DEF в баке превышает 10%, выполните принудительный тест давления воздуха для дозирования через диагностическое меню перед заменой узла насоса DEF.	Если телематический шлюз регистрирует кадры J1939 DM1, указывающие на неисправность системы SCR SPN 3364, при этом уровень в баке DEF показывает выше 10%, выполните принудительный тест давления воздуха дозирования через меню диагностики перед заменой узла насоса DEF.

Table 5: *NordicTrans Motors* practical stimuli: English source, synthetic human reference, and model outputs (Qwen3-VL-8B / GPT-4o) under **no-image** vs. **with-image** conditions. One row per scored text unit (matched vs. mispaired × FR / RU).

The MaTOS Pipeline for the Translation of Scientific Abstracts on the HAL Platform

Panagiotis Tsolakis[1] **Ziqian Peng**[1,2] **Laurent Romary**[1]
François Yvon[2] **Rachel Bawden**[1]
[1]Inria, Paris, France
[2]Sorbonne Université, CNRS, ISIR, Paris, France
firstname.lastname@inria.fr
lastname@isir.upmc.fr

Abstract

English dominates scientific publishing, which disadvantages researchers who are not native English speakers, especially those in the earlier stages of their careers. Being able to write and engage with scientific content written in their own language would clearly facilitate scientific production. The MaTOS project (Machine Translation for Open Science) seeks to reduce these barriers by developing machine translation tools for scientific documents in English and French. This article presents the design of the MaTOS pipeline for the HAL platform to automatically translate article abstracts, with author validation, to increase the number of bilingual abstracts on the platform. We also report preliminary experiments comparing translation of sentence, three-sentence chunks, and whole abstracts, evaluated using quality estimation metrics. We release all the code of the different stages of the pipeline.[1]

1 Introduction

A vast majority of academic publications are published in English (Gordin, 2015). While having a lingua franca has certain advantages, it also creates linguistic barriers for non-native speakers of English; they may feel less at ease both reading the literature and writing their own articles. In a world of ever-increasing publication, it is important to create a more inclusive research ecosystem that does not disadvantage non-native speakers and, on the contrary, embraces the diversity that comes with a multilingual community. Initiatives such as the *Helsinki Initiative on Multilingualism in Scholarly Communication*[2] recommend just that, encouraging the community to promote and facilitate writing in their own language, and for the Natural Language Processing (NLP) domain, the ACL 60-60 initiative[3] sought to break down linguistic barriers with their ambition to translate the ACL anthology into 60 different languages.

Machine translation (MT) is one way of both facilitating access to publications in English and enabling researchers to write in their own language. Progress in MT has been remarkable, thanks to neural architectures (Bahdanau et al., 2015), large-scale language models trained on large amounts of data (Lewis et al., 2020; Raffel et al., 2020), and, in more recent years, the use of large language models (LLMs) (Hendy et al., 2023; Zhu et al., 2024; Bawden and Yvon, 2023), which have provided new opportunities for longer document translation (beyond the sentence level) (Guo et al., 2025; O'Brien et al., 2025; Peng et al., 2025b) and the integration of domain-specific translation guidelines and terminologies (Lu et al., 2024a; Oncevay et al., 2025).

The MaTOS project[4] (Machine Translation for Open Science), a project funded by the French National Research agency, seeks to improve translation of academic documents between French and English (Bénard et al., 2023; Bawden et al., 2025). As part of the project, we have developed a pipeline to translate monolingual abstracts

[1]https://github.com/ANR-MaTOS/HAL_pipeline

[2]https://www.helsinki-initiative.org

[3]https://2022.aclweb.org/dispecialinitiative.html

[4]https://anr-matos.github.io

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 117–126

of scholarly publications submitted to the HAL platform from English into French and vice versa. The HAL platform is the French national scientific publication archive, hosting publications of a wide range of scientific domains. Its use is recommended, sometimes even mandatory, across all French public academic institutions.

As of February 2026, HAL hosts 4.5M references. Out of these, 1.7M references have a declared English abstract and no French abstract,[5] while 582k references have a declared French abstract and no English abstract. Through our translation pipeline, we aim to provide HAL users with an automatic translation of their monolingual abstracts in French or English, depending on the original source language, which the authors can then post-edit before it is associated with their publication. In this article, we present the design of the pipeline and also compare and analyse the translation quality of an LLM integrated into this pipeline when translating segments of varying lengths (isolated sentences, blocks of consecutive sentences and the whole abstract).

2 Related work

Traditionally, MT models have been trained to perform sentence-level translation, for computational reasons as well as the availability of large sentence-level parallel datasets. However, translating at the sentence level means that certain aspects of the translation cannot be ensured, notably those that span across several sentences or even whole documents such as lexical cohesion, which requires extra-sentential processing (Fernandes et al., 2023). Context-aware and document-level translation has therefore been an important area of study in MT from as far back as statistical MT (Hardmeier, 2012), and the task has been made more flexible with the use of neural MT models (Miculicich et al., 2018; Lopes et al., 2020) and nowadays with LLMs (Wang et al., 2023). The advent of LLMs that can handle large contexts, sometimes of thousands of tokens, has enabled the research community to envisage holistic document translation that can potentially ensure document-level consistency and coherence. Although theoretically whole-document translation would be ideal, it is less obvious that this is the optimal granularity for current models, particularly for longer documents. Despite the theoretical context window being big enough to accommodate the whole document and prompt, in practice it has been shown that additional problems can arise for the translation of longer documents, including duplicated segments, missing or hallucinated content, resulting in worse overall quality (O'Brien et al., 2025; Peng et al., 2025a). It has even been shown that sentences that appear far from the beginning of the input text are more affected by this decrease in translation quality than those at the beginning (Peng et al., 2025a). We therefore choose to test several granularities of translation in our pipeline to find the right balance. An additional advantage of LLMs for translation is that they can easily integrate additional information sources such as terminologies (Moslem et al., 2023; Lu et al., 2024a; Oncevay et al., 2025), style guidelines and domain information (Hu et al., 2024), as well as examples of the type of translation to be carried out (in the form of few-shot examples) (Tan et al., 2024) through prompting. We conduct experiments with both 0-shot and few-shot prompting for document-level translation.

A previous project had a similar aim to our own. The ANR COSMAT project (Lambert et al., 2012) aimed to translate entire PDF documents submitted to HAL, with the goal of having the authors post-edit, validate or reject the suggested translation. However, this pipeline was not integrated in the platform. Much has changed since, notably the MT technologies available for translation (and therefore the quality and usefulness of translation). Our ambition to translation abstracts initially rather than whole documents is partly due to it being more straightforward and more likely to be adopted by authors in this initial implementation of the pipeline, and abstracts are part of the publications' metadata and therefore belong to the public domain, making it possible to translate them for all articles, even those without permissive licences.

In a user-oriented setup, it is important to evaluate the quality of MT output before submitting it to the user for validation. While traditional MT evaluation scores like BLEU (Papineni et al., 2002) and chrF (Popović, 2015) require a reference translation (which we do not have in our setting, except in rare cases), it is possible to carry out reference-less evaluation, known as quality estimation (QE). Several neural QE metrics have been made available in the last years. COMET-

[5] HAL users may submit an abstract for their publication without specifying its language.

QE (Rei et al., 2021) and CometKiwi (Rei et al., 2022a), like their reference-based version, COMET (Rei et al., 2020), are based on a neural language model fine-tuned on human judgments of MT quality. We choose to use CometKiwi in this article as a measure of MT quality, and we also use it within the pipeline (alongside heuristic measures such as translation length ratio and language identification tools) to identify poor quality translations. More recently, Kocmi and Federmann (2023b) introduced GEMBA, an LLM-powered metric for assessing MT quality with or without a reference translation. GEMBA relies on zero-shot prompting of GPT-models to assess a translation and return a quality score. There have also been approaches to obtain fine-grained assessment of translation quality, including specific errors (Kocmi and Federmann, 2023a; Lu et al., 2024b). The one we use in this article for the assessment of MT quality is GEMBA-MQM (Kocmi and Federmann, 2023a), which emulates the multidimensional quality metrics (MQM) framework (Lommel et al., 2013; Freitag et al., 2021).

3 The MaTOS pipeline

The MaTOS pipeline involves the extraction and translation of abstracts from HAL and submitting the translated abstracts to the authors for validation and/or post-editing. Abstracts belong to the metadata of articles, and therefore we are able to process all submitted abstracts in compliance with the intellectual property code.[6] While it would be interesting in the future to work on the translation of whole articles, we believe that this initial first step (i.e. translating abstracts) is important.

As shown in Figure 1, the pipeline is composed of five steps, which are described in more detail below: (i) Extraction of publication metadata, (ii) language identification and filtering, (iii) translation using an LLM, (iv) filtering of the abstracts based on the translation to avoid insufficiently correct translations, and finally (v) submission of the translation to the authors for validation.

3.1 Extraction of submissions from HAL

This first step consists in extracting metadata of publications submitted to HAL. We are planning to initially deploy the pipeline only for publications submitted to the Inria portal of HAL (in computer science). However, in this paper we present results both from Inria portal publications (mostly computer science-related) and publications from all of HAL (from different institutions and scientific fields). We download the submissions on a daily basis using the HAL API.[7] Extracted metadata fields include the document ID, title, keywords, abstracts, author IDs and names, the day of submission, and the scientific field.[8] All kinds of publications are taken into account, including articles, theses, reports, images, videos, etc.

3.2 Language identification and source-side filtering/processing

This step has two aims: (i) identify English and French abstracts that are to be translated by the pipeline (i.e. those for which a translation does not already exist), and, as a by-product of the pipeline (ii) identify abstracts for which the translation is provided by the author(s), with a view to creating a parallel corpus of scientific publication abstracts.

The metadata fields specify the language of the abstract (i.e. English or French), but we carry out language identification to verify that the declared language matches the actual language of the abstract. We use an off-the-shelf fastText model (Joulin et al., 2016; Joulin et al., 2017) without a confidence threshold to predict the language. Any abstracts for which the predicted language does not match are excluded. We also exclude abstracts that contain fewer than 15 words or that end in an ellipsis (indicating an incomplete abstract).

We also collapse multiple spaces, replace subscripts and superscripts by their Unicode equivalents and remove HTML tags which are not part of the abstract content and posed problems for MT in preliminary experiments.

3.3 Machine Translation

We choose to use EuroLLM-9B-Instruct (Martins et al., 2025) for translation, as preliminary experiments showed the results to be good whilst minimising computational resources required. The model is an instruction-following model based on the 9B-parameter EuroLLM model. It has the advantage of being open-weight and trained on publicly available data in 35 languages, including parallel corpus to increase its translation abilities.

[6] https://doc.hal.science/en/legal-aspects/

[7] https://api.archives-ouvertes.fr/docs

[8] Although we do not currently use all extracted fields to help translation (e.g. keywords and the scientific field), we believe this could be useful information to include in the future.

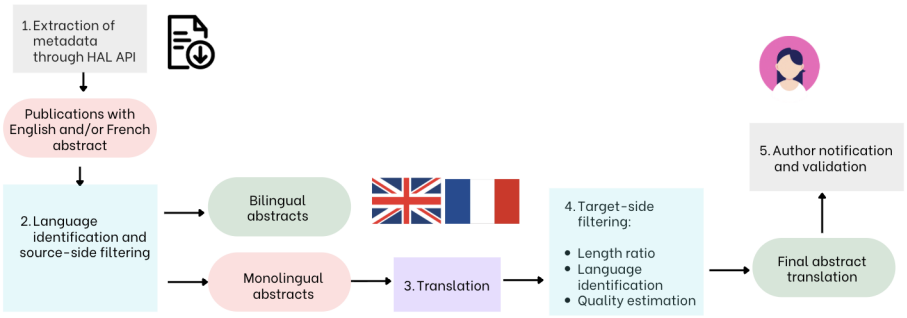


Figure 1: Illustration of the stages of the MaTOS pipeline.

We batch the translation requests to accelerate inference, and translate using vLLM (Kwon et al., 2023). The average time for processing a single request on an NVIDIA A100 GPU is 5.95 seconds. We only translate those abstracts for which the prompt and source sequence do not exceed the model's maximum context length of 4096 tokens.

In the experiments presented in this article, we test zero-shot, one-shot and two-shot MT. A zero-shot prompt example is displayed in Figure 2. We take few-shot examples from the ACAD-TRAIN and ACAD-BENCH subsets of the ACADATA parallel dataset of academic abstract (Lacunza et al., 2026) and use BM25 (Lù, 2024) to select the most similar texts, after all texts are lemmatised with SpaCy (Honnibal et al., 2020).

```
<|im_start|>system
You are a professional translator of scientific
documents. Translate the following text from
English into French. The target text must have the
same number of paragraphs as the source text.
Reply only with the translated text.
<|im_end|>

<|im_start|>user
English: Geological data show that, early in its
history, the Earth had a large-scale magnetic
field with an amplitude comparable to the one of
the present geomagnetic field.
French:
<|im_end|>

<|im_start|>assistant
```

Figure 2: Example prompt (English-French, sentence-level).

For the translation of scientific abstracts, which can be considered to be small documents, the size of the segment to be translated is important for the overall consistency of the output. In Section 4, we compare different translation granularities (isolated sentences, chunks of 3 sentences and whole abstracts), to test the trade-off between maximal access to context to aid translation and minimising potential deterioration in quality linked to having to translate longer sequences. As shown in Section 4, we find that document-level translation leads to fewer errors overall, notably concerning terminology and style. We therefore choose to translate at the document-level, but still envisage to split very long documents into smaller chunks to avoid processing issues and the degraded quality that can come with MT of very long texts.

3.4 Target-side filtering

It is important to ensure that the translated abstracts submitted to the authors are of sufficiently good quality to maximise the chances that the authors will accept to validate (or post-edit) the abstracts and to avoid any potential disillusion with the tools provided. We therefore apply the following filters on the output: (i) language identification to check that the output is in the expected language, (ii) verification of the length ratios between the source and target texts (rejecting any that differ by more than 35%)[9] and finally (iii) carrying out QE using CometKiwi (Rei et al., 2022b).

3.5 Author notification and validation

Only the translations of the abstracts that pass the previously described filters are submitted to authors. It is important for authors to maintain control over the information associated with their article deposits. We therefore implement a mechanism by which authors are notified of the translation of their abstract and are encouraged to validate and post-edit it, after which the translation will be added to the metadata of their article.

[9]This is based on the length ratios of the bilingual abstracts extracted from HAL in February 2026; French abstracts are 29% longer than their English translations, while English abstracts are 23% shorter than their French translations.

4 Experimenting with the translation segment size

As described above, we attempt to find the optimal granularity of the translation segment. In theory, more context should be useful (i.e. translating abstracts as a whole). However, in practice, it has been shown that this can have adverse effects, especially with longer texts, such as hallucinations and deleted sentences (Karpinska and Iyyer, 2023).

We compare three levels of granularity for translation: sentence-level, chunks of three sentences and whole abstracts. For the first two methods, we split abstracts into individual sentences using Trankit (Nguyen et al., 2021). 3-sentence chunks are the concatenation of three sentences, although there may be fewer than three sentences at the end of an abstract. For the document level, we also experiment with one-shot and two-shot prompting.

4.1 Data and evaluation

We experiment with two datasets collected from HAL and filtered as described in Section 3.4. The HAL-Inria corpus contains the first 500 abstracts submitted to the Inria collection of HAL starting from 1st February 2026,[10] while the HAL-all corpus contains the first 500 abstracts submitted to HAL (every discipline combined) starting from 23rd February 2026. Detailed statistics about the evaluation data can be found in Table 1.

	#docs	#sents	#aligned segs	#toks
HAL-Inria				
en–fr	500	3,775	3,429	121,692
fr–en	45	299	287	10,694
HAL-all				
en–fr	500	5,105	4,701	180,339
fr–en	500	2,831	2,479	120,210

Table 1: Statistics for the evaluation data.

For evaluation, we calculate QE scores with CometKiwi and with GEMBA-MQM to provide a finer-grained analysis of translation errors. We apply GEMBA-MQM to the complete document translations, i.e. where the sentence-level and 3-sentence-level translations are concatenated to form documents. However, CometKiwi's ability to faithfully evaluate longer text segments is subject to discussion, as the model is trained mainly on sentence-level judgments (Dahan et al., 2026). Therefore, we decide to apply it to shorter (sentence-like) segments. The same sentence alignments are not guaranteed for all granularities, so we create segments by aligning input and output sentences for each granularity using Bertalign (Liu and Zhu, 2022) and calculate the alignment that is compatible with all three, concatenating sentences in the case of many-to-1 alignments. We then apply CometKiwi to each aligned input-output pair.

[10]There are way fewer monolingual French abstracts in the INRIA collection, therewefore we only gathered 45 abstracts, submitted in all of February and March 2026

4.2 Quality estimation results and analysis

QE results can be found in Table 2 for HAL-all and in Table 3 for HAL-Inria.

Segment	#shots	CometKiwi	GEMBA-MQM	LR
		en–fr		
Sent	0	85.71	-6.29	1.419
3-sents	0	85.65	-4.87	1.370
Doc	0	85.24	-3.87	1.356
Doc	1	85.22	-3.98	1.349
Doc	2	85.05	-4.63	1.341
		fr–en		
Sent	0	83.31	-6.51	0.847
3-sents	0	83.39	-4.86	0.828
Doc	0	82.92	-4.44	0.822
Doc	1	82.74	-3.66	0.817
Doc	2	82.16	-4.15	0.816

Table 2: Comparison of MT performance for the HAL-all corpus. LR indicates the length ratio.

Segment	#shots	CometKiwi	GEMBA-MQM	LR
		en–fr		
Sent	0	84.48	-6.82	1.427
3-sents	0	84.34	-5.09	1.385
Doc	0	84.10	-5.09	1.380
Doc	1	83.87	-5.01	1.374
Doc	2	83.32	-6.07	1.366
		fr–en		
Sent	0	85.04	-6.93	0.797
3-sents	0	85.00	-5.68	0.798
Doc	0	84.73	-3.40	0.794
Doc	1	84.80	-3.84	0.794
Doc	2	84.85	-4.71	0.792

Table 3: Comparison of MT performance for the HAL-Inria corpus. LR indicates the length ratio.

CometKiwi and GEMBA-MQM CometKiwi scores are similar for all segment sizes. For both language pairs, the sentence-level and 3-sentence-level scores are almost identical and higher than document-level scores.

These results are not mirrored exactly by GEMBA-MQM scores, as the translation where

	Sent.	3-sents	Doc 0-shot	Doc 1-shot	Doc 2-shot
Acc./Mistranslation	4 (5)	8	4 (5)	4 (6)	2
Acc./Omission	0	2	0	1	0
Flu./Grammar	5 (6)	5 (4)	4	4	2 (3)
Flu./Punctuation	3 (4)	4	0	3	4
Term./Inappropriate for context	2	2	1	1	2
Term./Inconsistent use	2	4 (5)	2	1	0
Style/Awkward	2	3 (4)	3	2	1

Table 4: Number of occurrences of each GEMBA-MQM error type across different translation segment sizes. Macro error types are acc(uracy), flu(ency) and term(inology).

the text was translated sentence by sentence is scored lower than the other two granularities. This is maybe unsurprising, as we applied GEMBA-MQM to whole abstracts, while CometKiwi was applied to sentence-like segments, therefore possibly missing important errors concerning consistency and coherence. In our pipeline, we cannot use GEMBA-MQM for cost and reproducibility reasons (Kocmi and Federmann, 2023a); we therefore use CometKiwi, which is nevertheless sensitive to the most severe problems.

We also find that few-shot examples tend to decrease CometKiwi scores. One-shot prompting can have a light positive or negative impact on GEMBA-MQM scores, but two-shot prompting always deteriorates scores except for the fr–en direction for the HAL-Inria corpus.

Analysis based on GEMBA-MQM Gemba-MQM allows us to perform finer-grained evaluation, offering insights into translation errors in addition to providing QE scores. The detected errors are classified into three categories: critical, major and minor, and each category is weighted differently in the calculation of the final quality score.

We perform qualitative analysis by reviewing the translations and the Gemba-MQM error annotations for the first 10 en–fr abstracts of the HAL-all corpus. Table 4 shows the number of occurrences of the most important error types. We find that some errors were misclassified or hallucinated by Gemba-MQM and therefore indicate the true count of errors in parentheses after our re-annotation. We present some misclassified and false errors in Appendices A and B respectively.

The most prevalent error type is mistranslation, reflecting the difficulty for LLMs to translate scientific terms accurately. Inconsistent translation of terms becomes more common for smaller translation segments (i.e. worse at the 3-sents level), as consistent use of terminology often requires context beyond the sentence or chunk level (Xiao et al., 2011). We also find some basic grammatical errors. They mainly concern subject-verb agreement and gender agreement. We provide below two examples of errors annotated by GEMBA-MQM. In the first example, the English source term *bio-oil*, is inconsistently translated into French as *bio-huile* and *huile biologique*. In the second example, the French word *dimensionnement* 'sizing' is preceded by a feminine determiner, even though it is a masculine noun.

(1) terminology/inconsistent use: "bio-huile" vs "huile biologique" (both acceptable but inconsistent within the text)

(2) fluency/grammar: la dimensionnement optimal" should be "le dimensionnement optimal" (gender agreement error: "dimensionnement" is masculine)

Based on the above findings, we have decided to implement document-level MT in our pipeline. We will continue experimenting with few-shot prompting and different strategies for example selection.

5 Conclusion and future work

We have presented the pipeline of the MaTOS project for the translation of English and French publication abstracts submitted to the HAL publishing platform. Our aim is to break down language barriers in research (where English dominates) by encouraging publication in other languages, in our case French. The design of the pipeline takes into account the need to filter submitted abstracts and the resulting translations in order not to discourage users from using the technology. The idea is for users to validate or post-edit the translations, which will also be publicly available as a useful resource for the MT community.

In the near future, we intend to experiment with more sophisticated prompting techniques, integrating translation of terms using bilingual glossaries and other domain-specific information.

Acknowledgements

This work was supported by the French national research agency (ANR) as part of the MaTOS project (grant number: ANR-22-CE23-0033).[11] Rachel Bawden was also partly funded by her chair position in the PRAIRIE institute funded by ANR as part of the "Investissements d'avenir" programme under reference ANR19-P3IA-0001. The authors are grateful to the anonymous reviewers for their insightful comments and suggestions.

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the first International Conference on Learning Representations*, San Diego, CA.

Bawden, Rachel and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. 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 157–170, Tampere, Finland, June. European Association for Machine Translation.

Bawden, Rachel, Maud Bénard, Éric de la Clergerie, José Cornejo Cárcamo, Nicolas Dahan, Manon Delorme, Mathilde Huguin, Natalie Kübler, Paul Lerner, Alexandra Mestivier, Joachim Minder, Jean-François Nominé, Ziqian Peng, Laurent Romary, Panagiotis Tsolakis, Lichao Zhu, and François Yvon. 2025. MaTOS: Machine Translation for Open Science. In Bouillon, Pierrette, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Samuel Läubli, Martin Volk, Miquel Esplà-Gomis, Vincent Vandeghinste, Helena Moniz, and Sara Szoc, editors, *Proceedings of Machine Translation Summit XX: Volume 2*, pages 103–104, Geneva, Switzerland, June. European Association for Machine Translation.

Bénard, Maud, Alexandra Mestivier, Natalie Kubler, Lichao Zhu, Rachel Bawden, Eric De La Clergerie, Laurent Romary, Mathilde Huguin, Jean-François Nominé, Ziqian Peng, and François Yvon. 2023. MaTOS: Traduction automatique pour la science ouverte. In Boudin, Florian, Béatrice Daille, Richard Dufour, Oumaima El, Maël Houbre, Léane Jourdan, and Nihel Kooli, editors, *Actes de CORIA-TALN 2023. Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023*, pages 8–15, Paris, France, 6. ATALA.

Dahan, Nicolas, Rachel Bawden, and François Yvon. 2026. MetaDocEval: A Contrastive Framework for Evaluating Machine Translation Metrics at the Document-Level. In *Proceedings of the 26th Annual Conference of the European Association for Machine Translation*, Tilburg, the Netherlands, June. European Association for Machine Translation (EAMT).

Fernandes, Patrick, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When Does Translation Require Context? A Data-driven, Multilingual Exploration. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada, July. Association for Computational Linguistics.

Freitag, Markus, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Gordin, Michael D. 2015. *Scientific Babel*. University of Chicago Press.

Guo, Jiaxin, Yuanchang Luo, Daimeng Wei, Ling Zhang, Zongyao Li, Hengchao Shang, Zhiqiang Rao, Shaojun Li, Jinlong Yang, Zhanglin Wu, and Hao Yang. 2025. Doc-Guided Sent2Sent++: A Sent2Sent++ Agent with Doc-Guided memory for document-level Machine Translation. arXiv Preprint:2501.08523 [cs], January.

Hardmeier, Christian. 2012. Discourse in Statistical Machine Translation: A Survey and a Case Study. *Discours - Revue de linguistique, psycholinguistique et informatique*, (11).

Hendy, Amr, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? A comprehensive evaluation. *CoRR*, abs/2302.09210.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Hu, Tianxiang, Pei Zhang, Baosong Yang, Jun Xie, Derek F. Wong, and Rui Wang. 2024. Large language model for multi-domain translation: Benchmarking and domain CoT fine-tuning. In Al-Onaizan, Yaser, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5726–5746, Miami, Florida, USA, November. Association for Computational Linguistics.

[11] http://anr-matos.github.io/

Joulin, Armand, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomás Mikolov. 2016. Fasttext.zip: Compressing text classification models. *CoRR*, abs/1612.03651.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In Lapata, Mirella, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April. Association for Computational Linguistics.

Karpinska, Marzena and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore, December. Association for Computational Linguistics.

Kocmi, Tom and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore, December. Association for Computational Linguistics.

Kocmi, Tom and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. 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 193–203, Tampere, Finland, June. European Association for Machine Translation.

Kwon, Woosuk, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Lacunza, Iñaki, Javier Garcia Gilabert, Francesca De Luca Fornaciari, Javier Aula-Blasco, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. 2026. ACAData: Parallel dataset of academic data for machine translation. In Piperidis, Stelios, Núria Bel, Henk van den Heuvel, Nancy Ide, Simon Krek, and Antonio Toral, editors, *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 8498–8519, Palma, Mallorca, Spain, May. European Language Resources Association (ELRA).

Lambert, Patrik, Jean Senellart, Laurent Romary, Holger Schwenk, Florian Zipser, Patrice Lopez, and Frédéric Blain. 2012. Collaborative machine translation service for scientific texts. In Segond, Frédérique, editor, *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 11–15, Avignon, France, April. Association for Computational Linguistics.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Liu, Lei and Min Zhu. 2022. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634, 12.

Lommel, Arle Richard, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK, November 28-29. Aslib.

Lopes, António, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In Martins, André, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal, November. European Association for Machine Translation.

Lu, Hongyuan, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024a. Chain-of-dictionary prompting elicits translation in large language models. In Al-Onaizan, Yaser, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, Florida, USA, November. Association for Computational Linguistics.

Lu, Qingyu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024b. Error analysis prompting enables human-like translation evaluation in large language models. In Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816, Bangkok, Thailand, August. Association for Computational Linguistics.

Lù, Xing Han. 2024. BM25S: orders of magnitude faster lexical search via eager sparse scoring.

Martins, Pedro Henrique, João Alves, Patrick Fernandes, Nuno Miguel Guerreiro, Ricardo Rei, M. Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report. *CoRR*, abs/2506.04079.

Miculicich, Lesly, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In Riloff, Ellen, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium, October-November. Association for Computational Linguistics.

Moslem, Yasmin, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. 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 227–237, Tampere, Finland, June. European Association for Machine Translation.

Nguyen, Minh Van, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In Gkatzia, Dimitra and Djamé Seddah, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online, April. Association for Computational Linguistics.

O'Brien, Dayyán, Bhavitvya Malik, Ona de Gibert, Pinzhen Chen, Barry Haddow, and Jörg Tiedemann. 2025. DocHPLT: A massively multilingual document-level translation dataset. In *Proceedings of the Tenth Conference on Machine Translation*, pages 286–300, Stroudsburg, PA, USA, September. Association for Computational Linguistics.

Oncevay, Arturo, Charese Smiley, and Xiaomo Liu. 2025. The impact of domain-specific terminology on machine translation for finance in European languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2758–2775, Stroudsburg, PA, USA. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Peng, Ziqian, Rachel Bawden, and François Yvon. 2025a. Investigating length issues in document-level machine translation. In Bouillon, Pierrette, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Ana C. Farinha, Marco Gaido, Joke Daems, Dorothy Kenny, Helena Moniz, and Sara Szoc, editors, *Proceedings of Machine Translation Summit XX: Volume 1*, pages 4–23, Geneva, Switzerland, June. European Association for Machine Translation.

Peng, Ziqian, Rachel Bawden, and François Yvon. 2025b. Self-retrieval from distant contexts for document-level machine translation. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 220–240, Suzhou, China, November. Association for Computational Linguistics.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

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).

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, November. Association for Computational Linguistics.

Rei, Ricardo, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are References Really Needed? Unbabel-IST 2021 Submission for the Metrics Shared Task. In Barrault, Loic, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online, November. Association for Computational Linguistics.

Rei, Ricardo, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022a. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Rei, Ricardo, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Tan, Weiting, Haoran Xu, Lingfeng Shen, Shuyue Stella Li, Kenton Murray, Philipp Koehn, Benjamin Van Durme, and Yunmo Chen. 2024. Narrowing the gap between zero- and few-shot machine translation by matching styles. In Duh, Kevin, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 490–502, Mexico City, Mexico, June. Association for Computational Linguistics.

Wang, Longyue, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore, December. Association for Computational Linguistics.

Xiao, Tong, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level consistency verification in machine translation. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China, September 19-23.

Zhu, Wenhao, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In Duh, Kevin, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico, June. Association for Computational Linguistics.

A Misclassified errors

In this section we show some examples of errors that were detected by GEMBA-MQM in our evaluation sample but were classified in the wrong category.

(3) fluency/grammar: some minor punctuation issues in the abbreviation list (missing commas or semicolons in a few places)

This was annotated as a *fluency/grammar* error, but we re-annotated is as a *fluency/punctuation* error, since it concerns missing punctuation marks.

(4) accuracy/mistranslation: "la hybridation" should be "l'hybridation" (missing elision of the article "la" before a vowel)

This was annotated by GEMBA-MQM as an *accuracy/mistranslation* error, even though it is a grammar error.

B False errors

In this section we show some examples of errors that were detected by GEMBA-MQM in our evaluation sample but are actually correct translations.

(5) terminology: the terms "sphère biogéographique" for "biogeographic realm," "dissemblance" for "dissimilarity" and other scientific terms are consistent and appropriate.

This was counted as a terminology error, even though the model itself explains that scientific terms are translated in a consistent and appropriate manner.

(6) style: the translation is formal and appropriate for the scientific context.

This was counted as a style error, even though the model itself explains that the translation is formal and appropriate.

Automated Information Extraction and Template Filling from Client Style Guides

Leonor Graça
TransPerfect, Lisbon, Portugal
leonor.graca.int@unbabel.com

Vera Cabarrão
TransPerfect, Lisbon, Portugal
vera.cabarrao@unbabel.com

Helena Moniz
University of Lisbon, Portugal
CLUL, Lisbon, Portugal
INESC-ID, Lisbon, Portugal
helena.moniz@edu.ulisboa.pt

Abstract

Style guides are a centrepiece of professional translation workflows. Yet, their integration into automatic pipelines remains underexplored. This paper presents exploratory work on information extraction from client style guides and application to a templated style guide, developed to be a system prompt. This template is then applied during an LLM-based translation to automatically produce outputs that are compliant to client's requirements. The study focused on seven language pairs (LP), evaluating the automatic extraction, and translation quality and compliance with the style guide. The extraction demonstrated reliable performance across languages and file formats. Translation quality and adherence were evaluated using human preference annotation, comparing two Tower models (Tower Zen 9B and Tower+ 72B). The results indicate a modest advantage for Tower+, but with mutual acceptability in certain instances. These findings establish a viable semi-automatic framework for style guide integration in translation workflows, and motivate further investigation across broader domains, clients, and LPs.

1 Introduction

Language is never neutral. Every decision in a translation task carries meaning, even more so in a professional translation setting. Inconsistent usage of terminology can erode brand trust. An ambiguous tone can lead to miscommunications between the customer and the seller. Improper formatting of legal sections could lead to liability issues. The absence of clear rules, beyond language guidelines, forces the translator to make decisions without context, leading to inconsistent results.

Style guides exist to prevent exactly that. They are often described as reference documents, but that quite undersells them. Style guides are one of the most crucial tools in a translation company. When looking at different perspectives, from translators themselves to language providers (American Translators Association, 2019; Lingualinx, 2024; Ghosal, 2024), they all agree that without clear guidelines that determine tone, grammar, terminology, formatting, and audience definition, translators are left to make calls that should already have been resolved, leading to inconsistent messaging, costly revisions, and even legal consequences. ISO (2015) proves this point, creating a standard for translation processes. It is precisely the presence of guidelines that eliminate the guesswork by giving every linguist, reviewer, translator, etc., the same rulebook to work from.

However, for all their importance, they come with one significant challenge. As they are provided by the client, no two are alike, arriving with different formats, lengths, structures, and types of information. Some run for dozens of pages with detailed examples, while others are a page of bullet points. This inconsistency is difficult enough to handle with human translators. But, as the industry turns to Large Language Models (LLM), this becomes more acute. As powerful as they might be, LLMs tend to struggle in instruction-following (Heo et al., 2025; Young et al., 2025), even more so when parsing these style guides, having one of

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 127–132

the most valuable assets in the translation workflow becoming one of its biggest blind spots.

To circle around this, the research aims to develop an automated framework for generating machine-readable style guides that can be integrated directly into LLM-assisted translation workflows, enhancing translation quality through real-time application of client-specific style guides as system prompts while simultaneously improving operational efficiency by automating the style guide implementation process across diverse client requirements and linguistic specifications.

2 Literature Review

A translation style guide is a document that incorporates relevant information for translation tasks. Through these, translators can ensure consistency in format and language throughout all brand documentation (Aranbure et al., 2018).

However, a different challenge arises. Recent industry data underlines the rapid transformation toward AI-assisted workflows. The Business Research Company (2026) projects a Compound Annual Growth Rate (CAGR) of 25.2% for AI in translation, forecasting market expansion to $2.94 billion in 2025. Furthermore, CSA Research reports a $24.6 billion decline in the translation market since 2019, representing a 45% drop, suggesting industry restructuring alongside technological adoption. The question is no longer whether to adopt them, but rather how to integrate them effectively into established workflows.

Despite their capabilities, LLMs still exhibit notable limitations, particularly in adhering to instructions. Research has proven that LLMs struggle when processing complex and large guidelines (Jang et al., 2022; Jaroslawicz et al., 2025), suggesting that directly inputting extensive client style guides would likely result in unsatisfactory results. Furthermore, these style guides are developed for the human mind, and do not follow prompt engineering strategies, begging the question, how can we make LLMs follow style guides?

With that, a primary work was built upon this. A study conducted by the author (Graça, 2025), prior to this, presented initial research on how to properly develop style guides that are readable by an LLM, through its application as a system prompt. This was done with the development of a three-step methodology: a General Template — a one-fits-all approach, any document from any domain could be translated using it; a Domain Template — a style guide tailored for one specific domain in which any document within it could be translated using it; and a Document Template — one style guide for each type of document, within a domain.

The results showed its effectiveness in improving translation quality and compliance, proposing the Domain Template for simpler documents and the Document Template for more complex ones as the best approach. However, this was tailored to be more manual, as all the style guides are filled-in by a human before application. Additionally, it misses the specifications that certain clients have. Domains have generic rules. However, certain clients may wish to follow different parameters.

As such, the fourth-step of this methodology was thought out, the Client Template, to fit the needed industry use case, and tailor this system prompt approach to client specifications.

3 Methodology

Building upon Graça (2025), this research added to the framework by creating a fourth-step. This methodology brought two additions to the previous work, automating the process through a script integrated with the GPT API, while making it adaptable to the needs of individual clients. Data was extracted from the client's style guide and applied to a template structure, automatically creating an LLM-readable style guide that follows the clients needs and requirements. This template was later applied as a system prompt during translation, through TowerLLM.

3.1 Information Extraction and Application

3.1.1 Template

The previous work mentioned in Section 2 led to the development of AI-readable style guides and, subsequently, to its templated structure. As the study intends to be generally applicable to all clients, the initial step consisted of an adaptation of the General Template for this use-case.

However, this created a new challenge. The previous templates were filled-in by a linguist and/or translator. But, as the purpose was to automatise the first step, it was necessary to turn the template into a prompt to be adapted by the model. As such, the template itself was adapted through prompt engineering, adding instructions to the relevant sections, e.g. prompt *"add the relevant rules in the style guide for formal/informal usage and tone of*

voice." added to the section **Tone and Formality**. In this way, the model should incorporate the rules regarding formality and tone of voice in this section while employing the extracted data.

In order to create an appropriate template, it was necessary to analyse different client style guides so as to identify key elements. An aspect to consider was the importance of a section on one style guide, that was irrelevant for another. A marketing client may require the section **Humor**, while a legal client would not. Not all categories are intended for all clients. Nevertheless, testing revealed that the best approach was to have a prompt filter the necessary sections based on the client.

Furthermore, it was determined what rules must be unchangeable, e.g. *"Apply punctuation consistently, following the rules of the [Target Language]."*. For certain sections, without a fixed prompt, the translation would not adhere to the language norms. In other words, the model already knows the generic grammar rules, as such those are not added to the template. However, without that prompt in the template, the translation will only apply the exceptions, and lose quality because the generic grammar rules are not applied.

Lastly, the study showed that style guides do not always have an LP. Certain clients have a style guide per Target Language (TL), independent of the Source. To accommodate both, two templates were developed, one for LP and one for TL.

3.1.2 Model Instructing

The initial stage of the research was to automate the process, and the best way was through a Python script. Due to token restrictions, TowerLLM is unable to perform the extraction and application. As such, we proceeded to its experimentation with GPT API, more specifically GPT 4o. For future iterations, we intend to test Gemini on this task.

First, two input files were loaded: the client's style guide and the template, defining the output schema, whose sections are initialised as empty arrays in a dictionary structure. Second, the style guide is processed in chunks, depending on its size, and each chunk was submitted to the API with a prompt to extract and return information. The data was extracted from all chunks, one at a time, and rewritten as a clear and structured prompt, aggregated by category into the structured dictionary, ensuring that multiple relevant rules per section are retained. Lastly, a second prompt was applied, incorporating the template and the aggregated data, and directing the model to organise the extracted information and properly add it to the template, according to its rules. For any empty sections, it should add generic rules or discard. The completed template is written to a configurable output file.

3.2 Data

For the development stage, and as testing data, three different clients of different domains were used: legal, with banking data; marketing, with product description; and media, with promotional text. Each had style guides that vary in length, format, and structure. All the necessary materials — templates, scripts, and prompts — for the pipeline were tested using these clients, until satisfactory results were achieved across the board.

Following the validation of the methodology, a testing framework was established with a different client from the previous three as the primary case-study. This client was selected as it had both the resources and the interest in implementing this study. The previous three clients did not have the necessary data availability or the data was too simple to evaluate the style guide adherence.

The client's data encompassed multiple brands targeting distinct demographic segments, including adult and children's markets, with product categories spanning various domains. Translation tasks were centred on detailed product descriptions. For this purpose, we selected 28 products, each with four types of descriptions: title; consumer description; shopper description; and product description. Each product, with the four descriptions, had around 20 to 30 sentences.

This research was conducted in seven languages: Japanese (ja-JP), es-ES, Danish (da-DK), Korean (ko-KR), Polish (pl-PL), pt-PT, and Standard German (de-DE). Two of them were selected as they had two style guides each, PL and DE.

The style guides were similar to each other. They were *.docx* files, with similar structures, around 10 pages long, and with a list of rules for grammar, format, and tone of voice. The secondary files for PL and DE comprised a detailed list of formatting rules for specific scenarios, and they were longer and more dense. The processed style guides had around 1500 tokens.

The style guides were processed and post-edited by the author, needing minor annotations. For the translation step, two Tower models were tested in instruction-following, to determine the best when

applying the pipeline to a translation workflow. The models were TransPerfect's proprietary models, TowerZen 9B, the most recent model, and Tower+ 72B (Rei et al., 2025), tested in the three-tier approach assessed in Section 2. For unbiased answers, the models were labelled as A, corresponding to TowerZen 9B, and B, Tower+ 72B.

The annotations were performed by a research linguist or freelancer, with high proficiency in the TL, and prior experience with similar tasks. Detailed guidelines were prepared, covering task descriptions, rules, evaluation criteria, and key considerations. To ensure clarity, an open communication channel was maintained between the author and the annotators, and a dedicated step for comments was incorporated. Both models were given a score of 1 (worst) to 5 (best) for translation quality.

It is important to note that the quality scores were separate from the style guide application, one translation could receive a score of 3, and yet perfectly follow the style guide. To further distinguish between these aspects, two additional questions were included: "Best Style Guide adherence" and "Best quality". For each question, the annotator would select "A", "B", "Equally Bad" if both models failed, or "Equally Good" if both models presented good quality.

4 Results

This section presents the empirical findings from the two experimental phases: the automated information extraction from client style guides and the evaluation of translation quality achieved through the application of these extracted style guides.

4.1 Information Extraction and Application

Overall, this process had quite positive results, requiring post-edition, but to a short extent. The issue was in the system's tendency to omit more specific rules while maintaining accuracy in the information it did extract. On average, four rules were eliminated for being overly general or unnecessary, and 12 rules were missing and added in post-edition. In fact, no incorrect information was generated, suggesting that the extraction methodology prioritises precision. However, in opposition, the system would add information that an LLM would already know. While not incorrect, it was generic or irrelevant, making it unnecessary.

The extraction quality varied according to the specificity and clarity of the source material. In cases where style guides contained ambiguous or unclear rules, the system occasionally produced information that diverged from the intended specifications. Generic style conventions, such as punctuation and formatting rules, were well-captured, whereas domain-specific or client-particular requirements that opposed the general rules, frequently required manual edition, as seen by the higher quantity for addition as opposed to deletion.

Despite these limitations, the extraction framework produced outputs that were well-structured and accurately represented the original style guides provided by the client, indicating that the methodology serves as a viable foundation for semi-automated style guide generation, provided that human oversight is maintained for validation and completion of specialised requirements.

The performance of the script varied depending on the format of the style guide. The *.txt* files and *.docx* / *.doc* consistently yielded the most reliable results. When rules were missing, it did not appear to be due to formatting constraints. *PDF* files, with more complex *PDF* layouts — multiple columns, graphics, tables, and distinct text boxes — were more challenging. A substantial portion of the content was not captured. This suggests that while it generalises well across most common file formats, documents with highly complex visual layouts may benefit from a pre-processing step prior to extraction, in order to ensure a more comprehensive capture of stylistic content.

4.2 Translation and Instruction-Following

	Style Guide Adherence				Translation Quality		
	Equally Good	Equally Bad	A	B	A mean	B mean	Δ (B-A)
ES	4	13	4	7	3.28	3.6	+0.32
DE	1	4	14	9	2.78	2.96	+0.18
KO	0	7	0	21	2.57	3.14	+0.57
PT	4	8	5	11	3.36	3.18	-0.18
PL	11	0	9	8	3.8	3.78	-0.02
DA	22	4	0	2	2.21	2.21	0
JA	18	0	7	3	2.6	2.75	+0.15
	30.6%	18.4%	19.9%	31.1%	2.9	3.1	

Table 1: Style guide adherence and translation quality results

The evaluation results in Table 1 indicate a marginal difference between the two models, with A (TowerZen 9B) achieving a mean translation quality score of 2.9 and B (Tower+ 72B) a mean of 3.1. For four of the seven languages, Model B received a higher score, with the most significant difference in ko-KR (Δ = +0.57) and es-ES

(Δ = +0.32), although the gap between the two remains modest.

Across languages and models, the adherence produced a number of positive results. However, several recurrent patterns indicate areas for improvement. The most consistent shared challenge was trademarks (words tagged with ® or ™), and product names. Both models translated terms that ought to have been kept. Measurement and number formatting were flagged across many languages as well. Both models comprised non-localised units alongside the localised ones rather than replacing them, and did not modify the numbers, not localising thousand separators.

In es-ES, Model B was selected by the annotator nearly twice as often as Model A, (7 vs. 4), although a notably high number of "Equally Bad" in style guide adherence (13) indicates that neither model consistently satisfied the requirements. Literal translations occasionally occurred, e.g. "great value" translated to "expensive," and "walkways" translated as "pasarelas," a term more appropriate for a bridge or a fashion runway than the setting indicated in the source. However, most of the annotated errors were precisely the no conformity to rules, as the language itself was quite positive, and neither of the rules not followed were critical errors.

For pl-PL, it also stands out as a language without "Equally Bad", suggesting that both models produced outputs that the annotator deemed acceptable. Both pl-PL and pt-PT had a preference in translation quality for Model A or was rated in pair with B (Δ = -0.02 and Δ = -0.18, respectively). The de-DE accompanies this trend, but with a larger distinction between the models, with A demonstrating a higher overall adherence. However, both models had register issues, occasionally defaulting to formal structures when an informal tone was required.

For da-DK, it shows identical mean scores for both models in translation quality (2.21), with A having a preference for adherence, showing the uncertainty of the annotator for this language. It had the most individual errors, with major mistranslations in Model A ("gingerbread man" translated as "pejsmand") and cross-language interference in Model B due to Norwegian spellings. In contrast, ja-JP was a balanced language, with both models assessed similarly in 64% of situations, presenting mostly compliance difficulties largely limited to trademark and measurement norms that appeared consistently across assignments. However, these languages recorded the highest number of "Equally Good" in adherence (22 and 18, respectively), indicating satisfaction with both models.

The models performed well more often than not. Taking into account the responses in which both models were rated good (30.6%), Model A showed positive results in approximately 50% of cases (19.9% + 30.6%), while Model B reached around 62% (31.1% + 30.6%). In contrast, responses rated as "Equally Bad" represented 18.4% of the cases, indicating that clear failure was the least common outcome.

Across LPs, the two models show closely marched quality scores, with differences of less than 0.32 points in all cases. PL, pt-PT, and es-ES came out as the strongest performers, with both models achieving mean scores above 3 on a 5-point scale. DA consistently presents the lowest quality scores, which may reflect the relative scarcity of training data.

The co-occurrence of high quality and high "Equally Good" in pl-PL and pt-PT suggests that the style guides did not negatively impact translation quality, maintaining a high baseline of acceptability.

The results also show substantial cross-lingual variation in both adherence and quality, which constitutes the most salient finding of this evaluation. This variation likely reflects differences in the complexity and specificity of each language style guide, the typological distance between source and TL, and the relative volume of language-specific fine-tuning data available to each model. These findings motivate future work on the formulation of language-aware style guides and model selection strategies tailored to individual TL.

5 Conclusion and Future work

This research presents an exploratory evaluation of style guide information extraction through GPT 4o, and application as system prompts to MT translation across seven LPs, comparing two Tower models under a qualitative analysis annotation.

Building on the three-tier approach proposed in (Graça, 2025) and discussed in 2 — which itself yielded positive results — the present work introduces a fourth step that both extends and reinforces these findings. Together, they offer converging evidence that style guide prompting is a viable option

to improve translation quality. Model B demonstrated a consistent, if moderate, advantage in adherence across most languages examined. That said, the cross-lingual variation observed in the results serves as an important reminder that its effectiveness is shaped by language-specific factors, including the availability of training data, the structural complexity of the TL, and the degree of specificity of the style guide itself.

The main limitation of this study was the use of a single annotator per language, a constraint imposed by time and budget considerations. While this allowed us to conduct a broad, cross-lingual evaluation, it means that inter-annotator agreement was not assessed and that bias may influence the results. Future work will incorporate multiple annotators per language, and apply standard inter-rater measures to ensure better evaluations. Additionally, the evaluation was conducted in one text domain (product description), which limits the generalisability of the findings. Style guide compliance and model behaviour may differ across domains or other languages not tested here.

This work was a first step toward a systematic understanding on how to best optimise client style guides through a more automatic approach, extracting its information through a script that proved itself successful, and of how style guide can be operationalised as system prompts in LLM-based pipelines.

Due to the results of this work, this study is now part of a larger ongoing research that examines style guides across multiple client domains and text types. Subsequent evaluations will involve a pilot that will begin with multiple clients from different industries, using diverse style guides and source material. These studies will allow the assessment of whether the observed patterns generalise beyond the description of the product or not, and to assess the liability of this application in scale to the translation pipeline.

Together, these efforts aim to provide a more comprehensive picture of the conditions under which style guide system prompts are effective for production translation workflows.

6 Acknowledgments

This work was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI); Fundação para a Ciência e a Tecnologia (FCT), through Portuguese national funds under projects UID/50021/2025 (DOI:https://doi.org/10.54499/UID/50021/2025) and UID/PRR/50021/2025 (DOI: https://doi.org/10.54499/UID/PRR/50021/2025) and by CLUL, UID/214/2025 (https://doi.org/10.54499/UID/00214/2025)

References

American Translators Association. 2019. The whys and hows of translation style guides. A case study. *The Savvy Newcomer*. Accessed February 10, 2026.

Aranbure, Yaiza, Silvia Arenas, Javier Augusto, Ana Azuaje, Naiara Ortega, and Sonia Solana. 2018. The use of style guides in technical and scientific translations.

Ghosal, Pamela. 2024. Why you should work with a translation style guide. *Phrase*. Accessed February 10, 2026.

Graça, Leonor. 2025. Contextual parameters in domain-specific translation: A three-tier model. Master's thesis, School of Arts and Humanities, University of Lisbon, Lisbon, Portugal. `http://hdl.handle.net/10400.5/118388`.

Heo, Juyeon, Christina Heinze-Deml, Oussama Elachqar, Kwan Ho Ryan Chan, Shirley Ren, Udhay Nallasamy, Andy Miller, and Jaya Narain. 2025. Do llms "know" internally when they follow instructions?

International Organization for Standardization. 2015. ISO 17100:2015 Translation services — Requirements for translation services. Technical report, ISO. Accessed February 18, 2026.

Jang, Joel, Seonghyeon Ye, and Minjoon Seo. 2022. Can large language models truly follow your instructions? In *NeurIPS ML Safety Workshop*.

Jaroslawicz, Daniel, Brendan Whiting, Parth Shah, and Karime Maamari. 2025. How many instructions can llms follow at once? `https://arxiv.org/abs/2507.11538`.

Lingualinx. 2024. The crucial role of style guides in language translation. *News and Updates*. Accessed February 10, 2026.

Rei, Ricardo, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms.

The Business Research Company. 2026. Ai in lnaguage translation market report 2026. Technical report, The Business Research Company.

Young, Richard J., Brandon Gillins, and Alice M. Matthews. 2025. When models can't follow: Testing instruction adherence across 256 llms.

A Longitudinal Study of the Adoption of Specialized MT Systems in Canadian Parliamentary Translation

Michel Simard Jeniffer Leal-Wyss Gabriel Bernier-Colborne Rebecca Knowles
National Research Council Canada
{Michel.Simard, Jeniffer.Leal,
Gabriel.Bernier-Colborne, Rebecca.Knowles}@nrc-cnrc.gc.ca

Abstract

Since 2023, translators for the Parliament of Canada have had the option to use neural machine translation (NMT) technology provided by the National Research Council of Canada (NRC) to support their work in translating parliamentary publications between French and English. We present our analysis of an anonymized dataset of translators' interactions with our Hawkeye MT systems, collected since their introduction and covering a period of 2.5 years. This data provides a unique perspective on how translators interact with the systems, how their use evolved over time and how it impacts the nature of their translations.

1 Introduction

As machine translation is becoming a standard fixture in professional translators' work environments, we see a growing literature on translator's changing work habits (do Carmo and Moorkens, 2022; O'Brien, 2024), attitudes towards technology (Nunes Vieira and Alonso, 2019; Guerberof Arenas, 2025), motivations for adoption and non-adoption (Cadwell et al., 2018; Zaretskaya, 2015), etc. Such studies can be extremely rich in details and serve a useful function in understanding the impact of such a transformation of translation work environments on the professional practices and lives of translators. In most cases, however, these studies can only offer a snapshot of this rapidly-evolving landscape at a given point in time. We rarely have the opportunity to witness and document the transformation process itself, as it is happening. This paper is about one such opportunity.

Since 2023, the National Research Council of Canada (NRC) has been providing custom MT technology to the Government of Canada's Translation Bureau (TB), to assist the linguists of its parliamentary service in translating documents published by the Parliament of Canada, most notably the Hansard, i.e. the transcripts of the House of Commons debates, in both of Canada's official languages, French and English. The Hawkeye systems are neural machine translations (NMT) systems, created from Canadian parliamentary texts. Parliamentary translators have access to a number of MT tools, however Hawkeye is the only one that is currently directly integrated into Prism, the Parliament of Canada's publication workflow management application, which is the translators' main work tool (O'Brien, 2002). No translation memory tools are integrated into Prism. Unlike many work environments where postediting MT is mandatory, use of Hawkeye MT is optional: translators are actively encouraged to use it but translations are not automatically pre-populated with MT and translators who want to use it have several user interface options for inserting Hawkeye MT output into their editing window.

Since the official deployment of the systems in May 2023, user interactions with Prism and Hawkeye have been systematically recorded. This anonymized data offers us a unique view on how translators interact with the systems and how their use of Hawkeye MT evolved over time. Because the data include the text of the translations at different stages—initial MT, translator's draft and final revised version—we also have a view of how MT impacts the nature and quality of translators'

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 133–141

work (cf. shorter-term studies like Macken et al. (2020)). In the following pages, we present our analysis of over two and a half years of these user interaction logs.[1]

We first analyze how the use of Hawkeye MT globally evolved over time: we observe that, from 16% of texts when Hawkeye was first deployed in May 2023, the use of the system climbed to 60% in December 2025. We then look at individual (anonymous) translators' work habits with MT to see whether patterns of use emerge: we find that while users are free to decide for each paragraph whether or not to use Hawkeye, in practice, they either use it all the time or not at all. Finally, we examine how the use of Hawkeye affects the translations themselves: we see that when translators chose to insert Hawkeye MT into their translation window, they produce translations that are much more similar to the initial MT than those who don't. This suggests that Hawkeye indeed provides a useful initial draft for those translators.

2 Hawkeye and Prism

The Hawkeye systems are neural machine translations (NMT) systems, trained exclusively with Canadian parliamentary bilingual data. They are based on the Sockeye NMT toolkit (Hieber et al., 2022). Hawkeye systems are created and maintained for four different domains: House of Commons debates, House of Commons committees, Senate debates and Senate committees, but in this work we only focus on House of Commons systems. Distinct systems are produced for English-French and French-English translation. All systems for a given language direction are based on a common system, trained from scratch on all available parliamentary data. Each domain-specific system is then produced by fine-tuning this base system using only the domain-specific data. Table 1 summarizes the data used to produce the latest updated systems, in December 2025. The systems are trained using all parliamentary texts in the proprietary "House of Commons" XML format; as a result, systems handle structured XML input, producing translations with valid XML markup. All systems are updated three times a year, by recreating the systems from scratch, using all data available at that time. More details can be found in Knowles et al. (2024) and Knowles et al. (2023).

		Words	
Domain	**Segments**	**English**	**French**
HoC Debates	7.7 M	135.3 M	148.6 M
HoC Committees	16.6 M	227.1 M	245.7 M
Sen. Debates	0.6 M	11.1 M	12.5 M
Sen. Committees	1.9 M	29.6 M	32.4 M
Total	**26.8 M**	**403.1 M**	**439.3 M**

Table 1: Training data for Hawkeye systems (Dec. 2025).

Hawkeye systems are integrated into Prism (O'Brien, 2002), the workflow application that is the primary means of producing the Hansard and other publications of the Canadian parliament. Within this environment, linguists produce and revise translations using a dedicated three-paned user interface (UI) window: the left-hand pane is used to select from a list of documents that have been assigned to the translator or reviser, the middle pane displays the source-language text to be translated, and the right-hand pane is used to edit the target-language version. The source and target texts within each document are divided into text *chunks*: paragraph-like units that most often correspond to interventions from a single speaker. Both versions also contain structural XML tags, which in the UI are displayed as grey labels that the user can manipulate in various ways.

Translators are free to decide whether or not they use Hawkeye MT: the translation editing pane is not pre-populated with machine translation. Translators who want to use Hawkeye need to explicitly insert MT into their target translation, by activating various UI components. This operation can be performed for all chunks at the beginning of a new document, or individually for each chunk.

3 The Prism-Hawkeye Logs

Systematic recording of data on the use of Hawkeye MT within Prism began shortly after the first systems were deployed, in May 2023. In this work, we examine a dataset of 286,858 records that were collected over a period of 31 months, between 15 May 2023 and 12 December 2025.[2] In what follows, we refer to this dataset informally as the "logs". Table 2 presents general dataset statistics.

Each record in the logs corresponds to a text

[1]The described study design was reviewed and approved by the National Research Council of Canada's Research Ethics Board (NRC-REB #2024-22).

[2]This number of records is computed after filtering out texts with no translation: those that may have been incomplete at the moment of collection or which were translated outside of the Prism interface.

HoC Domain	EN-FR rec.	EN-FR words	FR-EN rec.	FR-EN words	Total rec.	Total words
Debates	104 k	10.9 M	365 k	3.8 M	141 k	14.7 M
Committees	949 k	6.7 M	51 k	3.6 M	146 k	10.3 M
Total	199 k	17.7 M	880 k	7.3 M	287 k	25.0 M

Table 2: Number of records and words in dataset.

chunk, i.e. a segment of text about the size of a paragraph (87.2 words per chunk on average). Each record contains the source-language version of that text (we refer to this field as *SrcText*), along with two or three target-language versions: 1) *MTText*: the machine translated version produced by Hawkeye (regardless of whether or not it was used by the translator), 2) *TrText*: the translation produced by a parliamentary translator, and 3) *RevText*: optionally, a revised translation, as produced by a reviser.

As mentioned above, Hawkeye systems are trained using data from both houses of the Canadian Parliament: the House of Commons and the Senate. However, only House of Commons systems were available to parliamentary translators during the period covered by our analysis. Therefore, in this study, we consider only the two House of Commons domains: *House of Commons Debates* and *House of Commons Committees*, in both translation directions.

We note that parliamentary translators are organized in two teams: the Committees team and the Debates team. Both teams translate documents from the House of Commons and the Senate, have both English and French translators (with some doing both directions), and also have revisers (most of whom also perform translation work). Translations of Debates are required to be systematically revised and to be ready for publication the next morning. Because House of Commons debates normally take place during the day (the House typically sits between 10am and 7pm), the Debates team usually works the evening shift. Translations of committees are not subject to the same tight deadlines and translators of that team normally work the day shift. However, they are often asked to also translate Debates on days with high volumes.

Regarding language, 69.3% of texts in the logs were translated from English into French and 30.7% from French into English. This is representative of the distribution of language interventions in Parliament. The source language of each chunk is also recorded in the logs.

A number of records have an empty *RevText*, i.e. the field containing the final version of the translation, after revisions: 138,557 (48.3%). Most of these are chunks that were ultimately not revised; in these cases, the translator's version *TrText* is the final version.[3] As noted above, translations of Debates are all revised; most Committees' translations in the logs are not. We made this information explicit by adding a Boolean field *Revised* to each record, whose value is based on whether *RevText* is instantiated or not.

Each record also contains anonymized identifiers of the translator and reviser who handled the translation, as well as various timestamps.

Finally, one particularly relevant piece of information is contained in the *HMTinserted* field, a Boolean value indicating whether the translator inserted the Hawkeye MT output into their translation window using dedicated UI components.

4 Analysis

4.1 Use of Hawkeye MT

As mentioned above, the *HMTinserted* field is a Boolean value indicating whether the translator inserted the Hawkeye MT output into their translation. We use that information to analyze the use of Hawkeye MT (sometimes abbreviated HMT below) by translators. Of course, just because a translator had HMT inserted into their translation for a given segment of text does not necessarily mean that they actually based their translation on it; they may have inserted it initially and then later chose to delete it. Conversely, the fact that a translator did not insert HMT does not mean that they did not use any MT at all: parliamentary translators have access to other MT tools, such as DeepL Pro[4] or GCtranslate,[5] which they sometimes use rather than Hawkeye. However, because these MT tools are not integrated into Prism, translators need to manually insert the MT into their translation, either by cutting and pasting from a different window or by typing. We currently have no way of detecting these events. Therefore, in what follows,

[3] A small number of records with empty *RevText* may be ones for which the revision was not yet completed at the moment of data collection.

[4] https://www.deepl.com/en/pro

[5] https://www.canada.ca/en/public-services-procurement/news/2025/09/gctranslate-using-artificial-intelligence-to-build-a-more-agile-modern-and-bilingual-public-service.html

when we use phrases such as "translated with the help of Hawkeye MT", we merely mean that the *HMTinserted* flag was set to *True*. Conversely, when that flag is set to *False*, we cannot assume that the translation was produced without the help of any MT; we simply do not know whether some MT was used or not.

Globally, 46.9% of all text chunks translated in Prism during the period covered by the logs were translated with the use of Hawkeye. Figure 1 shows global HMT use depending on translation direction and the origin of the texts. We observe that HMT was used more for translation into French than for translation into English, and more for the translation of Committee hearings than for Debates.

Figure 2 shows the evolution of Hawkeye use month-by-month between May 2023 and December 2025. Overall, from 16% in May 2023, HMT use gradually increased to over 50% in the fall of 2024 and appears to hover between 50 and 60% since then. Figure 2 also shows this evolution as a function of text origin (middle plot) and language (lower plot). We note that, while HMT was used more for Committees than for Debates overall, since the summer of 2025, HMT usage is comparable for both domains. This suggests that while translators of the Debates team were slower to adopt HMT, their usage later increased, to a point where they now use it equally often.

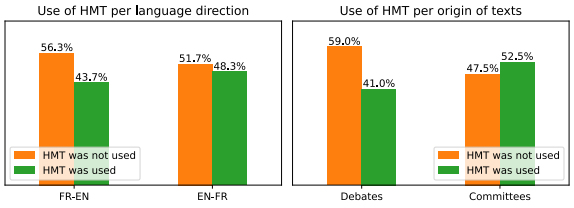


Figure 1: Hawkeye MT use as a function of translation direction (left) and origin of text (right).

4.2 Translators

During the period covered by the logs, a total of 145 linguists (as identified by distinct anonymized IDs) either translated or revised House of Commons texts. Table 3 gives a more detailed breakdown by task, language, and origin of texts. It is worth noting that most revisers also worked as translators during the covered period. Many translators worked on both debates and committees: this is true of 56 French and 25 English translators.

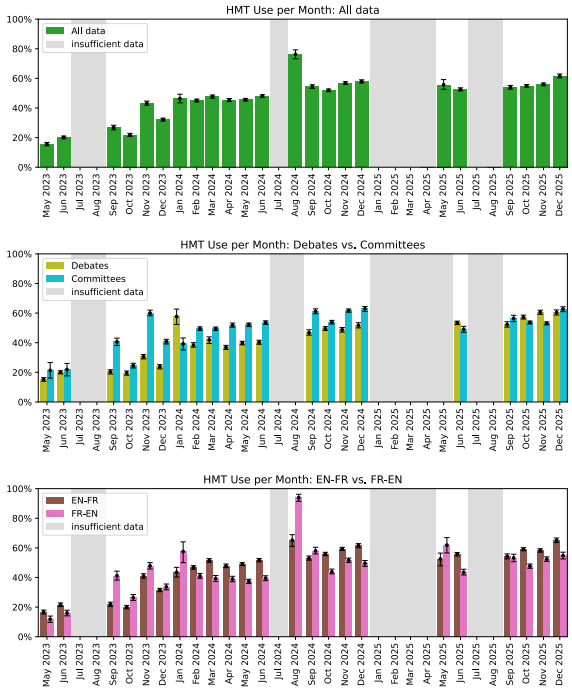


Figure 2: Monthly percentage of Hawkeye MT Use: global (top), compared per text origin (middle) and per language (bottom). Only months during which 20,000 words or more were translated for all conditions are reported; confidence intervals computed by bootstrap resampling (1000 iterations, $p = 0.01$). The gaps correspond to periods during which Parliament was not sitting, most notably the prorogation of January 2025 and the summer pauses.

	Translators		Revisers	
Language	**Debates**	**Comm.**	**Debates**	**Comm.**
EN-FR	75	88	21	11
FR-EN	33	41	11	6

Table 3: Numbers of linguists involved in each parliamentary task.

Finally, 12 translators did both English and French translation, 7 of whom also worked as revisers.

We saw in Section 4.1 that translators have used Hawkeye to help translate approximately 47% of all text between May 2023 and December 2025. But is it that 47% of all translators used HMT all of the time, or that all translators used HMT 47% of the time? To find out, we computed the proportion of text for which each translator used HMT, during different periods.

Figure 3 shows the distribution of translators based on their use of HMT, and how this distribution evolved over time: each shade of blue denotes a different time period. Within these histograms, each column shows the percentage of translators whose use of HMT fell within a given range during the given period. Looking at the lightest shade

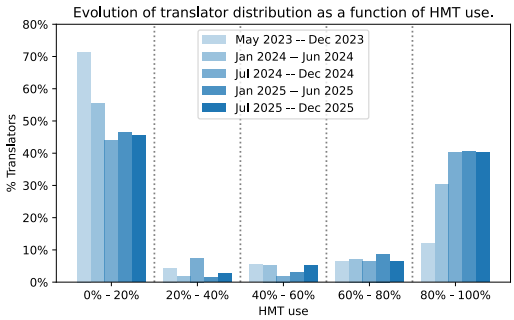


Figure 3: Evolution over time of the distribution of translators based on the proportion of documents for which they used Hawkeye MT.

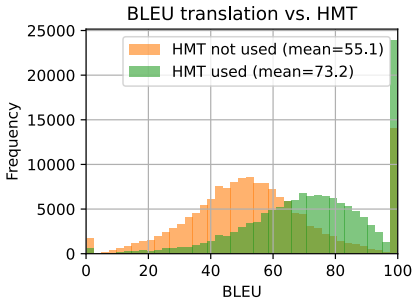


Figure 4: Distribution of textual similarities (BLEU metric) between machine translation output and draft translations, whether Hawkeye MT was used (green) or not (orange).

columns, we see that, between May and December 2023, the vast majority (more than 70%) of translators used HMT for less than 20% of the texts that they translated; for simplicity, we refer to this group as *non-users* of Hawkeye MT. During the same period, about 12% of translators used HMT for more than 80% of the texts that they translated; we call this group *regular users* of Hawkeye. In between these two extremes we find what we call *occasional users*: translators who used Hawkeye for anywhere between 20% and 80% of all their translations: between May and December 2023, these amounted to about 15% of all translators.

As time progresses (darker shades), the proportion of non-users of HMT diminishes, while the proportion of regular users increases. Both groups seem to stabilize around July 2024: non-users at about 45% and regular users at 40%. All this time, the size of the group of occasional users remains relatively stable. We performed a brief manual examination of how individual anonymous translators' use of Hawkeye changed month-to-month. In some cases, the translators who appear (in Figure 3) to be occasional users would be more accurately described as users who are in the process of migrating from one group (non-users or regular users) to the other. That is, they are increasing (or decreasing) their use of Hawkeye over the course of the log file collection time span. This suggests that the vast majority of parliamentary translators are either non-users of Hawkeye (meaning they never or very rarely use it) or regular users (meaning they use it systematically).

4.3 MT Usage and Impact on Revision

We finally examined whether translators actually make use of the Hawkeye machine translation output. This was done by measuring the textual similarity between the MT output (field *MTText*) and the translator's draft (field *TrText*); we then compared the distributions of these similarities under the two conditions: *Hawkeye MT was used* and *Hawkeye MT was not used*. The result of this comparison, using BLEU as similarity metric, is shown in Figure 4.[6] For cases where the *HMTinserted* flag is *True*, the translator's draft is much more similar to the MT output (average similarity is 73.2) than for cases where that same flag takes the value *False* (55.1). This indicates that when translators import MT into the editing pane, they do effectively tend to base their translation on that proposed MT.

A more important question perhaps is whether translations produced with the help of Hawkeye MT differ from those that do not use Hawkeye MT in terms of their quality. As a proxy, we measure the amount of revision edits that were performed by revisers on translators' drafts.[7] For this, we compute the Translation Edit Rate (TER) between the translator's draft (field *TrText*) and the final reviser's version (field *RevText*).[8] This metric can be interpreted as the percentage of words of the text chunk that were edited—inserted, deleted, modified or moved—by the reviser (higher values indicate more revision). We then compare the distributions of these rates under the two conditions: *Hawkeye MT was used* and *Hawkeye MT was not used*. The results are shown in Figure 5.

The dotted lines in this figure show the global

[6]Note that, in this context, BLEU is *not* used to measure MT quality, but merely textual similarity between the MT and the translator's draft – see Appendix A.

[7]This analysis is limited to chunks that were actually revised, i.e. which had non-empty *RevText* and were thus marked with the Boolean field *Revised* set to true.

[8]Note that, in this context, TER is *not* used to measure machine translation quality directly, but rather to measure the amount of revision work performed by revisers – see Appendix A.

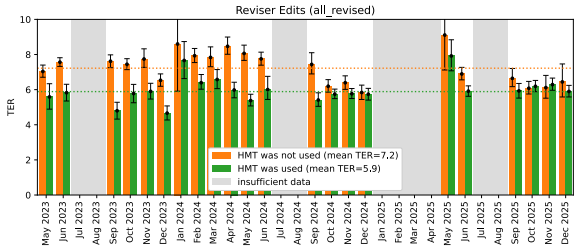


Figure 5: Monthly comparison of the average amount of revision per chunk (measured in TER) performed by revisers on translator drafts, depending on whether these were done with or without the help of Hawkeye MT. Only months during which at least 500 chunks were translated are reported; confidence intervals computed by bootstrap resampling (1000 iterations, $p = 0.05$).

average TER for translations done with and without the help of HMT. Globally, we observe lower TER values for translations produced with the help of Hawkeye MT than for those produced without: the average number of revision edits for chunks produced wthout Hawkeye MT is 7.22; for those produced with the help of Hawkeye MT, it is 5.89, yielding a statistically significant difference of 1.33 ± 0.16 between the two conditions (estimated through 1000 iterations of bootstrap resampling with $p = 0.01$). However, while this tendency is quite strong in 2023 and 2024, it seems to diminish substantially in 2025, to a point where most differences between the two conditions are not statistically significant. It is difficult to determine whether these variations over time carry any significance. At this point, we simply observe that, based on the amount of edits performed by revisers, translations produced with the help of Hawkeye do not appear to be of lesser quality than those produced without.

Note that this assessment reflects the amount of revisions in terms of surface-level edits, not the severity of errors that revisers observe in draft translations. Also worth noting: as far as we know, apart from the name of the translator, Prism does not provide revisers with information that would allow them to determine whether a given translation was produced with or without the help of Hawkeye.

5 Conclusions

We presented our analysis of user interactions of the translators and revisers of the Translation Bureau's parliamentary service with the Hawkeye MT components of the Prism parliamentary publication workflow management software, based on user interaction data collected between May 2023 and December 2025. Overall, we found that during the covered period, translators used Hawkeye machine translation for 47% of all the texts they translated using the Prism environment. This use increased steadily in 2023 and 2024, from 16% in May 2023 to approximately 55% in the fall of 2024, plateauing at about that level throughout 2025. During that year, Hawkeye was used more frequently for translation into French (57.8%) than into English (50.0%), and more frequently for the translation of Debates (56.4%) than for committee evidence (53.9%).

We observed that most translators either rarely use Hawkeye MT (*non-users*) or almost always use it (*regular users*). In the period between July and December 2025 (last six months of logged period), these two groups represented 44% and 41% of all translators, leaving only about 15% of translators qualifying as *occasional users* of Hawkeye MT.

In terms of quality, translations produced with the help of Hawkeye required fewer revisions on average than those produced without, although the difference between the two conditions seems to fluctuate over time. Again, we note that this analysis of quality was based solely on the amount of edits performed by revisers (in terms of the number of words and punctuation marks revised), not on the gravity of the errors they encountered.

Collection of user interaction data in Prism is ongoing, and we hope to repeat this analysis periodically and follow the evolution of the observed trends and better our understanding of how Hawkeye MT is used by parliamentary translators. We note in passing that the current data collection procedure does not allow us to reliably measure translator and reviser productivity. If measuring this is critical, then additional instrumentation of the Prism translation user interface and/or voluntary user disclosure would be required. Another thing that the current data collection procedure does not allow us to detect is situations where translators (or revisers) use other translation tools, such as DeepL Pro or GCtranslate, to produce translations. Again, this would require different instrumentation.

Finally, and as noted above, there seem to exist differences in the quality of the translations produced with and without the help of Hawkeye MT. This warrants a more in-depth analysis, in which we would examine the nature of the revisions performed on each kind of translation.

Acknowledgments

We wish to thank the parliamentary translation team and all our partners at the Canadian government's Translation Bureau, without whom this work would not have been possible.

References

Cadwell, Patrick, Sharon O'Brien, and Carlos Teixeira. 2018. Resistance and accommodation: factors for the (non-) adoption of machine translation among professional translators. *Perspectives*, 26(3):301–321.

do Carmo, Félix and Joss Moorkens. 2022. Translation's new high-tech clothes. In Huertas-Barros, Elsa, Gary Massey, and David Katan, editors, *The human translator in the 2020s*, pages 11–26. Routledge.

Guerberof Arenas, Ana. 2025. Perspectives on machine translation, post-editing, and automation. In Walker, Callum and Joseph Lambert, editors, *The Routledge handbook of the translation industry*, pages 186–204. Routledge.

Hieber, Felix, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, Surafel Lakew, Prashant Mathur, Anna Currey, and Marcello Federico. 2022. Sockeye 3: Fast neural machine translation with pytorch.

Knowles, Rebecca, Samuel Larkin, Marc Tessier, and Michel Simard. 2023. Terminology in neural machine translation: A case study of the Canadian Hansard. 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 481–488, Tampere, Finland, June. European Association for Machine Translation.

Knowles, Rebecca, Samuel Larkin, Michel Simard, Marc A Tessier, Gabriel Bernier-Colborne, Cyril Goutte, and Chi-kiu Lo. 2024. Some tradeoffs in continual learning for parliamentary neural machine translation systems. In Knowles, Rebecca, Akiko Eriguchi, and Shivali Goel, editors, *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 102–118, Chicago, USA, September. Association for Machine Translation in the Americas.

Macken, Lieve, Daniel Prou, and Arda Tezcan. 2020. Quantifying the effect of machine translation in a high-quality human translation production process. *Informatics*, 7(2).

Nunes Vieira, Lucas and Elisa Alonso. 2019. Translating perceptions and managing expectations: An analysis of management and production perspectives on machine translation. *Perspectives*, 28:1–22.

O'Brien, Audrey. 2002. Prism: The house of commons integrated technology project. *Canadian Parliamentary Review*, 25(2):20–22.

O'Brien, Sharon. 2024. Human-centered augmented translation: against antagonistic dualisms. *Perspectives*, 32(3):391–406.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Zaretskaya, Anna. 2015. The use of machine translation among professional translators. In *Proceedings of the EXPERT Scientific and Technological Workshop*, pages 1–12.

A Measures of Textual Similarity

In different parts of this study, we find the need to quantify the differences between different translations of the same text. We base these comparisons using *textual similarity metrics*: functions that produce values that can be interpreted as the degree of similarity between two strings of text.

In practice, the metrics that we use in this study are standard machine translation evaluation metrics: BLEU, ChrF and TER. Here, however, we use these MT evaluation metrics strictly as textual similarity metrics, and *not* to directly measure MT (or postedited text) quality against a reference. For example, in Section 4.3, we use BLEU to compare the similarity between Hawkeye outputs (hypotheses) and the translators' drafts (references) under two different conditions; in that same section, we use TER to measure the similarity between translators' drafts (hypotheses) and final, revised translations of the same texts (references). In both of these examples, these metrics' values should not be interpreted as denoting machine translation quality. In the case of BLEU, we are examining how much of the original MT appears in the translator draft. In the case of TER, we are examining

how many revisions were made to the translator's draft, conditioned on whether Hawkeye MT was selected at translation time or not (without knowing directly whether it was used or not).

We use the `sacrebleu`[9] implementation of BLEU, ChrF and TER machine translation evaluation metrics Post (2018), with signatures provided in Table 4.

[9] `https://github.com/mjpost/sacrebleu`

BLEU	`nrefs:1|case:mixed|eff:yes|tok:13a|smooth:add-k[1.00]|version:2.4.3`
ChrF	`nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.4.3`
TER	`nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no|version:2.4.3`

Table 4: Sacrebleu signatures for metrics used in this paper. Because we average similarities between relatively short segments of text, we need to compute BLEU scores with smoothing: we use *add-1* smoothing.