

NeTTIT'2026

**3rd International Conference on
New Trends in Translation and Technology**

**Proceedings of the International Conference
New Trends in Translation and Interpreting Technology**

Edited by Marie Escribe, Alicia Picazo Izquierdo, Constantin Orăsan, Tharindu Ranasinghe, Gloria
Corpas Pastor, Marko Tadić and Ruslan Mitkov

June 24 – 27, 2026
Dubrovnik, Croatia

New Trends in Translation and Technology
3rd International Conference on
New Trends in Translation and Interpreting Technology
NeTTIT'2026

PROCEEDINGS

June 24 – 27, 2026
Dubrovnik, Croatia

Series Online ISSN 2815-4711
Online ISBN 978-954-452-110-3
e-book site: www.acl-bg.org, www.nettt-conference.com/2026/

Designed by INCOMA Ltd.
Shoumen, BULGARIA

Preface

NeTTIT'2026 is the successor to the established New Trends in Translation and Technology (NeTTT) conference, whose first edition was held in Rhodes, Greece in June 2022 and whose second edition took place in Varna, Bulgaria in July 2024. Both editions of NeTTT were a resounding success, and during the Varna event we decided to slightly revise the title of the conference in order to encourage more submissions and discussions related to interpreting technology. NeTTIT'2026, held in Dubrovnik, marks the first edition under the new name New Trends in Translation and Interpreting Technology, which we intend to carry forward in order to attract submissions and attendees from both the translation technology and interpreting technology communities.

The conference has two main aims: (i) to bridge the gap between academia and industry in the fields of translation and interpreting by bringing together academics in linguistics, translation studies, machine translation and natural language processing, alongside developers, practitioners, language service providers and vendors who work on or are interested in different aspects of translation and interpreting technology; and (ii) to serve as a distinctive forum for discussing the latest developments and practices in the field. The conference provides an environment for professionals who wish to learn about new trends, present their latest work, share their experience, and establish business and research contacts, collaborations and new ventures. At the same time, it offers researchers the opportunity to disseminate cutting-edge work on translation and interpreting technologies and receive feedback from their peers. In addition to its traditional topics, NeTTIT'2026 features a special theme track titled "The Future of Translation and Interpreting Technologies in the Era of LLMs and Generative AI".

The programme of NeTTIT'2026 includes 31 presentations, among them two keynote speeches by Yves Champollion and Marko Grobelnik. The programme also features a pre-conference tutorial on quality estimation for Machine Translation by Tharindu Ranasinghe, a tutorial on post-editing by Marie Escribe, and a tutorial on speech-to-text for interpreting by Constantin Orasan.

For the first time, the proceedings of NeTTIT will be included in the ACL Anthology, ensuring extensive dissemination and broad citation impact.

We would like to thank the University of Zagreb and especially Prof. Marko Tadić for their logistical and financial support. We are also very grateful to Juremy for once again sponsoring this distinctive event.

A heartfelt thank you goes to all members of the Organising Committee, who worked tirelessly to make this event a success. We would particularly like to acknowledge the contributions of our early-career colleagues Marie Escribe, Alicia Picazo, Xiaojing Zhao and Laura Noriega.

Constantin Orasan, Tharindu Ranasinghe, Marie Escribe, Alicia Picazo Izquierdo, Marko Tadić and Ruslan Mitkov

Organisers

The International Conference NeTTIT'2026 is organised by:

- Bulgarian Association for Computational Linguistics
- Croatian Language Technologies Society
- University of Zagreb, Faculty of Humanities and Social Sciences
- HR-CLARIN

Sponsor

- Juremy



Conference Chairs

- Ruslan Mitkov, Lancaster University, UK and University of Alicante, Spain
- Gloria Corpas Pastor, University of Malaga, Spain
- Marko Tadić, University of Zagreb, Croatia

Programme Committee Chairs

- Constantin Orăsan, University of Surrey, UK
- Tharindu Ranasinghe, Lancaster University, UK

Publication Chairs

- Marie Escribe, LanguageWire and Polytechnic University of Valencia, Spain
- Alicia Picazo Izquierdo, University of Alicante, Spain

Publicity and Sponsorship Chair

- Vilelmini Sosoni, Ionian University, Greece

Conference Administrator

- Nikolai Nikolov, Association for Computational Linguistics, Bulgaria

Members of the Organising Committee

- Xiaojing Zhao, The Hong Kong Polytechnic University
- Laura Noriega, University of Malaga, Spain

Programme Committee

- Khetam Al Sharou, Dublin City University, Ireland
- Abdelalah Alsolami, University of Jeddah, Saudi Arabia
- Silvia Bernardini, University of Bologna, Italy
- Alba Bonet Jover, University of Alicante, Spain
- Pierrette Bouillon, University of Geneve, Switzerland
- Vicent Briva-Iglesias, Dublin City University, Ireland
- Parthena Charalampidou, Aristotle University of Thessaloniki, Greece
- Helle Dam Jensen, Aarhus University, Denmark
- Anna Beatriz Dimas Furtado, University of Galway
- Marie Escribe, LanguageWire and Polytechnic University of Valencia, Spain
- Michael Farrell, University Institute of Modern Languages, Italy
- Maria Fernandez-Parra, Swansea University, UK
- Łukasz Grabowski, University of Opole, Poland
- Manuel Herranz, Pangeanic, Spain
- Amal Haddad Haddad, University of Granada, Spain
- Hansi Hettiarachchi, Lancaster University, UK
- Kristian Hvelplund, University of Copenhagen, Denmark
- Miguel A. Jimenez Crespo, Rutgers University, US
- Valentini Kalfadopoulou, Ionian University, Greece
- Dorothy Kenny, Dublin City University, Ireland
- Ekaterina Lapshinova-Koltunski, Hildesheim University, Germany
- Raquel Lazaro Gutierrez, Alcala University, Spain
- Todor Lazarov, New Bulgarian University, Bulgaria
- Elpida Loupaki, Aristotle University of Thessaloniki, Greece
- Elizabeth Marshman, Ottawa University, Canada
- Núria Molines Galarza, Valencia University, Spain
- Johanna Monti, Naples University, Italy
- Ricardo Muñoz Martín, Bologna University, Italy
- Jean Nitzke, University of Adger, Norway
- Laura Noriega, University of Malaga, Spain
- Lucas Nunes Vieira, Bristol University, UK
- David Orrego-Carmona, Warwick University, UK
- John E. Ortega, Northeastern University, US
- Timea Palotai-Torzsas, Juremy, Hungary
- Damith Premasiri, Lancaster University, UK
- Rozane Rebechi, Universidade Federal do Rio Grande do Sul, Brazil
- Mark Shuttleworth, The Hang Seng University of Hong Kong, Hong Kong
- Diana Singureanu, University of Surrey, UK
- Elena Isabelle Tamba, Romanian Academy, Romania
- Shiyi Tan, University of Surrey, UK
- Irina Temnikova, “Big Data for Smart Society” Institute, Bulgaria
- Alina Trubnikova, University of Granada, Spain
- Eleni Tziafa, University of Athens, Greece

- Cecilia Yalangozian, Machine Translate, US
- Han Xu, Hong Kong Polytechnic University, Hong Kong
- Sudhansu Bala Das, University of Galway, Ireland
- Xiaojing Zhao, The Hong Kong Polytechnic University, Hong Kong
- Shuyin Zhang, The Chinese University of Hong Kong, Hong Kong

Table of Contents

<i>Do readers prefer AI-generated Italian short stories?</i>	
Michael Farrell	1
<i>Semantic typology and audio description. A corpus-based study of motion events</i>	
Teresa Molés Cases	9
<i>How Well Can Large Language Models Detect, Classify, and Correct Errors in Chinese-English Public Sign Translations?</i>	
Shuyin Zhang, Xindi Hao and Zhenhai Li	17
<i>Training Interpreting Learners to Collaborate with Captions and Subtitles during Simultaneous Interpreting</i>	
Huang Yujie and Han Xu	29
<i>Punctuation Patterns in Translated and Original Chinese-English Literary Texts: A Weibull Distribution Analysis</i>	
Jiawei SHI and Sihui HE	34
<i>From SOP to AI-Native Pipeline: Design and Evaluation of a Translation-Memory-Grounded LLM Workflow (Nixa AI)</i>	
Ana Dumančić Acar	41
<i>Building a European Multilingual Evaluation Dataset: The MMLU Localisation Project within the EMT Network</i>	
Pilar Sánchez-Gijón, Susana Valdez, Sofía Calvo Del Barrio, Florence Bellemont, Anna Kokkinidou and Mihai Cristian Brasoveanu	46
<i>A Comparative Study of LLMs for Idiomatic Expression Identification and Translation Equivalent Extraction in an English–Portuguese Subtitle Corpus</i>	
Rozane Rebechi and Diego Alves	54
<i>Navigating the "Fluency Trap": Cognitive Risks and Structural Challenges in Korean-English Patent Translation Post-Editing</i>	
Jinam Lee, Jaehee Ryu and Janet Sohlhee Yu	63
<i>A Corpus-based Analysis of Semantic Similarity in Machine Translation and Human Interpreting</i>	
Kaifusai Julaiti and Andrew K. F. Cheung	68
<i>Generated Voices: Examining Patterns and Shifts in AI Translation of Online Feminist Discourse</i>	
Ailin Li	79
<i>Challenges in Code translation using Large language models</i>	
Lukas Theuerkauf, Ksenia Neumann, Abdulrahman Nahhas, Damanpreet Singh Walia and Maria Chernigovskaya	84
<i>Exploring Sentiment Transfer in Cantonese–English Court Interpreting: A Computational Analysis Using Multilingual BERT</i>	
Sihui HE and Jiawei SHI	93
<i>How Decoding Settings Shape Lexical Variation in Iterative LLM-Translation Workflows: A Case Study of Chinese-English Literary Translation</i>	
Lulu Wang, Hengtian Zhang, Lieve Macken and Kanglong Liu	98

<i>Dual-Metric Compliance and Quality Evaluation of Knowledge Graph Mediated Translation in Regulated Domains: An Enhanced Architectural Framework</i>	
Viveta Gene and Vilelmini Sosoni	109
<i>From Diagnosis to Action: An Evidence Based Maturity Model for AI Integration in Translation</i>	
Viveta Gene and Valeria Cannavina	119
<i>Enhancing Persian-Arabic OCR for Translation Workflows: A Benchmark-Driven Optimization Approach</i>	
Masoud Shahnazari and Abied Alsulaiman	130
<i>A Corpus-based Approach to Translating Specialised Terminology in Literature: A Two-step Protocol</i>	
Laura Noriega and Gloria Corpas Pastor	138
<i>Fine-tuning Arabic-English NMT Engines for Investment Laws</i>	
Najat B. Al-Huthali, Yuan Zou and Constantin Orasan	146
<i>#MTFails to #MTWins: Investigating User Perspectives on Machine Translation in Social Media</i>	
Khetam Al Sharou	157
<i>Comparative analysis of the double-stage revision phase (English-Italian) with an AI chatbot assistance</i>	
Sofía Calvo del Barrio and Pilar Sánchez Gijón	167
<i>Does AI-INT Training Transfer Across Interpreting Modes? Evidence from Eye-Tracking</i>	
Jingyi Li and Yu Weng	189
<i>Using LLMs for Glossary-Constrained Terminology Correction in Machine Translation</i>	
Seyda Öztaş, Sergi Alvarez Vidal and Antoni Oliver Gonzalez	193
<i>Flow as Regulated Engagement in Chinese–English Translation: Evidence from Self-Report, Eye Tracking, and Physiological Measures</i>	
Yanze Li and Yu Weng	198
<i>Development of an Uzbek–English Parallel Corpus Search System: A Case Study of Literary Works</i>	
Shahlo Khamroyeva Mirdjonovna, Gulirano Baxtiyor qizi Nuriddinova, Mehriniso Kilicheva Kadirovna, Gulrukh Erkin kizi Boltaeva, Nodira Roziqovna Qurbonova and Aziza Islomovna Kamalova	203
<i>Benchmarking Multilingual Terminology Translation for UNDRR-ISC Hazard Information Profiles</i>	
Maria Carmen Staiano, Johanna Monti, Francesca Chiusaroli, Slava Tykhonov, Ken Hawkins, Hélène Jacot des Combes, Saini Yang, Urbano Fra Paleo and Virginia Murray	207
<i>Translation Quality Assessment of legal translations by MateCat and its use in future translation didactics</i>	
Edyta Żrałka	231
<i>COPECO-Speech: Integrating Speech and LLMs for Post-Editing in Translation Workflows</i>	
Jeevanthi Uthpala Liyanapathirana, Pierrette Bouillon, Jonathan David Mutal, Sabrina Girletti and Lise Volkart	238
<i>Translation Memory Systems: Evolution, Retrieval Methods, and Human-Centred Perspectives</i>	
Alicia Picazo-Izquierdo, Salima Lamsiyah and Ruslan Mitkov	245
<i>Project HRVOJE-M: integration of Croatian LLM with a TBMS</i>	
Marko Tadić and Mladen Stojak	254

Do readers prefer AI-generated Italian short stories?

Michael Farrell

IULM University

Milan

Italy

michael.farrell@iulm.it

Abstract

This study investigates whether readers prefer AI-generated short stories in Italian over one written by a renowned Italian author. In a blind setup, 20 participants read and evaluated three stories, two created with ChatGPT-4o and one by Alberto Moravia, without being informed of their origin. To explore potential influencing factors, reading habits and demographic data, comprising age, gender, education and first language, were also collected. The results showed that the AI-written texts received slightly higher average ratings and were more frequently preferred, although differences were modest. No statistically significant associations were found between text preference and demographic or reading-habit variables. These findings challenge assumptions about reader preference for human-authored fiction and raise questions about the necessity of synthetic-text editing in literary contexts.

1 Introduction

In a recent study, Farrell (2026) asked 69 professional translators to distinguish between AI-generated short stories in Italian and a human-written text. Nearly as many participants mistook the synthetic texts (SynTs) for the human-authored text (HAT) as correctly identified them. This may suggest that the participants perceived the SynTs as better written, more *human-like* and maybe even preferred them.

Similarly, a study by Zhang and Gosline (2023) found that advertising content generated in English by AI, as well as human-created English advertising content augmented by AI (i.e.,

automatically edited), was perceived as higher quality than content produced solely by human experts. Likewise, a study by Porter and Machery (2024) revealed that AI-generated English poetry is indistinguishable from human-written verse and is rated more favourably.

To test whether some people prefer reading Italian SynT, it was decided to conduct an experiment on a group of volunteer readers at a public library using the same short stories used in Farrell's previous experiment

2 Aims

The principal objective of this experiment was:

- To determine whether people prefer reading short stories in Italian generated by AI or a short story written by an eminent human author.

There were also two secondary aims:

- To determine whether age, gender, educational background, first language, reading frequency, kind of things people normally read and things they prefer to read affect their preference or otherwise for AI-generated text.
- To shed light on the need for synthetic-text editing (STE), which aims to make the text "more engaging" and give it "a more human voice" (Farrell, 2025b). If people prefer reading unedited SynT, this form of editing might be unnecessary, at least for the kind of texts examined in this study.

3 Method

The experiment was conducted at the public library (*Biblioteca Civica "F. Pezza"*) in Mortara, Pavia, Italy and the participants were recruited through

leaflets distributed on-site and a local newspaper article promoting the event a few days before it took place. The study was presented as an investigation into the factors influencing readers' preferences for particular texts.

The participants were asked to read three short stories in Italian of similar length (1807, 1626 and 1331 words). It was estimated that it would take about 24 minutes to read all three texts, considering an average reading speed of about 200 words per minute. Since they were also asked to answer questions and write assessments, in practice they were allowed about 45 minutes.

The participants were not told that any of the texts were generated by AI to prevent them from expressing their opinions on the basis that they might have suspected the texts to be synthetic. The titles were removed and replaced with geometric shapes (an oval, a hexagon and a five-pointed star) so that the texts could be identified during the experiment. These shapes were chosen since they lack a natural order and therefore do not suggest a ranking.

Each participant received an envelope containing the three short stories in random order, a questionnaire, a chart showing a few example opinions one might want to express when criticizing a short story and a pen. At the end of each short story, there was a form to be filled in with a score representing how much they had liked the story from 0 (I did not like it at all) to 10 (I liked it very much), together with a space in which they could explain the reason for the score they assigned (in free form). The participants were also encouraged to underline or mark the parts of the text which had informed their opinions. Moreover, they were asked whether they recognized the story or thought they had already read it, or whether they believed they could identify the style of a specific author.

The opinion expressed had to be personal. Therefore, consultation with other people, including other participants in the experiment, was not permitted. Internet consultation via laptop, smartphone or other devices was also prohibited.

The questionnaire asked for some demographic data (age, gender, educational background and first language), information on reading habits (reading frequency on a scale from 1 to 10 and kinds of texts read most frequently) and personal preferences (in

the case of a free choice, kinds of text read most willingly). Lastly, consent for the use and publication of anonymized data was requested on a separate form to ensure participant anonymity.

At the end of the experiment, the participants placed all the materials received back into the envelope, sealed it and returned it to the researcher.

After the envelopes were returned, there was a short debriefing to satisfy the curiosity of the participants regarding the names of the authors and the titles of the stories, during which it was revealed that two of the texts were generated by AI.

3.1 Texts used

The two AI-written texts used in this experiment were generated by ChatGPT-4o for two earlier studies (Farrell, 2025a; 2026), which examined whether postgraduate translation students and professional translators, respectively, were able to distinguish SynT from HAT. The stories were created using prompts designed to emulate the plot and style of the human-written text used in the same experiments and were not machine translations of an existing source text. Full details of the prompts and prompt-engineering techniques are provided in the first cited paper.

The texts were labelled with the same geometric shapes used in Farrell's second experiment: the SynT marked with a hexagon corresponded to ST5 in his first study, while the one marked with a five-pointed star corresponded to ST2. The human-written text was once again Alberto Moravia's short story *L'incosciente* (The Reckless Man), from *Racconti romani* (Roman Tales, 1954), and was marked with an oval. It was chosen for its brevity and because it predates machine translation and AI, which ensures that these technologies could not have played any role in its creation.

As in his second experiment, involving professional translators, the three short stories were not divided into consecutive short extracts but were presented in full to the participants.

All three texts were analysed using the Plagamme AI detector¹ to determine whether any objectively measurable differences existed between them.

¹ www.plagamme.com

Text	Kind	Length (words)	Mean score	Standard deviation	1st place rankings	AI detector score
Oval	Human	1807	6.83	± 1.71	6	17%
Hexagon	Synthetic	1626	7.33	± 2.32	9	94%
Star	Synthetic	1331	7.42	± 1.93	7	83%

Table 1: Scores assigned to short stories

3.2 Statistical analyses

Two exact tests were used to analyse the demographic and reading-habits data. The Fisher–Freeman–Halton exact test was computed using Dr. Daniel Soper’s calculator², while Fisher’s exact test was computed using Jeremy Stangroom’s calculator³, except for the comparison marked with an asterisk in Table 2, for which Navendu Vasavada’s calculator⁴ was used since it allows for zero cell values, unlike Stangroom’s. Spearman’s rank correlation coefficient (ρ) was used to examine correlations with the results of the previous experiments, while a scatterplot was used to compare the findings with AI detector data.

4 Results

4.1 Scores

Twenty volunteers chose to attend the public library for the experiment reported in this study. Two were excluded for not rating all three short stories. Table 1 shows the mean score and number of first-place⁵ rankings assigned to each short story.

4.2 Reasoning given for the scores assigned to the oval text (HAT)

4.2.1 Positive comments (common)

The readers frequently praised the narrative flow and dialogue, describing it as engaging, fluid and psychologically credible. Several noted the humour, original narrative perspective (first-person) and well-developed characters. Some appreciated the depth of themes like conscience and awareness, which were handled with sensitivity.

4.2.2 Negative comments (common)

Many criticized the slow or unengaging pacing and overly mechanical structure. A few found the language either too simplistic or mildly irritating,

despite being appropriate for the characters. Some found the text prolix, confusing or monotonous, lacking suspense or brilliance.

4.2.3 Neutral comments (occasional)

A couple of comments were descriptive rather than evaluative, e.g., noting the main character’s redemption or expressing curiosity about an unfamiliar word.

4.3 Reasoning given for the scores assigned to the hexagon text (SynT)

4.3.1 Positive comments (common)

Frequently described as well-written, clear and emotionally resonant. Multiple readers appreciated the structure, fluid style and the depth of emotional development. The text was often seen as engaging, with a strong sense of narrative progression.

4.3.2 Negative comments (occasional to common)

Some readers felt it lacked originality, describing it as clichéd, banal or too moralistic. A few pointed out superficial character psychology or found its themes predictable and unexciting.

4.3.3 Neutral comments (occasional)

A few remarks described the story’s structure or moral development in neutral or factual terms, without clear praise or criticism.

4.4 Reasoning given for the scores assigned to the star text (SynT)

4.4.1 Positive comments (common)

Many found the writing fluid, immersive and emotionally effective. Several highlighted the moral depth or psychological insight. The conclusion and introspective tone were often praised.

² www.danielsoper.com/statcalc/calculator.aspx?id=58

³ www.socscistatistics.com/tests/

⁴ astatsa.com/FisherTest/

⁵ Joint first place was permitted; therefore, the total exceeds the number of participants.

Demographic	Oval (p value / test)	Hexagon (p value / test)	Star (p value / test)
Age group	.413 / FFH	.242 / FFH	1.000 / FFH
Gender	.569 / F	1.000 / F	.057 / F*
Educational level	.510 / FFH	.615 / FFH	1.000 / FFH
First language	.200 / F	.242 / F	1.000 / F

Table 2: Demographic variables (FFH = Fisher–Freeman–Halton exact test; F = Fisher's exact test)

4.4.2 Negative comments (occasional to common)

Some criticized the predictability of the plot, stylistic repetition or emotional flatness. A few were clearly frustrated by tropes (e.g., the female temptress) or found the story preachy and uninspired.

4.4.3 Neutral comments (rare)

A single note mentioned similarity with one of the other short stories in the experiment, without a clear judgment.

4.5 Author or text recognized

When reading the Oval story, one reader was reminded of an English author, “the one who wrote about a boy obsessed with Tony Hawk who gets his girlfriend pregnant.” The description fits Nick Hornby. Cesare Pavese was named as the possible author of the Hexagon text, and both Cesare Pavese and Italo Calvino were associated with the Star text.

4.6 Demographic data

Despite clear instructions, only fifteen participants completed the demographic and reading-habits questionnaire. The respondents were divided into two roughly equal groups based on their ratings of the Oval story on a 0–10 scale (0 = strong dislike, 10 = strong appreciation). Scores below 6 were classified as low ($n = 7$), and scores of 6 or above as high ($n = 8$). These groups were then crossed with the categorical bands defined in the following subsections for each demographic variable, and contingency tables were built for each of the three stories. On the basis of these tables, as shown in Table 2, no statistically significant associations

were found between any demographic variables and the scores assigned to the short stories ($p > .05$).

4.6.1 Age group

The participants were grouped into three 20-year age bands.

4.6.2 Gender

All the women assigned high preference scores to the star text, compared with only half of the men; this difference approached statistical significance (Fisher's exact $p = .057$). Neither of the other two texts came close to showing a gender association.

4.6.3 Educational level

The various qualifications reported were grouped into lower secondary, upper secondary and tertiary level.

4.6.4 First language

All the participants reported Italian as their first language, with three exceptions: two bilingual speakers (Italian–Romanian and Italian–French) and one native Spanish speaker. For the purposes of analysis, first language was categorized as Italian versus non-Italian/bilingual. Although both bilingual participants and the native Spanish speaker assigned high scores to the Oval and Star texts, the small size of this group prevented this pattern from reaching statistical significance.

4.7 Reading habits

The same preference rating categories defined for the demographic data were applied to the reading-habits data. As shown in Table 3, no statistically significant associations were found between

Reading habit	Oval (p value / test)	Hexagon (p value / test)	Star (p value / test)
Reading frequency	1.000 / F	1.000 / F	.476 / F
Typical reading	1.000 / FFH	1.000 / FFH	1.000 / FFH
Preferred reading	.782 / FFH	1.000 / FFH	.352 / FFH

Table 3: Reading-habit variables (FFH = Fisher–Freeman–Halton exact test; F = Fisher's exact test)

reading-habit variables and the scores assigned to the short stories ($p > .05$).

4.7.1 Reading frequency

The participants rated their reading frequency on a 1–10 scale (1 = rarely, 10 = constantly). Since the lowest score reported was 5, responses were grouped into moderate (5–7) and high (8–10) frequency for the purposes of analysis.

4.7.2 Typical reading

The participants were asked about the kinds of materials they typically read. Given the limited sample size, the results were grouped into three mutually exclusive categories: fiction, non-fiction and both.

4.7.3 Preferred reading

The participants were also asked what they would prefer to read if given free choice. When no preference was stated, typical reading was used as a proxy, and responses were categorized using the same three groups.

5 Discussion

The experiment took place in the researcher's hometown (just under 15,750 inhabitants). Consequently, some of the participants knew him, and in at least two cases – but probably only two – were aware of his research into AI-generated texts. One of them noted on her feedback form that she believed some of the stories were AI-generated and correctly identified both SynTs. However, since she still gave them medium to high scores (5 and 7), her data was retained. No other participants mentioned AI in their feedback.

The mean scores assigned to each text fall within a narrow range: the HAT (Oval) received an average score of 6.83, Hexagon 7.33 and Star 7.42, all out of 10. While the SynTs slightly outperformed the human-authored one on average, the differences were modest. In terms of reader preference, Hexagon was most frequently ranked first (9 times), followed by Star (7 times) and Oval (6 times). These outcomes suggest no clear disadvantage for AI-generated Italian fiction under blind conditions.

A scatterplot comparing human mean scores with Plagiarism scores for the three texts (Chart 1) shows that while the human scores are tightly clustered, the AI-detector scores vary widely, from

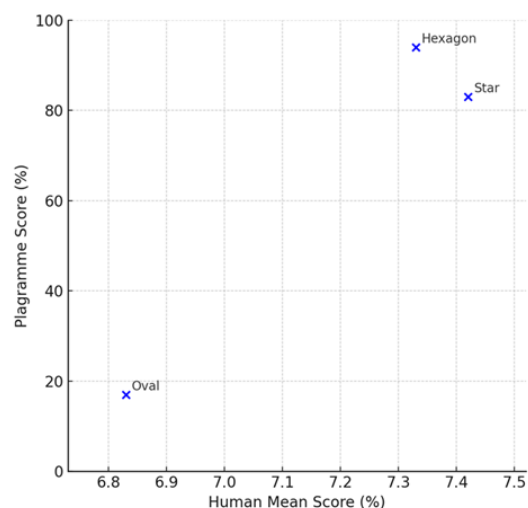


Chart 1: Scatterplot comparing human mean scores with Plagiarism scores

17% to 94%. Consequently, there is no real alignment or consistent relationship between the two scores.

In the qualitative comments, the HAT (Oval) was frequently praised for its narrative voice, humour and character development, though some criticized its pacing and verbosity. Hexagon was described as fluid and emotionally engaging but occasionally clichéd or overly moralistic. Star was noted for its immersive style and psychological insight yet drew some criticism for predictability and stylistic repetition. Despite these varied evaluations, no text received unequivocal acclaim or rejection. The balance of praise and criticism suggests that all three texts were assessed primarily on their perceived literary and emotional qualities.

The SynTs were associated with Cesare Pavese, and in one case also with Italo Calvino, who were contemporaries and editorial colleagues at Einaudi. Calvino's early neorealist work shares affinities with both Pavese and Moravia, particularly in its portrayal of wartime and youth. Interestingly, Pavese was linked to both AI-generated texts, perhaps due to a perceived tone of introspection, existential reflection or emotional restraint. The fact that the real Italian story was instead associated with a contemporary English novelist, Nick Hornby, suggests that the AI imitations may have appeared more "literary" or stereotypically Italian than the authentic text.

None of the readers in this study commented on the grammatical errors or the English syntactic and

lexical influence⁶ observed in Farrell’s previous experiments with the same texts (2025a; 2026). This difference may be attributed to the task: here, participants were simply asked to indicate their preferred text, whereas in the earlier studies they were explicitly asked to identify AI-generated writing. It may also suggest that the professional and trainee translators, who took part in the earlier experiments, pay attention to different textual features than casual readers, whose focus may be more on content and emotional resonance than on linguistic accuracy.

Regarding first language, it is interesting to observe that two participants identified as bilingual. In Farrell’s earlier study involving 69 professional translators (2026), none identified as bilingual. This may reflect a difference in how bilingualism is interpreted. While professional translators clearly speak more than one language, when asked to report their first language, they typically indicate their dominant one.

As reported in the results, both bilingual participants and the native Spanish speaker fell within the high-score group when assessing the Oval text (HAT), although this difference was not statistically significant. This may suggest a subtle trend. One possible explanation might lie in the calques, semantic loans and syntactic transfer from English observed in the SynTs (Farrell, 2026). It may be that bilingual people and native Spanish speakers are more appreciative of the distinctive features of an authentic Italian text. At the same time, however, this interpretation is contradicted by the high scores they also gave to the Star text.

There appears to be no clear explanation for the possible association between gender and preference for the Star story, which, in any case, did not reach the conventional threshold for significance ($p < .05$). No other demographic or reading-habit variables approached significance, suggesting they are unlikely to have influenced

preferences among the texts used in this experiment.

5.1 Comparison with the results of the previous experiments

As can be seen in Table 4, the HAT received the lowest preference rating from the readers and was rated as most likely to be AI-generated by the postgraduates in Farrell’s previous study (2025a). In contrast, one of the AI-generated texts (Star) received the highest preference rating and was judged by both the postgraduates and professional translators to be the most *human-like* (Farrell, 2025a; 2026). The postgraduates’ AI-likeness scores exhibit a perfect inverse correlation with the reader’s preference rankings (Spearman’s $\rho = -1.0$), while the correlation between reader and translator scores is moderately negative ($\rho = -0.5$, not statistically significant), and the correlation between postgraduates and translators is moderately positive ($\rho = 0.5$). These patterns suggest a potential alignment between reader preference and the perception of human authorship, regardless of whether the text was in fact human-written. However, the small sample size (only three texts) limits the statistical power, and the correlations – while mathematically strong – may reflect chance alignments rather than robust trends.

5.2 Need for synthetic-text editing

The rationale behind STE rests on two key assumptions: first, that readers can reliably distinguish between SynT and HAT; and second, that they prefer reading HAT.

Regarding the first assumption, one participant – whose identity was revealed indirectly through the content of her feedback – explicitly identified the AI-generated texts, suggesting that detection is possible. This participant taught French at a university in Milan and regularly assessed student work for unauthorized AI use. Her background may have contributed to her ability to detect the

Text	Kind	Reader ranking	Postgrad ranking	Translator ranking	AI detector ranking
Oval	HAT	3	1	2	3
Hexagon	SynT	2	2	1	1
Star	SynT	1	3	3	2

Table 4: Comparison of rankings from all three experiments and the AI detector

⁶ Although it may seem surprising to mention English syntactic and lexical influence in the context of AI-generated Italian text rather than machine translation,

Farrell (2026) hypothesizes that the observed effect may result from the well-documented predominance of English-language data in ChatGPT’s training.

stylistic and structural cues typical of synthetic writing. This observation, along with findings from previous studies (Dou et al., 2022; Farrell, 2025a; 2026), indicates that some people are probably able to identify AI-generated texts, particularly in professional or evaluative contexts.

The findings of this study challenge the second assumption. Both AI-generated texts received higher mean scores and more first-place rankings than the HAT, and the distribution of reader praise and criticism was relatively balanced across all three. These results are in line with those of Farrell's second study, which also found that some readers may actually prefer SynT when authorship is not disclosed.

The textual anomalies observed in the Italian SynTs during Farrell's earlier experiments (2025a, 2026) included low burstiness (the natural variation in sentence length and rhythm typical of human writing), as well as narrative contradictions and continuity errors. Similar issues have been reported in English SynTs (Dou et al., 2022). Farrell also noted influences from English syntax and lexical calques, along with Italian grammatical errors. However, this study shows that casual readers do not appear to readily detect these faults. Moreover, grammar tends to be less of an issue in English SynT, which exhibits a high degree of grammatical and orthographic accuracy (Radtke and Rummel, 2025).

In light of these findings, STE may not be critical for the kinds of texts analysed in this study. This does not imply, however, that STE is unnecessary in high-stakes or accuracy-sensitive domains, such as medicine, law or science, where textual errors could have serious consequences. It may also function much like the preference for handmade over mass-produced products: driven not by necessity, but by a desire for authenticity.

5.3 Limitations

This experiment was limited to a small selection of texts of a similar kind in a single language and was carried out on a small number of people. As a result, the findings and conclusions of this study may not be broadly generalizable.

6 Conclusion

The findings of this experiment suggest that, in blind conditions, readers may prefer AI-generated short stories in Italian to those written by established human authors. The slightly higher

ratings and first-place rankings for the SynTs indicate that such narratives can meet or even exceed the expectations of casual readers in terms of engagement, clarity and emotional resonance. Notably, the specific SynTs chosen for this study were the ones most often mistaken for human-written in previous experiments. In other words, we seem to like what we perceive to be human, even when that perception is mistaken. No demographic or reading-habit factors were found to influence these preferences significantly.

While STE may remain important in high-stakes or technical domains, its necessity for creative fiction – at least in Italian – appears less clear-cut. Nevertheless, given the small number of texts, these interpretations should be treated as tentative. Further study with more material and a broader reader base is needed to validate these observations and explore the robustness of the relationship between preference and perceived – though not necessarily actual – human authorship.

Acknowledgments

The author gratefully acknowledges the support of the staff at the *Biblioteca Civica 'F. Pezza', Civico.17*, in Mortara, Pavia, Italy, as well as the volunteer participants whose contributions made this study possible.

References

- Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith and Yejin Choi. 2022. *Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text*. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 7250–7274, Dublin, Ireland. Association for Computational Linguistics.
- Michael Farrell. 2025a. *Can postgraduate translation students identify machine-generated text?* Proceedings of Machine Translation Summit XX Volume 1, 432–441. Geneva, 23–27 June.
- Michael Farrell. 2025b. *Editing synthetic text from generative artificial intelligence: two exploratory case studies*. Proceedings of the 46th Conference Translating and the Computer, Luxembourg, 18–20 November, 2024.
- Michael Farrell. 2026. *Can professional translators identify machine-generated text?* In Proceedings of the 27th Annual Conference of the European Association for Machine Translation (EAMT 2026): Translators & Users Track, Tilburg, 15–18 June 2026.

- Brian Porter and Edouard Machery. 2024. *AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably*. Sci Rep 14, 26133.
- Anna Radtke and Nikol Rummel. 2025. *Generative AI in academic writing: Does information on authorship impact learners' revision behavior?* Computers and Education: Artificial Intelligence, 8, Article 100350.
- Yunhao Zhang and Renée Gosline. 2023. *Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human-GAI collaboration in persuasive content generation*. Cambridge University Press.

Semantic typology and audio description.

A corpus-based study of motion events

Teresa Molés-Cases

Universitat Politècnica de València /

Valencia (Spain)

temocal@upv.es

Abstract

Audio description for the blind and visually impaired has received little attention within the Thinking-for-speaking and the Thinking-for-translating hypotheses, which have been mainly tested through narrative texts and motion event lexicalization. This study examines how motion events (i.e., action scenes) are lexicalized in the English and Spanish audio descriptions of the Irish animated film *Wolfwalkers*. The contribution seeks to determine whether the Spanish audio description is the result of intersemiotic or interlinguistic transfer and it examines how the audio-describer's mother tongue, time constraints, and audio description guidelines jointly shape the film experience for the visually impaired.

1 Introduction

Audio description (AD) is a mode of audiovisual translation that provides access to visual information for blind and visually impaired audiences by means of verbal narration. Because AD must fit into the pauses between dialogue, film narration and sound effects, audio-describers are required to select and verbalize visual information under strict temporal constraints. Among the various types of information conveyed in AD, motion events are particularly salient, as action

constitutes a substantial part of audiovisual narratives (cf., Romero-Muñoz, 2023:26¹).

Cross-linguistic research has shown that languages differ systematically in how they lexicalize motion events. These typological differences raise important questions for audio description, especially when ADs are produced in typologically distinct languages or when they are translated from an existing AD. Differences in the encoding of motion may affect how vividly and accurately action scenes are reconstructed in the listener's mind, potentially leading to divergent reception experiences across languages (e.g., Bilbo scuttles off indoors vs. *Bilbo entra en su casa* 'Bilbo enters in his house', *The Hobbit: An Unexpected Journey*, Molés-Cases, 2024:291).

This contribution examines the lexicalization of motion events in ADs in English and Spanish. Building on previous research on motion events in multimodal texts, the study adopts a corpus-based approach to examine how manner, path and boundary-crossing are encoded in AD, and to what extent typological factors, medium-specific constraints, and transference interact in shaping the final product.

¹ According to Romero-Muñoz (2023:26), action, or the mobility code (Chaume, 2012), made up 44.05 % of the signs in four Spanish AD scripts from Netflix (Spain).

2 Theoretical background

2.1 Audio Description as audiovisual translation

Benecke (2004) defines audio description as “a technique used for making theatre, movies and TV programmes accessible to blind and visually impaired people [...] [it] describes the action, body language, facial expressions, scenery and costumes” (2004:78). Because AD must be inserted into the pauses between dialogue, narration and relevant sound effects, it is subject to severe temporal constraints that directly affect linguistic choice.

AD is typically conceived as a form of intersemiotic translation, since meaning is transferred from the visual to the verbal mode (Jakobson, 1959). However, this characterization does not fully capture the diversity of AD production practices. In some contexts, audio descriptions are created independently in each language from the audiovisual product, whereas in others they are translated or adapted from a pre-existing AD script in another language (Fryer, 2016; Matamala, 2019). This distinction is particularly relevant for cross-linguistic research, since translated ADs may exhibit interference effects or structural convergence resulting from the source script.

Audio description is commonly regarded as a mode of audiovisual translation (Díaz-Cintas & Remael, 2007; Fryer, 2016):

AD is increasingly recognized as a constituent part of audiovisual translation and complements subtitling for people who are deaf [...] by providing access to audiovisual media for people who are blind or partially sighted (Fryer, 2016:2)

Audio description practice is shaped by international (e.g., ADLAB project, Remael et al., 2015) and national guidelines (e.g., in Spain, UNE 153020, AENOR, 2005; in the United Kingdom, ITC Guidance on Standards for Audio Description, ITC, 2000; Ofcom’s Guidelines on the Provision of Television Access Services, 2021, among other). These guidelines generally promote clarity, conciseness, and syntactic simplicity, while encouraging vivid lexical choices that enhance mental imagery without introducing subjective interpretation or unnecessary redundancy (Bittner,

2014; Holsanova, 2016). Such recommendations interact with language-specific grammatical resources, and may therefore yield systematic differences in the way information is selected and encoded across languages.

2.2 Motion events and typology

Most AD guidelines agree on the factors that should be covered by audio description, one of which is motion, that is, “action, time or anything that moves or is indicated by movement” (Bittner, 2014:1). The analysis of motion events in this study is based on Talmy’s (1985, 2000) framework for motion event typology. According to this author, a motion event is “a situation containing movement or maintenance of a stationary location. The basic motion event consists of one object (the ‘Figure’) moving or located with respect to another object (the reference-object or ‘Ground’)” (Talmy, 1985:80-81). Talmy (1985) details the semantic components that together make up a motion event. Basic components include Figure, Motion, Path and Ground. In addition to these, motion events may incorporate optional components such as Manner and Cause. Among these semantic components, Talmy considers Path the “schematic core” of a motion event².

In an attempt to classify languages according to how they express path, Talmy (1985, 2000) refers to two big typological groups: satellite-framed languages and verb-framed languages. In satellite-framed languages such as English, path is commonly expressed outside the verb, for instance through particles or prepositional phrases (i.e. satellites), while the verb itself encodes manner (e.g., Ben limped in). This pattern enables the frequent and detailed expression of manner-of-motion distinctions. In contrast, verb-framed languages such as Spanish tend to encode path in the verb, relegating manner to optional adjuncts or leaving it unexpressed altogether (e.g., Ben *entró cojeando* ‘Ben entered limping’).

Boundary-crossing events constitute a particularly revealing context for cross-linguistic comparison. In English, boundary-crossing does not restrict the use of manner-of-motion verbs, which can freely co-occur with path particles (e.g., Robin [...] slips through the iron bars). Here, *slip* encodes manner, while the adverb *through*

² Although Talmy (2000) writes the semantic components of motion events with initial capital letters (Figure, Motion, Path, Ground, Manner, and Cause), lowercase letters will be

used from this point on for the sake of consistency and clarity.

specifies the trajectory and crossing of a boundary. The resulting description is dynamic and precise, yet remains relatively economical in terms of duration. In Spanish, however, boundary-crossing strongly constrains verb choice. As a result, manner is often neutralized or replaced by more general verbs of path (e.g., *Sale de la cueva* ‘He goes out of the cave’).

These typological tendencies have been shown to influence not only grammatical structure but also narrative or rhetorical style. Research on narrative translations has consistently found that satellite-framed languages display greater manner salience and lexical diversity, whereas verb-framed languages prioritize path and spatial configuration (c.f., [Alonso, 2018](#); [Cifuentes-Férez, 2013](#); [Ibarretxe-Antuñano, 2003](#)). Audio description, as a time-constrained and highly selective form of narration, constitutes a particularly revealing testing ground for these typological effects.

2.3 Thinking-for-speaking and Thinking-for-translating

[Slobin’s \(1987\)](#) notion of Thinking-for-speaking offers a cognitive explanation for these cross-linguistic regularities. According to this view, language production requires speakers to attend to those aspects of experience that are most readily expressible within the grammatical and lexical resources of their language. Consequently, speakers of different languages develop habitual patterns of event construal that reflect typological affordances.

[...] each language provides a limited set of options for the grammatical encoding of characteristics of objects and events. Thinking-for-speaking involves picking those characteristics that (a) fit some conceptualization of the event, and (b) are readily encodable in the language ([Slobin, 1987:435](#)).

In the domain of motion, speakers of satellite-framed languages tend to focus on manner, while speakers of verb-framed languages are more likely to foreground path. In later studies on original narrative texts in English and Spanish, [Slobin \(1996\)](#) confirmed that satellite-framed languages present a richer and more diverse and more frequently used, lexicon of manner-of-motion than verb-framed languages.

The relevance of Thinking-for-speaking extends beyond spontaneous speech to include planned and constrained forms of language use, such as written narration and audiovisual translation. Although

audio description is not produced spontaneously, describers must nevertheless make decisions about which aspects of a visual scene to encode linguistically, often under time constraints. As a result, typologically conditioned preferences are likely to surface in AD, even when describers aim to follow standardized guidelines.

In translation contexts, [Slobin \(1987, 2000\)](#) introduces the notion of Thinking-for-translating, which highlights the cognitive effort involved in re-conceptualizing events from a source language into a target language. Translators must negotiate between the conceptual patterns activated by the source text and the linguistic affordances of the target language.

The translator always faces the problems of thinking for reading and thinking for writing. It is necessary to go from the mental image produced by reading the source language to an appropriate rendering of that mental image in the target language. Translation reveals the extent to which the target language can accommodate itself to the source language, versus assimilating the original to its own patterns of thinking for writing ([Slobin, 2000:123](#)).

In the case of audio description, Thinking-for-translating is particularly relevant when an AD script is translated from another language, as describers may be influenced by the explicit encoding choices of the source AD while simultaneously having to comply with the constraints of the target language and the audiovisual medium.

Taken together, Thinking-for-speaking and Thinking-for-translating offer a complementary framework for analysing motion events in audio description. While the former accounts for cross-linguistic differences rooted in typology, the latter helps explain variation arising from translation practices, temporal limitations, and the multimodal nature of audiovisual products. These considerations led to the proposal of the Thinking-for-audio-describing phenomenon ([Molés-Cases, 2024](#); [Molés-Cases & Feist, 2025](#)). In [Molés-Cases \(2024\)](#) the English and Spanish ADs of the film *The Hobbit: An Unexpected Journey* (2012) are examined. And in [Molés-Cases & Feist \(2025\)](#) a German-Spanish corpus of films aimed at children and young adults is analysed. Overall the findings of these studies show that manner-of-motion information is more frequent and diverse in audio descriptions in satellite-framed languages than in verb-framed languages. It is confirmed that this difference stems from the combined influence of

the describer's native language and the constraints and guidelines governing audio description.

3 The study

3.1 Research questions

The study addresses the following research questions:

- Is the Spanish audio description analysed the result of intersemiotic or interlinguistic transfer?
- Do the English and Spanish audio descriptions differ in the amount and richness of motion information conveyed?
- If differences are observed, are they attributable to language typology, AD-specific constraints, or a combination of both?
- Do sound effects in the audiovisual product compensate for manner-of-motion information not included in the audio description?

3.2 Materials

The corpus analysed in this study consists of the English and Spanish audio descriptions of the Irish animated film *Wolfwalkers* (Moore & Stewart, 2020), both available on Apple TV (Spain). The film has a running time of 103 minutes and contains a high density of action.

3.3 Methodology

The English and Spanish audio-described versions of the film were transcribed using Riverside³, an AI-based app. After a manual review of both transcriptions, all the instances of self-agentive motion events (specifically, translational motion⁴, e.g., She hops clumsily from roof to roof, and local wander⁵, e.g., *Robin y su padre caminan por el pueblo* 'Robin and her father walk through town') present in the audio descriptions were identified and annotated by the author. Each event was analysed in terms of:

- inclusion or exclusion of manner information;
- semantic component(s) encoded in the verb (manner, path, manner + path, or motion);
- presence or absence of boundary-crossing;
- temporal constraints, measured as the amount of time available between AD segments and dialogue/narration or sound (to this end, the speech rate parameters validated by Pellegrino et al., 2011 were adopted and Time Code Reading was annotated).

In addition, diagnostic criteria were applied to determine whether the Spanish AD was the product of intersemiotic or interlinguistic transfer. Based on Romero-Muñoz (2024), these criteria included the alignment of silence gaps, structural similarity, and the presence of calques or interference phenomena.

4. Analysis and results

4.1. Source of transfer

In order to determine whether the Spanish audio description had been produced directly from the image or from the original English audio description, the descriptions at the beginning of each audio description (English and Spanish) were compared, and the differentiation parameters proposed by Romero Muñoz (2024, 2025) were then applied. The three parameters indicated by this author were confirmed:

- a) Similarity in silence gaps, and more specifically, similar times of beginning and end of each description (e.g., TCR 00:02:12, A huge figure approaches him, *Una gran figura se le acerca*; TCR 00:02:55, He sprints towards the nearby town, *Corre hacia un pueblo cercano*).
- b) Similarity of linguistic structures (e.g., A buck stops in the middle of a path and stares into the distance. An axe swings into a tree. An enormous tree falls to the ground and three men cut it into pieces. *Un ciervo se detiene un medio de un camino y mira hacia un lado. Un hacha corta a un árbol. Un gran*

³ <https://riverside.com/>

⁴ According to Talmy (2000: 35), translational motion refers to "motion in which the location of the Figure changes in the time period under consideration".

⁵ According to Talmy (1991: 489), local wander is "motion in which the average position of the Figure does not vary in relation to the ground".

árbol cae al suelo. Tres hombres lo cortan en trozos más pequeños ‘A deer stops in the middle of a path and looks to one side. An axe cuts into a tree. A large tree falls to the ground. Three men cut it into smaller pieces’).

- c) Potential calques (e.g., lack of indefinite article at the beginning of sentence in Spanish, *Conejos grises saltan por el suelo* ‘Grey rabbits hop on the ground’; *Soldados salen corriendo del pueblo*, ‘Soldiers exit the town running’).

The study confirms thus that the Spanish audio description shows evidence of interlinguistic transfer from the original English audio description.

4.2. Amount of motion information

The quantitative analysis shows that the English audio description contains 269 motion event descriptions, while the Spanish audio description includes 301. However, while the English AD tends to condense multiple semantic components into a single clause, the Spanish AD more frequently distributes information across simpler structures or prioritizes only the most essential component of the motion event, typically path. This difference becomes particularly salient in fast-paced action scenes, where temporal constraints are strongest.

4.3. Manner-of-motion information

Clear differences emerge when examining the richness of motion descriptions, especially with regard to the inclusion of manner. A percentage of 66.91% of the English motion event descriptions include explicit manner-of-motion information (e.g., *The wolves swarm around the large man*), whereas only 35.88% of the Spanish descriptions do so (e.g., *Corre hacia un pueblo cercano* ‘He runs to a nearby village’). This asymmetry is consistent with the typological distinction between satellite-framed English and verb-framed Spanish.

4.4. Verb choice

Regarding the verb lexicon, as expected according to the Thinking-for-speaking hypothesis, the English version includes many more manner verbs and multiword expressions than the Spanish one (49 versus 6 types, e.g., *slink*, *caminar* ‘walk’). However, the Spanish version includes a similar

number of path verb types than the English AD (27 versus 26 types, e.g., *regresar* ‘come back’, *come*). This result is unexpected according to the Thinking-for-speaking hypothesis, but it should be borne in mind that only a single film was analyzed here and that it has been confirmed that the Spanish AD has been created from the English AD. In the analysis verbs and expressions including both manner and path information in English and Spanish (6 versus 5 types, e.g., *escape*, *trepár* ‘climb’) and a general motion verb in English (1 type, e.g., *move*) were identified too.

4.5. Boundary-crossing

Boundary-crossing events were identified in the two versions (61 cases in English and 37 in Spanish). Manner-of-motion was encoded in the majority of these in English (75.40%) (e.g., Robin steps through her bedroom window), but it was encoded in the minority of cases in Spanish (18.91%) (e.g., *Ruedan juntas hacia el interior de la cueva* ‘[They] roll together into the cave’), being just the lexicalization of path more frequent (e.g., *Sale de la cueva* ‘He goes out of the cave’).

4.6. Temporal constraints and their impact

Temporal limitation emerges as a key explanatory factor in the Spanish data. A total of 89.11% of Spanish motion event descriptions that do not include manner occur in segments where one second or less is available for narration. Given the relatively high syllabic rate of Spanish (cf. [Pellegrino et al., 2011](#) found that Spanish speakers produce approximately 8 syllables per second), describers appear to favour structurally simple and semantically economical constructions. This suggests that this restriction inherent in the medium of AD may have influenced describers’ likelihood of including manner information.

4.7. Compensation of manner through sound effects

It was also examined whether the audiovisual product included sound effects that could compensate for missing manner-of-motion information in the Spanish audio description. This was observed in 10 cases, resulting in an overall compensation of manner through sound effects of 5.18% (e.g., *Avanza entre los árboles* ‘He advances through the trees’, sound effect: walking in the forest; *Sube al escenario* ‘He goes on stage’, sound effect: panting and straining).

4.8. Discussion

Taken together, the results suggest that differences between the English and Spanish audio descriptions cannot be attributed solely to typological differences. While Thinking-for-speaking accounts for the higher salience of manner in English, medium-specific constraints and translation-related factors play a crucial role in shaping the Spanish AD too.

The Spanish audio description reflects a convergence of verb-framed lexicalization patterns, temporal restrictions inherent to AD, and the influence of an English source text. This combination leads to reductions in manner information, which may result in less vivid mental representations of action scenes for Spanish-speaking audiences.

At the same time, the findings highlight the adaptive nature of audio description as a practice situated at the intersection of language, cognition, and multimodality.

5. Conclusions

This study set out to examine how motion events are conceptualized in audio description, focusing on the comparison between English and Spanish and drawing on the frameworks of Thinking-for-speaking and Thinking-for-translating. The results demonstrate that audio description is a cognitively and linguistically complex activity in which typological preferences, medium-specific constraints, and translation-related processes interact in different ways.

The contrastive analysis shows that the English audio description consistently encodes manner-of-motion more frequently and more diversely than the Spanish audio description. This finding aligns with well-established typological distinctions between satellite-framed and verb-framed languages and supports the applicability of Thinking-for-speaking to multimodal texts such as AD. English describers tend to exploit the compactness of verb-particle constructions to integrate manner and path within tight temporal windows, resulting in vivid and information-dense motion descriptions.

In Spanish, by contrast, manner information is significantly reduced, particularly in boundary-crossing events and under severe time constraints. Besides, the strict timing requirements of audio description, combined with the higher syllabic rate

of Spanish, seem to encourage describers to prioritize path and core motion over more fine-grained manner distinctions. From a Thinking-for-translating perspective, some degree of interference has been observed in the Spanish AD.

These findings suggest that differences in motion event encoding across languages should not be interpreted simply as losses or deficiencies. Rather, they reflect strategic adaptations to the affordances and constraints of both the target language and the audiovisual medium. Audio description emerges as a situated practice in which describers negotiate between linguistic norms, cognitive habits, and accessibility guidelines in order to ensure narrative coherence and listener comprehension.

In addition, the study confirms that the lexicon of manner verbs was both more diverse and more frequent in the English audio description than in the Spanish one. The analysis also identified a) cases with time constraints and b) cases where the presence of sound effects in the audiovisual product compensated for the effects of Thinking-for-speaking.

Finally, the study is not without limitations. The analysis is based on a single audiovisual product, and future research should extend the scope to additional languages, films, genres, and production contexts in order to assess the generalizability of the findings. Further work could also integrate reception studies to explore how differences in manner encoding affect listeners' mental representations and overall reception experience.

Overall, this research contributes to the growing body of work on motion events in audiovisual translation and highlights the relevance of cognitive-typological approaches for understanding audio description.

Acknowledgments

This study was supported by the research project CIGE/2022/153, funded by the Conselleria d'Innovació, Universitats, Ciència i Societat Digital, Generalitat Valenciana (Spain). <https://teresamolescases.com/pagines/project1.htm>

References

AENOR. 2005. *UNE 153020, Audiodescripción para personas con discapacidad visual. Requisitos para la audiodescripción y elaboración de audioguías.*

<https://tienda.aenor.com/norma-une-153020-2005-n0032787>.

Alonso Alonso, Rosa. 2018. Translating motion events into typologically distinct languages. *Perspectives*, 26(3): 357-376. <https://doi.org/10.1080/0907676X.2017.1387578>

Benecke, Bernd. 2004. Audio description. *Meta: Journal des traducteurs*, 49(1):78-80. <https://doi.org/10.7202/009022ar>

Bittner, Hansjörg. 2014. *Audio description guidelines – A comparison*. Unpublished manuscript. https://www.uni-hildesheim.de/media/_migrated/content_uploads/AD_Guidelines_Comparison_-_Read.pdf

Cifuentes-Férez, Paula. 2013. El tratamiento de los verbos de manera de movimiento y de los caminos en la traducción inglés-español de textos narrativos. *Miscelánea: A Journal of English and American Studies*, 47: 53-80. https://doi.org/10.26754/ojs_misc/mj.20139226

Díaz-Cintas, Jorge and Aline Remael. 2007. *Audiovisual Translation. Subtitling*. St. Jerome, Manchester.

Fryer, Louise. 2016. *An introduction to audio description: A practical guide*. Routledge, London.

Holsanova, Jana. 2016. A cognitive approach to audio description. In Anna Matamala and Pilar Orero (Eds.), *Researching audio description*, 49-73. Palgrave Macmillan, London.

Ibarretxe-Antuñano, Iraide. 2003. What translation tells us about motion: A contrastive study of typologically different Languages. *IJES, International Journal of English Studies*, 3(2): 151-175.

ITC, 2000. *ITC Guidance on standards for audio description*. <https://dcmp.org/learn/355-the-acb-audio-description-project-guidelines>.

Jakobson, Roman. 1959. On linguistic aspects of translation. In Reuben A. Brower (Ed.), *On translation*, 232-239. Harvard University Press, Cambridge, Massachusetts.

Matamala, Anna. 2019. *Accessibilitat i traducció audiovisual*. Eumo, Barcelona.

Molés-Cases, Teresa. 2024. Thinking-for-audio-describing: motion events in the film *The Hobbit: An Unexpected Journey*. *Acta Linguistica Hafniensia* 56(2): 280-297. <https://doi.org/10.1080/03740463.2024.2413789>

Molés-Cases, Teresa and Michele Feist. 2025. Feeding the imagination: Linguistic features of motion descriptions in audio-described movies. *Language*

Sciences

<https://doi.org/10.1016/j.langsci.2025.101728>

110.

Ofcom. 2021. *Guidelines on the provision of television access Services*. https://www.ofcom.org.uk/_data/assets/pdf_file/0025/212776/provision-of-tvaccess-services-guidelines.pdf

Pellegrino, François, Christophe Coupé and Egidio Marsico. 2011. A cross-language perspective on speech information rate. *Language*, 87(3):539-558.

Remael, Aline, Nina Reviers and Gert Vercauteren. (Eds.). 2015. *Pictures painted in words: ADLAB audio description guidelines*. Edizioni Università di Trieste, Trieste. https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.openstarts.units.it/entities/publication/be626243-8361-4b94-a74e-05224e6120ae&ved=2ahUKEwiG8d6cw72SAxWohv0HHRAFCOYQFnoECCcQAQ&usg=AOvVaw2FJMvSwrJsE7Vnqo_WV5cY

Romero-Muñoz, Alejandro. 2023. Multimodal analysis as a way to operationalise objectivity in audio description: a corpus-based study of Spanish series on Netflix. *J. Audiov. Transl.* 6(2): 8-32. <https://doi.org/10.47476/jat.v6i2.2023.251>

Romero-Muñoz, Alejandro. 2024. La traducción de audiodescripciones en las plataformas de streaming. ¿Una posible herramienta para la enseñanza de lenguas? *Proceedings of 3rd International Congress: Education and Knowledge*. Octaedro, Barcelona.

Romero-Muñoz, Alejandro. 2025. *Hacia una audiodescripción inmersiva. Estudio descriptivo, comparativo, multimodal y de recepción experimental sobre la objetividad y subjetividad en audiodescripción*. Doctoral Thesis. Universitat Jaume I. <https://www.tdx.cat/handle/10803/695910#page=1>

Slobin, Dan I. 1987. Thinking for speaking. *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society*, 435-445.

Slobin, Dan I. 1996. Two ways to travel: Verbs of motion in English and Spanish. In M. Shibatani and S. Thompson (Eds.), *Grammatical Constructions: their form and meaning*, 195-220. Clarendon Press, Oxford.

Slobin, Dan I. 2000. Verbalized events: A dynamic approach to linguistic relativity and determinism. In Susanne Niemeier and René Dirven (Eds.), *Evidence for linguistic relativity*, 107-138. Mouton de Gruyter, Berlin.

Talmy, Leonard. 1985. Lexicalization patterns: Semantic structure in lexical forms. In Timothy Shopen (Ed.), *Language typology and syntactic description*,

Vol. 3, 57-149). Cambridge University Press, Cambridge.

Talmy, Leonard. 1991. Path to realization: A typology of event conflation. In *Proceedings of the Seventeenth Annual Meeting of Berkeley Linguistics Society*, 480-519, Berkeley.

Talmy, Leonard. 2000. *Toward a cognitive semantics*. MIT Press, Cambridge, Massachusetts.

How Well Can Large Language Models Detect, Classify, and Correct Errors in Chinese-English Public Sign Translations?

Shuyin Zhang

The Chinese University of Hong
Kong, Shenzhen
zhangshuyin@cuhk.edu.cn

Xindi Hao

The Chinese University of Hong
Kong, Shenzhen
xindihao@link.cuhk.edu.cn

Zhenhai Li

Hong Kong Polytechnic University
25145728r@connect.polyu.hk

Abstract

This study evaluates six frontier large language models on their ability to detect, classify, and correct errors in Chinese-English public sign translations. Model performance is assessed through both automatic and human evaluation, using a 100-item test set constructed from authentic bilingual signage in China. Results show that all models perform strongly in error detection but exhibit complementary strengths in classification and correction, with no single model consistently outperforming others across all error types. This suggests that a multi-model workflow may be more effective than relying on any single model.

1 Introduction

Public signage, the language displayed on public road signs, place names, institutional signs, and government notices in a given territory (Landry & Bourhis, 1997), is an important component of urban governance, international communication, and public services of a city (Shenzhen Foreign Affairs, 2026). In China, Chinese-English bilingual public signs are widely used and generally fall into two categories: the names of public places and institutions, and public service information (SAC & AQSIQ, 2013). Previous studies have summarized the features of Chinese-English public signs—concise, standard, straightforward, and sometimes culture-specific (Amenador & Wang, 2022), which, to a certain

extent, points to the need for explicit guidelines to ensure such qualities are consistently achieved. In Shenzhen, China's most vibrant international metropolis, bilingual public signs are particularly prevalent, and the regulation of public signage translation has been in place for more than a decade. Nevertheless, rapid urban development and increasing international communication have both made existing expressions outdated and led to the emergence of new types of public signs. In response, Shenzhen has recently issued updated guidelines for public signage translation. However, a large number of existing signs across the city remain non-compliant and require review and revision to meet the new standards.

The need to standardize public signage translation and the availability of relevant guidelines provide a basis for exploring the feasibility of partly automating the review and revision process. This study therefore explores whether large language models (LLMs) can be used to pre-screen translations and provide suggestions. To this end, it evaluates the performance of six LLMs—three Chinese models and three English models—in handling errors in public sign translations. Based on a dataset of 100 bilingual public signs, comprising translations with intralingual, interlingual, intercultural, and multiple errors, as well as correct translations, the study assesses the models' abilities across three tasks: error detection, error classification, and error correction. Error detection and error classification are assessed through automatic evaluation. For error correction, since intralingual errors such as

misspellings and capitalization mistakes are well within the capabilities of current LLMs, evaluating correction of these errors would be uninformative; the study therefore focuses correction evaluation on interlingual and intercultural errors. Specifically, automatic evaluation is employed to check whether LLM-corrected translations exactly match the reference translations; human evaluation, as the “gold standard” (Graham et al., 2015, p. 1183), is then conducted for cases without an exact match, providing a more comprehensive assessment of correction quality. By identifying well-performing LLMs, this study aims to support the standardization of bilingual public signage and provide empirical evidence for the broader application of LLMs in translation evaluation.

2 Related Work

Recent studies have demonstrated the potential of LLMs for machine translation and translation evaluation. Kocmi and Federmann (2023) introduced GEMBA, which uses GPT models for reference-free evaluation. Fernandes et al. (2023) applied LLMs to MQM error annotation, and Lu et al. (2024) proposed error analysis prompting, bringing LLM evaluations closer to human judgments. On a broader level, Zheng et al. (2024) established the LLM-as-a-Judge framework, investigating its reliability and potential biases, which provides a theoretical foundation for applying LLMs to translation quality assessment. Moreover, Raunak et al. (2023) showed that GPT-4 can carry out effective automatic post-editing even without large-scale triplet data. Research on public signage translation has primarily focused on error taxonomy and textual analysis, yet few studies have applied LLMs to reviewing public signage translations. The present study connects these lines of inquiry by addressing the underexplored intersection of LLM-based quality assessment and correction in the context of public signage translation.

3 Experimental Setup

3.1 Dataset Construction

The test set of 100 Chinese–English public sign pairs was constructed from an annotated corpus of authentic public signage in Shenzhen, China. The corpus was developed through the routine correction and revision work: errors were

identified and corrected by translators, reviewed by editors and validated by domain experts, and a comprehensive error typology was derived inductively from this process. The full typology also covers categories such as inconsistent translation across signs, formatting errors, source-text errors, missing translations, and redundant signage. These were excluded from the present study, as they concern source-text quality, signage management, or visual presentation rather than translation quality, and cannot be reliably assessed when the model receives only an isolated sign pair without visual or contextual information.

The study therefore focuses on three categories most relevant to evaluating how LLMs diagnose and correct errors in public sign translations: Intralingual Errors (target-language issues identifiable without reference to the source text), Interlingual Errors (problematic transfer between the source and target languages), and Intercultural Errors (inadequate adaptation to the target audience’s cultural and cognitive expectations). These broadly correspond to the MQM dimensions of Linguistic Conventions, Accuracy, and Audience Appropriateness (Lommel et al., 2014). The test set also includes a Multiple Errors category, containing signs exhibiting two or more of the primary error types, and a Correct category, which tests whether models can recognize proper translations without over-correcting them. The final set is evenly distributed across the five categories (20 instances each). Table A presents the five categories.

Category	Description
Intralingual Errors	misspellings, capitalization errors, punctuation and spacing problems, and grammatical errors
Interlingual Errors	awkward literal translations, semantic distortion, register mismatches, functional inequivalence, and terminology errors
Intercultural Errors	cultural schema conflicts, cultural taboos, and identity sensitivities
Multiple Errors	Two or more of the above error types in the same sign
Correct	No identifiable errors

Table A. Categories in the test set

3.2 Prompt Design

The prompt was designed to test three core abilities of LLMs: error detection (judging whether a translation is compliant), error classification (categorizing identified errors), and error correction (providing an appropriate recommended translation). It comprised five components—role and task definitions, a classification taxonomy, working rules, five output examples, and the *General Rules*—that together operationalized these abilities. The prompt was written in Chinese because the *General Rules* on which all judgments were based are themselves in Chinese; the other four components were translated into English for presentation in this paper (see Appendix A).

The prompt began with the role and task definitions that cast the model as an expert bilingual public signage proofreader tasked with identifying non-compliance with the *Shenzhen Guidelines for the Use of English in Public Service Areas—Part 1: General Rules* (Foreign Affairs Office of Shenzhen Municipal People’s Government, 2026) and either providing a recommended translation or confirming correctness. It then presented a classification taxonomy requiring the model to assign a category to each identified issue.

The working rules operationalized the three abilities through a structured JSON output format. Each issue had to include four fields: “error_type” (testing classification); “location” and “reason” (grounding detection judgments in specific clauses); and “suggestion” (testing error correction). Errors were listed individually rather than aggregated, because a model might correctly identify one error while missing or misclassifying others within the same test item; independent evaluation of each error therefore provided a more accurate measure of model performance. An “overall_category” field provided a summary label for each test item, which enabled batch-level analysis. Five few-shot examples were included, covering all possible output scenarios: three single-type errors, one multiple-error case, and one correct case. Few-shot prompting was adopted because the task required a specific JSON output format and a domain-specific error taxonomy that the model might misinterpret from instructions alone. Finally, the core provisions of

the *General Rules* were supplied, covering translation principles, translation methods, and writing conventions, thereby providing the standard against which compliance was judged.

3.3 Model Selection

The models to be tested were selected from the Chatbot Arena leaderboards (<https://arena.ai/leaderboard>), a crowdsourcing platform where users rate the performance of LLMs via pairwise comparisons in real-world cases. The large volume of user ratings, combined with advanced ranking algorithms, makes the leaderboards a more reliable reflection of human preferences (Chiang et al. 2024a: 1-3). The present study drew specifically from the Writing, Literature & Language sub-board of the Text Arena leaderboards under the Occupational category (accessed on March 15, 2026), which captures user assessments of model performance on tasks central to writing, literature, and language occupations – including translation – making it a fitting basis for model selection (Chiang et al. 2024b; Arena Team 2025). The top three models from organizations based in English-speaking countries (Claude Opus 4.6, Gemini 3.1 Pro Preview, and Grok 4.20 Beta) and in China (GLM-5, Qwen-3.5-397B-A17B, and Kimi-K2.5) were selected to compare how frontier LLMs from the two groups perform across the three abilities assessed in this study.

3.4 Process

3.4.1 Data Collection

Each message sent to the LLMs was constructed by appending the paired source and translated texts below the prompt, with brief labels indicating the source text and the translation (see Appendix B for a sample input and response). For error detection, items with a non-empty issues array were treated as the model having detected at least one error, while an empty issues array and an “overall_category” of “correct” were treated as error-free judgments; the detection score was computed across all 100 items by comparing these judgments against the human references. For error classification, each “error_type” value (e.g., “interlingual”) was extracted from the issues array and compared against the human reference. For error correction, the suggested revised translation was extracted

from the “suggestion” field. Approximate time and cost for each model were also recorded.

3.4.2 Evaluation of LLM Performance

Three scoring metrics were used for evaluation. Error detection awards one point per item if the model correctly identifies it as containing or not containing a translation error, regardless of whether the error categories match those assigned by human translators. Error classification awards one point if all identified error types exactly match the human-assigned categories, 0.5 points if at least one category matches, and zero points otherwise; for items in the correct category, a model receives one point if it likewise identifies no error. Error correction compares the model’s suggested translation against the human reference, awarding one point only if any of the suggested revisions matches the reference translation identically. Error correction scores were computed only for the interlingual and intercultural categories (40 items), since the correct category contains no errors to correct, intralingual errors involve surface-level issues whose correction is well within the capabilities of all tested models, and the multiple errors category involves overlapping types that make consistent scoring difficult to define.

Each of the first two metrics was computed separately for items within each category and then aggregated into an overall score ranging from 0 to 100; the error correction score ranged from 0 to 40. Error detection and error classification were evaluated automatically. Error correction was first evaluated by automatic exact-match comparison and then supplemented by human judgment. As a single Chinese public sign may have multiple proper translations, two of the authors independently reviewed the LLM-corrected translations for items in the interlingual and intercultural error categories that did not pass the automatic exact-match evaluation, each assessing whether the revised translations conformed to the *General Rules* without prior discussion. Inter-annotator agreement was then calculated using Cohen’s Kappa before any disagreements were resolved. Disagreements were subsequently discussed until a consensus on the final judgment was reached. Both authors have participated in the review of public signage in Shenzhen for more than ten months and have been accredited as “professional public signage translators” by the Translation Association of Shenzhen. For each

revised translation, the model received one point if the translation was considered proper and zero points if it was improper.

3.4.3 Data Analysis

Given the limited test set size, non-parametric and permutation-based statistical tests were adopted, which remain valid without large-sample assumptions (Dror et al. 2018). To examine whether the six models exhibited significant performance differences overall on each metric ($n = 100$ for detection and classification; $n = 40$ for correction) and within each error type sub-category, Friedman omnibus tests were conducted. As the omnibus test indicates only whether the models differ overall without identifying which ones differ, it was followed by paired sign-flip permutation tests between models, with Holm correction applied to control the error rate and identify the differentiating pairs. To examine whether the model’s performance differed across the three tasks, the same procedure was followed. For every model, a Friedman omnibus test was conducted on the 40 items for which all three metrics were defined; Holm-corrected paired sign-flip permutation tests were then performed to identify which task pairs differed.

4 Results

Tables B and C present the scores and corresponding rankings of the six models across the first two metrics, and Table D presents the automatic exact-match score for error correction. The results in Table D are preliminary, as the final error correction scores can only be obtained after incorporating human evaluation.

Error detection scores were high across all models and in the interlingual, intralingual, intercultural, and multiple-error categories, with most models achieving perfect or near-perfect scores. The primary source of variation was the correct category, where scores ranged from 11 to 14. These differences drove the overall detection rankings, in which Gemini 3.1 Pro Preview led with 92 and Claude Opus 4.6 trailed with 87. Error classification revealed sharper differences. In the intralingual category, the three English models occupied the top three positions (Claude Opus 4.6 17.5, Gemini 3.1 Pro Preview 16.5, Grok 4.20 Beta 16.5), while the two Chinese models, Kimi-K2.5 and Qwen-3.5-397B-A17B, scored

Overall (100/100)		
Ranking	Model	Score
1	Gemini 3.1 Pro Preview	92
2	Qwen 3.5 397B A17B	91
3	GLM-5	90
4	Grok 4.20 Beta	90
5	Kimi-K2.5	89
6	Claude Opus 4.6	87
Intralingual Errors (20/20)		
Ranking	Model	Score
1	Grok 4.20 Beta	20
2	Qwen 3.5 397B A17B	20
3	Claude Opus 4.6	19
4	Gemini 3.1 Pro Preview	19
5	GLM-5	18
6	Kimi-K2.5	18
Interlingual Errors (20/20)		
Ranking	Model	Score
1	Claude Opus 4.6	20
2	GLM-5	20
3	Gemini 3.1 Pro Preview	20
4	Grok 4.20 Beta	20
5	Kimi-K2.5	20
6	Qwen 3.5 397B A17B	20
Intercultural Errors (20/20)		
Ranking	Model	Score
1	Kimi-K2.5	20
2	GLM-5	19
3	Gemini 3.1 Pro Preview	19
4	Grok 4.20 Beta	18
5	Qwen 3.5 397B A17B	18
6	Claude Opus 4.6	17
Multiple Errors (20/20)		
Ranking	Model	Score
1	Claude Opus 4.6	20
2	GLM-5	20
3	Gemini 3.1 Pro Preview	20
4	Grok 4.20 Beta	20
5	Qwen 3.5 397B A17B	20
6	Kimi-K2.5	19
Correct (20/20)		
Ranking	Model	Score
1	Gemini 3.1 Pro Preview	14
2	GLM-5	13
3	Qwen 3.5 397B A17B	13
4	Grok 4.20 Beta	12
5	Kimi-K2.5	12
6	Claude Opus 4.6	11

Table B: Rankings of error detection score of the six LLMs by error category

Overall (100/100)		
Ranking	Model	Score
1	Gemini 3.1 Pro Preview	65.5
2	Grok 4.20 Beta	65
3	GLM-5	64.5
4	Qwen 3.5 397B A17B	64
5	Claude Opus 4.6	63.5
6	Kimi-K2.5	62
Intralingual Errors (20/20)		
Ranking	Model	Score
1	Claude Opus 4.6	17.5
2	Gemini 3.1 Pro Preview	16.5
3	Grok 4.20 Beta	16.5
4	GLM-5	16
5	Kimi-K2.5	14.5
6	Qwen 3.5 397B A17B	14.5
Interlingual Errors (20/20)		
Ranking	Model	Score
1	GLM-5	18
2	Claude Opus 4.6	17.5
3	Gemini 3.1 Pro Preview	17.5
4	Grok 4.20 Beta	17.5
5	Kimi-K2.5	17.5
6	Qwen 3.5 397B A17B	17
Intercultural Errors (20/20)		
Ranking	Model	Score
1	Kimi-K2.5	4
2	Claude Opus 4.6	1.5
3	GLM-5	1.5
4	Grok 4.20 Beta	1.5
5	Qwen 3.5 397B A17B	1.5
6	Gemini 3.1 Pro Preview	1
Multiple Errors (20/20)		
Ranking	Model	Score
1	Qwen 3.5 397B A17B	18
2	Grok 4.20 Beta	17.5
3	Gemini 3.1 Pro Preview	16.5
4	Claude Opus 4.6	16
5	GLM-5	16
6	Kimi-K2.5	14
Correct (20/20)		
Ranking	Model	Score
1	Gemini 3.1 Pro Preview	14
2	GLM-5	13
3	Qwen 3.5 397B A17B	13
4	Grok 4.20 Beta	12
5	Kimi-K2.5	12
6	Claude Opus 4.6	11

Table C: Rankings of error classification score of the six LLMs by error category

noticeably lower (14.5 each). The intercultural category produced the most striking contrast: Kimi-K2.5 scored 4, roughly triple the scores of the remaining five models (1 to 1.5), though all models performed poorly in absolute terms. Overall classification rankings placed Gemini 3.1 Pro Preview first (65.5) and Kimi-K2.5 last (62). The automatic exact-match scores for error correction were markedly lower than those for detection and classification across all models for both interlingual errors and intercultural errors. In the interlingual error category, GLM-5 achieved the highest score (10), while Claude Opus 4.6 and Kimi-K2.5 scored the lowest (5). In the intercultural error category, Claude Opus 4.6, GLM-5, and Gemini 3.1 Pro Preview scored 6, while the remaining three models scored 4.

Significance analysis, however, provides limited support for distinguishing among the models overall. No metric-level omnibus distinguished the six models on detection, classification, or correction (Friedman $p = 0.47$, 0.85 , and 0.13 , respectively), so the overall orderings should be read descriptively. Disaggregating by error type, the only reliable difference was in intercultural classification (Holm-adjusted $p = 0.016$), reflecting Kimi-K2.5’s descriptively higher score relative to all other models. The task-level gradient, by contrast, was strongly supported: every model’s performance differed significantly across the three metrics (per-model Friedman $p < 0.001$), with detection exceeding both classification and correction for all six (all $p < 0.001$). Grok 4.20 Beta and Kimi-K2.5 also performed significantly better on classification than on correction (Holm-adjusted $p = 0.029$ and 0.003).

Model	Interlingual Error Score	Intercultural Error Score
GLM-5	10	6
Gemini 3.1 Pro Preview	8	6
Qwen 3.5 397B	8	4
Grok 4.20 Beta	6	4
Claude Opus 4.6	5	6
Kimi-K2.5	5	4

Table D. Automatic exact-match scores for the error correction scores

Overall (40/40)		
Ranking	Model	Score
1	Gemini 3.1 Pro Preview	33
2	Kimi-K2.5	32
3	GLM-5	30
4	Grok 4.20 Beta	26
5	Qwen 3.5 397B A17B	24
6	Claude Opus 4.6	23
Interlingual Errors (20/20)		
Ranking	Model	Score
1	Gemini 3.1 Pro Preview	18
2	Kimi-K2.5	17
3	GLM-5	15
4	Grok 4.20 Beta	15
5	Qwen-3.5-397B-A17B	15
6	Claude Opus 4.6	11
Intercultural Errors (20/20)		
Ranking	Model	Score
1	Gemini 3.1 Pro Preview	15
2	GLM-5	15
3	Kimi-K2.5	15
4	Claude Opus 4.6	12
5	Grok 4.20 Beta	11
6	Qwen-3.5-397B-A17B	9

Table E. The sums of the automatically-evaluated and human-evaluated proper translations of the six LLMs under the interlingual and intercultural error categories

For non-matching translations in the interlingual and intercultural error categories, the two raters showed high inter-annotator agreement, with Cohen’s Kappa of 0.90 for interlingual errors, 0.89 for intercultural errors, and 0.90 overall. Table E presents the numbers of LLM-revised translations judged to be proper translations in the two categories, combining the results of automatic and human evaluation. Compared with the automatic exact-match results in Table D, the number of revised translations deemed proper increased for all six models in both categories after human evaluation was incorporated. For interlingual errors, Gemini 3.1 Pro Preview correctly revised 18 of the 20 translations, ranking first among the six LLMs, while Claude Opus 4.6 produced only 11 revised translations deemed proper, the lowest number among the six models. For intercultural errors, Gemini 3.1 Pro Preview, GLM-5, and Kimi-K2.5

Model	Approx. Time	Credit Spent
Claude Opus 4.6	1.5 h	\$8.84
Gemini 3.1 Pro Preview	1 h	\$5.03
Qwen 3.5 397B A17B	3.5 h	\$4.06
GLM-5	3.5 h	\$3.87
Kimi-K2.5	1.5 h	\$3.01
Grok 4.20 Beta	1 h	\$2.13

Table F. Approximate evaluation time and credit spent

each produced 15 properly revised translations out of 20, while Qwen-3.5-397B-A17B generated only 9, ranking last. Table F summarizes the approximate time and cost of the evaluation task for each model. Grok 4.20 Beta incurred the lowest cost at \$2.13, while Claude Opus 4.6 incurred the highest at \$8.84. In terms of time, Qwen 3.5 397B A17B and GLM-5 were the slowest at 3.5 hours each, while Grok 4.20 Beta and Gemini 3.1 Pro Preview were the fastest at approximately one hour.

5 Discussion

5.1 Error Detection Ability

According to Table B, though the margin is narrow, Gemini 3.1 Pro Preview achieved the highest score among the six LLMs, indicating a tendency to demonstrate the strongest overall performance in error detection in this exploratory study, whereas Claude Opus 4.6 gained the lowest score—suggesting a trend to exhibit the weakest performance in this category. All six LLMs identified the 20 translations with interlingual errors as containing errors, and five of the six, all except Kimi-K2.5, also identified all 20 translations with multiple errors as improper. These results indicate that top-ranking models developed in both China and English-speaking countries are generally capable of identifying public sign translations with interlingual or multiple errors as improper. However, all six LLMs achieved their lowest scores in the correct category, suggesting a tendency to flag correct translations as erroneous. Within this category, Gemini 3.1 Pro Preview obtained the highest score, while Claude Opus 4.6 achieved the lowest. Although the margin is only three points, it may partly explain the

comparatively weaker performance of the latter in overall error detection.

5.2 Error Classification Ability

As shown in Table C, Gemini 3.1 Pro Preview achieved the highest score among the six LLMs; in contrast, Kimi-K2.5 obtained the lowest score, suggesting their relatively strongest and weakest overall ability in error classification respectively. With regard to the classification of translations with intralingual errors, Claude Opus 4.6 tended to demonstrate the strongest performance, while Kimi-K2.5 and Qwen-3.5-397B-A17B, which obtained the lowest scores among the six LLMs, inclined to demonstrate the weakest performance. Notably, the three English models occupied the top three positions, outperforming the Chinese models. This suggests that English models tend to perform better than Chinese models in classifying intralingual errors, such as misspellings, capitalization errors, and punctuation or spacing issues. Regarding the classification of translations with interlingual errors, the six LLMs showed broadly comparable performance. The highest-ranked model, GLM-5, scored only one point higher than the lowest-ranked model, Qwen-3.5-397B-A17B, and the remaining four models obtained identical scores. As for intercultural errors, Kimi-K2.5 achieved a higher score than all the other five LLMs, while Gemini 3.1 Pro Preview gained the lowest score among the six models, with a margin of three points. More broadly, all six LLMs obtained their lowest scores in this category, and these scores were much lower than the corresponding detection scores, indicating that the models can detect that translations containing intercultural errors are erroneous but have difficulty classifying this specific type of error. For translations with multiple errors, Qwen-3.5-397B-A17B achieved the highest score in error classification, whereas Kimi-K2.5 gained the lowest score among the six LLMs, with a 4-point discrepancy. Moreover, all six LLMs obtained slightly lower scores in error classification than for error detection scores in this category. This indicates that although LLMs are highly capable of recognizing translations with multiple errors as improper, they may struggle to accurately classify all error types within a given translation.

5.3 Error Correction Ability

As shown in Table E, Gemini 3.1 Pro Preview achieved the highest score in correcting translations with interlingual errors, indicating that it may be suitable for providing revision suggestions for public sign translations in this category. In contrast, Claude Opus 4.6 gained the lowest score—seven points lower than Gemini 3.1 Pro Preview—suggesting a relatively limited ability to correct interlingual errors. For translations with intercultural errors, Gemini 3.1 Pro Preview, GLM-5, and Kimi-K2.5 gained the identically highest score, whereas Qwen-3.5-397B-A17B obtained a comparatively lowest score in revising such translations. Moreover, when error correction performance and the time and cost incurred by the six LLMs throughout the process are considered together, Gemini 3.1 Pro Preview not only corrected the highest number of translations with interlingual and intercultural errors, but also required the least time, suggesting it may be a relatively efficient choice for the error correction of public sign translations in this exploratory study. However, Claude Opus 4.6 incurred the highest cost among the six models, while Qwen-3.5-397B-A17B required considerably more time to generate outputs. Therefore, these two models may be less suitable choices for the error correction of public sign translations.

5.4 Implications

The results present a statistically supported gradient of difficulty across the three abilities for the six models progressing from detection through classification to correction. Several patterns merit further interpretation.

The over-detection observed across all models could potentially be mitigated through prompt engineering, for example, by adding explicit instructions discouraging unnecessary corrections. Further research could compare the effectiveness of different prompting strategies in improving detection precision.

The contrasting classification strengths of English and Chinese models point to a broader practical opportunity. English models outperformed Chinese models in classifying intralingual errors, likely because their training on larger and more diverse English-language corpora gives them stronger command of English orthographic and typographic conventions. That a

Chinese model achieved the strongest intercultural classification performance suggests complementary strengths across models from different linguistic backgrounds. A multi-model workflow, in which each model handles the error types it classifies most accurately, could exploit these complementarities. Such a workflow could also address the difficulty all models exhibited with multi-error translations, where they tended to identify the most salient error while overlooking co-occurring ones; assigning each error type to a dedicated model might reduce the risk of less prominent errors being overlooked.

The pronounced gap between intercultural detection and classification scores may reflect not an inherent architectural limitation but rather insufficient culturally annotated training data. Future research could explore whether fine-tuning models on datasets that explicitly encode cross-cultural translation norms leads to improved intercultural error classification. If effective, such models might complement the kind of multi-model approach suggested above. By contrast, interlingual error classification appears to be an area where the tested models performed comparably in this study, meaning that practical considerations such as processing speed and cost may serve as more relevant criteria for model selection.

Finally, strong detection or classification performance does not guarantee strong correction ability, as identifying what is wrong in a translation and producing an improved version draw on different capabilities. For public sign translation, which often involves large volumes of text across an entire city or region, time and cost are critical factors as well.

6 Conclusion

This study evaluated six frontier LLMs on their ability to detect, classify, and correct errors in Chinese–English public sign translations, based on a 100-item test set constructed from authentic bilingual signage in Shenzhen, China. A key observation is that the six models showed complementary strengths, which suggests that a multi-model workflow may be more useful than relying on a single model. At the detection stage, all six models performed strongly enough to serve as pre-screening tools for identifying non-compliant translations. At the classification and correction stages, however, performance varied

across models and error types, suggesting that model selection should be task-specific and that practical factors such as processing time and cost should also be considered. A combined workflow—in which designated models handle specific error types and human reviewers make final decisions—warrants empirical investigation as a potentially more feasible approach.

Limitations

The few-shot examples included in the evaluation prompt were selected based on the researchers' judgment, and their influence on model performance has not been systematically examined; future work could explore how different example sets or numbers of examples affect evaluation outcomes. In addition, the test set consists of only 100 items drawn from a single city, which limits the generalizability of the findings; a larger dataset would allow for more robust conclusions. Furthermore, each error category contains only 20 samples, and score differences between models are often as small as 1–2 points; a sample of this size affords very limited statistical power, so these null results should be read as insufficient evidence to separate the models rather than as evidence of genuine equivalence. A larger dataset, together with multiple experimental runs, would provide the power needed to detect true differences should they exist and to place the present rankings on firmer footing. Finally, of the six models evaluated, only GLM-5 and Kimi-K2.5 are open-source. Closed-source models undergo frequent updates that can render findings less durable over time and generally do not support fine-tuning. Since public signage translation is a specialized domain, fine-tuning on domain-specific data represents a promising direction that can only be pursued with open-source models. Future work should therefore evaluate a broader range of open-source LLMs.

References

- Arena Team. 2025. Arena expert and occupational categories. <https://arena.ai/blog/arena-expert/>
- Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 12.
- Foreign Affairs Office of Shenzhen Municipal People's Government. 2026. 深圳市公共服务领域英文译写规范通则. Shenzhen Municipal People's Government.
- Kofi B. Amenador and Zhiqiang Wang. 2022. A systematic review of the Chinese-English translation of public signs. *SAGE Open*, 12(2).
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Curran Associates Inc., New York.
- Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*. Association for Computational Linguistics, Singapore.
- Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, Bangkok, Thailand.
- Rodrigue Landry and Richard Y. Bourhis. 1997. Linguistic landscape and ethnolinguistic vitality: An empirical study. *Journal of Language and Social Psychology*, 16(1):23–49.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Melbourne, Australia.
- SAC and AQSIQ. 2013. 公共服务领域英文译写规范 第1部分：通则 Guidelines for the use of English in public service areas—Part 1: General rules (GB/T 30240.1-2013). Standards Press of China.
- Shenzhen Foreign Affairs. 2026. 《深圳市公共服务领域英文译写规范通则解读》第一讲——为何要规范公共服务英文译写? . Retrieved March 15, 2026 from https://mp.weixin.qq.com/s/AK_Jw-h0ZrqTtHpHx7A9hQ?poc_token=HLORj2mjJPBqY7p1xEVBdkKwePBYAFRhibfJ2u9G
- Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of

translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, Tampere, Finland.

Vikas Raunak, 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*. Association for Computational Linguistics, Singapore.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024a. Chatbot Arena: An open platform for evaluating LLMs by human preference. *arXiv preprint arXiv:2403.04132*.
<https://ui.adsabs.harvard.edu/abs/2024arXiv240304132C>

Wei-Lin Chiang, Tianle Li, Naman Jain, Yifan Song, Lisa Dunlap, Dacheng Li, Evan Frick, and Anastasios N. Angelopoulos. 2024b. Chatbot Arena categories: Definitions, methods, and insights.
<https://blog.lmarena.ai/blog/2024/arena-category/>

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado.

Acknowledgments

This work was supported by the 2025 Guangdong Tertiary Education Teaching Quality & Reform Project: “LLMs + Language Technology” Course Development and Industry–Academia Integration for International Communication; and The Chinese University of Hong Kong, Shenzhen 2024 Teaching Innovation Grant: “Enhancing Translation Education with AI: Leveraging Large Language Models (LLMs) to Align Teaching with Industry Demands”.

Appendix A. Prompts

The following prompt was sent to the models to generate error annotations:

你是公共服务领域中英双语公示语的译写与书写规范审校专家。你的任务是：根据给定的“译写原则/译写方法和要求/书写要求”，对输入的中文公示语的英文译文进行合规性检查，指出问题并给出推荐译文；若译文完全合规，则判定为正确。

错误分类体系

对每一处问题，按以下三类进行分类：

- **Intralingual Error**（语内错误）：目标语言自身的问题，无需参照源文即可识别。包括拼写错误、大小写错误、标点与空格问题、语法错误等。

- **Interlingual Error**（语际错误）：源语言与目标语言转换过程中产生的问题。包括机械直译、语义偏离、语体不匹配、功能不对等、术语错误等。

- **Intercultural Error**（跨文化错误）：未能充分适应目标受众的文化认知与期待。包括文化图式冲突、涉及文化禁忌或身份敏感性等。

工作规则

1. 你只依据输入中提供的规范与信息判断，不要编造不存在的依据。

2. 每条结果输出一行 JSON 对象，不要输出解释、Markdown 或代码框。

3. JSON 对象包含以下字段：

- ``issues``：问题列表（数组）
- ``overall_category``：总体分类

4. 若译文完全合规：

- ``"issues": []``
- ``"overall_category": "correct"``

5. 若不合规，逐条在 ``issues`` 中列出问题，每条必须包含：

- ``error_type``（错误类型）：intralingual / interlingual / intercultural

- ``location``（定位）：英文译文中的具体问题片段。仅当英文译文存在未译出内容时，可引用中文原文的片段。

- ``reason``（原因）：结合“译写原则/译写方法和要求/书写要求”的要点，简要说明为何违规

- ``suggestion``（推荐译文）

6. 所有问题逐条列出后，输出 ``overall_category``：

- 若所有问题属于同一类型，``overall_category`` 为该类型（intralingual/interlingual/intercultural）

- 若问题涉及两种及以上类型，
`"overall_category": "multiple"

7. 推荐译文必须简洁、符合英语公示语习惯、与中文意图一致；必要时用 // 表示换行（仅用于 WARNING/CAUTION/DANGER 等建议分行的情况）。

8. 不要输出与任务无关的内容，不要讨论训练过程。

示例

示例 1: Intralingual Error

输入：

中文：停车场

英文：Parkng Lot

输出：
{"issues":[{"error_type":"intralingual","location":"Parkng","reason":"违反规范 1.2.1，英文译写应符合英文使用规范，\"Parking\"拼写缺少字母 i。","suggestion":"Parking"}],"overall_category":"intralingual"}

示例 2: Interlingual Error

输入：

中文：来宾止步

英文：Guests Stop

输出：
{"issues":[{"error_type":"interlingual","location":"Guests Stop","reason":"违反规范 1.3.3 和 2.2.3，公共服务信息应根据内容和意图意译，使用英语同类信息的习惯用语；\"来宾止步\"属于限制进入类信息，机械直译未传达其功能。","suggestion":"Staff Only"}],"overall_category":"interlingual"}

示例 3: Intercultural Error

输入：

中文：老弱病残孕专座

英文：Seats for the Old, Weak, Sick, Disabled and Pregnant

输出：
{"issues":[{"error_type":"intercultural","location":"Old, Weak, Sick, Disabled","reason":"违反规范 1.4 文明性要求，逐一罗列身体状况带有歧视色彩，在英语文化中构成文化图式冲突，易引起冒犯。","suggestion":"Priority Seats"}],"overall_category":"intercultural"}

示例 4: Multiple Errors

输入：

中文：禁止攀爬

英文：Do Not Clime the Wall

输出：
{"issues":[{"error_type":"intralingual","location":"Clime","reason":"拼写错误，climb 误拼为 clime。","suggestion":"Climb"},{"error_type":"interlingual","location":"the Wall","reason":"违反规范 1.2.2 和 1.3.2，原文\"禁止攀爬\"未指定具体对象，译文添加了源文不存在的信息，不符合简洁明了的要求。","suggestion":"No Climbing"}],"overall_category":"multiple"}

示例 5: Correct

输入：

中文：紧急出口

英文：Emergency Exit

输出：
{"issues":[],"overall_category":"correct"}

译写原则/译写方法和要求/书写要求
(followed by General Rules)

Appendix B. Sample Input and Response

Input:

中文：爱心座椅
英文：Love Seat

Response (Claude Opus 4.6):

<pre>{ "issues": [{ "error_type": "interlingual", "location": "Love Seat", "reason": "违反规范 1.3.3 和 1.2.2，'爱心座椅'指为有需要的乘客（如老人、孕妇、残障人士等）设置的优先座位，应使用英语同类信息的习惯用语。机械直译为'Love Seat'在英文中指双人小沙发（情侣座），语义完全偏离原文功能含义。", "suggestion": "Priority Seat" }], "overall_category": "interlingual" }</pre>
--

Training Interpreting Learners to Collaborate with Captions and Subtitles during Simultaneous Interpreting

Huang Yujie

Department of Language Science and
Technology, The Hong Kong Polytechnic
University / Hong Kong Special
Administrative Region, China

24041413r@connect.polyu.hk

Xu Han

Department of Language Science and
Technology, The Hong Kong
Polytechnic University / Hong Kong
Special Administrative Region,
China

han12.xu@polyu.edu.hk

Abstract

Captions and subtitles are two promising AI-enhanced tools that can be integrated into the simultaneous interpreting (SI) process. Although previous studies have shown that these tools can improve interpreters' performance, their effective use requires specialized training. To help interpreting learners make better use of these tools during SI, the present study developed a specialized training based on learners' needs and investigated its impact on learners' interaction with captions and subtitles using eye-tracking technology. The findings reveal that 1) compared with captions, subtitles attracted less visual attention; 2) the specialized training helped learners foster a more flexible attention switching rather than heavier dependence on visual assistance.

1 Introduction and background

In the age of artificial intelligence (AI), the practice of simultaneous interpreting has undergone significant transformation ([Fantinuoli, 2023; Pöhhacker and Liu, 2024](#)). With the assistance of AI-enhanced captions and subtitles, the real-time display of source texts and translated texts, interpreters have access to more than just auditory input for information retention. The evolving technological landscape has generated considerable attention to examine the effects of

these two tools on interpreters' cognitive processing and performance quality, with some employing eye tracking technology ([Cheung and Li, 2022; Yuan and Wang, 2023, 2024](#)). While various factors, such as the presenting forms, the accuracy and latency of the captions and subtitles can influence the overall effectiveness of integration ([Li and Chmiel, 2024](#)), how interpreters interact with these tools remains a crucial determinant ([Wang et al., 2025](#)). To achieve effective and seamless engagement with captions and subtitles, interpreters must enhance skills in critical AI output evaluation and cognitive resource management. Existing studies have constantly noted the pressing need to equip interpreting learners with these skills, they only offered preliminary empirical implications for curriculum design ([Guo et al., 2025](#)), leaving the impact of specialized training largely underexplored.

The present study aims to address this important pedagogical gap. Drawing on learners' needs identified through pre-training interviews, a specialized training was developed. The training was designed to provide learners with foundational knowledge and practical experience in using captions and subtitles during SI. As part of the author's ongoing PhD project, this study focused on learners' interactions with captions and subtitles, specifically how learners view the captions and subtitles during SI before and after the training in order to assess how the training intervention influenced attention allocation nad

coordination in technology-mediated interpreting environments.

Specifically, the study is going to answer the following two research questions:

RQ1. How do interpreting learners interact with captions and subtitles during simultaneous interpreting?

RQ2. How do training affect interpreting learners' interaction with the captions and subtitles during simultaneous interpreting?

The findings of this study are expected to provide valuable insights into how educational interventions can facilitate the effective integration of AI technologies into professional interpreting practice.

2 Methodology

2.1 Experiment design

To address the research questions, two laboratory experiments (one before and one after the training with 1.5 month interval) and three training seminars were administered. 19 postgraduate students (15 female, 4 male) with interpreting as their major were recruited. Their mean age was 24.2 (SD = 2.09). All the participants were L1 Chinese and L2 English users. Their language proficiency was demonstrated by a mean IELTS score of 7.48 (SD = 0.34). In terms of prior experience in using captions and subtitles for SI, no participants reported having previously utilized these tools to assist SI.

Considering all participants are interpreting learners whose capability may not ensure them to perform SI directly without any preparation. In each lab experiment, participants were first given 15 minutes to familiarize themselves with the source texts and then required to perform the English-Chinese SI task in front of an EyeLink 1000 Plus eye-tracker, using a head-mounted setup to ensure precise tracking. The source texts used in the pre- and post- training were identical in terms of length, theme, and task difficulty. Each was segmented into three chunks, and each chunk was controlled in terms of word count, lexical density, lexical sophistication, syntactic complexity and readability to maximize comparability and then recorded. Captions and subtitles were provided by iTour Voice Translator, a platform enables real-time automatic

translation and bilingual subtitles with a low Word Error Rate (WER) of 2%. As shown in figure 1, the stimuli were presented to participants in three conditions: (1) without captions or subtitles, (2) with captions only, and (3) with subtitles only. The order of these conditions was counter-balanced across participants to mitigate order effects. Prior to each task condition, a 13-point calibration was conducted to ensure measurement accuracy. After the main tasks, a semi-structured interview was conducted to explore participants' perceptions of the effectiveness and challenges associated with using subtitles and captions during SI.

Upon finishing the first lab experiment, a semi-structured interview which involved 15 open-ended questions organized around five themes (1) preference for the three SI modes and in perceived usefulness of captions and subtitles, (2) perceived influence of captions and subtitles on interpretation quality, (3) perceived influence of captions and subtitles on cognitive load, (4) perceived influence of captions and subtitles on stress level, (5) attitudes and expectation for specialized training was administered to obtain participants' subjective experiences when collaborating with captions and subtitles for SI



Figure 1: Interfaces of interpreting tasks

2.2 Specialized training

The semi-structured interviews conducted after the first lab experiment identified three major difficulties learners encountered when using captions and subtitles during SI:

- They were unsure how to allocate their attention appropriately so as to minimize visual interference.

- They did not know how to cope with the constantly changing captions and subtitles.
- They experienced heightened stress when processing multimodal inputs.

To address these issues, three training sessions were designed and conducted before the second lab experiment. Overall, the sessions were consisted of lectures and discussions. Following Prandi's (2023) recommendations, each session combined theoretical input with guided practice. The theoretical component introduced the black box of SI, redundancy effect, and principles of attention management. The practical component encompassed structured SI exercises under captioned and subtitled conditions, with targeted feedback on learners' visual attention and informed strategies for evaluating and selectively using AI-generated outputs. In addition, an expert interpreter with over 15 years of professional experience was invited to share his experience and strategies for integrating captions and subtitles into SI. Furthermore, the training encouraged learners to identify optimal strategies for processing captions and subtitles, and to share these insights with their peers.

2.3 Eye-tracking indicators

To examine how learners allocate and coordinate their attention during SI with the assistance of captions and subtitles, we employed eye-tracking to trace their gaze in the task. Drawn inspiration from existing studies (Yuan and Wang, 2023, 2024), three eye-tracking indicators, namely dwell time, fixation count, and run count on the visual assisted area were analyzed. The proportions of dwell time and fixation count on the visual-aid areas were used to index the overall amount and intensity of visual attention allocated to the captions or subtitles, indicating how extensively learners relied on these visual aids during simultaneous interpreting. Run count measured the frequency with which learners' eyes moved to and from the captions or subtitles area, reflecting how flexibly they shifted their attention between the visual aid and the speaker. To ensure the precision of eye movement data, we followed White (2008),

removing fixations shorter than 80 ms or longer than 1,200 ms for analyses.

For each indicator, a linear mixed-effects model was fitted with tool type (subtitles vs. captions), training (pre- vs. post-training), and their interaction as fixed effects.

3 Preliminary results

As shown in Table 1, the main effect of assisted tool type on dwell time was statistically significant and negative (Estimate = -0.40, SE = 0.10, $t = -4.02$, $p < 0.001$), indicating that learners spent a smaller proportion of time viewing the subtitle area compared to the caption area. The main effect of training was not statistically significant (Estimate = -0.11, SE = 0.06, $t = -1.85$, $p = 0.07$), nor was the interaction between tool type and training (Estimate = 0.11, SE = 0.06, $t = 1.91$, $p = 0.07$). However, a downward trend in captions dwell time following training was apparent in the visualized data (see figure 2), suggesting that, after training, learners tended to focus less on the captions.

Fixed effects	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	0.91	0.09	9.96	<0.001***
AI-assisted tool	-0.40	0.10	-4.02	<0.001***
Training	-0.11	0.06	-1.85	0.07
AI-assisted tool * Training	0.11	0.06	1.91	0.07

Table 1: Effects of AI-assisted tool type and training on visual aid dwell time

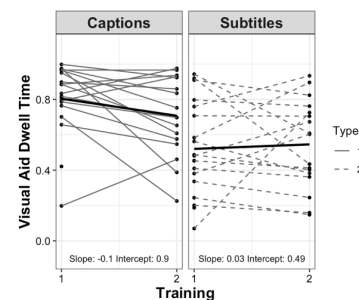


Figure 2: Variation trend in visual aid dwell time before and after training: captions vs. subtitles

Similarly, for fixation count (see Table 2), the main effect of AI-assisted tool type was statistically significant and negative (Estimate = -0.30, SE =

0.09, $t = -3.17$, $p = 0.003$), suggesting that learners made fewer fixations on the subtitles area than on the captions area. But neither the main effect of training nor the interaction effect reached statistical significance, although a downward trend of fixation count on captions can be observed in Figure 3. These findings are consistent with the dwell time results, indicating that subtitles are associated with reduced visual attention to the visual aid area compared to captions

Fixed effects	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	0.92	0.08	11.83	<0.001***
AI-assisted tool	-0.30	0.09	-3.17	0.003**
Training	-0.08	0.05	-1.61	0.12
AI-assisted tool * Training	0.08	0.06	1.34	0.19

Table 2: Effects of AI-assisted tool type and training on visual aid fixation count

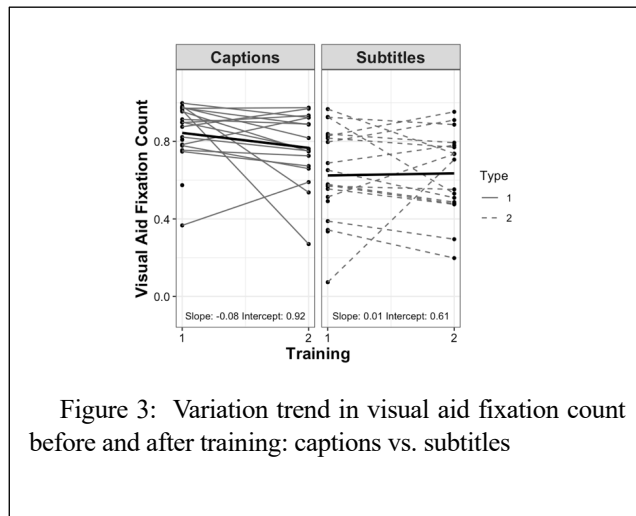


Figure 3: Variation trend in visual aid fixation count before and after training: captions vs. subtitles

As for run count (see Table 3), the main effect of training was statistically significant and positive (Estimate = 0.18, SE = 0.06, $t = 3.04$, $p = 0.005$), indicating that training led to an increased number of runs to the subtitles or captions area. The increased trend is evident as shown in Figure 3. However, the main effect of AI-assisted tool type and the interaction between tool type and training were not statistically significant (Estimate = 0.22, SE = 0.13, $t = 1.70$, $p = 0.10$), although a trend toward more runs in the subtitles condition was observable in Figure 4.

Overall, these results suggest that, in comparison to captions, subtitles are associated with less dwell time and fewer fixations on the

visual aid area, reflecting a lower visual attention allocation. Meanwhile, training did not significantly affect dwell time or fixation count on visual aid areas, but it did lead to a significant increase in run count, indicating that after training, learners were more likely to alternate their attention between the visual aid area and the speaker, rather than fixating for longer periods or making more fixations within the visual aid area itself.

Fixed effects	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	0.13	0.01	1.35	0.19
AI-assisted tool	0.22	0.13	1.70	0.10
Training	0.18	0.06	3.04	0.005**
AI-assisted tool * Training	-0.04	0.08	-0.53	0.61

Table 3: Effects of AI-assisted tool type and training on visual aid run count

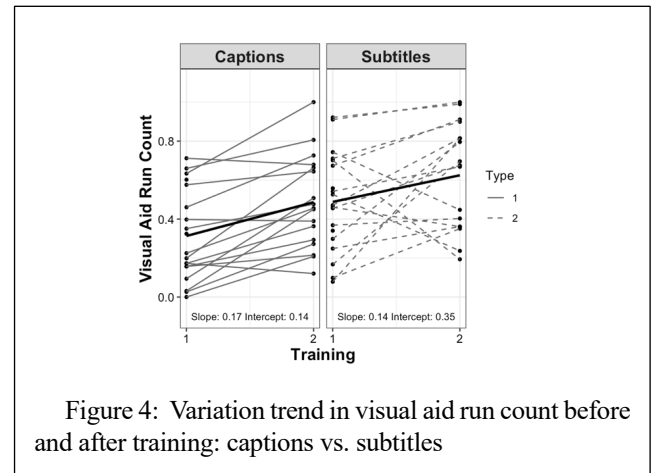


Figure 4: Variation trend in visual aid run count before and after training: captions vs. subtitles

4 Conclusion

This study investigated how the proposed training based on interpreting learners' own needs shaped their interaction with AI-enhanced captions and subtitles during SI. According to the Eye-tracking data, subtitles attracted less visual attention than captions, as measured by shorter dwell time and fewer fixations on the visual-aid are, suggesting that captions may serve as a more useful aid during SI. However, this interpretation requires further verification given many factors such as the display form of captions and subtitles may interplay.

Moreover, the study did not consider interpreting performance, ensuing work would analyze it and see how it correlates with the attention allocation to obtain a more comprehensive picture. While the training did not substantially modify the overall amount of visual attention devoted to captions or subtitles, it significantly increased run counts, indicating that learners became more adept at alternating their attention between the visual aid and the speaker. These findings suggest that targeted training can foster more flexible and strategic coordination of multimodal inputs in SI, rather than encouraging overreliance on AI-generated texts. However, given the small sample size, further research with larger cohorts and different language pairs is needed to confirm and extend these preliminary results.

Acknowledgments

The study reported in this paper forms part of the first author's ongoing PhD research, which is supervised by the second author.

References

- Andrew K.F. Cheung and Tianyun Li. 2022. Machine aided interpreting: An experiment of automatic speech recognition in simultaneous interpreting. *Translation Quarterly*, 104(2):1–20.
- Claudio Fantinuoli. 2023. Towards AI-enhanced computer-assisted interpreting. In *Interpreting Technologies – Current and Future Trends*, volume 37, pages 46–72. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Meng Guo, Yuxin Xie, Lili Han, Victoria Lai Cheng Lei, and Defeng Li. 2025. Bridging traditional and AI-assisted simultaneous interpreting: Empirical insights for curriculum design. *The Interpreter and Translator Trainer*, 19(3–4):425–443.
- Tianyun Li and Agnieszka Chmiel. 2024. Automatic subtitles increase accuracy and decrease cognitive load in simultaneous interpreting. *Interpreting*, 26(2):253–281.
- Franz Pöchhacker and Minhua Liu. 2024. Interpreting technologized: Distance and assistance. *Interpreting*, 26(2):157–177.
- Bianca Prandi. 2023. *Computer-assisted simultaneous interpreting: A cognitive-experimental study on terminology*. Berlin: Language Science Press.
- Xiaoman Wang, Binhua Wang, and Lu Yuan. 2025. The function of ASR-generated live transcription in simultaneous interpreting: Trainee interpreters' perceptions from post-task interviews. *Humanities and Social Sciences Communications*, 12(1):166.
- Lu Yuan and Binhua Wang. 2023. Cognitive processing of the extra visual layer of live captioning in simultaneous interpreting: Triangulation of eye-tracking and performance data. *Ampersand*, 11:100131. <https://doi.org/10.1016/j.amper.2023.100131>.
- Lu Yuan and Binhua Wang. 2024. Eye-tracking the processing of visual input of live transcripts in remote simultaneous interpreting: Preliminary findings. *The Forum*, 22(1):119–145. <https://doi.org/10.1075/forum.00038.yua>.

Punctuation Patterns in Translated and Original Chinese-English Literary Texts: A Weibull Distribution Analysis

Jiawei Shi

The Hong Kong Polytechnic University
jiawei.shi@connect.polyu.hk

Sihui He

The Hong Kong Polytechnic University
sihui.he@connect.polyu.hk

Abstract

Recent research reveals that punctuation patterns exhibit cross-linguistic universal features, as the intervals between consecutive punctuation marks follow a discrete Weibull distribution. While this mathematical regularity provides a valuable tool for quantitative linguistics, its application to translated texts remains underexplored. Translation alters punctuation through the tension between preserving source text style and adhering to target language structural rules. To investigate this, we constructed a bidirectional corpus of twenty-eight classic twentieth-century novels and their authoritative translations, comprising Original Chinese (OC), Original English (OE), Translated Chinese (TC), and Translated English (TE). We modeled punctuation intervals across these subcorpora using the discrete Weibull distribution. Results confirm a good fit across all subcorpora, reinforcing cross-linguistic universality. Parameter analysis reveals distinct signatures. Chinese texts exhibit lower punctuation frequency and higher rhythmic regularity, whereas English texts show the inverse. Crucially, translation effects are directionally asymmetric. Both translation directions show movement toward the respective target-language norms, but the extent of movement is asymmetric. TC shifts strongly toward Chinese original norms, whereas TE shows only a partial shift toward English norms and remains shaped by Chinese source-language rhythm. These lightweight and interpretable parameters may also serve as useful features for automatic translationese detection, thereby bridging quantitative linguistics and practical translation technology.

1 Introduction

Punctuation is a fundamental aspect of written language that enhances clarity, meaning, and expression. It consists of various symbols and marks used to organize and structure sentences, indicating pauses, emphasis, and relationships between words and ideas (Chafe, 1988). The proper use of punctuation is important for effective communication, as it helps convey the intended message accurately and ensures that readers interpret the text as intended.

Beyond its functional role in grammar and style, this organizational capacity makes punctuation a measurable architectural feature of texts. Such structural complexity aligns with the broader view of language as a system where “more is different” (Anderson, 1972: 393). A small inventory of phonemes and letters generates words, which in turn form sentences, allowing the full richness of semantics to emerge. Quantitative analyses have identified a family of statistical laws governing these structures. For example, Zipf’s law describes word frequency distributions, Heaps’ law relates vocabulary size to text length, and the Menzerath and Altmann law captures relationships across hierarchical linguistic units (Altmann, 1980; Heaps, 1978; Zipf, 1949). Within this quantitative landscape, punctuation has also attracted attention. Although punctuation marks differ in nature from words, they obey Zipf’s law, and their inclusion in frequency analyses restores the power-law scaling of word frequency distributions (Kulig et al., 2017). Beyond frequency, full stops are the principal carriers of long-range correlations in narrative texts (Drożdż et al., 2016), confirming that punctuation functions not merely as an orthographic convention but as a structurally integral component of written language.

An informative approach is to study inter-punctuation intervals, that is, the number of words between consecutive punctuation marks. Literary

texts provide an ideal domain for these quantitative investigations. Authors possess stylistic freedom to manipulate punctuation for rhythmic and expressive purposes, generating a rich diversity of interval lengths. This variance effectively captures the full statistical spectrum of punctuation behavior. Indeed, analyses of literary prose across seven major Western languages reveal that these interval distributions follow the discrete Weibull distribution (Nakagawa and Osaki, 1975) with near universal regularity (Stanisz et al., 2023). In Western languages, this regularity typically emerges when all punctuation marks are considered. The two Weibull parameters, p and β , encode demonstrably language-specific information. English has been reported to be among the least constrained languages in placing consecutive punctuation marks, with Spanish showing a similar tendency, while other languages occupy distinct regions of the parameter space (Stanisz et al., 2023). This language specificity persists even in experimental literary works, though texts such as *Finnegans Wake* present thicker distribution tails and a decreasing hazard function not observed in typical literary prose (Stanisz et al., 2024a).

Extending this Weibull framework to Chinese requires attention to the historical and structural specificity of its punctuation system. Unlike Western punctuation, which developed organically within an alphabetic tradition, modern Chinese punctuation was largely imported from the West. Systematic adoption occurred only in the 1920s and 1930s, and many canonical pre-modern works were retroactively punctuated during this period (Dolina et al., 2025). This backdrop creates quantitative challenges because sentence segmentation in Chinese remains unsettled in natural language processing (Zhou et al., 2025). The historical ambiguity between commas and periods often blurs the line between clauses and complete sentences, as the Chinese comma can function like an English period in many contexts (Kong and Zhou, 2014). Furthermore, the minimal unit of measurement may be either the individual character or the compound word, yielding different interval calculations (Liu et al., 2023).

Despite these structural challenges, quantitative analyses confirm that Chinese participates in cross-linguistic punctuation universality. Dolina et al. (2025) indicate that inter-punctuation distances in Chinese literary works successfully follow the

discrete Weibull model. However, larger intervals are less frequent than in English translations, which is consistent with the tendency of Chinese toward longer comma-separated clausal units. Beyond the Weibull pattern, Chinese sentence length shows high flexibility and supports complex multifractal structures, as observed in Gao Xingjian's *Soul Mountain*. The large-scale study by Liu et al. (2023) corroborates that Chinese novels exhibit fractal patterns analogous to Western literature, though with weaker long-range correlations and a lower incidence of multifractality.

Because the Weibull parameters p and β successfully capture these typological profiles, they provide a sensitive diagnostic tool for translationese. In translation, two opposing tendencies operate. Translators tend to preserve source language punctuation patterns, while the target language imposes its own structural constraints (Stanisz et al., 2024b). They found that these tensions resolve in favor of the target language, shifting a text's position in the (p and β) parameter space toward the region occupied by target language originals. However, corpus-based studies of specific language pairs reveal that source language interference frequently persists. Rodrí guez-Castro (2011) found that English-to-Spanish journalistic translations preserve English period and dash patterns, producing a residual source language effect. Similarly, Zuo and Dai (2014) documented systematic Anglicization in translated Chinese, demonstrated by the suppressed frequency of the Chinese-specific enumerative comma which has no direct English counterpart.

Taken together, existing studies establish that punctuation distributional patterns are universal in form yet language specific in parameter values. More importantly, they reveal a dynamic tension in translated texts between target language normalization and source language interference. Yet, a bidirectional comparison to untangle these conflicting forces remains absent. Prior work has been largely unidirectional and preliminary (Dolina et al., 2025). Whether English to Chinese and Chinese to English translations converge asymmetrically toward their respective target norms has not been formally modeled. Given the vast structural distance between the two punctuation systems, these source and target pressures likely operate with different magnitudes depending on the translation direction.

In addition to these insights, the Weibull parameters p and β offer lightweight features for natural language processing because they encode text rhythm independent of vocabulary. They can serve as robust features for automatic translationese detection and provide novel metrics for evaluating rhythmic fluency and source-language interference in machine translation outputs. To address the current research gaps, the present study fits discrete Weibull distributions to a bidirectional Chinese and English literary translation corpus. This corpus comprises four subcorpora including English originals, Chinese originals, English to Chinese translations, and Chinese to English translations. We address two primary research questions. First (RQ1), do all four subcorpora conform to the discrete Weibull distribution? Second (RQ2), if so, do the two translation directions exhibit asymmetrical convergence in the resulting parameters (p and β)?

2 Methods

2.1 Corpus construction

A bidirectional Chinese-English literary translation corpus was compiled comprising 56 texts organized into four subcorpora: Original Chinese (OC), Original English (OE), Translated Chinese (TC, English-to-Chinese), and Translated English (TE, Chinese-to-English) (see Appendix A). English source texts were drawn from the Modern Library’s *100 Best Novels of the 20th Century*, while Chinese source texts were drawn from Asia Weekly’s *Top 100 Chinese Novels of the 20th Century*, two widely recognized canonical lists that ensure the representativeness and literary prestige of the selected works. All original novels are accompanied by established, authoritative translations.

2.2 Data extraction

Inter-punctuation intervals (IPIs) are defined in this study as the number of words between consecutive sentence-terminating punctuation marks. The extraction procedure involved three steps. First, sentence delimiters were identified. Because the linguistic definition of a sentence remains disputed, we used the intervals between specific punctuation marks as the measurement unit for cross-linguistic statistics (Zhou et al., 2025). Following the procedure of Dolina et al. (2025), the delimiter set was standardized across both languages to include

the full stop, exclamation mark, question mark, semicolon, and ellipsis. Second, word segmentation was performed. For Chinese texts, the Jieba word segmenter was applied, with all punctuation marks excluded from word counts, ensuring that IPIs reflect only the segmented word tokens between consecutive major punctuation delimiters. For English texts, regular expression-based tokenization was used, which automatically excludes non-alphabetic tokens. Lastly, IPIs were calculated as the word count between each pair of adjacent sentence-terminating marks. Table 1 presents the corpus composition and the total IPI counts for each subcorpus.

Subcorpus	Texts	Tokens (Chars. for ZH)	IPIs
OC	14	1,237,642 (2,201,356)	89,083
OE	14	1,375,450	113,282
TC	14	1,387,244 (2,515,899)	110,914
TE	14	1,484,728	111,861
Total	56	5,485,064	425,140

Table 1: Corpus composition.

2.3 Discrete Weibull distribution

The discrete Weibull distribution (Nakagawa and Osaki, 1975) has the probability mass function:

$$P(X = k) = (1 - p)^{(k-1)^\beta} - (1 - p)^{k^\beta}, k = 1, 2, 3, \dots$$

where p ($0 < p < 1$) is the scale parameter and β ($\beta > 0$) is the shape parameter. In linguistic terms:

- p is positively correlated with punctuation density: higher p values indicate shorter IPIs and more frequent punctuation.
- β reflects rhythmic regularity: higher β values indicate more concentrated distributions and more predictable sentence rhythms.

When the probability of the first success p depends on the number of already performed trials k , this dependence is called hazard function. For the discrete Weibull distribution, the hazard function is given by:

$$h(k) = 1 - (1 - p)^{k^\beta - (k-1)^\beta}$$

It represents the conditional probability of punctuation after the k -th word, given no punctuation in the preceding $k-1$ positions. When $\beta > 1$, $h(k)$ increases with k , indicating “sentence length pressure”—longer sentences create greater urgency for termination. When $\beta = 1$, the model reduces to the memoryless geometric distribution.

Parameters were estimated by maximum likelihood estimation (MLE), with optimization performed using the L-BFGS-B algorithm under the natural constraints $0 < p < 1$ and $\beta > 0$ (Byrd et al., 1995). Goodness of fit was assessed mainly by the root mean square error (RMSE) between the empirical and theoretical probability mass functions. Because formal goodness-of-fit tests are often overly sensitive in very large samples (D’Agostino and Stephens, 1986), we interpreted fit quality primarily through RMSE, the closeness of empirical and theoretical means, and visual agreement between observed and fitted distributions.

3 Results

3.1 RQ1: Distributional fit

The discrete Weibull distribution provided good fits across all four subcorpora. As shown in Figure 1, the fitted curves closely track the empirical distributions throughout. RMSE values were uniformly low (range: 0.0016 – 0.0036), and theoretical means deviated from empirical means by no more than 0.07 words in any subcorpus, confirming the model’s adequacy for capturing IPI distributions across both languages and translation conditions. At the individual text level, all 56 texts achieved satisfactory fits (mean RMSE = 0.0039), showing robustness across texts of varying lengths and styles. This high goodness-of-fit indicates that the rhythmic chunking of both original and translated texts obeys consistent mathematical patterns, validating the use of the Weibull parameters p and β for further analysis.

3.2 RQ2: Translation effects and asymmetry

To understand how the translation process affects structural rhythms, we first established the baseline differences between the two languages. Mann-Whitney U tests confirmed that OE and OC differ significantly in both parameters. They differ in p , which reflects punctuation density and average

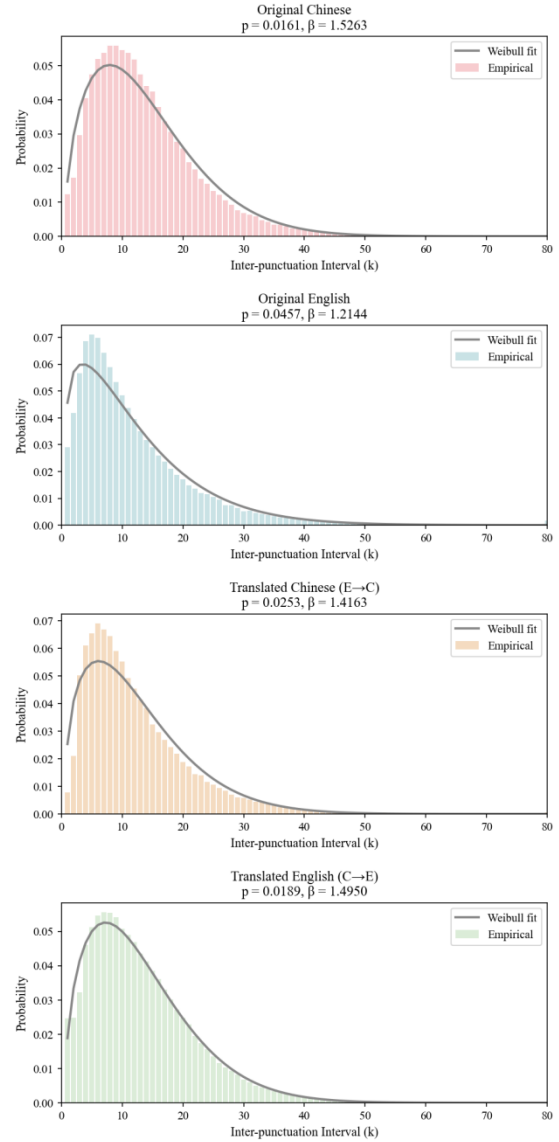


Figure 1: Weibull fits of punctuation intervals across four subcorpora.

chunk length ($U = 33.0, p = .003, d = 1.26$), and in β , which captures the regularity and predictability of rhythmic pacing ($U = 170.0, p = .001, d = 1.48$). This reflects typological differences, as the Chinese originals examined here are characterized by longer and more regular punctuation-bounded units, whereas the English originals show denser punctuation and greater rhythmic flexibility.

Then, the act of translation influences these natural patterns. Within the English corpora, TE moves substantially away from OE norms, showing significant differences for p ($U = 161.0, p = .004$) and for β ($U = 30.0, p = .002$). By contrast, TC remains statistically similar to OC for p ($U = 59.0, p = .077$) and for β ($U = 125.0, p = .223$). Moreover, TC and TE do not differ significantly

from each other in p ($U = 107.0, p = .696$) or in β ($U = 94.0, p = .872$), suggesting a tendency for translations from opposite directions to occupy a shared intermediate rhythmic space, neither as sparse and rigidly paced as native Chinese nor as densely punctuated as native English. This convergence is consistent with Baker’s (1993) leveling-out hypotheses, insofar as the translation process appears to neutralize extreme source-language punctuation profiles and pull texts toward a more conservative middle ground.

While this rhythmic convergence occurs in both directions, the underlying translation mechanisms driving it are asymmetrical. Wilcoxon signed-rank tests on paired source-translation texts reveal a directional contrast (Table 2): in the C→E direction, translation produces non-significant changes with negligible-to-small effect sizes ($d = 0.08$ and 0.31), whereas in the E→C direction it produces highly significant, large-effect shifts in both parameters ($W = 0.0, p < .001, d > 1.2$).

Direction	Para.	W	p -value	d
C→E (OC vs. TE)	p	28.0	.135	0.08
C→E (OC vs. TE)	β	23.0	.068	0.31
E→C (OE vs. TC)	p	0.0	< .001	1.35
E→C (OE vs. TC)	β	0.0	< .001	1.27

Table 2: Wilcoxon signed-rank tests by translation direction.

This asymmetry is visually captured in Figure 2. Arrows from OE source texts to their TC translations are markedly longer and strongly directed toward the OC norm, representing active structural adaptation. In contrast, arrows from OC source texts to their TE translations are consistently short and loosely directed, visualizing the retention of source features. Quantitatively, E→C translations shift 87.2% on p and 69.9% on β toward Chinese norms at the aggregate level. In practical terms, this massive shift means translators are adopting the sparser punctuation rhythm of Chinese, extending text into longer rhythmic units with more rigid pacing. Meanwhile, C→E shifts are far weaker, with aggregate shifts of 5.1% and 21.4%, meaning translators tend not to fully

restructure the source-language rhythmic pattern toward English original norms. A 2×2 ANOVA with text language (Chinese vs. English) and translation status (original vs. translated) as factors further supported this interpretation. The interaction between text language and translation status was significant for both p ($F = 10.05, p$ -value = .003) and β ($F = 10.66, p$ -value = .002), indicating that the effect of translation differs across the two text languages.

At the process level, hazard function profiles corroborate these findings. The hazard ratio $h(30)/h(10)$ quantifies sentence-length pressure, namely the increasing conditional probability of punctuation as an interval grows. Chinese texts inherently impose steeper hazard growth than English texts, with a ratio of 1.76 in OC versus 1.26 in OE. During E→C translation, TC shifts toward the higher Chinese hazard profile with a ratio of 1.56, confirming active structural adaptation to the Chinese target-language rhythm, in which

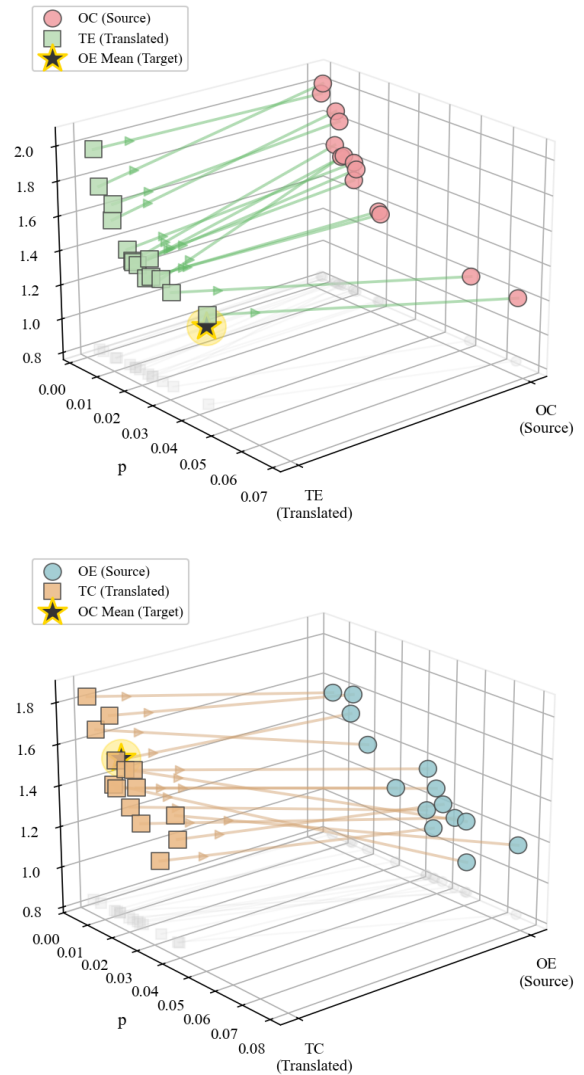


Figure 2: Parameter distributions by subcorpus.

punctuation is less frequent overall but the probability of termination rises more steeply as intervals grow longer. Conversely, in C→E translation, TE largely retains the high-pressure source-language profile with a ratio of 1.70, remaining far closer to Chinese than to English norms. This limited relaxation of punctuation rhythm reflects strong source-language interference and structural persistence, where the sustained rhythmic pattern of the Chinese source text is largely preserved in the English translation.

4 Conclusion

This study has shown that the discrete Weibull distribution provides a principled and statistically robust method for quantifying punctuation rhythm in literary translation. Beyond model validity, the central finding reveals that translation from English to Chinese adapts strongly toward target language norms, whereas translation from Chinese to English tends to preserve source language features. This offers quantitative grounding for directional asymmetry in structural adaptation and source language interference.

To ensure methodological rigor and control for potential confounding variables, the present study focused specifically on twentieth-century literary novels. Having established this quantitative baseline, future research can expand this analytical approach to explore other text genres, historical periods, and diverse language pairs. Additional potential directions include comparing multiple translations of the same source text to isolate variations specific to individual translators, as well as correlating parameter shifts with translator background and training.

Moreover, the Weibull parameters p and β provide a reproducible basis for characterizing translation style while offering concrete utility for computational applications. For automatic translationese detection, these parameters serve as lightweight and domain-independent features that capture rhythmic deviation without relying on surface semantics. Furthermore, this approach presents a novel mechanism for text generation in large language models. The hazard function $h(k)$ can be integrated into the decoding phase as a control signal to adjust punctuation probability dynamically. This may help models to overcome unnatural mechanical rhythms and generate text that matches the natural breathing rhythm of human writers across specific parameter spaces.

Ultimately, this analytical approach provides a quantitative tool that supports both empirical translation research and natural language processing tasks.

References

- Wallace Chafe. 1988. Punctuation and the prosody of written language. *Written Communication*, 5(4):395–426.
- Philip W. Anderson. 1972. More is different: Broken symmetry and the nature of the hierarchical structure of science. *Science*, 177(4047):393–396.
- Gabriel Altmann. 1980. Prolegomena to Menzerath’s law. *Glottometrika*, 2:1–10.
- Harold S. Heaps. 1978. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press.
- George K. Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press.
- Andrzej Kulig, Jarosław Kwapien, Tomasz Stanis, and Stanisław Drożdż. 2017. In narrative texts punctuation marks obey the same statistics as words. *Information Sciences*, 375:98–113.
- Stanisław Drożdż, Paweł Oświęcimka, Andrzej Kulig, Jarosław Kwapien, Katarzyna Bazarnik, Iwona Grabska-Gradzińska, Jan Rybicki, and Marek Stanuszek. 2016. Quantifying origin and character of long-range correlations in narrative texts. *Information Sciences*, 331:32–44.
- Toshio Nakagawa and Shunji Osaki. 1975. The discrete Weibull distribution. *IEEE Transactions on Reliability*, 24(5):300–301.
- Tomasz Stanis, Stanisław Drożdż, and Jarosław Kwapien. 2023. Universal versus system-specific features of punctuation usage patterns in major Western languages. *Chaos, Solitons & Fractals*, 168:113183.
- Tomasz Stanis, Stanisław Drożdż, and Jarosław Kwapien. 2024a. Statistics of punctuation in experimental literature—The remarkable case of *Finnegans Wake* by James Joyce. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 34(8).
- Michał Dolina, Jakub Dec, Stanisław Drożdż, Jarosław Kwapien, Jin Liu, and Tomasz Stanis. 2025. Quantifying patterns of punctuation in modern Chinese prose. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 35(2).
- Yikai Zhou, Jingyang Jiang, and Haitao Liu. 2025. Language universals in sentence length: Comparing sentence length distributions of 10 languages.

Cognitive Science, 49(9):e70115.

Fang Kong and Guodong Zhou. 2014. Chinese comma disambiguation on K-best parse trees. In *Natural Language Processing and Chinese Computing*, ed. Chengqing Zong, Jian-Yun Nie, Dongyan Zhao, and Yang Feng. Communications in Computer and Information Science, Vol. 496, 13–22. Springer, Berlin.

Jin Liu, Edward Gunn, Fuad Youssef, Jacob Tharayil, Wyatt Lansford, and Ying Zeng. 2023. Fractality in Chinese prose. *Digital Scholarship in the Humanities*, 38(2):604–620.

Tomasz Stanisław, Stanisław Drożdż, and Jarosław Kwapien. 2024b. Complex systems approach to natural language. *Physics Reports*, 1053:1–84.

Mónica Rodríguez-Castro. 2011. Translationese and punctuation: An empirical study of translated and non-translated international newspaper articles (English and Spanish). *Translation and Interpreting Studies*, 6(1):40–61.

Shangjun Zuo and Guangrong Dai. 2014. Corpus-based study of punctuations in translated Chinese. *Journal of Fujian University of Technology*, 12(2):121–126.

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyu Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

Ralph B. D’Agostino and Michael A. Stephens (Eds.). 1986. *Goodness-of-Fit Techniques*. Marcel Dekker, New York.

Mona Baker. 1993. Corpus linguistics and translation studies: implications and applications. In *Text and Technology: In Honour of John Sinclair*, ed. Mona Baker, G. Francis, and E. Tognini-Bonelli, 233–250. John Benjamins, Amsterdam/Philadelphia.

C→E Direction	E→C Direction
Mulberry and Peach	愤怒的葡萄
活着 To Live	The Great Gatsby 了不起的盖茨比
红高粱家族 Red Sorghum	The Heart is a Lonely Hunter 心是孤独的猎手
边城 Border Town	The Postman Always Rings Twice 邮差总按两遍铃
酒徒 The Drunkard	The Sun Also Rises 太阳照常升起
长恨歌 The Song of Everlasting Sorrow	To the Lighthouse 到灯塔去
骆驼祥子 Rickshaw Boy	Tropic of Cancer 北回归线
黄金时代 Golden Age	Under the Volcano 火山下

Appendix A: Corpus Text List

C→E Direction	E→C Direction
半生缘 Half a Lifelong Romance	Animal Farm 动物庄园
古船 The Ancient Ship	Brave New World 美丽新世界
呐喊 Call to Arms	Lord of the Flies 蝇王
围城 Fortress Besieged	Of Human Bondage 人性的枷锁
城南旧事 Memories of Peking	The Age of Innocence 纯真年代
家 Family	The Catcher in the Rye 麦田里的守望者
桑青与桃红	The Grapes of Wrath

From SOP to AI-Native Pipeline: Design and Evaluation of a Translation-Memory-Grounded LLM Workflow (Nixa AI)

Ana Dumančić Acar

Studio Nixa Prijevodi d.o.o., Zagreb, Croatia

ana@studionixaprijevodi.hr

Abstract

Recent advances in Large Language Models make end-to-end AI-driven translation pipelines technically feasible. However, professional deployment requires more than raw translation quality: it demands terminological consistency, reuse of prior approved work, and auditable quality control. This paper presents **Nixa AI**, an internal AI-native system developed within a professional Language Service Provider (LSP). Unlike integrations that layer LLMs onto existing CAT tools, Nixa AI is built from the ground up around a translation-memory-grounded LLM pipeline, with a cascade retrieval strategy that prefers previously approved translations over model generation. Internal evaluation on real client documents shows that output quality is not the primary bottleneck. The principal barrier to full automation is **epistemic**: in domains with legal or financial consequences, accountability still requires a human validator. The system therefore positions the translator as the final layer of confidence rather than attempting to remove them from the loop.

1. Introduction

Neural Machine Translation and, more recently, Large Language Models (LLMs) have made it

plausible to design translation pipelines in which the machine performs the bulk of the linguistic work. In practice, Language Service Providers (LSPs) have been slower to adopt such systems than the underlying capability would suggest. The gap is rarely in raw translation quality. It lies in terminology control, consistency across long documents, reuse of prior corrections, and — most importantly — in who answers for the final text.

Off-the-shelf LLM translation, deployed without a memory and terminology layer, produces output that is locally fluent but globally inconsistent: the same source term is rendered differently in successive segments, approved phrasing from prior projects is not reused, and small but consequential elements — numbers, proper nouns, tagged formatting — are reworded or quietly dropped. In a one-off consumer setting these are minor annoyances. In a professional LSP setting they are the work. The LSP carries the legal and financial responsibility for the delivered document, and that responsibility is attached to a named human reviewer. This changes what "good enough" means.

This paper describes the design, implementation, and evaluation of **Nixa AI**, an internal system built by a professional LSP to examine whether an AI-native workflow can meet the operational standards of paid client work. The emphasis throughout is practitioner-oriented: the report describes what was built, what the system does well, where it does not, and what we learned about trust as a deployment constraint.

2. System Overview

Nixa AI was designed as a modular pipeline that mirrors the structure of a professional translation workflow while re-implementing its stages in an AI-native architecture. The system is not a plug-in to an existing CAT tool; the segment editor, translation memory, glossary, and QA surface are all built into the platform.

A standard job proceeds through the following stages:

1. **Document ingestion and segmentation.**

Source documents in common office formats are converted to a standards-based bilingual intermediate representation, preserving inline formatting and non-translatable elements.

2. **Translation Memory cascade.** For each segment, the system searches the client's Translation Memory in order of decreasing confidence: in-context exact matches, exact matches, fuzzy matches, and semantic matches. Exact matches above a trust threshold (quality and provenance) are reused directly as the final target, bypassing the LLM entirely. Lower-confidence hits — in-context exact, fuzzy, and semantic — are not applied as the final translation; instead, they are passed to the LLM as strong reference context alongside the source segment, so that the model's output stays aligned with prior approved work.

3. **Glossary grounding.** The client's workspace-level glossary is attached to the translation prompt as a constraint, with mandatory terms enforced and suggested terms provided as guidance; the list is prioritised and capped to fit the prompt budget. In parallel, per-segment source-side term detection drives two additional layers: an in-editor annotation that highlights glossary terminology to the reviewer, and a deterministic post-translation pass that checks and repairs missing or inconsistent mandatory terms in the output.

4. **LLM translation for uncovered segments.**

Segments that the cascade does not resolve with sufficient confidence are translated by an LLM in small batches, with tag-simplified input to minimise format corruption.

5. **Automated quality assurance and targeted repair.**

Translated segments are checked against the glossary, source structure, tag integrity, and a set of content-level rules. Segments flagged by this audit may be subjected to a *targeted* LLM repair step — not a full re-translation pass — in which only the flagged segments and their specific issues are returned to the model for correction. The outcome is surfaced to the reviewer together with the provenance of each segment.

6. **Human review and confirmation.** A translator reviews the output segment-by-segment. Confirmed segments are written back to the Translation Memory, so that the client's corrections compound over time.

7. **Document repackaging.** The approved translations are injected back into the bilingual intermediate and merged into the original document format.

The cascade structure is important. For repetitive and well-memorised content, the LLM is not involved at all. For novel content, the LLM operates in a tightly constrained environment — given source, glossary, and surrounding context, and expected to return format-preserving output. For ambiguous or high-stakes content, the translator remains the decision-maker.

3. Alignment with Existing SOPs

The pipeline was built on top of Studio Nixa's established operational procedures rather than in opposition to them. The traditional SOP — document preparation, glossary creation, termbase conversion, CAT tool setup, pre-translation, human translation and revision, quality control, delivery — was retained as the structural backbone. The AI-native system compresses several of these stages (CAT setup, pre-

translation, much of first-pass translation) into automated steps, but preserves their purpose. In particular, glossary creation and quality assurance remain first-class stages rather than implicit side-effects of the model.

Traditional SOP Workflow	Nixa AI Pipeline
Document Preparation	Source Document Ingestion
Glossary Creation	Glossary (client-owned, progressively enriched)
Termbase Conversion	<i>(absorbed into platform — no external termbase)</i>
CAT Tool Setup	<i>(absorbed — the platform is the CAT tool)</i>
Pre-translation (TM + MT)	Translation Memory Cascade + LLM for gaps
Human Translation & Revision	Human Review & Confirmation (TM learns on confirm)
Quality Control	Automated QA + Targeted LLM Repair + Human Review
Delivery	Document Repackaging & Delivery

Figure 1: Mapping of Traditional Workflow to Nixa AI

Rather than replacing the workflow, the pipeline reconfigures it into an AI-native system in which the stages most amenable to automation are compressed, and the stages that require judgement are preserved and supported with better tooling.

4. Methodology and Evaluation

The system was evaluated on several hundred real client documents comprising, in aggregate, on the order of hundreds of thousands of translation segments. The dominant language pair is English ↔ Croatian, with a smaller share of Croatian into other target languages. The evaluation is not a controlled benchmark: it is paid client work, delivered through the system, across the domains routinely handled by the LSP.

Input content varies substantially. It ranges from short, highly structured forms and recurring clauses to long, narrative documents. A notable share of incoming documents originates as scanned PDFs that are passed through an upstream OCR and layout-reconstruction step before entering the pipeline as editable input; the pipeline itself does not attempt to process raw scans. Over time the client's workspace glossary has accumulated tens of thousands of entries, providing a substantive terminology base against which each job is grounded.

Internal translators used the system for paid production work and concurrently documented, at segment level:

- terminological errors and inconsistencies with prior approved work;
- structural and tag-level issues, including dropped or mis-rendered inline formatting;
- fluency issues subtle enough to survive initial review;
- domain-specific inaccuracies, especially in content where correctness is not obvious to a generalist reader;
- fidelity issues in which the output was fluent but did not preserve key elements of the source (numbers, named entities, quantifiers).

These observations, together with the automated QA signals captured by the system, drove three kinds of refinement: adjustments to the translation prompt and its surrounding context window, expansion of the deterministic post-translation layer (the rule-based step that repairs predictable surface failures before LLM repair is considered), and tightening of cascade thresholds governing when a TM hit is trusted as final. This iterative process was substantially more time-consuming than initially anticipated. In effect, translators became co-developers of the system, and their structured feedback — rather than any synthetic benchmark — was the primary driver of improvement.

5. Results

Output quality is domain-dependent and can be summarised as follows:

- **Structured and repetitive content** (forms, boilerplate, recurring clauses): the TM cascade handles the majority of segments, LLM involvement is minimal, and quality is excellent.
- **Controlled domains with a mature glossary**: LLM involvement is higher, but terminology is consistent and fluency is strong.
- **Novel, unstructured, or legally consequential content**: quality is generally good, but residual errors appear and they are often subtle.

The categories of failure that the automated QA layer is designed to surface fall into five broad families, each arising recurrently during internal evaluation:

8. **Fidelity** — segments that are fluent but do not preserve something from the source: missing numbers, dropped named entities, untranslated passages that look plausible because the model returned the source unchanged, or output whose length ratio against the source is anomalous enough to suggest content loss.
9. **Terminology** — deviations from the client-specific glossary, including missing mandatory terms and inconsistent renderings of the same source term across segments.
10. **Format** — inline tag corruption, leaked placeholder artefacts, hallucinated tags not present in the source, and tag-structure mismatches that would break the round-trip back into the original document format.
11. **Surface** — capitalisation, punctuation, and whitespace deviations that are locally minor but accumulate into a professional polish problem at document scale.

12. **Stylistic** — phrasing that is grammatical but unnatural or over-literal for the target variety.

The most operationally significant observation is qualitative. The system's fluency is high enough to **mask** the fidelity and terminology failures it does produce. A reviewer who is not actively looking for these errors will miss them. This shifts the reviewer's task from correction to vigilance, and it has direct implications for workflow design and review-time estimation. It also shaped the system itself: the automated QA layer was iterated on alongside the translation layer, because surfacing errors to the reviewer turned out to be at least as important as reducing their rate.

6. Discussion: The Trust Problem

The central finding is not about capability. Modern LLMs, augmented with a disciplined Translation Memory and glossary layer, can produce translations that are, on average, good enough to ship. The open question is whether that output can be trusted *without* human verification in domains where errors carry legal or financial consequences.

In the LSP's professional context, the answer is currently no. Responsibility for the final text sits with the LSP, and therefore with a named human. An output that is 98% correct, but where the 2% is invisible, is from a liability standpoint harder to defend than an output that is visibly imperfect. The issue is not quality; it is **attributable accountability**.

This reframes what AI-native translation systems should optimise for. Raw quality is necessary but not sufficient. The system must also produce outputs whose errors are *discoverable* — through QA signals, match-provenance indicators, and a workflow that directs reviewer attention to the segments most likely to be wrong. Nixa AI invests heavily in this surface area: segment-level status, the match type and score of the winning Translation Memory candidate, glossary-conformance indicators, and a review flow in

which each confirmed segment becomes a reusable asset for future work.

7. Conclusion

Nixa AI demonstrates that an AI-native translation pipeline can meet professional standards for a meaningful share of real production work. The technology is ready. The adoption constraint is not performance but trust, and trust is earned not by removing the translator but by giving them the tools to validate the system's output efficiently and auditably.

The translator's role is therefore evolving. In an AI-native pipeline, translators spend less time producing first drafts and more time as evaluators, validators, and the named human behind each delivered document. Far from being displaced, they become the final — and, in regulated domains, indispensable — layer of confidence in the workflow.

Keywords

Translation Memory, Large Language Models, AI-native systems, Translation workflows, Human-in-the-loop, Glossary control, Trust in AI, Language Service Provider

Building a European Multilingual Evaluation Dataset: The MMLU Localisation Project within the EMT Network

Pilar Sánchez-Gijón

UAB/ Spain

pilar.sanchez.gijon@uab.cat

Susana Valdez

Leiden University/ the
Netherlands

s.valdez@hum.leidenuniv.nl

Anna Kokkinidou

Directorate-General for
Translation of the European
Commission / Belgium

anna.kokkinidou@ec.europa.eu

Florence Bellemont

Leiden University/ the
Netherlands

f.m.f.e.l.g.bellemont@umail.lei
denuniv.nl

Sofia Calvo Del Barrio

UAB/ Spain

Sofia.Calvo@autonoma.cat

Mihai Cristian Brasoveanu

Directorate-General for
Translation of the European
Commission / Belgium

Mihai.BRASOVEANU@ec.eur
opa.eu

Abstract

This paper reports on a collaboration between the Directorate-General for Translation (DGT) and the European Master's in Translation (EMT) to localise the MMLU dataset into 11 European languages. Beyond creating a more inclusive benchmark for LLM evaluation, the project offers master's students authentic, project-based professional training in translation, revision, project management, and multilingual coordination, while highlighting key methodological, administrative, and workflow challenges.

1 Introduction

The growing integration of artificial intelligence (AI) into language services has intensified demand for high-quality multilingual resources capable of evaluating AI systems across diverse linguistic contexts. One prominent example is the Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2021), a widely used dataset for assessing the reasoning and knowledge capabilities of large language models across 57 domains. While the MMLU dataset has become one of the leading benchmarks for evaluating Large Language Models (LLMs), several efforts have sought to create translated and culturally relevant versions into multiple languages (e.g., Global-

MMLU (Singh et al. 2024) or MMMLU from OpenAI). These initiatives reflect a growing recognition that English-centric evaluation may pose significant limitations; yet most existing multilingual versions rely primarily on machine translation (MT) or crowdsourced post-editing, raising persistent questions about linguistic quality, terminological consistency, and cultural appropriateness in the target languages.

This paper reports on a large-scale localisation initiative that pursues a dual aim: to produce a human-translated version of the MMLU dataset into European languages, and to do so through an authentic, project-based training environment in which translators in training engage with an authentic institutional workflow. The initiative is carried out in collaboration between the network of European Master's in Translation (EMT) programmes and the Directorate-General for Translation (DGT) of the European Commission, and is coordinated within the EMT framework by the Working Group dedicated to Linguistic Data and Terminology. Beyond its technical objectives, the initiative serves as a structured, practice-based training environment in which master's students from participating universities engage in authentic multilingual workflows. In doing so, they contribute to the creation of a high-impact European language technology resource while simultaneously developing translation

competences aligned with real-world translation scenarios.

This paper presents the design and pedagogical rationale of this ongoing initiative, and reflects on what this model of academia-institution collaboration may offer for translation training in the context of an increasingly AI-mediated language industry.

This paper is structured as follows: Section 2 reviews relevant literature on project-based learning. Section 3 reports on the DGT-EMT initiative, followed by a discussion of key challenges in section 4.

2 Didactical and Theoretical Background

In contemporary translation training, pedagogical approaches are increasingly grounded in holistic, learner-centred principles that closely mirror professional practice. Models such as Authentic Experiential Learning and Project-Based Learning (PjBL) emphasise real assignments and "learning by doing" (Korda, 2023, Kiraly 2005), while Situated Learning recreates professional environments—such as Simulated Translation Bureaus (STBs) (Van Egdom *et al.*, 2020)—in a structured setting. In this sense, these approaches operationalise the EMT Competence Framework by embedding professional competences into authentic tasks, while also drawing on social constructivist principles that position students as active participants engaged in collaborative problem-solving under guided facilitation.

The practical application of these pedagogical approaches can be seen in a number of successful projects. For instance, the INSTB Network (International Network of Simulated Translation Bureaus), established in 2015, is a strategic partnership that includes 19 institutions across 10 countries (Van Egdom *et al.*, 2020).

Project management within these training simulations mirrors professional practices and involves several critical steps. Students often take on key roles, such as project manager, translator, terminologist, QA manager, and reviser, all of which are vital to the functioning of the simulation. The creation of business plans, market assessments, and pricing strategies is also part of the learning process. Professional tools, such as Slack, are used to manage internal workflows, while students also engage in activities like

"storytelling" for branding and creating websites to attract potential clients.

In authentic projects, students negotiate deadlines and clarify project briefs with external clients. In simulated projects, the instructor may take on the role of a "mock client," providing students with a controlled yet practical environment for learning client interaction.

The incorporation of real-world elements into these projects offers numerous benefits. Notably, these projects enhance employability by providing students with market-oriented experience, thus improving their ability to navigate the job market upon graduation (Uribe de Kellet, 2022). Additionally, the self-efficacy of students is significantly strengthened, as they develop confidence in managing professional risks and tasks. Moreover, soft skills such as teamwork, leadership, interpersonal communication, and negotiation are fostered through collaborative project work. Lastly, the international nature of these projects promotes internationalization, allowing students to engage in cross-cultural networking and develop intercultural competence.

However, these projects also present several challenges. External pressure from working on authentic projects can introduce unpredictability and the stress of real-world consequences. Aligning the logistics of academic calendars across different international partner universities can also prove to be a significant obstacle, hindering smooth collaboration. Additionally, ethical concerns arise in the context of STBs, as they must ensure that they do not undercut the market rates of professional translators. These simulations often focus on non-commercial or non-profit projects to avoid this issue. Furthermore, initial anxiety is common among students at the start of a project, given the demanding nature of professional standards. Finally, complexity is heightened when engaging in international collaborations, which may discourage some students, particularly when deadlines are tight.

3 The MMLU Dataset

The Massive Multitask Language Understanding (MMLU) benchmark is a widely used evaluation framework designed to assess the breadth of knowledge and reasoning capabilities of large language models across 57 academic subjects, ranging from STEM and humanities to social sciences and professional domains. Originally

introduced by Hendrycks et al. (2021), MMLU has become a standard reference for measuring model performance on diverse, real-world tasks, making it a critical tool for advancing AI transparency, robustness, and multilingual alignment.

Figure 1 shows a sample question from the MMLU benchmark, provided here as an illustrative example.

Q: Which of the following mental disorders is most prevalent in the world?

- Option 1: Bipolar disorder
- Option 2: Schizophrenia
- Option 3: Alcohol use disorders
- Option 4: Anxiety disorders

Figure 1. List of questions and choices 99-global facts, question 97

3.1. Translation Project: The MMLU Dataset, setting the ground for the DGT-EMT project

Emerging from the need for a European benchmark for evaluating LLMs, this project was initiated by the AI Unit of the DGT in November 2025. It was developed within the framework of an ongoing collaboration between the DGT and the EMT network, bringing together institutional expertise in multilingual communication and professional translator training across participating European universities. In November 2025, the AI Unit of the DGT proposed a new collaboration to Working Group 5 on Linguistic Data and Terminology of the EMT network: the localisation of an existing LLM evaluation dataset into the European languages, to be carried out by master's students across EMT member universities as part of their professional training. This design, embedding a high-impact language technology task within an authentic academic workflow, defines the dual aim of the initiative (see Introduction). The first phase of the project is scheduled for completion by 30 June 2026, in alignment with the spring academic semester.

3.2. The MMLU Dataset Translation Project

The MMLU dataset (Hendrycks et al., 2021) is a widely used benchmark for evaluating LLMs through multiple-choice questions spanning diverse domains (e.g., history, law, mathematics,

medicine, engineering). As noted earlier, existing multilingual extensions of MMLU have relied predominantly on MT or crowdsourced MTPE, approaches that have been shown to introduce translation artefacts, terminological inconsistencies, and culturally inappropriate renderings that can distort evaluation outcomes (Singh et al., 2024). While such approaches offer scalability, their suitability for high-stakes benchmarking contexts remains an open question.

Extending the MMLU dataset to European languages addresses gaps in linguistic representation and supports more equitable evaluation in the context of the European Commission's and DGT's active role in the evaluation of LLMs in all of the EU, focusing on ensuring linguistic and cultural diversity across Europe pursuant to the overarching principle of multilingualism and the principle of 'no language left behind'. The goal of the European Commission is aligned with the Digital Decade Policy Programme 2030, in actively pursuing "Digital Language Equality" (DLE). The strategy focuses on using LLMs and AI to bridge the widening gap between English and other European languages. In this vein, there has been growing evidence that such multilingual LLMs exhibit strong capabilities for some or many of the under-resourced languages they have been trained in (Rehm, Grützner-Zahn & Barth, 2025:1). The EMT network substantially backs up this EU-wide effort by actively participating in this project with significant impact given that it has members in almost all EU member states representing many under-resourced languages in its programmes.

The dataset comprises around 15,000–16,000 items (over 1,400,000 words in English). Its concise, structured format makes it well suited to collaborative translation workflows at scale.

3.3 Project Design and Participation Model

In this initiative, we opted for full human translation, without MT as a basis, carried out by master's students in translation under expert supervision. Human translation remains the 'golden standard' for multilingual benchmarking and cross-lingual evaluations as MT generated data can introduce biased unnatural phrasing (Artetxe, Ruder & Yogatama, 2020:5). This methodological choice reflects a commitment to semantic fidelity, terminological consistency, and cultural appropriateness as preconditions for the valid

benchmarking of multilingual systems. In addition, by not relying on neural machine translation (NMT) for this purpose—even when followed by human post-editing—we avoid a recursive feedback loop in which automatically generated content is reused to train or evaluate systems, sometimes described as the ‘ouroboros effect’ (Shumailov et al., 2024).

The EMT network is composed of 81 master’s programmes across 24 countries. All EMT programmes and their students were invited to participate in the localisation of the dataset. Each programme can engage in the project according to the modality that best fits its academic planning and constraints.

Three main forms of participation have been established:

- Integration within a course or module, where the entire class group contributes to the project as part of structured teaching activities introduced and supervised by the instructor.
- Participation through remote cooperation agreements coordinated between the university and the DGT.
- Voluntary participation as an extracurricular activity, allowing students to contribute outside the formal curriculum.

The call for participation has been positively received, with 21 EMT university programmes joining, 230 students, among whom 2 student project managers, from 11 EMT member countries actively translating and revising into 11 target languages currently in the project. No formal access test or proficiency assessment was conducted for the project. All students whom the supervising professors deemed academically appropriate were admitted to participate. The volume of words to be translated into each target language is not fixed, but rather dynamically determined by the number of participating institutions and the specific mode of student involvement adopted in each case. This results in a flexible and scalable distribution of tasks across languages and institutions (see Table 1).

Language Pair	No. of Students	Expected value (in words)	Universities
EN-CZ	14	94000	Charles University
EN-DE	3	70000	University of Vienna; University College Cork
EN-EL	4	80000	Aristotle University of Thessaloniki
EN-ES	10	82000	Universitat Autònoma de Barcelona; Universidad de Salamanca; University of Lisbon; University College Cork
EN-FR	27	149300	Europe; Université de Liège; University College Cork; University of Lisbon
EN-HU	11	78200	Eötvös Loránd Tudományegyetem
EN-IR	1	20000	University College Cork
EN-IT	100	704000	University of Bologna; Università degli Studi Internazionali di Roma - UNINT; Katholieke Universiteit Leuven; University of triete; Università degli Studi di Udine; Université Bourgogne Europe; University College Cork; University of Lisbon; University of Naples "L'Orientale"
EN-NL	3	50000	Leiden University
EN-PL	39	202000	University of Warsaw; University of Krakow (UKEN); University of Gdańsk; University of the National Education Commission; Krakow University
EN-PT	18	44010	University of Lisbon; Université Bourgogne Europe
TOTAL	230	1574910	

Table 1: Overview of language pairs, number of participating students, expected word volume, and participating universities

In parallel, the DGT has formalised remote cooperation agreements with those students who commit to translating and in turn revising at least total of 20,000 words, thereby aligning the project with recognised forms of professional training and ensuring adequate institutional support and supervision. These remote cooperation agreements

between DGT and the involved EMT programmes—which constitute standard practice for other academic projects involving an agreement between DGT and a university and carried out remotely—contribute to strengthening the link between academic training and professional practice within the EMT framework. In total 11 agreements were signed with participating universities, involving 35 students. This level of participation reflects both the interest generated by the initiative and its capacity to mobilise a substantial number of remote trainees across different educational and linguistic contexts.

3.4 Project Management and Coordination

Project management is coordinated within the EMT framework by Work Group 5 on Linguistic Data and Terminology, led by Pilar Sánchez-Gijón (Universitat Autònoma de Barcelona) and Susana Valdez (Leiden University). In close collaboration with the AI Unit of the Directorate-General for Translation (DGT) and EMT officials, the general conditions for project execution were defined, including workflow design, quality standards, and coordination mechanisms across participating institutions.

At the operational level, responsibility for managing the translation process is assigned to two student project managers (PMs), each representing one of the coordinating master's programmes (Leiden University and Universitat Autònoma de Barcelona). These students act as project managers, creating and assigning tasks, ensuring consistency and quality, facilitating communication by addressing queries and resolving technical issues and serving as the main point of contact between their local teams and the central coordination structure. This model reinforces both project management skills and accountability, while ensuring alignment across languages and institutions.

In addition, the project benefits from the systematic support of the DGT's Language Units, which act as expert advisors throughout the workflow. They provide authoritative guidance on terminological decisions, domain-specific conventions, and language-specific nuances, ensuring that translations align with interinstitutional standards and remain consistent across languages. To operationalise this alignment, an ad hoc style guide was developed for the project, drawing on the DGT's interinstitutional style

guides and selectively consolidating the conventions most relevant to the MMLU dataset. The guide, prepared by a student from the Universitat Autònoma de Barcelona, serves as a practical reference that harmonises terminology, punctuation, formatting, and QA criteria across teams, thereby enhancing usability and coherence. Beyond its immediate function, this framework exposes students to professional revision practices and establishes a structured interface between academic training and institutional expertise.

The project was formally launched on 16 February 2026 with a kick-off meeting bringing together all participating universities, marking the official start of activities and coordination across institutions.

3.5 Workflow and Task Execution

Following the initial kick-off meeting, detailed information has been collected regarding the volume of words that each student commits to translating and revising. Based on this allocation, participants were granted access to a dedicated RWS Trados Enterprise tenant managed by the Universitat Autònoma de Barcelona, through which all project management and translation activities are carried out.

Each student undertakes a dual role, committing to both the translation and revision of an equivalent volume of text. PMs create specific projects for each language pair and assign each file to one student to be translated or revised. These files typically contain a variable volume of text, ranging from just over 100 words to approximately 1,800 words, allowing for manageable yet substantial units of work.

Participants receive task notifications directly through the platform, which are also reinforced via email communication to ensure timely awareness and engagement. In parallel, faculty members who have integrated the project into their courses are provided with PM-level access, enabling them to monitor progress, retrieve assignments, and support their students throughout the process.

To facilitate communication and problem-solving, a dedicated Microsoft Teams environment was set up within the DGT tenant, bringing together all project participants. Within this space, a shared Q&A spreadsheet is maintained, where students systematically record doubts and issues encountered during translation. The PMs oversee this process, providing initial responses and, when

necessary, escalating queries to the relevant DGT Language Units. This structured communication channel ensures efficient knowledge sharing, traceability of decisions, and consistency across languages.

4 Key Challenges and Operational Constraints

This section outlines the main challenges encountered in the design and implementation of the project, which stem from its scale, multilingual scope, and institutional complexity. These challenges are not only operational but also methodological and pedagogical, as they directly impact the reliability of the dataset as an evaluation resource, the coordination across a distributed academic network, and the learning experience of participating students. The following subsections discuss these challenges from three complementary perspectives: the construction of a European multilingual evaluation dataset, the administrative and coordination framework required to sustain the initiative, and the practical execution of large-scale collaborative translation workflows. Since the project is ongoing, this section reflects an initial reflection on the topic.

4.1 The challenge of building a European evaluation dataset

A central methodological issue lies in the US-centric bias present in parts of the MMLU dataset, which becomes apparent in a number of its questions. Direct adaptation of these items to a European context during localisation would risk modifying their difficulty and, in turn, compromising cross-lingual comparability and construct validity. To address this, a selective two-step approach has been adopted. Items with a global scope or without culturally bound references will be translated directly without content adaptation, while those containing local-specific elements undergo a prior revision process to neutralise or adapt such references to a European setting before translation. This strategy aims to preserve both semantic equivalence and item difficulty across languages.

Another methodological issue lies in the fact that some parts of the MMLU dataset feature translated quotes from sources in European languages, yielding the question whether to find the source text or to retranslate the quote from English to the relevant language (French, for example). In this

case, the team decided to ask the translators to find the source of the quote to ensure accuracy of the testing material.

Even with these measures in place, the development of a multilingual evaluation dataset at the European scale entails further methodological constraints. Maintaining comparability requires careful control of semantic precision, difficulty calibration, and domain coverage.

4.2 Administrative and coordination challenges

The administrative dimension follows a staged, adaptive coordination model. In a first phase, the coordinators compiled a consolidated overview of participating institutions, student contributors, projected translation volumes, and the designated academic representatives authorised to sign on behalf of each university. This mapping supports a calibrated engagement strategy. Formal agreements (when in place) have been subsequently established with universities whose projected contribution approximates the workload of a standard remote internship, thereby ensuring alignment between academic participation and recognised professional training frameworks. Upon completion, each student will be issued an official certificate of collaboration specifying the number of words translated and revised, providing formal recognition of their contribution for both academic and professional purposes.

4.3 Execution and workflow challenges

At the operational level, project execution raises challenges related to scale, quality assurance, and the coordination of geographically distributed teams. The workflow follows a structured, student-led translation–revision model in which each participant translates a set of files and reviews an equivalent volume produced by peers. For language combinations with particularly high workload capacities—most notably Polish, French and Italian—a second revision layer has been introduced to strengthen quality control and ensure greater consistency.

A key responsibility of PMs is the fine-grained calibration of task allocation, taking into account individual student availability as well as, where relevant, the constraints imposed by course-based participation. This requires ongoing adjustments to maintain both feasibility and an equitable distribution of work across participants.

Finally, the management of Q&A processes constitutes a critical coordination layer. It demands continuous monitoring and rapid response times to avoid delays in the workflow. A centralised Q&A system is shared across all language projects, enabling the systematic identification of recurring issues and promoting the efficient dissemination of solutions across language pairs. This approach reinforces cross-lingual consistency and decision traceability.

5 Final Remarks

This paper has reported on a large-scale, collaborative localisation project aimed at extending the MMLU benchmark to 11 European languages within the EMT–DGT framework. Beyond its technical objectives, the initiative functions as a structured training environment that brings together academic institutions, students, and EU language professionals in a shared workflow. The project demonstrates the potential of coordinated, multilingual efforts to contribute to more inclusive and representative evaluation resources for LLMs, while simultaneously fostering professional competences aligned with the EMT framework (EMT Expert Group 2022).

At the same time, the project is conceived as a learning process in itself. For the purposes of this presentation, feedback will be collected from both participating students and faculty members, as well as from the DGT’s translation units involved in the project, in order to better understand its impact and identify areas for improvement. These perspectives will provide valuable insights into the pedagogical value of the initiative, the effectiveness of the workflow design, and the challenges encountered in practice, thereby informing future iterations and potential scaling of the model.

Beyond its immediate scope, the project also highlights its potential for replicability and transferability to other domains and datasets. The creation of evaluation resources that reflect European cultural, social, and linguistic realities is of clear strategic interest, particularly in the context of developing more inclusive and representative language technologies. Equally significant is the collaborative model underpinning this initiative, which leverages the distributed effort of a large number of participants to achieve outcomes that would be difficult to attain individually. This collective approach not only enhances efficiency and coverage, but also creates valuable

opportunities for training, knowledge exchange, and experiential learning within a real-world, multilingual setting.

Acknowledgments

The authors would like to express their sincere gratitude to all participating students, as well as to the professors and programme coordinators from the EMT network, for their commitment and active involvement in this initiative. Their collaboration has been essential to the development and success of the project.

References

- Artetxe, M., Ruder, S., & Yogamata, D. (2020). *A Call for More Rigor in Unsupervised Cross-Lingual Learning*. arXiv:2004.09696. <https://arxiv.org/abs/2004.09696>
- EMT Expert Group. (2022). European master’s in translation—EMT competence Framework. https://commission.europa.eu/system/files/2022-11/emt_competence_fw_2022_en.pdf
- European Language Equality Project. (2022). *Towards digital language equality: Language technology support for European languages*. <https://european-language-equality.eu/>
- European Master’s in Translation network, Directorate-General for Translation, European Commission. (2022). *European Master’s in Translation Competence Framework* https://commission.europa.eu/system/files/202211/emt_competence_fw_2022_en.pdf
- Hendrycks, Dan, Collin Burns, Steven Basart, et al. 2021. “Measuring Massive Multitask Language Understanding.” Paper presented at International Conference on Learning Representations. October 2. <https://openreview.net/forum?id=d7KBjmI3GmQ>.
- Kiraly, D. 2005. “Project-based Learning: A Case for Situated Translation.” *Meta* 50 (4): 1098–1111. <https://doi.org/10.7202/012063ar>
- Korda, D. (2023). Translation training: The use of authentic projects. *Current Trends in Translation Teaching and Learning E*, 10, 302 – 337. <https://doi.org/10.51287/ctl202310>
- Rehm, G., Grützner-Zahn, A., & Barth, F. (2025). Are Multilingual Language Models an Off-ramp for Under-resourced Languages? Will we arrive at Digital Language Equality in Europe in 2030?. <https://arxiv.org/abs/2502.12886>.
- Shumailov, I., Shumaylov, Z., Zhao, Y. et al. AI models collapse when trained on recursively generated data. *Nature* 631, 755–759 (2024). <https://doi.org/10.1038/s41586-024-07566-y>

- Singh, Shivalika, Angelika Romanou, Clémentine Fourier, et al. 2024. "Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation." arXiv.Org, December 4. <https://arxiv.org/abs/2412.03304v2>
- Uribe de Kellett, A., (2022). Real-world translation: learning through engagement. In c. Hampton & S. Salin (Eds.), *Innovative Language Teaching and Learning at University: Facilitating Transition from and to Higher Education*, Research-Publishing.net, 133-142. <https://doi.org/10.14705/rpnet.2022.56.1380>
- Van Egdom, G.-W., Konttinen, K., Vandepitte, S., Fernández-Parra, M., Loock, R., & Bindels, J. (2020). Empowering Translators through Entrepreneurship in Simulated Translation Bureaus. *HERMES - Journal of Language and Communication in Business*, 60, 81–95. <https://doi.org/10.7146/hjlc.v60i0.121312>

A Comparative Study of LLMs for Idiomatic Expression Identification and Translation Equivalent Extraction in an English–Portuguese Subtitle Corpus

Rozane Rebechi

Federal University of Rio Grande do Sul
Porto Alegre - RS, Brazil
rozane.rebechi@ufrgs.br

Diego Alves

Saarland University
Saarbrücken, Germany
diego.alves@uni-saarland.de

Abstract

Idiomatic expressions (IEs) pose challenges for natural language processing due to their non-compositionality and dependence on conventionalized usage. This study investigates the ability of large language models (LLMs) to identify IEs in English and to locate their corresponding translations in Brazilian Portuguese within a parallel subtitle corpus. Using a subset of 10,000 aligned sentence pairs from the OpenSubtitles corpus, two multilingual LLMs were evaluated through zero-shot prompting in two tasks: (i) idiom detection in English sentences and (ii) extraction of their translated equivalents in Portuguese. Model outputs were manually annotated by an expert linguist to assess accuracy and to classify translation strategies based on degrees of idiomaticity and conventionality. Results reveal substantial variation in how models operationalize idiomaticity, with openai/gpt-oss-20b achieving higher precision in idiom identification, while meta-llama/Llama-3.3-70B-Instruct adopts a broader and more permissive interpretation. In the translation identification task, both models demonstrate strong performance in identifying corresponding translations, with most choices classified as either idiomatic and conventional or non-idiomatic but acceptable. Qualitative analysis further shows that LLMs rely on contextual and intertextual cues, sometimes inferring idiomatic meanings beyond the surface form, but also producing false positives driven by lexical familiarity or stylistic features.

1 Introduction

Idiomatic expressions (IEs) pose a well-known challenge for translation, largely due to their non-compositional meanings and cultural specificity (Donthi et al., 2025). Moreover, the definition of what constitutes an IE varies across authors, further complicating their identification and treatment in NLP and translation studies (Ramisch, 2012).

Although LLMs have shown strong performance in many language tasks, their treatment of figurative and conventionalized language is still not fully understood. Idiomatic expressions therefore provide a useful test case for evaluating how these models interpret meaning beyond literal word combinations.

The objective of this paper is to examine how large language models (LLMs) understand the concept of IEs in English by performing two tasks: (i) identifying IEs in sentences extracted from the OpenSubtitles corpus, and (ii) identifying their translations into Portuguese using parallel corpus, in order to assess whether the idiomaticity is preserved across languages. The aim is to evaluate the overall performance of LLMs against the annotations of an experienced human annotator, thereby establishing a baseline for improving prompting strategies for this task as well as refining human annotation guidelines.

Finally, the data extracted from the parallel corpus is used in a case study that analyses the translation strategies applied in the subtitles, providing insights into how IEs are rendered in practice.

2 Related Work

According to Alisoy (2025), idiomatic expressions are best understood within the broader category of phraseological units, defined as conventionalized multiword expressions that function as single semantic units and whose overall meaning cannot be straightforwardly derived from their individual components.

Structurally, idiomatic expressions can take different syntactic forms, including nominal, verbal, adjectival or adverbial, and sentential units, such as proverbs and fixed sayings that function as complete utterances. Semantically, idiomatic expressions are organized along a continuum of com-

positionality. At one end they are fully transparent expressions, whose meanings are largely inferable from their components; in the middle are semi-transparent (motivated) idioms, where figurative meaning is partially grounded in metaphor or metonymy; and at the other end are opaque idioms, whose meanings are not predictable from their individual elements [Alisoy \(2025\)](#).

Within NLP, the automatic identification of idiomaticity represents a well-known challenge due to the idiosyncratic and context-dependent nature of idiomatic expressions ([Sag et al., 2002](#)). Recent advances in LLMs have opened new possibilities for addressing this issue, as these models can capture complex contextual dependencies from large-scale data ([Jurafsky and Martin, 2023](#); [Naveed et al., 2025](#)).

From a Corpus Linguistics perspective, idiomaticity is closely associated with frequency, patterning, and conventionality in authentic language use. Following the definition of [Tagnin \(2005\)](#), conventionality refers to the fact that certain expressions are established through usage within a speech community rather than being freely created by speakers. The idiom principle ([Sinclair, 1991](#)) posits that speakers tend to rely on semi-preconstructed units rather than generating language word by word. Parallel corpora enable the investigation of how expressions are realized across translated contexts. In this sense, they provide a productive interface between Corpus Linguistics and Translation Studies, allowing researchers to examine how idiomatic meanings are preserved, adapted, or neutralized in translation. Existing evidence suggests that, although LLMs can recognize – and translate – figurative language, they often rely on surface-level cues or lexical familiarity, which may lead to false positives and overgeneralization ([Castaldo and Monti, 2024](#)).

Finally, studies in Translation Studies have shown that translated texts tend to exhibit systematic features, often described as translationese, including tendencies toward simplification and explicitation ([Rubino et al., 2016](#)). These features are particularly relevant in the analysis of subtitles, which are typically produced under temporal and spatial constraints and may vary in quality and conventionality, especially in the case of fan-generated content.

3 Methodology

In this section, we present the parallel corpus, the LLMs selected for the tasks, the prompting strategies, and the evaluation methodology.

3.1 Parallel Corpus

To assess whether LLMs can correctly identify IEs in English and the tokens corresponding to their translations in Brazilian Portuguese, we used the OpenSubtitles corpus v2024 ([Lison and Tiedemann, 2016](#)), a large multilingual parallel corpus compiled from movie and television subtitles. The English–Brazilian Portuguese portion of the corpus contains 115,123,153 aligned sentence pairs. As this study is intended as a preliminary experiment to examine the feasibility of the task for the human annotation and compare differences in performance across LLMs, we processed the first 10,000 sentence pairs of the corpus, assigning a unique sentence ID to each aligned pair in both the English and Brazilian Portuguese subsets.

3.2 Large Language Models

Two open-weight multilingual LLMs were chosen for this study:

1. `openai/gpt-oss-20b`¹ is an open-weight large language model released by OpenAI under the Apache 2.0 license. It has approximately 21 billion parameters and is designed for low-latency inference, local deployment, and specialized tasks, with support for reasoning and tool use ([OpenAI, 2025](#)).
2. `meta-llama/Llama-3.3-70B-Instruct`² is a multilingual instruction-tuned large language model developed by Meta. It contains 70 billion parameters, supports a 128k-token context window, and is optimized for multilingual dialogue, reasoning, and text generation tasks ([Meta, 2024](#)).

Thus, we compare two state-of-the-art models with different architectures and sizes in order to establish a baseline for further LLM comparisons.

3.3 Prompting Strategies

For idiomatic expression identification, we adopted a zero-shot prompting strategy in which each sen-

¹<https://huggingface.co/openai/gpt-oss-20b>

²<https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>

tence was presented individually to the model together with a system prompt defining the notion of idiomatic expression and providing illustrative examples. The prompt instructed the model to determine whether the sentence contained an idiom and to return a structured one-line output consisting of a binary label (1 = idiom present; 0 = no idiom present) followed by the detected expression, or NA when no idiom was identified. This constrained output format was designed to facilitate automatic parsing and large-scale evaluation. All generations were produced using the models’ default decoding configuration, with no modifications to parameters such as temperature or sampling settings. The full prompt is provided in Appendix A.

For the identification of idiomatic expression translated from English to Brazilian Portuguese, we applied a similar zero-shot prompting strategy. The full prompt is provided in Appendix B. In this setting, the model was provided with an English sentence containing a pre-identified idiomatic expression, its corresponding idiom span, and the aligned Portuguese translation. Importantly, only sentences that were classified as containing an idiomatic expression in the first task (i.e., those assigned the label 1) were considered for this step. The system prompt instructed the model to extract the Portuguese tokens that correspond to the translation of the idiomatic expression, returning them in a strictly formatted single-line output to ensure consistency and facilitate automatic evaluation.

In this study, we implemented our experiments using the Hugging Face Transformers Python library³. The models were loaded through the high-level pipeline API, enabling straightforward deployment of instruction-tuned chat models. This framework was used for both the idiomatic expression detection and translation identification tasks.

3.4 Evaluation

The LLMs outputs were first pre-processed to identify the answer regarding each task.

For the IE identification task, we selected the sentence IDs for which the LLMs identified an expression, generating a list for each model consisting of the sentence ID, the original sentence, and the extracted idiomatic expression. These lists were manually annotated to verify whether the extracted expression corresponds to a true idiomatic

expression in English, according to the following labels:

- **Yes** — the identified expression is a conventionalized idiomatic expression in English. For example: *poster child* in *You, convincing yourself you’re in control, how lucky to be so delusional. The poster child of fortuitous living. Fortuitous?*
- **No** — the expression is not an idiomatic expression in English or is not used with an idiomatic meaning. For example: *white elephant* in *The white elephant of the King of Siam, the sacred animal,... the supreme incarnation of Buddha, has vanished.*
- **Inconclusive** — the identified expression may be idiomatic, but there is insufficient context to make a definitive judgment in the given example. Example: *throw to the lions* in *Throw them to the lions!*
- **Error** — LLM error or hallucination (e.g., incorrect or non-existent extraction). For example, gpt-oss-20b model identified *go going nowhere* as an IE extracted from *We’re too tired... we’re not going anywhere!*

For the translation evaluation task, we manually annotated the Portuguese outputs produced by the LLMs (i.e., the words in the Portuguese sentences corresponding to the translation of the English IE) to assess how the English idiomatic expressions were rendered in Brazilian Portuguese. Each identified translation was classified according to the following labels:

- **Idiomatic and conventional** — the English expression was translated using an existing and conventional idiomatic expression in Brazilian Portuguese. For example, *She’s making fun of you right now.* was translated as *Ela está tirando sarro de você agora.* [lit. She is making fun of you now].
- **Non-idiomatic but conventional** — the English expression was translated using a non-idiomatic expression, but the translation is correct and conventional in usage. For example, *We’re combing through manifests.* was translated as *Estamos verificando registros.* [lit. We are checking registers].

³<https://huggingface.co/docs/transformers/index>

- **Idiomatic but non-conventional** — the English expression was translated using an idiomatic expression that is not commonly used in Brazilian Portuguese, but is still considered acceptable as it preserves the intended meaning. For example, *I can't find anything in the mine, and so I must drown my sorrows in this bottle.* was translated as *Eu não consigo encontrar nada na mina, e então eu devo afogar minha dores nesta garrafa.* [lit. I cannot find anything in the mine, and so I must drown my pains in this bottle] (instead of *afogar as mágoas*) [lit. drown the sorrows].
- **Inaccurate** — the translation does not correctly convey the meaning of the source idiomatic expression. For example, *Look, for this to hold water with a grand jury, there can be no other person who could have touched it.* was literally translated as *olha, para isso segurar a água com um grande júri, não pode haver outra pessoa quem poderia ter tocado nele.*, although hold water could be idiomatically translated as *colar* [lit. glue] or *parar em pé* [lit. stand on foot].
- **Error** — LLM error or hallucination (e.g., incorrect or irrelevant output). For example, the translation of the EI in *You have to make it.* was identified as *faça-lo" but with correct accent? The phrase.*

The annotations were performed by a linguist with many years of expertise in translation from English to Brazilian Portuguese.

4 Results

In the following subsections, we present the results regarding the identification of English idiomatic expressions and the analysis of the main translation strategies in the Brazilian Portuguese corpus.

4.1 Idiomatic Expression Identification Using LLMs

From the LLM outputs, we observe substantial differences in the number of idiomatic expressions identified within the first 10,000 lines of the Open-Subtitles corpus. While gpt-oss-20b identified 928 instances, Llama-3.3-70B-Instruct exhibited a broader interpretation of idiomatic expressions, identifying 1,387 cases.

Figure 1 presents the distribution of results for the IE identification task.

Considering the objective of the task (i.e., correctly identifying sentences containing idiomatic expressions), we define accuracy as the proportion of cases in which the label **Yes** was assigned. Under this definition, gpt-oss-20b achieves a substantially higher accuracy (83.5%) compared to the other model (43%).

The proportion of cases in which the identified expression does not correspond to an idiom is also higher for LLaMA-3.3-70B-Instruct. However, in 209 cases (15.1%), this model identifies sequences of tokens that may correspond to idiomatic expressions, but the label is marked as **Inconclusive**, due to insufficient context for a definitive judgment. This suggests that LLaMA-3.3-70B-Instruct adopts a broader and more permissive interpretation of what constitutes an idiomatic expression.

Finally, the number of errors is also higher for LLaMA-3.3-70B-Instruct, further supporting the overall superior performance of gpt-oss-20b under this evaluation setting.

The intersection analysis shows that the highest agreement between models occurs in the **Yes** category (298 shared cases), indicating consistent identification of idiomatic expressions. In contrast, overlap is lower for **No** (26 cases) and **Inconclusive** (15 cases), suggesting greater variability in non-idiomatic and borderline decisions. No overlap is observed for error cases, indicating model-specific failures. Overall, agreement is stronger for clear idiomatic cases, while divergence increases in more ambiguous contexts.

To examine whether broader sentential context facilitates the identification of idiomatic expressions, we compared the mean length (in number of words) of English source sentences across classification outcomes (**Yes**, **No**, **Inconclusive**, and **Error**). Sentence length was used as a proxy for the amount of contextual information available to the models.

For each model, pairwise comparisons between annotation groups were conducted using Welch's two-sample t-tests, which do not assume equal variances. Because multiple comparisons were performed, Bonferroni correction was applied to control the family-wise error rate. Table 1 reports only statistically significant contrasts involving the **Yes** category.

Results indicate that, for both models, correctly

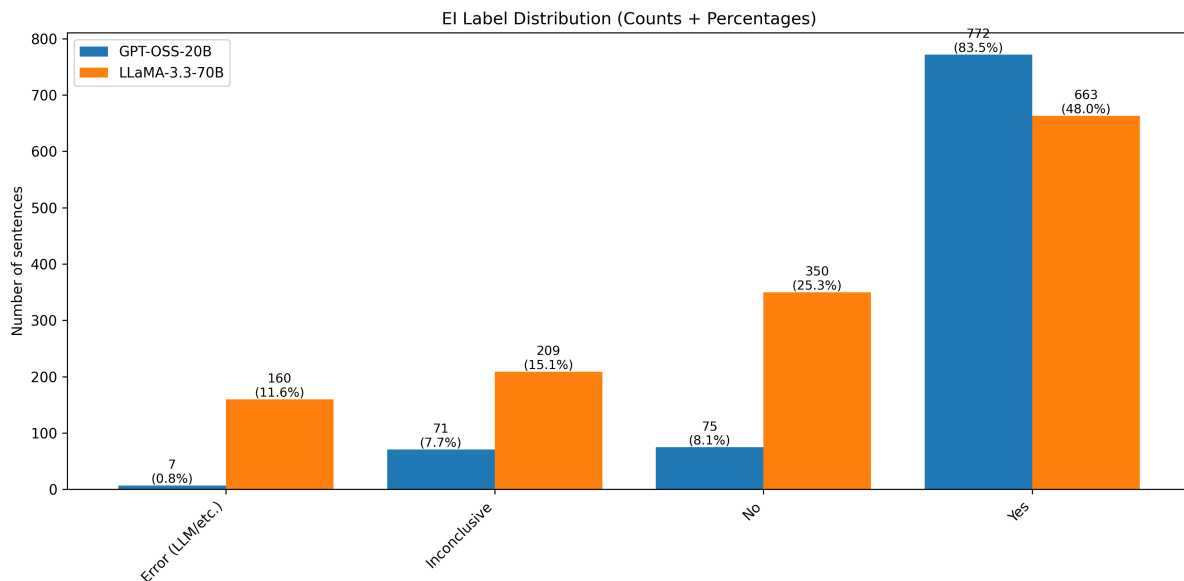


Figure 1: Distribution of idiomatic expression (IE) detection labels for gpt-oss-20b and LLaMA-3.3-70B-Instruct. Percentages are computed relative to the total number of IEs identified by each model over the first 10,000 sentences of the OpenSubtitles corpus.

identified idiomatic expressions tended to occur in longer sentences. For gpt-oss-20b, sentences classified as *Yes* were significantly longer than those classified as *Inconclusive*. For Llama-3.3-70B-Instruct, *Yes* sentences were significantly longer than both *Inconclusive* and *Error* cases. No significant differences were found between *Yes* and *No* in either model after correction. This suggests that additional context within the same sentence appears to help reduce uncertainty (*Inconclusive*) and errors. However, sentence length alone does not fully explain successful idiom recognition, as no significant differences were observed between *Yes* and *No*.

In this study, we focused on evaluating LLM performance only for sentences in which the models indicated the presence of an idiomatic expression. Accordingly, our analysis emphasized precision/accuracy rather than recall. One limitation of this approach is the absence of recall estimates. The larger number of idiomatic expressions identified by the LLaMA model suggests that its recall may be higher; however, this hypothesis requires a more detailed investigation and is left for future work.

4.2 Translation Strategies in OpenSubtitles

With the idiomatic expressions (IEs) correctly identified by the LLMs (i.e., considering the correct outputs from both models, while excluding over-

lapping cases to avoid duplication), we can analyse the most common translation strategies used in the Brazilian Portuguese corpus translated from English as shown in Figure 2.

The results indicate that the majority of expressions are translated using valid and conventional idiomatic equivalents in Brazilian Portuguese, for example, *none of their business*, translated as *não são da sua conta*. This is followed by non-idiomatic but accurate translations, for example, *madly in love*, literally translated as *loucamente apaixonado* instead of *perdidamente apaixonado* [lit. holplessly in love], which is more conventional. The number of inaccurate translations is higher than the number of cases classified as idiomatic but non-conventional. An example of an inaccurate translation is the literal rendering of *This is loving our neighbor as ourselves* as *Isto ama nosso vizinho como nós mesmos*, instead of the conventionalized expression *amar ao próximo como a ti mesmo* [lit. to love the one next to you as you love yourself].

In Figure 3, we present the distribution of translation strategies for each LLM separately.

The per-model analysis reveals different output patterns across the LLMs. Although the category idiomatic and conventional is the most frequent for both models, the gap between this class and non-idiomatic but conventional is larger for gpt-oss-20b. The remaining categories show broadly similar distributions, with gpt-oss-20b presenting a relatively

Model	Yes	No	Inconcl.	Error	Significant contrasts
GPT-OSS-20B	10.25 (6.87)	8.52 (6.48)	6.69 (6.71)	6.43 (3.36)	Yes > Inconclusive ($p < .001$)
Llama-3.3-70B-Instruct	10.51 (7.25)	9.24 (6.80)	7.05 (5.44)	8.63 (5.53)	Yes > Inconclusive ($p < .001$); Yes > Error ($p = .003$)

Table 1: Mean sentence length (words) by classification outcome and significant pairwise contrasts

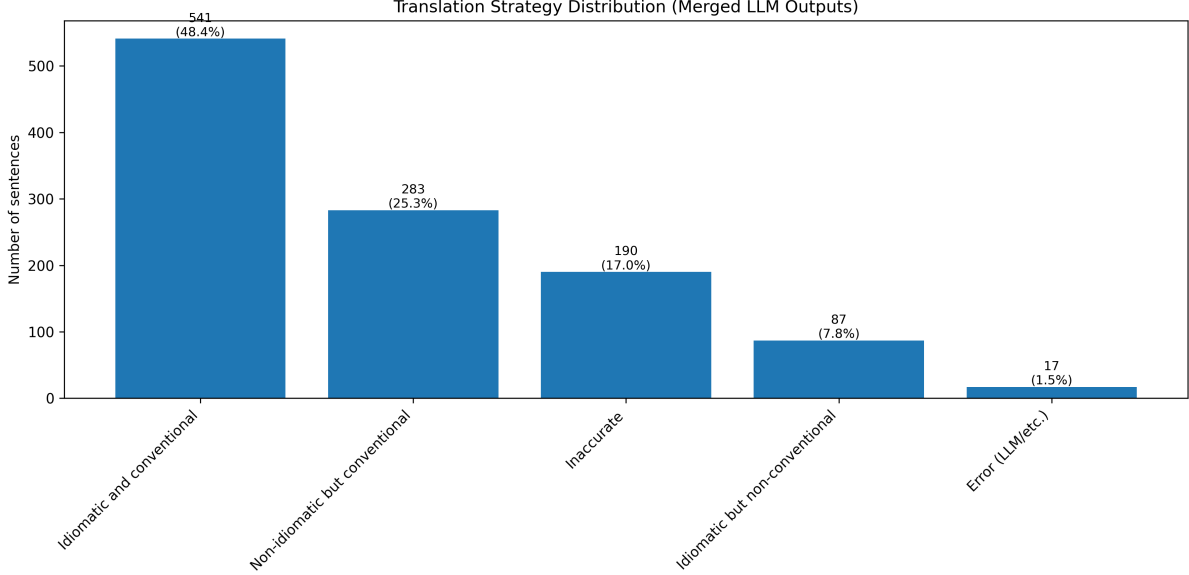


Figure 2: Distribution of translation strategy labels for the merged unique set of correctly identified idiomatic expressions (duplicates removed). Percentages are calculated relative to the total number of valid translation cases for each model and for the merged dataset.

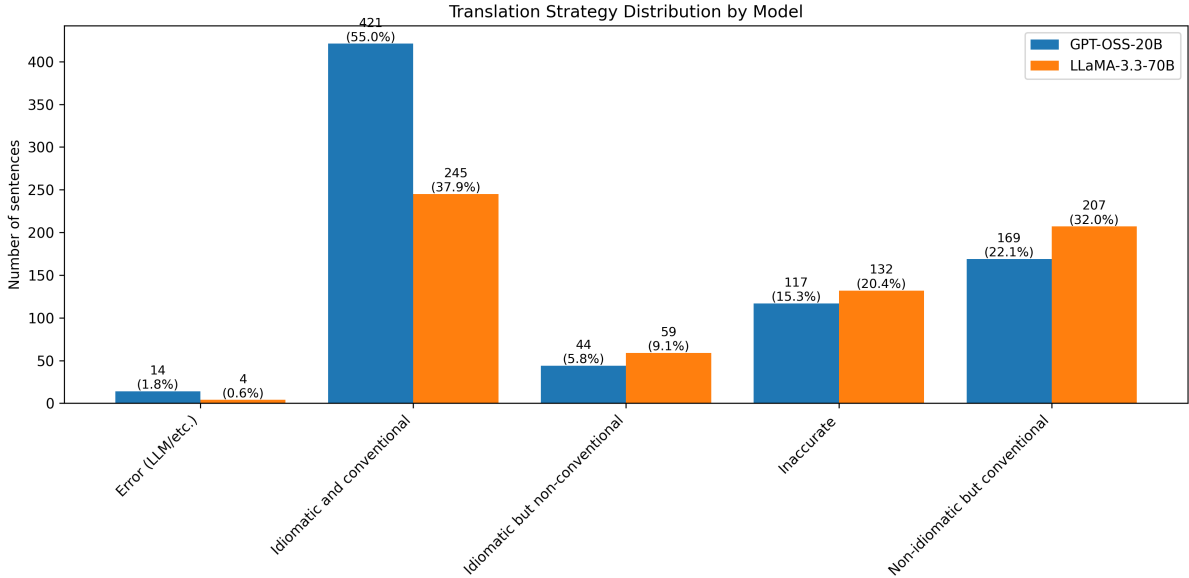


Figure 3: Distribution of translation strategy labels for gpt-oss-20b and Llama-3.3-70B-Instruc models. Percentages are calculated relative to the total number of valid translation cases for each model.

higher number of errors in the identification of the translated tokens.

The merged analysis reduces the impact of potential idiomatic expressions missed by one model

but identified by the other. This suggests a possible strategy for future work: combining the outputs of several LLMs that perform well in the initial IE identification task and applying a voting threshold,

whereby an expression is accepted as idiomatic when identified by the majority of models.

Still, when considering the overall percentages in the merged set, we observe that the class distribution remains closer to that of the gpt-oss-20b model, thus yielding broadly similar material for discussion and further supporting the stronger performance of this model compared with Llama-3.3-70B-Instruct under the present evaluation setting.

Finally, we observe that performance on the translation identification task is relatively high for both models, with error rates of only 1.8% for gpt-oss-20b and 0.6% for Llama-3.3-70B-Instruct.

5 Qualitative Analysis

A qualitative investigation of the LLM outputs reveals important tendencies in how models interpret idiomaticity, process contextual information, and identifies translation equivalents.

A first interesting pattern concerns the extent to which LLMs rely on broader contextual and intertextual knowledge when identifying idiomatic expressions. In several cases, the models—particularly Llama-3.3-70B-Instruct—appeared to infer idiomatic expressions that were not explicitly present in the input sentence but were strongly evoked by the context. For instance, in the subtitle *The wall has been torn down; the horse in which the soldiers are hidden, is in the city.*, the model inferred the expression *Trojan horse*, despite its absence in the surface form. A similar phenomenon was observed in the sentence *And you, too ... my son Brutus?*, where the model identified the Latin expression *Et tu, Brute*. These cases suggest that LLMs may draw on intertextual associations, going beyond the immediate linguistic input. While this may reflect a form of enriched semantic processing, it also raises questions regarding overgeneration and the boundaries between inference and identification in annotation tasks. Human annotators, by contrast, may lack sufficient context to draw conclusions.

Another recurring issue involves false positives driven by the recognition of familiar or conventionalized lexical patterns rather than contextual meaning. For example, the expression *white elephant* was identified by both LLMs as idiomatic in contexts where it clearly referred to a literal animal. Similarly, sequences such as *PURSUIING FATE* were often misclassified as idiomatic expressions, although they appear to function as titles or stylistic

elements within subtitle conventions, such as capitalization.

Finally, the qualitative analysis of the translation data highlights issues that are not directly attributable to the models themselves but to the nature of the OpenSubtitles corpus, whose subtitles are not always produced by professional translators and may contain non-standard or literal renderings. This is illustrated by cases such as the literal translation of *We will not be run out.* as *Nós não seremos corridos para fora.*, instead of more conventional alternatives such as *Não seremos expulsos* or *Não seremos escoraçados*. An inaccurate translation in OpenSubtitles, however, does not seem to compromise the identification of the translation by the LLMs.

6 Conclusion and Future Work

This study examined the ability of two multilingual LLMs to identify idiomatic expressions in English subtitle data and to locate their translations in Brazilian Portuguese within a parallel corpus. The results show clear differences between the models in the first task. While gpt-oss-20b achieved higher precision in idiom identification, Llama-3.3-70B-Instruct adopted a broader interpretation of idiomaticity, leading to a larger number of candidates but also more false positives. In the translation identification task, both models performed well overall. Most of the translations found in OpenSubtitles correspond to acceptable or conventional renderings in Portuguese; however, a considerable number of inaccurate translations is still present.

The findings also suggest that idiomatic expressions remain a useful benchmark for evaluating semantic sensitivity in LLMs, since they require models to go beyond literal word combinations and attend to contextualized meaning. At the same time, the qualitative analysis showed that models may overgenerate idioms on the basis of lexical familiarity or broader cultural associations.

Future work may expand this study in several directions. A first line of research concerns larger datasets, additional models, and alternative prompting strategies. Another possibility is to investigate the relation between idiomaticity, conventionality, and processing effort through Information Theory. Less conventional translation choices may involve higher surprisal and therefore greater cognitive or post-editing cost.

A further direction is the systematic compari-

son between human and machine translation in idiomatic contexts. Human translators may produce more idiomatic and contextually appropriate solutions, especially in texts with a high density of figurative language (Rebechi et al., 2024). Comparing human and machine outputs could therefore offer useful insights into the strengths and limitations of each approach.

7 Limitations

This study presents several limitations that should be acknowledged. First, we used a limited subset of the OpenSubtitles corpus for the analysis, consisting of the first 10,000 sentences. Therefore, the selected sentences are representative of only a small number of series and movies in the dataset. In future work, we plan to randomize the sentence selection in order to obtain a more representative sample.

Regarding the LLMs, we evaluated only two models and tested a single prompting strategy for each task. Since the results for IE identification were not satisfactory, future experiments should include additional models and alternative prompting strategies (e.g., few-shot prompting).

Finally, we also acknowledge a limitation related to the annotation process, as it was conducted by only one expert. We are currently preparing specific annotation guidelines for future annotation campaigns with more annotators, which will also focus on calculating recall for the IE identification task.

References

- Hasan Alisoy. 2025. A Taxonomic Approach to Structural and Semantic Dimensions in English Phraseology. *Porta Universorum*, 1(4):72–79.
- Antonio Castaldo and Johanna Monti. 2024. Prompting large language models for idiomatic translation. In *Proceedings of the 1st Workshop on Creative-text Translation and Technology*, pages 32–39.
- Sundesh Donthi, Maximilian Spencer, Om B Patel, Joon Young Doh, Eid Rodan, Kevin Zhu, and Sean O’Brien. 2025. Improving llm abilities in idiomatic translation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 175–181.
- Daniel Jurafsky and James H. Martin. 2023. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Online manuscript. Draft version.
- Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 923–929.
- Meta. 2024. Llama-3.3-70b-instruct. <https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>. Hugging Face model card, accessed April 11, 2026.
- Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.
- OpenAI. 2025. *gpt-oss-120b gpt-oss-20b model card*.
- Carlos Ramisch. 2012. A generic framework for multiword expressions treatment: from acquisition to applications. In *Proceedings of ACL 2012 Student Research Workshop*, pages 61–66.
- Rozane Rebechi, Nathalia Marcon, Guilherme Fallier, and Aline Villavicencio. 2024. Machine translation of multiword expressions: Double Dutch or crystal clear. *Proceedings of New Trends in Translation and Technology*, pages 116–137.
- Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef Van Genabith. 2016. Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 960–970.
- Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15. Springer.
- John Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.
- Stella Esther Ortweiler Tagnin. 2005. *O jeito que a gente diz: expressões convencionais e idiomáticas*. Disal.

A Idiomatic Expression Identification Prompt

SYSTEM_PROMPT = """ You are a linguistics expert.

Objective: Your task is to read and analyze a sentence and determine whether it contains an idiomatic expression. An idiomatic expression (or idiom) is a phrase whose meaning is different from the literal meanings of the individual words that

make it up. Instead of being interpreted word-for-word, the phrase has a figurative meaning commonly recognized by native speakers. Idioms are frequently used in everyday language and often reflect cultural habits and patterns of speech. For example, the expression "break the ice" does not literally mean to break ice, but instead means to start a conversation or make people feel more comfortable in a social situation.

For each sentence, you must respond with EXACTLY ONE LINE containing: 1. A number (0 or 1) followed by a tab, output 1 if an idiomatic expression is present and output 0 if no idiomatic expression is present. 2. The idiomatic(s) expression(s) if present, separated by —, or NA if not

IMPORTANT RULES: - Your entire response must be a single line - Start with either "0" or "1" - Follow it with a tab character - Then the exact idiomatic expression text, or NA if label is 0 - Do not include any other text, explanations, or analysis

Example input: "To start the meeting, the teacher told a joke to break the ice."

Example output: 1\tbreak the ice

Example input: "The cat sat on the mat."

Example output: 0\tNA

Remember: PLEASE, the information must be given in a separate line, starting with "FINAL OUTPUT:". For example, "FINAL OUTPUT:0". ""

B Translation Identification Prompt

SYSTEM_PROMPT = "" You are a linguist expert in translation from English to Portuguese.

Objective: Your task is to analyse an English sentence (EN_SENT) containing an idiomatic expression (IE) and check a Portuguese translated sentence to identify which tokens in the Portuguese sentence (PT_SENT) correspond to the translation of the idiomatic expression from English.

For group of sentence in English, idiomatic expression in English, translated sentence in Portuguese, you must respond with EXACTLY ONE LINE containing the words in the Portuguese sentence that corresponds to the translation of the idiomatic expression in English.

IMPORTANT RULES: - Your entire response must be a single line - Do not include any other text, explanations, or analysis

Example input: EN_SENT: Watch your back Watch your back. Right, K-6 we're going to K-6. Hate to say I told you so. IE: Watch your back PT_SENT: tome cuidado, certo, k-6, vamos para k-

6, odeio diga que eu disse a você, bem, nós vamos conseguir para testá-lo imediatamente, saberemos em breve,

Example output:

EI TRANSLATED TOKENS ARE: tome cuidado

Remember: PLEASE, the translated token information must be given in a separate line, starting with "EI TRANSLATED TOKENS ARE:". For example, "EI TRANSLATED TOKENS ARE: tome cuidado". ""

Navigating the "Fluency Trap": Cognitive Risks and Structural Challenges in Korean-English Patent Translation Post-Editing

Jinam Lee

Korea Institute of Patent Information
/ Seoul, Republic of Korea
thekeytoyou@kipi.or.kr

Jaehee Ryu

Korea Institute of Patent Information
/ Seoul, Republic of Korea
kittyhawk1903@kipi.or.kr

Janet Sohlhee Yu

Korea Institute of Patent Information
/ Seoul, Republic of Korea
janetsyu@kipi.or.kr

Abstract

This study investigates the fluency trap in Korean (KR) – English (EN) patent translation, where advanced NMT/LLM outputs mask critical technical errors through deceptive syntactic coherence. We analyze how KR-specific structural disparities, such as zero-anaphora and head-final dependencies, lead to silent omissions that fatally distort a claim's legal scope. Our findings suggest a potential cognitive halo effect, where linguistic smoothness erodes the analytical scrutiny required to detect misattributed technical agency. To mitigate this risk, we propose the Critical Revision Model (CRM), shifting the post-editing paradigm from stylistic refinement to adversarial technical verification. We conclude that the legal integrity of intellectual property depends not on linguistic naturalness, but on the rigor of human-led syntactic and technical auditing.

1 Introduction

The global landscape of intellectual property (IP) protection has undergone a seismic shift, driven by surging cross-border patent filings and reliance on automated linguistic technologies. Within the framework of the Patent Cooperation Treaty (PCT), the demand for high-fidelity translation between linguistically distant pairs, specifically Korean (KR) and English (EN), has pressured both institutional infrastructure and human expert reviewers. As the volume of the Korean PCT applications surpassed 23,000 in 2024 (WIPO, 2025), integrating Neural Machine Translation (NMT) and, more recently, Large Language Models (LLMs) into the pipeline became essential.

Yet this integration carries an underappreciated risk. As NMT and, more recently, LLMs produce increasingly fluent output, translation errors became harder to detect. In patent examination where a single omitted technical modifier can alter the entire legal scope of a claim, this represents a systemic vulnerability.

This study investigates the fluency trap in technical document translation drawing on professional practice in KR-EN patent revision and supported by existing literature. We analyze its structural and cognitive origins, and propose the Critical Revision Model (CRM) as a practical countermeasure for professional post-editors.

2 Evolution of NMT and Structural Risks in KR-EN Patent Translation

2.1 From SMT to LLMs: A Shifting Risk Profile

Statistical Machine Translation (SMT) output was often aesthetically poor and fragmented, yet these very flaws functioned as transparent markers of translation failure. In contrast, the NMT systems currently deployed in platforms such as KIPRIS of the MOIP of Korea, WIPO PATENTSCOPE and the EPO's ESPACENET, produce syntactically coherent sentences but made errors invisible within fluent prose.

By late 2024, the integration of LLMs within secured and closed-loop environments offers superior document-level consistency and terminological coherence, but introduce what we term technical hallucinations—content that is linguistically flawless yet factually or legally unsupported. Revisers must now move from managing visibly broken output to auditing deceptively perfect English.

2.2 Structural Disparities in KR-EN Translation

The linguistic distance between Korean and English is the primary driver of structural translation error. Korean follows Subject-Object-Verb (SOV) order, while English follows Subject-Verb-Object (SVO); this typological divergence requires extensive syntactic reordering—a transformation that MT systems handle inconsistently (Tamaoka et al., 2024; Cho et al., 2014). Within this broader cross-linguistic gap, two Korean-specific structural features are particularly problematic in KR-EN translation contexts and specifically critical in the patent domain.

- Zero-Anaphora (Pro-Drop)

Korean frequently omits subjects or objects when they can be inferred from context—a feature known as pro-drop (Kim et al., 2020). Human translators resolve these omissions through contextual inference; MT systems must perform complex discourse reasoning to identify the correct antecedent (Park et al., 2015). Failure at this step produces meaning errors; grammatically sound but technical agency, which component performs which function, is misattributed.

In patent documents, patent specifications and claims involve multiple agents intertwined in complex causal relationships. In this context, misattributing agency due to zero-anaphora is exceptionally critical, as it directly distorts the legal scope of the invention, leading to fatal defects in determining novelty or non-obviousness.

- Long-Distance Dependencies and Modifier Omission

Due to Korean’s head-final syntactic structure, modifying clauses typically precede the nouns they describe, and main verbs are placed at the very end of the sentence. This creates long-distance dependencies that force NMT systems to retain large amounts of context before generating the corresponding English structures.

Independent claims in US patents average around 130 words in a single sentence (Kuhn et al., 2019), while Korean claims typically span 40–90 *eo-jeol*, a Korean spacing unit where 1 *eo-jeol* typically converts to 2.5 to 3 English words, creating exceptionally long sentences. NMT performance degrades significantly as sentence length increases; specifically, BLEU scores approach zero for sentences exceeding 50 words

due to fixed context limitations (Koehn & Knowles, 2017). This degradation often results in the silent omission of critical technical modifiers.

Because these modifiers, such as specific temperature ranges, material properties, or spatial constraints, act as legal delimiters, their omission "expands" the claim scope beyond its intended limit.

3 Qualitative Findings: Two Compounding Failure Modes

The central finding of this study is that advanced NMT creates two mutually reinforcing failure modes in patent revision: a technical layer of structural omission, and a cognitive layer of attentional erosion. Together, these constitute the fluency trap.

To comply with confidentiality constraints on MTPE workflows, we grounded our qualitative methodology in over a decade of KR-EN patent revision experience rather than a quantitative corpus. We purposively sampled recurrent error patterns exclusively from publicly available Korean patent applications. The corresponding English outputs, generated by either built-in NMTs of open databases such as KIPRIS and PATENTSCOPE, or Large Language Models (LLMs), were then rigorously verified by the authors, acting as expert revisers.

3.1 Technical Layer: Structural Omission and Distortion

NMT prioritizes sentence-level cohesion, often truncating or distorting technical parameters embedded in pre-verbal modifier chains.

- Unresolved Pro-drop of Technical Processes

Type	Description
Application #	1020200023515
MT Source	NMT (KIPRIS Built-in NMT, 2026)
Source (in KR)	5cm×5cm× 5cm 아크릴 몰드에 5ml 를 넣고(5ml was added)
MT (in EN)	5 ml (of [mixture] is pro-dropped) was added to a 5 cm×5 cm×5 cm acrylic mold
MTPE (Human)	5 ml of “the resulting mixture” is poured into a 5 cm × 5 cm × 5 cm acrylic mold

Table 1: Example of Unresolved Pro-Drop.

In this description of a chemical process, the Korean source naturally drops the object “the (resulting) mixture” based on the context. The NMT system fails to resolve this zero-object, outputting a literal and ambiguous “5 ml was added.” This omission creates a critical ambiguity in a patent claim where material precision and reproducibility are paramount.

- Ambiguity of Modifier Scope in Long-Distance Dependencies

Type	Description
Application #	PCT/KR2022/015680
MT Source	NMT (WIPO Translate, 2026)
Source (in KR)	상기 단계(<i>step s200</i>)에서 정렬된(<i>aligned</i>) 상기 잉크방울 측정패드(<i>ink drop measurement pad</i>)의 제 1, 2 전극
MT (in EN)	the first and second electrodes of the ink drop measurement pad aligned in step s200
MTPE (Human)	the first and second electrodes of the aligned ink droplet measurement pad in step s200

Table 2: Example of Ambiguity of Modifier Scope in Long-Distance Dependencies.

Due to head-final syntax, the NMT misplaced “aligned” at the end, leaving it unclear whether it modifies the “pad” or the “electrodes.”

- Distortion of Logical Connections in Nested Structures

Type	Description
Application #	1020240190725
MT Source	NMT (KIPRIS Built-in NMT, 2026)
Source (in KR)	하이브리드 열교환기는 전열부 (<i>heat transfer unit</i>)가 반복 적층되는 (<i>repeatedly stacked</i>)_판 형상의 제 1 내지 제 4 플레이트 (<i>plates</i>) 포함하는 것으로,
Literal Translation	the hybrid heat exchanger [heat transfer unit-Subj] [repeatedly stacked-Mod] ... [plates-Obj] [<i>includes-Verb</i>]
MT (in EN)	the hybrid heat exchanger includes plate-shaped first and fourth plates in which heat transfer units are repeatedly stacked
MTPE (Human)	the hybrid heat exchanger in which a heat transfer unit includes first and fourth plate-shaped plates that are repeatedly stacked

Table 3: Example of Distortion of Logical Connections in Nested Structures.

The NMT output in Table 3 is grammatically flawless but completely distorts the structural logic of the invention. It claims that the heat transfer unit(s) are stacked in the plates, whereas the source text indicates that the plates are stacked to form the unit.

3.2 Cognitive Layer: Cognitive Ease and Erosion of Attention

The second failure mode is rooted in cognitive ease. High-fluency MT output may induce what Kahneman (2011) describes as System 1 (Quick and little or no effort thinking) processing—a fast, instinctive and emotional mode of thinking where syntactic smoothness is subconsciously equated with factual accuracy. This potential halo effect, particularly in the context of patent documents, may divert a reviser’s attention from technical precision to superficial stylistic satisfaction.

- Example 1: Reinforcing the Halo Effect in LLMs

Source (KR): “(*missing agency*) 상하거리를 주기적으로 측정하여...제어하는 **컨트롤러**” (Application No.: PCT/KR2024/020466)

→ Technical context: A separate sensor unit, which is pro-dropped technical agency here, measures the distance while the controller merely receives the data.

LLM (EN, GPT-5.5, 2026): “A controller that periodically **measures** a vertical distance and controls ...”

→ Halo effect: Assigning the verb, “measures” directly to the “controller” to satisfy English SVO requirement. In fact, the technical agency is distorted.

- Example 2: Fluency Trap in Modifier Attachment

Source (KR): “상기 연료전지 스택을 냉각시키고(*cooling the fuel cell stack*) 온도가 상승한 냉각수(*the cooling water whose temperature increases*)를 재냉각(*re-cooling*)하기 위해 순환시키는 복수의 냉각경로(*a plurality of cooling paths*)를 포함하는 냉각수 순환계통” (Application No.: PCT/KR2023/016469)

→ Technical Context: The phrase “cooling the fuel cell stack” modifies the cooling water whose temperature subsequently increases, and a plurality of cooling paths are for re-cooling the same.

NMT (EN, WIPO Translate, 2026): “A coolant circulation system including a plurality of cooling paths configured to cool the fuel cell stack and circulate to re-cool the coolant of which the temperature has risen”

→ Fluency Trap: The MT output exhibits a high degree of linguistic fluency and technical plausibility. Because it presents a seemingly logical interpretation in which the cooling paths “cool” the fuel cell stack, the mistranslated dependency structure is easily masked, reducing the likelihood of detection during post-editing.

Although appearing in different technical contexts, both examples illustrate how structural challenges - whether unsolved pro-drop (Example 1) or mistranslated modifier attachments (Example 2) can lead to deceptive agency distortion. Whether generated by an up to date LLM or an NMT system, this deceptively fluent output (Way, 2018) systematically appears to erode the analytical scrutiny essential for patent translation accuracy. The human-like fluency of these advanced models masks underlying factual errors, which can create a dangerous illusion of completeness and accuracy.

4 Proposed Strategy: The Critical Revision Model (CRM)

To counteract the technical and cognitive problems described above, we propose the Critical Revision Model (CRM)—a framework that shifts the reviser’s role from stylistic editing to adversarial technical verification. The CRM treats NMT output not as a reliable draft, but as a technically suspect hypothesis that must be stress-tested through the following pillars:

- Pre-emptive Agency Mapping: To neutralize the NMT’s halo effect, the reviser must first deconstruct the Korean source to map the technical agency—identifying exactly which component performs which action. By establishing this independent mental anchor before reading the English output, the reviser creates a cognitive firewall against “Subject-Object Swapping” often caused by the NMT’s re-ordering process.
- Adversarial Dual-Pass Strategy: A linear review is insufficient to break the trance of fluent NMT. The CRM utilizes a bifurcated approach:

Deconstructive Review on Technical Integrity: Ignoring grammar entirely to focus exclusively on “Silent Omissions.” Every modifier and numerical delimiter such as “via a sensor” is cross-referenced to ensure technical parameters were not sacrificed for sentence flow.

Reconstructive Review on Structural Alignment: Only after technical accuracy is locked, ensuring that English long-distance dependencies and nested clauses correctly mirror the Korean legal intent.

5 Conclusion

This study identifies a fundamental paradox in the evolution of patent MT: as machine-generated output becomes more fluent, the technical risks embedded within it become harder to detect. The fluency trap—comprising structural omission at the technical layer and attentional erosion at the cognitive layer—poses a systemic risk that is not mitigated, but amplified, by improvements in linguistic quality.

For patent revision, this demands a paradigm shift: from stylistic editing to rigorous technical

verification. The Critical Revision Model (CRM) provides a practical framework for this shift, equipping professional revisers to counteract the cognitive biases that high-fluency MT predictably induces.

In an increasingly automated translation landscape, fluency can no longer be equated with reliability. The legal integrity of intellectual property depends not on how natural the output sounds, but on how rigorously it is examined.

Discussion on Limitations

This study is based on qualitative analysis of KR-EN patent pairs from public databases. Future research should include quantitative error-frequency analysis across different technical domains such as biotechnology vs. semiconductors. Additionally, the impact of LLM-based prompting on reducing "technical hallucinations" warrants further investigation.

Ethics Statement

This study utilizes publicly available patent documents from PATENTSCOPE and KIPRIS, ensuring no infringement on data privacy or intellectual property rights. We acknowledge that while NMT models improve fluency, they may introduce structural errors that affect the legal scope of patents. Our analysis aims to mitigate these risks by highlighting the necessity of professional human post-editing.

Acknowledgments

The authors would like to acknowledge the support of our institution in facilitating the use of these professional resources for academic research.

References

- Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Association for Computational Linguistics, Doha, Qatar, pages 103–111. <https://doi.org/10.3115/v1/W14-4012>
- Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, NY.
- Yongtae Kim, Dongyul Ra and Soojung Lim. 2020. Zero-anaphora resolution in Korean based on deep language representation model: BERT. *ETRI*

Journal, 43(2):299-312. <https://doi.org/10.4218/etrij.2019-0441>

- Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics, pages 28–39. <https://doi.org/10.48550/arXiv.1706.03872>
- Jeffrey M. Kuhn & Neil C. Thompson, 2019. How to Measure and Draw Causal Inferences with Patent Scope, *International Journal of the Economics of Business*, Taylor & Francis Journals, 26(1): 5-38. <https://dx.doi.org/10.2139/ssrn.2977273>
- Arum Park, Seunghye Lim, and Munpyo Hong. 2015. Zero Object Resolution in Korean. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 439–448. <https://aclanthology.org/Y15-1050/>
- Katsuo Tamaoka, Shanshan Yu, Jing Zhang, Yuko Otsuka, Hyunjung Lim, Masatoshi Koizumi, and Rinus G. Verdonchot. 2024. Syntactic structures in motion: Investigating word order variations in verb-final (Korean) and verb-initial (Tongan) languages. *Frontiers in Psychology*, 15: 1360191. <https://doi.org/10.3389/fpsyg.2024.1360191>
- Andy Way. 2018. Quality expectations of machine translation. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation Quality Assessment: From Principles to Practice* (pages 159–178). Springer. https://doi.org/10.1007/978-3-319-91241-7_8
- WIPO, *PCT Yearly Review 2025: Executive Summary* (World Intellectual Property Organization, Geneva, 2025), <https://doi.org/10.34667/tind.58378>

A Corpus-based Analysis of Semantic Similarity in Machine Translation and Human Interpreting

Kaifusai Julaiti

The Hong Kong Polytechnic University

julaiti-

katelyn.kaifusai@connect.polyu.hk

Andrew K. F. Cheung

The Hong Kong Polytechnic University

andrew.cheung@polyu.edu.hk

Abstract

The incorporation of machine translation (MT) in interpreting activities has drawn increasing research attention. However, MT has often been examined as an assistant to interpreters within the framework of computer- or AI-assisted interpreting. Its independent performance in rendering original meaning in interpreting context, particularly in comparison to interpreter performance, has been relatively unclear. Within the Chinese-English language pair, this study examines the performance of MT systems and simultaneous interpreters in conveying semantic meaning, measuring it through semantic similarity using LaBSE, COMET, and partial human assessment. The findings suggest that though MT systems may deliver sufficient semantic meaning, they have yet to reach human parity in conveying implicit semantic cohesion and nuanced language use, which thus requires further fine-tuning.

1 Introduction

Product-oriented research on interpreting has long focused on whether the target speech preserves the original meaning and whether it is delivered with undiminished linguistic quality and fluency. In simultaneous interpreting (SI), where the target speech is produced in real time as the source speech unfolds, achieving this is cognitively demanding for human interpreters (Gile, 2009). As a result, SI output often features omissions or simplified linguistic structures (Kajzer-Wietrzny, 2015; Liu et al., 2022; Xu and Liu, 2023a, 2023b; Pym, 2019), which can affect how meaning is conveyed.

In such a challenging task, interpreters can seek assistance from their partners or various toolkits

like glossaries. In recent years, they have begun to leverage AI technologies such as machine translation (MT) and automatic speech recognition (ASR). MT systems often generate target texts from ASR-transcribed source speeches. The renditions can serve as a reference for interpreters or be displayed to audiences via subtitles. Related research has focused primarily on the use of MT as an assistive tool, leaving its independent ability to convey meaning in spoken contexts relatively underexplored. Although MT outputs have often been compared to human translations, the findings may not be generalized to interpreting scenario due to modality differences. Against the backdrop of increasing population of MT in interpreting practices, it raises a question that whether MT can achieve human parity in interpreting meaning. This calls for examinations on MT and interpreter performance in conveying the same original meaning.

Meaning delivery is usually operationalized through semantic meaning and quantified by semantic similarity. Prior work in translation studies has investigated mono- and cross-lingual semantic similarity to measure meaning perseverance (Yang and Zhou, 2022; Yang et al., 2025). Yet, this issue has not been sufficiently explored in the interpreting practice, where real-time constraints and cognitive load may increase semantic loss. Nor has it been extensively investigated in MT scenario, where a similar concept, n-gram overlaps with human translations, has been the major focus.

This study therefore examines the performance of MT systems and simultaneous interpreters in delivering the original meaning, using semantic similarity as a measure. The corpus consists of 250 Chinese speeches delivered at United Nations Security Council (UNSC) meetings, along with their English simultaneous interpreting (SI) and

machine translation (MT) renditions produced by three neural MT systems. Semantic similarity was measured using two automatic toolkits and partially evaluated by human raters. This study aims to provide insights into MT performance in interpreting contexts and to help stakeholders assess the use of MT in their services.

2 Literature review

2.1 Meaning delivery in interpreting

Meaning delivery is a common topic in linguistic research (Vandepitte, 2007), while in interpreting studies, similar concepts have been investigated under the topic of quality assessment, operationalized through information accuracy and linguistic conversion (Han, 2022; Su, 2021). It should be noted that, however, accuracy usually measures the formulation of information at a more superficial level, while meaning conveyance gauges more at the semantic, content, or pragmatic levels. In this regard, Gieshoff and Albl-Mikasa (2022) proposed enriching accuracy assessment by incorporating cohesion and secondary meaning units, but their manual detection on small datasets may introduce subjectivity and limit the validity of their findings. In other words, a metric set that can also represent non-superficial meaning, along with large-scale data analysis, may evaluate interpreter or machine performance more effectively, thereby informing MT-integrated interpreting practices.

Meaning is encoded within language signs, commonly words and sentences. These linguistic units have been largely explored in product-oriented interpreting studies, where lexical and syntactic complexity are popular metrics for performance evaluation. However, linguistic structures may be superficial, unable to directly represent meaning delivery. Nevertheless, those findings showed that interpreters tend to produce simpler linguistic structures probably due to cognitive overload (Liu et al., 2022; Lv and Liang, 2019; Xu and Li, 2022; Xu and Liu, 2023a). It could raise questions of whether the meaning can survive the cognitive pressure and simplified linguistic structures and whether MT systems that are immune to cognitive pressure can outperform interpreters in this respect. Addressing these concerns may enhance our understanding of interpreters' competence as well as MT's strengths and limitations, thereby providing

insights into interpreting practices where the application of various MT systems is increasing.

2.2 Semantic meaning and semantic similarity

Semantic meaning is a major metric to operationalize meaning delivery. It is defined as "possible words referred to by the utterance" (Vandepitte, 2007), namely propositional meaning that could be denoted by sentences. Semantic meaning has been mainly examined in translation products, mostly under the theoretical framework of "semantic translation" proposed by Newmark (1981), which refers to translating semantic structure and meaning. Within this framework, studies explored semantic equivalence in translations, often employing qualitative approaches (Rietveld and Van Hormelen, 2019). With the development of computational tools, researchers started to examine semantic equivalents through more labor-free methods such as Word2Vec and other semantic similarity toolkits (Liu et al., 2025; Yang and Zhou, 2022; Yang et al., 2025). While only a limited number of works have focused on semantics in interpreting renditions.

From a quantitative perspective, semantic similarity is a common metric of semantic meaning. It has been applied in both monolingual and cross-lingual contexts and quantified through an embedding-based method using measurements such as BLEU (Bilingual Evaluation Understudy) and BERT (Bidirectional Encoder Representations from Transformers). Yang and Zhou (2022) examined monolingual semantic similarity of several English translations of the Chinese classic *The Analects*, measuring the similarity scale through metrics like BERT and Word2Vec. They found that these renditions achieved semantic accuracy of more than 92% and the variations may stem from translation annotations and styles. Plus, the reliability of these metrics was confirmed as they showed the same trend of similarity scale. Later, from a cross-lingual perspective, Yang et al. (2025) compared semantic similarity between different translations of *The Analects* and the source texts using large language models (LLMs). Within the Chinese-English language pair, these studies have contributed to quantifying semantic meaning in translation-mediated texts. While regarding machine-produced translations, scholars have

proposed several semantic-based assessment metrics, including SentSim (Song et al., 2021), SemMT (Cao et al., 2022), SaMT (Li et al., 2023), and SemSyn (Chauhan et al., 2023). These self-proposed metrics are developed based on underlying algorithms such as BLEU and BERT. Notably, BLEU analysis is limited to parallel comparison of n-grams between machine translations and human translations, unable to disclose deeper semantic similarity between the source and target texts. Concerning this limitation, recent MT-related research has adopted another automatic evaluation metric, namely COMET (Cross-lingual Optimized Metric for Evaluation of Translation). Unlike the somewhat surface-level BLEU, COMET predicts human rating of MT quality in terms of semantic fidelity and fluency (Rei et al., 2020), reporting stronger correlations with human judgements (Rei et al., 2022). Overall, the above metrics were applied to examine cross-lingual semantic similarity in translation outputs. Extending such evaluations to interpreting-mediated outputs may enrich this body of research, further supporting the effectiveness of these tools in reflecting semantic similarity in different genres.

2.3 Machine translation in interpreting

Most research on MT usage in interpreting has been conducted within the field of machine interpreting (MI), where the focus is primarily on interpreter performance with the assistance of MT rather than on MT's independent performance in interpreting contexts. MI to date has two approaches, the cascading approach and the end-to-end approach (Fantinuoli, 2023). The cascading approach, which consists of automatic speech recognition (ASR), machine translation (MT), and text-to-speech synthesis (TTS), is more widely used because it is easier to implement. In contrast, end-to-end approaches avoid the intermediate transcription step by mapping spoken input directly to spoken output. However, the end-to-end approach has yet to provide renditions as good as the traditional cascading approach (ibid., 2023). At present, machine interpreting (MI) largely relies on machine translation (MT), with ASR and TTS serving as supporting components in the processing pipeline. As a result, the ability of MT to convey meaning is central when comparing MI with human interpreting.

Related studies have investigated the role of MT in assisting consecutive interpreting (CI). Chen and Kruger (2023) examined the effectiveness of ASR and MT in CI, reporting reductions in pauses and cognitive load. In subsequent work (2024), they introduced a computer-assisted interpreting (CAI) workflow, termed Computer Assisted Consecutive Interpreting (CACI), and found that it outperformed traditional CI in terms of overall quality and accuracy. These studies have contributed to understanding how MT systems assist interpreters in Chinese–English contexts, although their primary focus remains on interpreter performance. The independent performance of MT systems has rarely been investigated in interpreting settings, despite the potential insights it may offer to real-world interpreting practices assisted or dominated by MT.

Existing research provides limited insight into the performance of different MT systems in interpreting contexts. Liu and Liang (2024) offered some perspectives on the correlation between MTs' and interpreters' outputs. By comparing neural MT (NMT) translations with SI renditions, they found that the lexical complexity of machine translations was lower than that of human interpretations. Nevertheless, linguistic performance represents only one facet of MT quality, and investigating it alone cannot directly reveal whether MT has conveyed the essential meaning. Some translation studies have examined MT quality in rendering verbal texts, though only based on one or two speeches. They compared NMT renditions with LLM-generated ones, concluding that LLMs did not outperform NMT models across written and verbal texts (Ding, 2024; Li, 2024). While assessing certain pairs of NMTs (Google Translate, Baidu Translate, and iFLYTEK), Lei and Li (2023) found that Google Translate outperformed the others with fewer errors in translating a keynote speech at the Boao Forum. Nevertheless, the investigated data size limits the reliability of their findings. Further research leveraging large datasets may produce more reliable findings to enrich the understanding of MT performance in interpreting settings, providing insights for MT-incorporated interpreting practices or training.

2.4 Research questions

The comparison of interpreters’ and MT systems’ performance in conveying meaning remains underexplored in interpreting studies. Such examinations may help stakeholders of interpreting services assess the application of MT, thereby facilitating more multilingual communications. Thus, this study examines the semantic similarity of human interpretations and MT renditions, aiming to provide insights into MT performance in the investigated language pair and genre. Two research questions were addressed:

RQ1: To what extent do human interpretations and MT renditions achieve semantic similarity with the source texts?

RQ2: Do their semantic similarity levels differ significantly? If so, what potential factors may contribute to that variation?

3 Methods

3.1 Corpus and MT systems

The study adopts a corpus-based approach. A corpus was compiled with Chinese source texts and English SI outputs from UNSC meetings as well as English MT outputs of the same source texts. The source speeches were delivered by Chinese delegates at daily meetings, dated from 2023 to 2024. The topics of these speeches vary, though not significantly due to the Security Council’s central focus on international security issues. Specifically, the topics encompass international security emergencies, long-lasting security issues, and UN peacekeeping missions. The language style in the UN is often consistent under the homogeneous topics and standardized terminologies (United Nations, n.d.). Thus, the influence of register of the source texts can be factored out to a certain extent. Interpreters in the Chinese interpreting booth render Chinese (their native language) into English (their second language). Transcriptions of the SI renditions and the source speeches comprised the first two sub-corpora.

The other sub-corpora are the translations of the same source texts produced by three NMT systems (Google Translate, DeepL, and iFLYTEK). Although LLM-based MTs seem to be popular nowadays, previous studies suggest that NMT systems still outperform LLM-based

ones (Boyd and Mitkov, 2025), especially in the Chinese-English language pair (Ding, 2024; Lin, 2023). Thus, LLM-based MTs are not considered in the present study. Google Translate and DeepL are among the most commonly examined NMT systems in T&I studies, particularly those involving the English-Chinese language pair (Gao et al., 2024; Li, 2024; Liu and Liang, 2024). iFLYTEK, a well-known Chinese homegrown company specializing in language solutions, offers MT and SI subtitling services powered by AI technology. As a Chinese-developed NMT system, iFLYTEK was selected for this study since previous research found that it produces higher quality outputs in Chinese-English political translations compared to models primarily trained on European languages (Liu and Zhu, 2023). Along with Google Translate and DeepL, they provided multiple references to MT performance.

Overall, the corpus includes five sub-corpora, with each containing 250 speeches of varied word counts (see Table 1). After data collection, the source texts (ST) and four types of target texts (TT) were first aligned manually at the document level. Then, 250 ST-TT pairs were segmented into sentences using NLTK (for English) and a regex-based splitter (for Chinese) and aligned with Bertalign (many-to-many) in Python.

Sub-corpora	Word Count	Text size
Chinese source texts (ST)	222,201	250
Human interpreted texts (HIT)	145,058	250
Google translated texts (GTT)	139,142	250
DeepL translated texts (DTT)	147,967	250
iFLYTEK translated texts (ITT)	151,521	250

Table 1: Corpus information

3.2 Semantic similarity evaluators

After alignment, LaBSE (Language-agnostic BERT Sentence Embedding) and COMET analyses were conducted to calculate semantic similarity. LaBSE focuses primarily on semantic meaning at the sentence level and supports cross-lingual comparison. Developed by Google Research, it uses a dual-encoder architecture with BERT-based modules, enabling it to produce

semantically meaningful and comparable sentence embeddings across languages (Feng et al., 2022; Yang and Feng, 2020). It has been reported to achieve higher accuracy than original BERT models, even for low-resource languages (Tang et al., 2024). COMET evaluates MT quality by predicting human assessments. It usually leverages three information sources, including source texts, MT outputs, and reference translations (Rei et al., 2020). Diverging from traditional n-gram-based metrics such as BLEU, it can provide insights into correlations between MT and human outputs, as well as semantic fidelity of all renditions to the sources.

Overall, LaBSE was employed to evaluate semantic embeddings transferred into the target texts, with COMET serving as a triangulation method to increase the validity of potential findings. Following the analyses, pairwise comparisons were conducted to compare the resulting scores, examining differences between the pairs.

The study also involved two human raters to assess semantic similarity on random samples. Their assessments were used to examine human preferences between human and MT products, instead of offering scores comparable to LaBSE and COMET results. Both raters hold a master’s degree in translation and interpreting and have been working as translators in Hong Kong for years, specializing in the Chinese-English language pair. The raters were first briefed and trained using sample data. They were then introduced to the concept of semantic similarity and asked to evaluate the four types of renditions (HIT, GTT, ITT, DTT) against the source texts, scoring each on a 1-5 Likert scale. The rendition types in all rating materials were anonymized to eliminate preference bias. After training, the rating details were discussed for further refining. Then based on the refurbished principles, each rater independently evaluated 15 samples among 250 (6%). Sentences with notable score differences were discussed to reach a consensus. Finally, their scores were collected and normalized to match the scales of LaBSE and COMET results for easier reference.

4 Results

4.1 LaBSE scores

The descriptive figures of LaBSE analysis showed that MT systems reached higher semantic similarity scores in their products compared to human interpretations (see Table 2 and Figure 1). Among the MT systems, Google Translate performed the best, followed by DeepL and iFLYTEK. This indicates that MT systems seem to deliver more semantic embeddings than interpreters in the investigated context.

The aggregated LaBSE scores per document were compared among groups using a Friedman test ($N = 250$), $p < .001$, indicating significant differences among groups. Then Bonferroni-corrected Wilcoxon signed-rank tests were performed to indicate the specific pairwise differences (see Table 3). For HIT ($M = 0.835$, $SD = 0.023$) and GTT ($M = 0.857$, $SD = 0.016$), the LaBSE scores differed significantly, $W = 456$, $p < .001$, with GTT outperforming HIT. The similarity scores for ITT ($M = 0.855$, $SD = 0.016$) were also significantly higher than those for HIT, with $W = 865$, $p < .001$. Last, the difference between HIT and DTT ($M = 0.856$, $SD = 0.018$) was statistically significant, $W = 483$, $p < .001$, with DTT reaching higher similarity scores.

	Mean	Median	Std. Deviation
HIT	0.835	0.8374	0.0225
GTT	0.857	0.8563	0.0164
ITT	0.855	0.8559	0.0161
DTT	0.856	0.856	0.0177

Table 2: Descriptive statistics for LaBSE scores across all renditions

Pair	Statistic	p value
HIT-GTT	$W = 456$	$< .001^*$
HIT-ITT	$W = 865$	$< .001^*$
HIT-DTT	$W = 483$	$< .001^*$

Table 3: Pairwise Wilcoxon Signed-rank tests for LaBSE scores

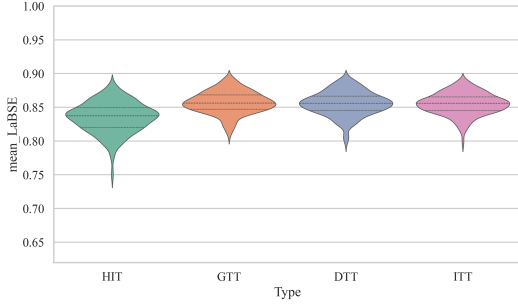


Figure 1: LaBSE scores for human interpreting and machine translations

4.2 COMET scores

Both reference-free (QE, quality estimation) and reference-based (REF) COMET were scored in Python with HIT serving as the reference source. For reference-free COMET, the descriptive statistics in Table 4 indicate that all MT outputs received numerically higher mean QE scores than HIT. A Friedman test indicated significant differences among the MT renditions, $\chi^2(3) = 461.82, p < .001$. Wilcoxon signed-rank tests with Bonferroni correction were conducted for six comparisons (see Table 5). HIT ($M = 0.827, SD = 0.0193$) significantly varied from all MT renditions with lower QE scores (see Figure 2). Among the MT sub-corpora, GTT ($M = 0.848, SD = 0.0091$) achieved prominently higher scores than ITT ($M = 0.845, SD = 0.0097$), which then differed significantly from DTT ($M = 0.847, SD = 0.0092$) with lower scores. While the variation between GTT and DTT was not statistically significant. These results aligned with those of LaBSE, supporting the reliability of the analysis.

For reference-based COMET, GTT ($M = 0.877, SD = 0.0169$) achieved the highest similarity to HIT, followed by ITT ($M = 0.874, SD = 0.0185$) and DTT ($M = 0.869, SD = 0.0169$). The Friedman test showed significant differences among MT renditions, $\chi^2(3) = 173.17, p < .001$. Table 6 summarizes the Bonferroni-corrected Wilcoxon signed-rank tests of three comparisons and Figure 3 visualizes the scoring variations.

	Mean	Median	Std Deviation
COMET-QE			
HIT	0.827	0.8315	0.0193
GTT	0.848	0.8498	0.0091
ITT	0.845	0.8455	0.0097
DTT	0.847	0.8497	0.0092
COMET-REF			
GTT	0.877	0.8794	0.0169
ITT	0.874	0.9713	0.0185
DTT	0.869	0.8757	0.0169

Table 4: Descriptive statistics for COMET scores of all renditions

Pair	Statistic	<i>p</i> value
HIT-GTT	$W = 250$	$< .001^*$
HIT-ITT	$W = 713$	$< .001^*$
HIT-DTT	$W = 350$	$< .001^*$
GTT-ITT	$W = 4734$	$< .001^*$
GTT-DTT	$W = 15118$	1.000
ITT-DTT	$W = 719$	$< .001^*$

Table 5: Pairwise Wilcoxon Signed-rank tests for reference-free COMET scores

Pair	Statistic	<i>p</i> value
GTT-ITT	$W = 1970$	$< .001^*$
GTT-DTT	$W = 7951$	$< .001^*$
ITT-DTT	$W = 5123$	$< .001^*$

Table 6: Pairwise Wilcoxon Signed-rank tests for reference-based COMET scores

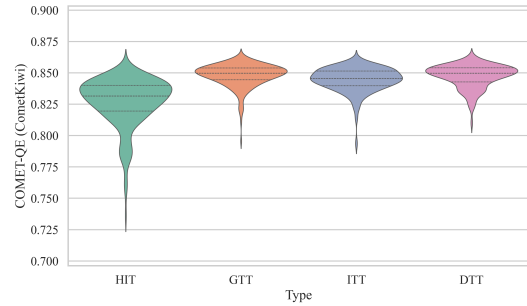


Figure 2: Reference-free COMET scores of different renditions

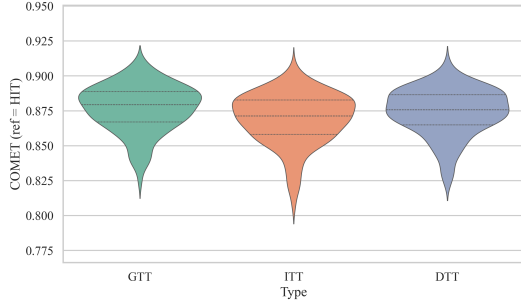


Figure 3: Reference-based COMET scores of different MT renditions

4.3 Human rating

The inter-rater reliability test reported a high reliability (over 94 percent for all) of the raters’ scoring (see Table 7). To align with the LaBSE and COMET scoring scales, the raters’ scores were normalized to a range between 0 and 1.

	HIT	GTT	ITT	DTT
Cohen’s <i>K</i>	0.980	0.951	0.948	0.970
Rater 1	0.967	0.939	0.922	0.913
Rater 2	0.969	0.937	0.920	0.912

Table 7: IRR and normalized mean scores of human ratings

The rating results were partly opposite to LaBSE and COMET scores. As Table 7 revealed, both raters scored HIT the best at delivering original semantic meaning, followed by GTT, ITT, and DTT. However, LaBSE and COMET-QE rated machine outputs higher than human output, with GTT receiving higher scores than DTT and ITT. Notably, human assessments on different machine outputs were consistent with the COMET-REF scores, as both ranked GTT over ITT and DTT.

5 Discussion

5.1 Interpreters vs. MT systems

LaBSE and COMET-QE results suggest that MT renditions delivered more semantic embeddings than human interpreters in the UNSC context. Possible reasons may lie in the underlying interpreting/translation mechanism and the varied capability in comprehending nuanced meaning units between interpreters and MT systems.

Due to processing pressure and limited cognitive resources, interpreters employ coping tactics during simultaneous interpreting (Gile, 2009). Previous studies based on large-scale corpora

have pointed out that interpreters may tend to simplify the structure or omit certain information due to factors such as interpreting directionality, time constraints, and cognitive demands (Kajzer-Wietrzny, 2015; Pym, 2009; Xu and Li, 2022; Xu and Liu, 2023a). In this context, interpreters may simplify their output or omit certain elements in order to maintain fluency and manage time constraints. This may, in turn, be reflected in lower levels of semantic similarity in human interpretations. By contrast, MT systems are not subject to the same cognitive constraints and tend to produce more complete translations, which may result in higher levels of semantic similarity. These differences are also relevant to how modality is realised, given its role in conveying meaning in discourse.

The raters preferred human to machines as the latter often failed to capture the correct semantic meaning in several aspects. Firstly, MT systems sometimes stuck to a surface-level translation, ignoring the exact meaning of the word in the designated context. For instance, they rendered 炮火之下 (literal translation: under the gun and fire) as ‘under fire’, while the interpreter’s rendition ‘under bombardment’ was assessed as a better option that clarifies the signified. Meanwhile, MT systems occasionally confused the semantic scope of words. For example, DeepL translated 战火 (literal translation: war fire) as ‘conflict’, which was considered to narrow the semantic range of the source term. The ‘war flame’ translated by interpreters, ‘flames’ by Google Translate, and ‘war’ by iFLYTEK were evaluated as more appropriate renditions that better preserved the original meaning. For some stance expressions, interpreters’ versions were often preferred over machines’ linguistically equivalent ones. In a conclusion statement expressing China’s willingness to cooperate, all MT systems translated the original verb 愿 into ‘is willing to,’ which is a correct equivalent; while the interpreter rendered it as ‘stands ready to.’ Both raters thought that the MT version does not fully deliver the firm stance and readiness of China in making efforts with all parties, as compared to the human interpretation. The above examples suggest that, although MT systems may preserve more of the source content by avoiding simplification or omission, they are less effective at capturing context-dependent meanings, such as stance and

intention. In other words, they often fall short of professional interpreters in conveying nuanced meaning in context.

At the sentence level, interpreters often outperformed MT systems when rendering compound sentences. Chinese compound sentences sometimes have no explicit inter-clausal conjunctions (Zhan and Bai, 2016), a structure that occurred frequently in the current UNSC corpus. In Example 1, the Chinese source speech contains a compound sentence with two clauses calling for action and a final clause stating its purpose, without the use of conjunctions:

Example 1:

ST: 中方重申，有关国家应立即解除对叙利亚的单边制裁，立即停止非法掠夺叙利亚资源，为叙利亚经济恢复发展创造有利条件。

*HIT: China reiterates that the unilateral sanctions imposed against Syria by certain countries should be lifted without delay, and the illegal plundering of Syrian resources should stop forthwith, **in order to** create favorable conditions for the Syrian economy to recover and develop.*

*GTT: China reiterates that relevant countries should immediately lift unilateral sanctions on Syria, immediately cease the illegal plundering of Syrian resources, **and** create favorable conditions for Syria's economic recovery and development.*

The human rendition makes the implicit purposive relation explicit, corresponding to “in order to,” whereas all MT systems link the final clause with “and,” as seen in Google’s version, thereby failing to convey the purposive relation..

This finding suggests that MT systems are not yet able to adequately capture deeper, non-explicit semantic relations in Chinese texts. By contrast, human interpretations more often make such relations explicit or convey them through alternative linguistic means. Accordingly, domain-specific training focused on common syntactic features or discourses may help improve MT performance in this regard. Several studies have reported that MT trained with deep-learning models and in-domain texts outperformed its originals in conveying meaning in certain language pairs (Moslem et al., 2022; Popel et al., 2020). This could point out a potential solution for the above issues of MT outputs.

Nevertheless, MTs’ overall delivery of semantic meaning did not seem to be significantly affected by their poor performance in rendering those nuanced and deeper-level meanings. Although the raters scored MT renditions lower than human interpretations, the normalized scores and automatic evaluation results showed that MT renditions still achieved high levels of semantic similarity. MT performance in translating nuanced language use may be improved by further training.

5.2 Varied performance among MT systems

Among the MT renditions, Google translations conveyed the most adequate semantic meaning as scored by LaBSE and COMET-QE. In the reference-based analysis, Google’s versions achieved the highest similarity to human interpretations. Human raters also gave Google translations the highest scores, only second to human interpretations. These results suggest that, compared to other systems, Google Translate tends to provide semantically sufficient translations that also share surface-level features with professionals’ renditions in the Chinese-English political interpreting context. This makes it a potential rival to interpreters in the investigated setting. On the other hand, DeepL and iFLYTEK shared no consistent patterns among LaBSE and COMET evaluations. DeepL outperformed iFLYTEK in semantic embedding comparisons but failed to do so in its proximity to human interpretations. This may suggest that DeepL tends to convey sufficient semantic meaning through somewhat restructured expressions, whereas iFLYTEK tends to produce outputs that resemble the expressions of professional interpreters but with relatively less semantic sufficiency.

Notably, human ratings were inconsistent with automatic evaluations in terms of interpreter performance, pointing to a fundamental comprehension problem. Although the raters only evaluated a small sample of data, providing limited insights into the preference between human and machine performance, it may suggest that humans and machines seem to comprehend semantic meaning in slightly different ways, especially in terms of nuanced meaning units. Nevertheless, human and automatic evaluations of MT outputs generally align, which may indicate that the automatic tools work effectively

when applied to MT outputs alone. Further research examining a comparable size of human and automatic evaluations may better untangle the effectiveness of those automatic tools.

6 Conclusion

This study examined MT and interpreter performance in conveying the original semantic meaning. The automatic analyses showed that machines achieved higher levels of semantic similarity than interpreters. While the raters seemed to prefer interpreters over machines in terms of conveying nuanced semantic units. The findings may suggest that interpreters may miss some semantic embeddings probably due to simplification or omission under processing challenges, whereas MT systems, free from such constraints, tend to preserve at least adequate surface-level semantic meaning. However, MT systems sometimes were incapable of correctly capturing nuanced meaning units or non-surface semantic cohesions, falling short of human parity in details. Future projects can train MT systems with such cases and fine-tune their performance to enhance the feasibility of MT in wider SI practices, giving one more option to organizers and audiences.

Nevertheless, this study has potential limitations. The MT reference is relatively limited. More balanced or diverse MT choices (such as Chinese-developed Baidu and LLM-based MT) can be included to provide a more balanced and comprehensive view of MT performance. Additionally, it must be acknowledged that, due to feasibility and cost-efficiency considerations, human rating was conducted only based on 6% of the corpus, which limits its role to a rough reference rather than a valid benchmark to compare with automatic ratings. Examining a similar data size of human and automatic evaluations through experiments may provide more effective insights into MT performance. Moreover, future studies can consider different meaning carriers, including pronouns, cohesive markers, and prosodic elements such as tone and pitch. Such examinations under different contexts and directionalities can enrich this body of research, providing suggestions for improving MT quality to facilitate more multilingual communications.

References

- Silvia Bernardini, Adriano Ferraresi, and Maja Miličević. 2016. From EPIC to EPTIC — Exploring simplification in interpreting and translation from an intermodal perspective. *Target. International Journal of Translation Studies*, 28(1): 61–86. <https://doi.org/10.1075/target.28.1.03ber>
- Boyd, W. J. and Mitkov, R. 2025. Machine Translation in the AI Era: Comparing previous methods of machine translation with large language models. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts*, pages 38-51.
- Jialun Cao, Meiziniu Li, Yeting Li, Ming Wen, Shing-chi Cheung, and Haiming Chen. 2022. SemMT: A semantic-based testing approach for machine translation systems. *ACM Transactions on Software Engineering and Methodology*, 31(2): 1–36. <https://doi.org/10.1145/3490488>
- Shweta Chauhan, Rahul Kumar, Shefali Saxena, Amandeep Kaur, and Philemon Daniel. 2023. SemSyn: Semantic-syntactic similarity based automatic machine translation evaluation metric. *IETE Journal of Research*, 70(4): 3823–3834. <https://doi.org/10.1080/03772063.2023.2195819>
- Sijia Chen and Jan-Louis Kruger. 2023. The effectiveness of computer-assisted interpreting: A preliminary study based on English-Chinese consecutive interpreting. *Translation and Interpreting Studies*, 18(3): 399–420. <https://doi.org/10.1075/tis.21036.che>
- Sijia Chen and Jan-Louis Kruger. 2024. A computer-assisted consecutive interpreting workflow: Training and evaluation. *The Interpreter and Translator Trainer*, 1–20. <https://doi.org/10.1080/1750399X.2024.2373553>
- Lijie Ding. 2024. A comparative study on the quality of English-Chinese translation of legal texts between ChatGPT and neural machine translation systems. *Theory and Practice in Language Studies*, 14(9): 2823–2833. <https://doi.org/10.17507/tpls.1409.18>
- Claudio Fantinuoli. 2023. Towards AI-enhanced computer-assisted interpreting. In G. Corpas Pastor and B. Defrancq (Eds.), *IVITRA Research in Linguistics and Literature (Volume 37)*, pages 46–71. John Benjamins Publishing Company. <https://doi.org/10.1075/ivitra.37.03fan>
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022, May.

- Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878-891.
- Ruiyao Gao, Yumeng Lin, Nan Zhao, and Zhenguang G. Cai. 2024. Machine translation of Chinese classical poetry: A comparison among ChatGPT, Google Translate, and DeepL Translator. *Humanities and Social Sciences Communications*, 11(1): 835. <https://doi.org/10.1057/s41599-024-03363-0>
- Anne C. Gieshoff and Michaela Albl-Mikasa. 2022. Interpreting accuracy revisited: A refined approach to interpreting performance analysis. *Perspectives*, page 1–19. <https://doi.org/10.1080/0907676X.2022.2088296>
- Daniel Gile. 2009. *Basic Concepts and Models for Interpreter and Translator Training* (Rev. ed., Vol. 8). John Benjamins Publishing Company.
- Chao Han. 2022. Interpreting testing and assessment: A state-of-the-art review. *Language Testing*, 39(1): 30–55. <https://doi.org/10.1177/02655322211036100>
- Marta Kajzer-Wietrzny. 2015. Simplification in interpreting and translation. *Across Languages and Cultures*, 16(2): 233–255. <https://doi.org/10.1556/084.2015.16.2.5>
- Suxia Lei and You Li. 2023. English machine translation system based on neural network algorithm. *Procedia Computer Science*, 228: 409–420. <https://doi.org/10.1016/j.procs.2023.11.047>
- Xinchen Li. 2024. Comparison of translation quality between large language models and neural machine translation systems: A case study of Chinese-English language pair. *International Journal of Education and Humanities*, 4(2): 121–128. <https://doi.org/10.58557/ijeh.v4i2.213>
- Jiaxin Li, Rize Jin, Joon-Yong Paik, and Tae-Sun Chung. 2023, November. Neural machine translation with an awareness of semantic similarity. In *Pacific Rim International Conference on Artificial Intelligence*, pages 223-235. Singapore: Springer Nature Singapore.
- Yiqing Lin. 2023, October. Transformer-based a automatic scoring model for translation jobs. In *Proceedings of the 2023 4th International Conference on Machine Learning and Computer Application*, page 413-417.
- Yufeng Liu, Shifa Chen, and Yi Yang. 2025. Semantic alignment: A measure to quantify the degree of semantic equivalence for English–Chinese translation equivalents based on distributional semantics. *Behavior Research Methods*, 57(1): 51.
- Yiyang Liu and Junying Liang. 2024. Multidimensional comparison of Chinese-English interpreting outputs from human and machine: Implications for interpreting education in the machine-translation age. *Linguistics and Education*, 80, 101273. <https://doi.org/10.1016/j.linged.2024.101273>
- Kanglong Liu, Zhongzhu Liu, and Lei Lei. 2022. Simplification in translated Chinese: An entropy-based approach. *Lingua*, 275, 103364. <https://doi.org/10.1016/j.lingua.2022.103364>
- Shanshan Liu and Wenxiao Zhu. 2023. An analysis of the evaluation of the translation quality of neural machine translation application systems. *Applied Artificial Intelligence*, 37(1). <https://doi.org/10.1080/08839514.2023.2214460>
- Qianxi Lv and Junying Liang. 2019. Is consecutive interpreting easier than simultaneous interpreting? – A corpus-based study of lexical simplification in interpretation. *Perspectives*, 27(1): 91–106. <https://doi.org/10.1080/0907676X.2018.1498531>
- Yasmin Moslem, Rejwanul Haque, John Kelleher, and Andy Way. 2022. Domain-specific text generation for machine translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Vol. 1)*, pages 14-30.
- Peter Newmark. 1981. Approaches to translation (Language Teaching methodology series). *Studies in Second Language Acquisition*, 7(1): 114.
- Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: A deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(1): 4381. <https://doi.org/10.1038/s41467-020-18073-9>
- Anthony Pym. 2009. On omission in simultaneous interpreting: Risk analysis of a hidden effort. In G. Hansen, A. Chesterman, and H. Gerzymisch-Arbogast (Eds.), *Benjamins Translation Library (Vol. 80)*, pages 83–105. John Benjamins Publishing Company. <https://doi.org/10.1075/btl.80.08pym>
- Ricardo Rei, José G. C de Souza, Duarte M. Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578-585.

- Ricardo Rei, 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. <https://doi.org/10.18653/v1/2020.emnlp-main.213>
- Laurent Rietveld and Frank V. Hormelen. 2019. Use of vocabulary translation strategies: A semantic translation analysis. *Applied Translation*, 13(2): 1–7.
- Yurun Song, Junchen Zhao, and Lucia Specia. 2021. SentSim: Crosslingual semantic evaluation of machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3143–3156. <https://doi.org/10.18653/v1/2021.naacl-main.252>
- Wei Su. 2021. Rubric-based self-assessment of Chinese-English interpreting. In *Testing and Assessment of Interpreting: Recent Developments in China*, pages 67–84. Singapore: Springer Singapore.
- Gongbo Tang, Oreen Yousuf, and Zeying Jin. 2024. Improving BERTScore for machine translation evaluation through contrastive learning. *IEEE Access*, 12: 77739–77749. <https://doi.org/10.1109/ACCESS.2024.3406993>
- United Nations. n.d.. *UNTERM: United Nations terminology database*. Retrieved January 6, 2026, from <https://unterm.un.org/unterm2/en/about>
- Sonia Vandepitte, S. 2007. Semantic and pragmatic meanings in translation. *Belgian Journal of Linguistics*, 21: 185–200. <https://doi.org/10.1075/bjl.21.13van>
- Cui Xu and Dechao Li. 2022. Exploring genre variation and simplification in interpreted language from comparable and intermodal perspectives. *Babel*, 68(5): 742–770.
- Han Xu and Kanglong Liu. 2023a. Syntactic simplification in interpreted English: Dependency distance and direction measures. *Lingua*, 294, 103607. <https://doi.org/10.1016/j.lingua.2023.103607>
- Han Xu and Kanglong Liu. 2023b. Investigating lexical simplification. In A. K. F. Cheung, K. Liu, and R. Moratto, *Corpora in Interpreting Studies (1st ed.)*, pages 197–215. Routledge. <https://doi.org/10.4324/9781003377931-13>
- Liwei Yang and Guijun Zhou. 2022. A semantic similarity analysis of multiple English translations of The Analects: Based on a natural language processing algorithm. *Frontiers in Psychology*, 13. <https://doi.org/10.3389/fpsyg.2022.992890>
- Liwei Yang, Guijun Zhou, and Lin Lin. 2025. From Confucius to computational linguistics: Quantifying cross-linguistic semantic similarity and semantic fidelity using large language models. *Digital Scholarship in the Humanities*, fqaf052. <https://doi.org/10.1093/llc/fqaf052>
- Yinfei Yang and Fangxiaoyu Feng. 2020, August 18. *Language-agnostic BERT sentence embedding*. Google Research. <https://research.google/blog/language-agnostic-bert-sentence-embedding/>
- Weidong Zhan and Xiaojing Bai. 2016. Sentence types. In C.-R. Huang and D. Shi (Eds.), *A Reference Grammar of Chinese 中文参考语法*, pages 401–450. Cambridge University Press. <https://doi.org/10.1017/CBO9781139028462.015>

Generated Voices:

Examining Patterns and Shifts in AI Translation of Online Feminist Discourse

Ailin Li

Newcastle University

A.Li31@newcastle.ac.uk

Abstract

A decade of research into machine bias has illustrated how gendered assumptions are embedded in MT technology, yet how these systems handle gendered ideological discourse remains largely unexplored. By observing the performance of ChatGPT, Claude, DeepSeek and Qwen on the translation of feminist discourse, this study investigates the way LLMs mediate and rewrite gendered ideological texts. It aims to uncover the dynamics between gender power, bias and censorship in LLM translation while reflecting on the ideological consequences of the invisible technical design choices behind AI tools.

Keywords: Large Language Models, Machine Bias, Machine Translation

1 Introduction

In 2014, when Professor Londa Schiebinger convened a Gendered Innovations workshop to confront gender bias in machine translation (MT), engineers from Google listened for twenty minutes and declared: ‘We can fix that!’ (Schiebinger, 2014). With a decade passing by, however, a definite solution seems to remain elusive according to Savoldi et al.’s (2025) review of over 100 studies.

How gendered power perpetuates MT that has been integrated in every aspect of life has led a promising track in translation studies since the emergence of the various algorithmic translation tools. So far, many have examined the gender bias in dedicated Neural Machine Translation (NMT)

systems such as Google Translate and DeepL (Piazzolla et al., 2023; Smacchia et al., 2024), yet much is to be researched regarding the gender issues in translation tasks completed by Large Language Models (LLM). Compared to NMT tools built on large corpora of paired texts in two languages, LLM models, with access to open internet data, are considered more impressionable to whatever biases and misinformation that data contains (Ullmann, 2022). As feminist translation scholars have long argued, translation can be an act of ideological manipulation (Von Flotow, 1997). If so, can such manipulation be found in LLM translation, which is constrained by culturally contextualised content restrictions and affected by biased data drawn from open online sources?

The current paper presents parts of an ongoing empirical study that seeks to grapple with the following research questions: 1) What strategies can be found in the translation of feminist discourse by LLMs? 2) Do models trained in different cultural and regulatory contexts exhibit different patterns in translating gender ideology discourse? The study experimented with two Western LLMs (ChatGPT and Claude) and two Chinese native LLMs (Deepseek and Qwen), which were requested under the same conditioned prompt to translate real-user posts in the Chinese online feminist debate. Given the limited scale, the paper primarily addresses RQ1 while offering preliminary findings for the second. The study sheds new light on the way contemporary translation technology interacts with gender ideology alongside the cultural contexts within which it develops.

2 Literature Review

Within scholarship regarding machine bias, earlier findings revealed that when translating sentences with gender pronouns and occupation descriptions, MT outputs tend to assign male defaults to STEM occupations while associating female pronouns with traditionally feminine roles such as nurse, nanny, and secretary (Prates et al., 2019; Fitria, 2021). The causes for such bias can be multifold. Particularly, Ullmann (2022, p.7) offered a thorough technical account of how bias enters through training data via crossing pathways including selection bias (the training data not representing the deployment context), label bias (human annotators applying stereotyped judgements), semantic bias (word co-occurrence patterns encoding stereotypes) and model over-amplification. Concluded by Savoldi et al. (2025), machine bias is a ‘sociotechnical’ issue rooted in normative discrimination and asymmetries in languages and the sociocultural background where they are used.

The studies in the field of MT, however, primarily approach machine bias as a technical or lexical phenomenon, leaving the ideological aspect of translation relatively underexplored. Nevertheless, the concept of translation as an ideological act has long been developed by many other branches of translation studies especially feminist translation theories. According to feminist translation scholars, translation can be deemed as an intervention, where the gendered element in the ST may be amplified, supplemented, or reframed with a political purpose rather than transferred neutrally (Simon, 1996; Flotow, 1997). Though this paper does not directly borrow the categorisation of feminist translation strategies, it builds on the insight of feminist translation theories that gendered content is where ideological shifts in translation constantly occur.

The possibility that translation can reshape the ideological force of gendered content proves especially relevant to the discussion of LLM in the current work. Notably, studies on LLM only account for not up to one fifth of the overall research body in machine bias by 2025 (Savoldi et al., 2025). While NMT and LLM are both regarded as powerful MT technologies, there is significant specificity in the development and function process of the two. In NMT, each source sentence is mapped to its target-language equivalent existing in the bilingual corpora (Smacchia et al., 2024). An

LLM, by contrast, is trained primarily on monolingual text on a massive scale. When it translates, it deploys general language generation capabilities on a cross-lingual task while consulting a much wider database and discourse context than NMT (ibid.). Such technical differences, in essence, lead to different features in coping with gender content. On the one hand, LLMs possibly possess a stronger ability in resolving ambiguity from contextual cues scattered across a much larger text span (ibid.). On the other hand, the open web data that LLMs obtain their knowledge from, dominantly released by privileged and heteronormative internet users, is sometimes problematic and reflected in the LLM output (Ullmann, 2022). To put briefly, though operating in different mechanisms from NMT, LLMs still absorb and reinforce the gender stereotypes in their TT generation. Therefore, standing on previous literature which points to the existence of gender bias, this project further explores through what translation strategies the bias is enacted by the models.

As mentioned above, former research in MT gender bias has paid most attention to the correct disambiguation and rendering of gendered pronouns in synthetic test sentences. This paradigm proves productive for showcasing the gender bias and its scale quantitatively, yet does not extend to the rich ideological context of MT bias. In Piazzolla et al.’s (2023) words, the analysis of ‘natural, gender translation phenomena’ of real texts is needed. Similarly, O’Connor and Liu (2023) advocate that engagement with the broader literature on gender and language would ‘go beyond pointing out errors or biases on a technological level and link the research to a wider socio-cultural context’ (p. 2054). Such a call is significant in the case of LLM-based translation since what distinguishes it from an NMT product is its generative ability that constantly engages with its latest ideological environment. To this end, this study goes beyond pronoun translation to investigate the renditions of culturally-specific expressions and rewriting on a sentence level conducted by LLM.

3 Method

This study takes a mixed-method approach to examine how AI translation systems handle gender-charged discourse. 200 samples were first selected from the Sina Weibo Sexism Review

(SWSR) dataset (Jiang et al., 2022) as the source texts (ST). SWSR is the first annotated Chinese-language corpus designed for computational sexism detection. Composed of posts spanning both sexist assertions and feminist counter-speech, it is openly shared on Zenodo under a Creative Commons license, which permits its public use for academic research purposes. The 200 comments were selected for this study based on the criteria below:

1) The comment has to directly engage with social gender issues or explicitly invoke the keywords 'feminist' or 'feminism' (女权).

2) The original SWSR has coded with each comment with G (Generic, which targets a broader group, such as women as a category) and I (Individual, which targets a specific identifiable person — a named celebrity or a particular user). Since uneven knowledge of Chinese public figures across models might complicate the scope of cross-model comparison, only comments under the G group were considered.

The STs were submitted to ChatGPT, Claude, DeepSeek, and Qwen for translation tasks. More

Model	Developer	Version	Date of Usage
ChatGPT	OpenAI	GPT-5.3	5-7 March 2026
Claude	Anthropic	Sonnet 4.6	1-3 March 2026
DeepSeek	DeepSeek	V3	8-10 March 2026
Qwen	Alibaba	3.6 Plus	5-7 April 2026

Table 1: Details of Examined LLMs

details about the models are listed below.

As two of the most widely deployed AI systems in the English-speaking world, ChatGPT and Claude are developed within Western regulatory frameworks and content policies where gender equity has been an explicit concern (O'Connor & Liu, 2023). DeepSeek and Qwen, in comparison, are Chinese-developed models shaped by a different cultural and institutional environment focusing on political sensitivity to socialist core values and social harmony (Pan & Xu, 2026). The TTs produced by each model were compared against the STs on a sentence basis. To control the data more evenly, the prompt is designed to be straightforward without additional instructions for the LLMs:

Prompt: Please translate the following sentence into English: ST.

Each post was translated once per model with the output retained as produced.

Subsequently, different strategies that LLMs employed when translating feminist discourse (from both supportive and critical stances) were assigned their respective codes. It should be stressed that the coding scheme was not set up prescriptively in advance but established gradually along the comparison process, with categories added only after a new strategy was observed in the sample. As such, the coding scheme in this project is grounded in the data rather than imported from a pre-existing framework. The coding process was conducted by a single researcher as the aim is to identify the strategies made by LLMs and to compare their relative tendencies across models. It should also be noted that sometimes more than one translation strategy was simultaneously adopted by an LLM. To address such ambiguity, the sample was assigned all applicable codes rather than being forced into a single category.

4 Results

The STs from SWSR constantly involve culture-specific expressions and sensitive content which pose an array of challenges to the AI models trained under different regulatory guidance. My comparison of the STs and the four versions of TTs identified the following strategies most frequently utilised by the examined models:

- **Amplification of Contested Tone (Coded as AM):** Wherever the ST reads as argumentative, all four AI models often strengthen the certainty or assertiveness of the ST. Specifically, a hedged modal 'may' would be replaced with a stronger 'will', while intensifiers such as 'definitely' or 'always' are added to reinforce the statement. Though not necessarily changing its essential content, such amplification makes the TT sound more ideologically locked and forceful than its ST. This finding echoes with previous observations that neural systems don't merely reproduce gendered patterns but exaggerate certain 'controversial associations' (Savoldi et al., 2025).
- **Explicitation of Gendered Culture-specific Items (EX):** This strategy is found most commonly in the translations of culturally

encoded metaphors and idioms, which are rendered explicit and transparent in the TTs. The effect of explicitation depends on context, but the translated phrases sometimes impose a narrower and more sexualised interpretation targeting women as Table 2 shows.

ST	潜规则
Back Translation	Unspoken rules (an euphemism in Chinese for bribe)
TT (Claude)	(Female celebrities) sleep their way to the top

Table 2: Example of Explicitation (SWSR No.154)

- **Implication and Omission of Sensitive Content (IO):** All four models have shown a tendency to generalise or remove sensitive content in the STs. These texts include explicit insults, discussions on women's sexual rights, and radical feminist criticism on social systems and gender hierarchy. Consequently, some gender issues mentioned in the posts might become vague with sacrifice in the original intention of the post to convey women's voices and experiences. What causes such softening can be AIs' tendency to avoid producing harmful language in their training (Ullmann, 2022). Omission, though, turns out to be more arbitrary with seemingly neutral content occasionally deleted. This random deletion may stem from the common MT's training data patterns in compressing information.
- **Addition of Implied Message (AD):** Apart from DeepSeek, all other examined models have performed the addition of the information that has no equivalent in some STs based on their interpretation of the text. The scale of added texts in the test sample ranges from a short phrase to a whole extra paragraph. The question of whether such intervention is in any sense intentional goes beyond the scope of this study and is left to future work. What can be observed, however, is that this strategy resembles what feminist translators term 'supplementing' (Von Flotow, 1997) and reinforces the source's feminist stance as exemplified in Table 3 (see opposite column).

ST	他们连父权还在戕害女性这个进行时都看不到。
Back Translation	They can't even see that patriarchy is still harming women right now.
TT (Claude)	They can't even acknowledge the ongoing reality that patriarchy is still harming women. <u>It's a present-tense injustice, yet they remain blind to it.</u>

Table 3: Example of Addition (SWSR No.528)

Apart from adjustments made on the basis of the STs, two additional strategies are also detected in the AI translation sample:

- **Explanatory Note (EN):** This refers to the attachment of extra translation notes made by AI models. These notes are added most frequently when there are gendered concepts in the Chinese language and culture that AI models consider necessary to explain further. Besides, part of the explanatory notes also serve as a disclaimer of AIs to caution their readers of derogatory and sexist content involved in the STs.
- **Rejection (RJ):** When the ST is so sensitive that it triggers content restriction, the AI model would decline the translation task. Though only found in a small proportion of the sample, rejection represents a significant aspect to investigate the cultural taboos and censorship threshold in the training data of an AI model.

Finally, as preliminary findings for RQ2, the coding results at this stage has already revealed several intriguing relative tendencies. For instance, compared to the other models, Qwen showed a particularly high frequency in adding implied messages to pro-feminist speeches in its translation. A clearer contrast also emerged for explanatory notes: whereas ChatGPT and DeepSeek rarely provided them, Claude and Qwen took this almost as a constant translation habit. The different focuses of these two LLMs' notes are planned to be investigated in the later stage of this research. Finally, both Chinese models, Deepseek and Qwen, answered with refusals to translate the text when encountering the STs that denounce the institution of marriage. One possible explanation might be that such content clashes with the state demographic agenda of promoting marriage and childbearing in contemporary China (Cai and Liu, 2024).

5 Conclusion

It can be concluded that LLM tools act as flexible rewriters in translating gender discourse with divergent effects: some reproduce bias and sexism or silence feminist calls, yet others strengthen calls for gender equality. The findings suggest that translation products of LLM translation, far from being neutral, can carry intervention in ideologically contested content similar to human translators. Building on existing work on machine bias, it extends the role of ideological mediator in translation studies from the human translator to this new kind of digital agent. The data presented in this paper remains qualitative, which is recognised as a limit. However, a more comprehensive analysis with quantitative data on the coded strategies is currently underway. As a next step, this study will quantify how frequently each strategy appears across different models to see whether a model's cultural developmental background leaves a traceable mark on how it handles gender discourse. As AI-powered translation becomes increasingly implemented on social media where gender debates arise frequently, the study is also expected to inspire reflection on how AI can be enhanced as a facilitator of daily equality and inclusivity.

Acknowledgments

The author wishes to gratefully acknowledge Jiang et al. (2022) for making the SWSR dataset publicly available.

References

- Cai, Y. and Liu, B. 2024. From birth control to pronatalism: Population policy and women's reproductive rights in China since the 1980s. *Jindal Global Law Review*, 15(2):267–289.
- Fitria, T.N. 2021. Gender Bias in Translation Using Google Translate: Problems and Solution. *Language Circle: Journal of Language and Literature*, 15(2):285–292.
- Jiang, A., Yang, X., Liu, T. and Zubiaga, A. 2022. SWSR: A Chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.
- O'Connor, S. and Liu, H. 2024. Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities. *AI & SOCIETY*, 39(4):2045–2057.
- Pan, J. & Xu, X. (2026). Political censorship in large language models originating from China. *PNAS Nexus*, 5(2), pgag013.
- Piazzolla, S.A., Savoldi, B. and Bentivogli, L. 2023. Good, but not always Fair: An Evaluation of Gender Bias for three commercial Machine Translation Systems. *HERMES – Journal of Language and Communication in Business*, 63:209–225.
- Prates, M.O.R, Avelar, P.H. and Lamb L.C. 2020. Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications*, 32(10):6363–6381.
- Savoldi, B., Bastings, J., Bentivogli, L. and Vanmassenhove, E. 2025. A decade of gender bias in machine translation. *Patterns*, 6(6):101257.
- Schiebinger, L. 2014. Scientific research must take gender into account. *Nature*, 507(7490):9.
- Simon, S. (1996) *Gender in Translation*. London: Routledge.
- Smacchia, M., Za, S. and Arenas, A. 2024. Does AI Reflect Human Behaviour? Exploring the Presence of Gender Bias in AI Translation Tools. In A.M. Braccini, et al., editors, *Digital (Eco) Systems and Societal Challenges, Lecture Notes in Information Systems and Organisation*, volume 72, pages 355–374. Springer Nature, Cham.
- Ullmann, S. 2022. Gender Bias in Machine Translation Systems. In Hanemaayer, A. editor, *Artificial Intelligence and Its Discontents: Critiques from the Social Sciences and Humanities, Social and Cultural Studies of Robots and AI*, pages 123–144. Palgrave Macmillan, Cham.
- Von Flotow, L. 1997. *Translation and Gender: Translating in the 'Era of Feminism'*. St. Jerome Publishing, Manchester, and University of Ottawa Press, Ottawa.

Challenges in Code Translation Using Large Language Models

Lukas Theuerkauf

Ksenia Neumann

Abdulrahman Nahhas

Damanpreet Singh Walia

Maria Chernigovskaya

Otto-von-Guericke University Magdeburg

{lukas.theuerkauf, ksenia.neumann, abdulrahman.nahhas,
damanpreet.walia, maria.chernigovskaya}@ovgu.de

Abstract

This paper examines major challenges in code translation (CT) using Large Language Models (LLMs). Furthermore, it proposes potential solution approaches to address these translation challenges. The applied method is a systematic literature review (SLR), which summarizes current knowledge from various experiments and investigates observable patterns. The results indicate that code conversion between programming languages (PLs) can be significantly simplified by optimizing each step in the LLM workflow. The deployment of standardization frameworks plays a central role in this process. This study contributes to the existing literature by providing a structured overview of the challenges in code translation, as well as offering several best-practice solutions to guide readers.

1 Introduction

The rising popularity of LLMs has motivated extensive exploration of code-related tasks, particularly code generation and translation. Recent advancements position LLMs as powerful tools for code-to-code translation, drawing parallels to their success in natural language translation (Ibrahimzada, 2024; Geng et al., 2024). CT between PLs is a long-standing task in software engineering, supporting modernization of legacy systems, cross-platform compatibility and performance optimization. Automated CT offers substantial time savings and reduces risks associated with manual translation, such as errors and inconsistencies (Macedo et al., 2024). However, code-specific LLMs face critical limitations, most notably restricted context window sizes that hinder processing of industry-scale projects (Ibrahimzada, 2024). This paper investigates primary challenges encountered by LLMs in CT including syntax differences, contextual understanding, and library nuances through a systematic literature analysis. We further outline approaches to mitigate these challenges.

2 Methodology

The SLR was conducted for a targeted identification of challenges in CT using LLMs. The SLR is appropriate for this study as it allows for an objective and thorough examination of the challenges presented in the literature. In addition, possible solutions are proposed to overcome these barriers. Given the recent nature of this topic, it also provides a foundation for future research.

The SLR was executed following the Design Science Research (DSR) methodology for information systems, which encompasses four major phases (Österle et al., 2011). First, the research question and the papers objectives were defined. The research plan in this case is driven by the literature scope (keywords, database selection, exclusion criteria). A comprehensive search strategy was developed using keywords such as “LLM”, “code,” “software”, “code translation”, “code conversion,” and “software migration”. To ensure coverage of diverse sources, multiple academic databases were explored on 24/11/2024, including Scopus (32 initial results), SpringerLink (18 initial results), ACM Digital Library (135 initial results), ScienceDirect (17 initial results) and IEEE Xplore (9 initial results).

The second DSR phase involves the artifact design (Österle et al., 2011). This systematic review was conducted according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to ensure transparency and reproducibility. The PRISMA flowchart depicts the synthesis approach of the selected literature. By removing 18 duplicates and 5 sources with closed access as irrelevant literature (182) the initial literature selection of 211 shrunk to 18 finally selected papers.

Guideline	Adherence
Design as an artifact	Artifact = Literature synthesis
Problem relevance	Overcoming of IT system legacy
Design evaluation	SLR according to PRISMA
Research contributions	Gap identification and suggestions
Research rigor	Reproducible and unbiased findings
Design as search process	SLR incl. proposed search strategy
Communication of research	Clear representation of key findings

Table 1: DSR Guidelines based on Hevner et al. (2004)

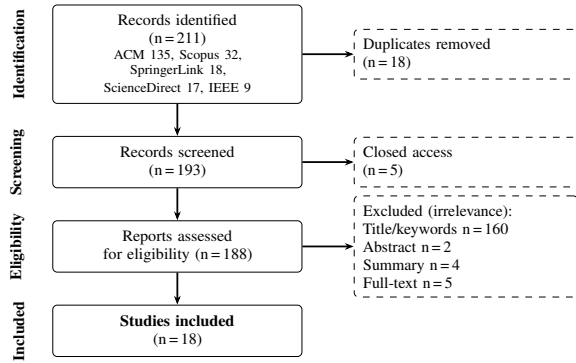


Figure 1: PRISMA flowchart of literature selection process.

Evaluation is the third phase of Design Science Research (DSR), focusing on validating artifacts produced against the specified objectives using the methods outlined in the research plan (Österle et al., 2011). In this study, the evaluation involves conducting a content analysis and comparing results with existing literature. The goal is to identify similarities, gaps, and limitations, ultimately aiming to answer the research question.

Lastly, the diffusion process involves the literature review and highlighting its key findings. Based on this summary, practical implications will be drawn and an outlook for future research will be provided (Österle et al., 2011).

To maintain a rigorous research approach, this study follows the Design Science Research (DSR) guidelines outlined by Hevner et al. (Hevner et al., 2004). The key aspects of the guideline application are summarized in Table 1.

Although DSR is conventionally associated with artefact construction, (Österle et al., 2011) explicitly position literature synthesis as a valid DSR artefact when it produces actionable design knowledge — a criterion this review satisfies through its best-practice guidelines for CT workflows.

To structure the findings, this paper employs three distinct categories throughout Section 4. **Challenges** denote recurring obstacles identified across multiple studies. **Findings** report empirical results or patterns observed in the reviewed literature. **Use Cases** illustrate concrete application scenarios described in individual papers.

3 Systematic Analysis

The systematic analysis represents the DSR artifact design, structuring and justifying the previously extracted literature. The research question subdivides into two perspectives: PL relationships (CT itself) and LLM technology. First, we examine interdependent difficulties in CT via a bug taxonomy. Next, we address the transition from low-resource (LRPL) to high-resource programming languages (HRPL), driven by industrial security, maintenance, and update requirements (Nitin, 2024). Table 2 provides quantitative justification through topic frequencies across the selected literature. Table 3 (Appendix A) classifies the selected literature across three dimensions: **what** is transformed (PL pairs and training data), **which** technology is used (model selection), and **how** it was achieved (success factors). Column abbreviations correspond to the topics in Table 2.

Topic	#Papers
HRPL	18
LRPL	9
Code-specific LLM	9
General purpose LLM	5
Prompting	7
Data training	18
Fine-tuning	2
Testing	2
Evaluation	5
Platform/Framework	6

Table 2: Topics covered in paper selection

4 Systematic Literature Comparison

Across the 18 reviewed studies, three structural patterns emerge. Data scarcity acts as a root cause that amplifies most other challenges: insufficient training corpora limit model specialisation, which in turn degrades both syntactic accuracy and semantic correctness. Second, the absence of standardised evaluation protocols makes cross-study comparisons unreliable, as different papers use CA, CodeBLEU, or DSR@K in isolation. Third, framework-level interventions (UniTrans, call graphs, transpiler integration) consistently outperform prompt

engineering alone, suggesting that architectural support is a more robust strategy.

In the following subsections, we examine the challenges in CT throughout the modeling process (data training, fine-tuning, prompting, testing, evaluation). The obstacles are distinguished within the scope of a bug taxonomy. Additionally, LRPL as well as platforms and frameworks will be emphasized.

4.1 Bug Taxonomy

Common industry use cases for CT include modernizing enterprise applications, migrating legacy software to cloud-native PLs, improving code synthesis, and ensuring production compatibility (Pan et al., 2024; Yan et al., 2023).

Challenge - Syntax & Semantics: CT methods range from compilers and rule-based approaches to neural networks, with the latter becoming state of the art (Yan et al., 2023). A key challenge is simultaneously understanding code syntax and semantics, as models must go beyond token prediction to handle target language violations (24%), API mismatches, and incorrect operator use (Pan et al., 2024).

Challenge - Dependency, Logic & Data Bugs: Dependency and logic bugs include missing imports, undefined elements, and incorrectly transformed logic. Data-related bugs account for over 33% of errors, driven by incorrect input parsing (18%) and data type mismatches (11%) caused by differing type systems across PLs (Pan et al., 2024).

Finding - Code Readability: Readability encompasses abstraction, modularization, complexity, naming conventions, and testability (Froimovich et al., 2025). LLMs still struggle with accuracy due to insufficient source comprehension and unclear input-output specifications (Yang et al., 2024).

4.2 Model Type

Code LLMs have gained significant attention in the industry due to their wide applications in the full lifecycle of software engineering (Di et al., 2024). As a preliminary point, we compare advantages and disadvantages of the code-specific and general-purpose models. Code-LLMs are tailored for programming tasks and detailed code insights, but they have limited general knowledge outside programming contexts. On the other hand, general-purpose LLMs have a broad knowledge base and are flexible in handling diverse topics. They lack

depth in specialized domains and are harder to scale due to extensive computational costs.

Finding - Specialization leads to improved performance in code-related tasks. Since specialized models are specifically trained for translation, general LLMs tend to produce more property violations than code-specific models (Eniser et al., 2024). Also, it was shown in practical multilingual scenarios, that multitask, fine-tuned version of LLMs outperform other models (Di et al., 2024).

Finding - Prompt simplicity achieves better results. This may be because simple prompts provide clear, concise, and unambiguous instructions. In contrast, more complex prompting strategies can introduce unnecessary noise and confuse the model. Additionally, the Divide-and-Conquer approach may hinder ChatGPT’s effectiveness due to a lack of global context. This limitation prevents the model from considering the overall structure and variable configurations of the code during translation (Aljaghtami et al., 2025).

Finding - FT is mandatory for general purpose LLMs. A Code-LLM processes source code as input and generates embeddings that encapsulate the semantic meaning of the code. However, without knowledge of the target language, CT cannot be performed effectively. Therefore, FT is essential for general-purpose LLMs to achieve satisfactory performance in CT tasks. Notably, fine-tuned models trained on appropriately tailored datasets often outperform even GPT-4 (Lei et al., 2023). GPT-4 and StarCoder outperformed explicitly trained CT models, primarily due to the exclusion of certain elements in the prompt, such as imports, class declarations, and method signatures. However, these models struggle with translating real-world projects because benchmark programs often lack complex dependencies and advanced language features. Furthermore, even when translation at the file or method level succeeds, it may still fail at the project level, where a broader contextual understanding is essential for accurate and coherent translation (Pan et al., 2024).

Notably, all 18 papers use training data as a primary variable, while fine-tuning and testing remain underrepresented (2 papers each), pointing to a methodological gap in the field.

4.3 Low Resource Programming Language

Challenge - Limited user bases and data availability of LRPLs. Early software systems were

developed using outdated PLs like Fortran and COBOL, which have a significantly smaller user base compared to modern languages such as Python and Java (Yan et al., 2023). Additionally, younger professionals typically lack experience with these older languages, making maintenance and modernization costly. The readability and compatibility challenges posed by the mix of PLs further complicate efforts to migrate legacy systems to new technology ecosystems or integrate software written in different languages (Yan et al., 2023).

Data scarcity presents a major challenge in CT, especially for less popular PLs, as mentioned above. Existing CT datasets are limited due to insufficient coverage of PLs. They primarily focus on a single pair of popular PLs, have a limited scale, and suffer from uneven data distribution (Yan et al., 2023).

Another potential solution to bridge the gap between an outdated LRPL and a current HRPL is the creation of a novel dataset specifically designed for training and evaluating LLMs. In the paper cited, the authors deal with CT between OpenMP Fortran and C++, significantly improving the model’s capacities (Lei et al., 2023).

In addition to the insufficient database, LLMs face structural differences between languages such as the shift from procedural to imperative or to object-oriented paradigm (e.g., COBOL to Java). Enhancing correctness and readability in CT (LRPL to HRPL) involves several key actions. Temperature sampling helps balance exploration and determinism. Refining translated Java code after translation, rather than during the process, leads to more effective improvements. Independent Java refinement, without referencing the original LRPL code, prevents unnecessary constraints. Since LRPL can distract from effective refinement, focusing solely on the Java code is beneficial. Finally, a sequential approach—first ensuring logical correctness, then enhancing readability—yields better results than addressing both simultaneously (Froimovich et al., 2025).

Finding - Semi-synthetic data improves CT.

First a code LLM synthesizes unit tests for commented code in a source HRPL, filtering out faulty tests and low-coverage code. Then the model translates the source code to a targeted LRPL, creating a candidate training dataset. At last a lightweight compiler compiles test cases from the source to the target language, filtering out obviously incorrect translations. This results in a validated training

corpus in the target language (Cassano et al., 2024).

Finding - Transpiler principles augment LLM translation. Another potential approach involves a system designed to enhance the CT capabilities by integrating the principles of transpilers into the LLM framework. A self-consistency filter is employed to generate a specified source code in static, dynamic, and natural language modalities, embedding them in LLM prompts. This ensures adherence to functional requirements, combining readability of LLMs with correctness of transpilers, thereby improving accuracy and utility in CT (Nitin, 2024).

Use Case - From ArkTS to Java. A key example of CT from LRPL to HRPL is converting ArkTS (Huawei’s OS language) to Java. Efficient CT relies on ArkTS architecture mimicking Java’s shared memory model as well as leveraged TypeScript (derivation of JavaScript) and prompt engineering. The Java2ArkTS framework breaks conversion into three thought chains: Converting Java to ArkTS-specific TypeScript, Mapping Java concurrency to the specific API, translates synchronization code via API. The most frequent compilation errors occur due to struggles with the API, misplacing functions or using invalid parameters. The converter also struggles with multi-library dependencies, limiting its real-world effectiveness (Liu et al., 2024).

Use Case - From C to Rust through type migration. Type migration involves managing types from the original PL and introducing new ones. This cannot be done by merely translating type names. Function bodies must also be restructured post-migration to ensure correct behavior. To address these challenges, LLMs are used for their deep understanding of program semantics and common patterns. LLMs assist in type migration by generating candidate signatures and mitigating type errors through iterative feedback from compiler error resolution. These techniques effectively increase the number of successfully migrated types while reducing errors (Hong and Ryu, 2025).

4.4 Training Data

Challenge - High data quality for effective CT.

The use of a high-quality pre-training dataset that is carefully filtered by program analyzers and optimized during the training process to significantly enhance model’s effectiveness (Di et al., 2024). A similarity test can be deployed to assess the qual-

ity of a dataset by measuring the syntactic and/or semantic similarity between pairs of code snippets. This analysis offers advantages across various applications, including but not limited to, plagiarism detection, code reuse and refactoring, bug detection and repair, licensing compliance, and malware detection (Lei et al., 2023). Nevertheless, the simplest way to boost the performance of a code-LLM in a PL is to train it further on natural data (code written by human programmers) (Cassano et al., 2024).

Challenge - Processing complex input source code and long output documents. Removing comments typically improves precision and F1 scores (evaluation metrics based on recall and precision) for embedding methods, as comments can cause LLMs to focus on comment similarity rather than code similarity. Therefore, we recommend removing comments by default to ensure that the analysis centered solely on code structure and semantics (Sekender et al., 2024).

If syntax is token-related to the PL, it attracts more attention. In a study analysing transformer attention distributions across code tasks, Mohammadkhani et al. (2023) find that models are less effective with samples that have long, complex source code. Poor model performance is often reflected in its attention distribution. Whenever the model fails to achieve a good output, it indicates that insufficient attention was paid to the corresponding token types for the respective downstream task. Even though pre-trained models have shown great results on software engineering tasks, none can be considered a fully solved problem. There are certain aspects of these models that still require further investigation (Mohammadkhani et al., 2023).

4.5 Fine-Tuning and Prompting

Challenge - Low comment accuracy. Fine-tuning (FT) should be considered to boost the CT performance of one PL into another. Regardless of dataset’s prior knowledge the accuracy could be improved either way (Lei et al., 2023). One of the key tasks for code LLMs is understanding code semantics, often guided by comments explaining the code. After code extraction, a supervised FT process is performed, selecting meaningful code-comment pairs based on rules such as discarding uninformative comments, applying quantitative limits, and excluding methods with less than 60% effective code (Di et al., 2024).

Challenge - Lack of parallel code data. FT models face significant challenges due to the limited availability of parallel code. Few-shot CT addresses this challenge by efficiently reusing prior knowledge from pre-training. It deploys task-adapted prompt tuning, freezing the pre-trained code and only updating parameters in a small prompt module. This enables parameter-efficient FT using only a small number of examples. The main goal of FSCTrans is to adapt a code PLM to the downstream task, allowing the CT model to reuse knowledge from the PLM while requiring only a small number of examples (Luo et al., 2026).

Finding - Fine-tuning improves tokenization performance and data augmentation. Tokenization performance is significantly enhanced due to larger vocabulary size and dedicated code-specific vocabulary of a model. The collection of question-answer pairs for code-related tasks leverages existing datasets to generate informative and contextually relevant outputs with the help of off-the-shelf models like ChatGPT. This approach is well-suited for divergent data augmentation (Di et al., 2024).

A prompt refers to a specific input text or a set of instructions provided to an LLM to trigger a particular response or action. Prompt crafting is an iterative approach composed of five phases: providing source code and original prompt, identifying incorrect translation, giving instructions for translation (bug mitigation and avoidance of natural language) as well as using model-specific keywords (Pan et al., 2024). Furthermore, the prompting technique chain-of-thought (guided argumentation) allows the model to simulate an orderly and structured way of thinking by sorting out the thinking process (Yan et al., 2023).

4.6 Frameworks

Challenge - Differences in syntax and dependency libraries between frameworks. Modernizing legacy software systems often faces barriers due to incompatibilities between different deep learning frameworks (DLFs). To address these challenges, a unified platform is needed, to evaluate models on diverse CT tasks. CodeTransOcean serves as a large-scale, comprehensive benchmark incorporating three multilingual datasets. Multilingual modeling significantly enhances translation quality for both HRPL and LRPL pairs while also improving training efficiency. A cross-DLF translation task is proposed to enable code migration between

DLF through automated CT. This approach offers several advantages, including improved code readability, support for secondary development and enhanced flexibility (Yan et al., 2023).

Finding - Call graphs and bridged debuggers enable the mitigation of compilation and runtime errors. Compilation and runtime errors can occur due to insufficient context awareness, especially without relying on test cases. The use of call graphs aims to optimize LLM input and integrate bridged debuggers and dynamic tests for error handling. This approach boosts the accuracy and practicality of CT. Additionally, the deployment of type inference and variable scope analysis provides further context for LLMs (Luo, 2025).

Finding - Harmonization eliminates the need for large-scale fine-tuning of LLMs. The unified CT framework UniTrans encompasses three key phases: test case generation, translation augmentation, translation repairs. UniTrans consistently improves performance across various LLMs, PLs, parameter sizes and datasets. It enables LLMs to generate valid inputs with over 90% success in just a few attempts. Despite its reliance on test cases, UniTrans remains efficient for real-world use, eliminating the need for large-scale fine-tuning. Instead, it leverages structured prompts and auto-generated test cases to significantly enhance LLM performance (Yang et al., 2024).

Use Case - LLM-based smart contract. In this particular use case, existing solidity smart contracts were translated into new smart contracts. This approach encodes concepts into an LLM and demonstrates how it can aid CT for a non-trained source language. However, the success of the CT-framework is driven by the use of an extremely LRPL. The iterative compiler error feedback loop mitigates bugs in the translated code (Karanjai et al., 2024).

4.7 Testing and Evaluation

Finding - Debugging Success Rate@K (DSR@K) reflects real-world software development scenarios. DSR@K measures whether translated code can be compiled and executed with the same behavior as the source code within K debugging rounds. DSR@K represents the percentage of samples that successfully execute and produce expected results out of all samples (Yan et al., 2023).

Finding - Property-guided search enhances

CT models by generating alternative translations ensuring specific functional properties.

This approach focuses on ensuring correct translations, regardless of the overall model quality. The procedure iteratively tests the model with mutated parameters to minimize property violations, continuing until all properties are satisfied within a given search budget. It effectively improves translation quality, particularly by addressing properties that are difficult to satisfy, and optimizes model output by minimizing violations (Eniser et al., 2024).

Challenge - Compilation errors due to inconsistent output format. Inconsistent output format refers to a problem where the translated code is presented in various formats, sometimes interspersed with natural language. There are two ways to approach this problem: text-oriented metrics (e.g., CodeBLEU, BLEU (Ren et al., 2021; Papineni et al., 2002)) and execution-based metrics (e.g., computational accuracy (CA)). It is important to distinguish between errors arising from the output format itself and those stemming from the quality of the translated code (Macedo et al., 2024).

CodeBLEU is well-suited for CT between Fortran and C++ codes (Ren et al., 2021). It enhances the traditional BLEU (Bilingual Evaluation Understudy) score, commonly used in natural language machine translation (Papineni et al., 2002), by incorporating code-specific features that capture both syntactic and semantic elements (Lei et al., 2023).

Challenge - Varying output formats and distributions. Variability in output complicates model comparison, output distribution analysis and metric reliability. Controlling output format through prompt engineering and lightweight regex-based post-processing improves code extraction. High match success, extraction success and CA validate this approach.

Output format control is crucial for CA. Direct LLM compilation yields the lowest CA, while Controlled Prompt with Regex achieves the highest across all models. This highlights the need for careful source code extraction to ensure fair comparisons. Differences in prompt sensitivity can significantly impact CA and Compilation Rate, making proper output handling essential to avoid misinterpretations (Macedo et al., 2024).

5 Conclusion

LLMs face numerous challenges in CT, requiring comprehension of both syntax and semantics be-

yond pure token prediction. LRPLs further complicate this due to limited data availability, while scarce parallel code data hinders model training generally.

Processing complex inputs, generating coherent outputs, and handling misleading code comments remain persistent issues. Syntax differences, dependency libraries, and framework inconsistencies hinder reusability, causing compilation errors. Simpler prompts tend to yield more consistent results by minimizing ambiguity.

A well-structured codebase enhances readability for both humans and LLMs. Incorporating transpiler principles standardizes conversion, while FT improves tokenization and data augmentation, particularly for general-purpose LLMs. CT frameworks address inconsistencies, and tools like call graphs and bridged debuggers provide execution insights to reduce runtime errors.

Ultimately, LLMs specializing in code-related tasks outperform generalized models, making FT a mandatory step for CT alignment. Future progress will depend on advances in model architecture, training data refinement and structured methodologies to achieve reliable CT across diverse PLs.

References

- Aamer Aljaghtami, Mohammed Banabila, Musab Alshehri, Mohammed Kabini, and Mohammad D. Alahmadi. 2025. Evaluating large language models for code translation: Effects of prompt language and prompt design. *ArXiv:2509.12973*.
- Federico Cassano, John Gouwar, Francesca Lucchetti, Claire Schlesinger, Anders Freeman, Carolyn Jane Anderson, Molly Q Feldman, Michael Greenberg, Abhinav Jangda, and Arjun Guha. 2024. [Knowledge transfer from high-resource to low-resource programming languages for code LLMs](#). *Proceedings of the ACM on Programming Languages*, 8:677–708.
- Peng Di, Jianguo Li, Hang Yu, Wei Jiang, Wenting Cai, Yang Cao, Chaoyu Chen, Dajun Chen, Hongwei Chen, Liang Chen, Gang Fan, Jie Gong, Zi Gong, Wen Hu, Tingting Guo, Zhichao Lei, Ting Li, Zheng Li, Ming Liang, Cong Liao, Bingchang Liu, Jiachen Liu, Zhiwei Liu, Shaojun Lu, Min Shen, Guangpei Wang, Huan Wang, Zhi Wang, Zhaogui Xu, Jiawei Yang, Qing Ye, Gehao Zhang, Yu Zhang, Zelin Zhao, Xunjin Zheng, Hailian Zhou, Lifu Zhu, and Xianying Zhu. 2024. [CodeFuse-13B: A pretrained multi-lingual code large language model](#). In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice*, pages 418–429, Lisbon, Portugal. ACM.
- Hasan Ferit Eniser, Valentin Wüstholtz, and Maria Christakis. 2024. [Automatically testing functional properties of code translation models](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21055–21062.
- Shmulik Froimovich, Raviv Gal, Wesam Ibraheem, and Avi Ziv. 2025. Quality evaluation of COBOL to Java code transformation. *ArXiv:2507.23356*.
- Mingyang Geng, Shangwen Wang, Dezun Dong, Hao-tian Wang, Ge Li, Zhi Jin, Xiaoguang Mao, and Xiangke Liao. 2024. [Large language models are few-shot summarizers: Multi-intent comment generation via in-context learning](#). In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13, Lisbon, Portugal. ACM.
- Alan R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. 2004. Design science in information systems research. *MIS Quarterly*, 28(1):75–105.
- Jaemin Hong and Sukyoung Ryu. 2025. [Type-migrating C-to-Rust translation using a large language model](#). *Empirical Software Engineering*, 30(1):3.
- Ali Reza Ibrahimzada. 2024. [Program decomposition and translation with static analysis](#). In *Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings*, pages 453–455, Lisbon, Portugal. ACM.
- Rabimba Karanjai, Lei Xudagger, and Weidong Shi. 2024. [SolMover: Feasibility of using LLMs for translating smart contracts](#). In *2024 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, pages 1–3, Dublin, Ireland. IEEE.
- Bin Lei, Caiwen Ding, Le Chen, Pei-Hung Lin, and Chunhua Liao. 2023. [Creating a dataset for high-performance computing code translation using LLMs: A bridge between OpenMP Fortran and C++](#). In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–7, Boston, MA, USA. IEEE.
- Runlin Liu, Yuhang Lin, Yunge Hu, Zhe Zhang, and Xiang Gao. 2024. [LLM-based Java concurrent program to ArkTS converter](#). In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, pages 2403–2406, Sacramento, CA, USA. ACM.
- Wenqiang Luo, Jacky Wai Keung, Boyang Yang, Jacques Klein, Tegawende F Bissyande, Haoye Tian, and Bach Le. 2026. Unlocking LLM repair capabilities through cross-language translation and multi-agent refinement. *ArXiv:2503.22512*.
- Yang Luo. 2025. Can we translate code better with LLMs and call graph analysis? Preprint.
- Marcos Macedo, Yuan Tian, Filipe Cogo, and Bram Adams. 2024. [Exploring the impact of the output format on the evaluation of large language models](#).

- for code translation. In *Proceedings of the 2024 IEEE/ACM First International Conference on AI Foundation Models and Software Engineering*, pages 57–68, Lisbon, Portugal. ACM.
- Ahmad Haji Mohammadkhani, Chakkrit Tantithamthavorn, and Hadi Hemmatif. 2023. [Explaining transformer-based code models: What do they learn? when they do not work?](#) In *2023 IEEE 23rd International Working Conference on Source Code Analysis and Manipulation (SCAM)*, pages 96–106, Bogotá, Colombia. IEEE.
- Vikram Nitin. 2024. [Using AI to automate the modernization of legacy software applications.](#) In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, pages 2514–2517, Sacramento, CA, USA. ACM.
- Hubert Österle, Jörg Becker, Ulrich Frank, Thomas Hess, Dimitris Karagiannis, Helmut Krcmar, Peter Loos, Peter Mertens, Andreas Oberweis, and Elmar J Sinz. 2011. [Memorandum on design-oriented information systems research.](#) *European Journal of Information Systems*, 20(1):7–10.
- Rangeet Pan, Ali Reza Ibrahimzada, Rahul Krishna, Divya Sankar, Lambert Pouguem Wassi, Michele Merler, Boris Sobolev, Raju Pavuluri, Saurabh Sinha, and Reyhaneh Jabbarvand. 2024. [Lost in translation: A study of bugs introduced by large language models while translating code.](#) In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13.
- Kishore Papineni, 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, Philadelphia, PA, USA. Association for Computational Linguistics.
- Shuo Ren, Daya Guo, Shuai Lu, Long Zhou, Shujie Liu, Duyu Tang, Neel Sundaresan, Ming Zhou, Ambrosio Blanco, and Shuai Ma. 2021. CodeBLEU: A method for automatic evaluation of code synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*. ArXiv:2009.10297.
- Naveed Sekender, Pei-Hung Lin, and Chunhua Liao. 2024. [Finding equivalent OpenMP Fortran and C/C++ code snippets using large language models.](#) In Alexis Espinosa, Michael Klemm, Bronis R. De Supinski, Maciej Cytowski, and Jannis Klinkenberg, editors, *Advancing OpenMP for Future Accelerators*, volume 15195, pages 143–160. Springer Nature Switzerland, Cham.
- Bo Wang, Ruishi Li, Mingkai Li, and Prateek Saxena. 2023. [TransMap: Pinpointing mistakes in neural code translation.](#) In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 999–1011, San Francisco, CA, USA. ACM.
- Weixiang Yan, Yuchen Tian, Yunzhe Li, Qian Chen, and Wen Wang. 2023. [CodeTransOcean: A comprehensive multilingual benchmark for code translation.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5067–5089, Singapore. Association for Computational Linguistics.
- Zhen Yang, Fang Liu, Zhongxing Yu, Jacky Wai Keung, Jia Li, Shuo Liu, Yifan Hong, Xiaoxue Ma, Zhi Jin, and Ge Li. 2024. [Exploring and unleashing the power of large language models in automated code translation.](#) *Proceedings of the ACM on Software Engineering*, 1:1585–1608.

A Systematic Literature Matrix

Table 3 classifies the 18 selected papers across three dimensions: what is transformed (PL pairs and training data), which technology is used (model selection), and how it was achieved (success factors). Column abbreviations correspond to Table 2 in the main paper: Data = training data focus; FT = fine-tuning; Pr = prompting; FMW = framework/platform; Test = testing; Eval = evaluation.

B Limitations and Future Work

Future research should prioritize enhancing models’ code comprehension rather than focusing solely on next-token prediction. A lack of deep code understanding remains a key factor contributing to reduced accuracy. Given the increasing industrial demand, particular attention should be directed toward LRPL translation to address legacy software challenges in enterprises. As demonstrated by the SLR, the integration of frameworks, fine-tuning techniques, and specialized datasets play a crucial role in managing the complexities of CT. One notable trade-off in code-specific LLMs is their performance gains through specialization, which often come at the cost of reduced generalization. Further exploration is needed to optimize this balance. Additional limitations include suboptimal attention distribution, inconsistencies in naming conventions and variability in output formatting. Addressing these issues is essential for improving reliability and consistency in code generation.

Techniques such as chain-of-thought prompting have been shown to enhance CT performance with relatively low effort. A potential research avenue is the compilation of best-practice guidelines to standardize effective prompting strategies. Furthermore, frameworks that support the CT process can significantly improve performance. Future studies could focus on developing a comprehensive “toolbox” through meta-analyses, which may also help

Paper	PL	Model type	Data	FT	Pr	FMW	Test	Eval
Eniser et al.	LR	Code-specific	x				x	
Di et al.	LR	Code-specific	x	x				
Yan et al.	LR	General purpose	x			x		x
Lei et al.	LR,HR	General purpose	x			x		x
Mohammadkhani et al.	HR	Code-specific	x					x
Macedo et al.	HR	Code-specific	x		x			
Luo et al.	LR,HR	Code-specific	x	x	x			
Pan et al.	HR	Code-specific	x					
Karanjai et al.	HR	General purpose	x			x		
Froimovich et al.	LR,HR	Code-specific	x			x		x
Hong and Ryu	LR,HR	General purpose	x	x				
Luo	HR	General purpose	x	x				
Yang et al.	HR	General purpose	x		x	x	x	
Cassano et al.	LR,HR	Code-specific	x		x			
Liu et al.	LR,HR	Code-specific	x			x		
Nitin	LR,HR	Code-specific	x		x	x		
Sekender et al.	LR,HR	General purpose	x		x			x
Aljagthami et al.	HR	General purpose	x		x			x

Table 3: Systematic literature matrix.

identify gaps and unresolved challenges in current methodologies. Compared to traditional CT (non-LLM-based), the primary limitations of LLM stem from constraints in context length, lack of determinism, and reasoning limitations (Pan et al., 2024). While LLMs can generalize from a limited number of examples, their effectiveness is restricted by the size of the context window, limiting the complexity of tasks they can process in a single prompt. Additionally, their reasoning capability is constrained by token limits, leading to weaker performance in certain few-shot reasoning tasks. However, techniques such as chain-of-thought prompting can improve results by facilitating more structured and coherent outputs (Wang et al., 2023).

While frameworks provide essential support for LRPL code translation, they may not be effective for languages that have not been included in a model’s training data. However, strategic prompting techniques may enable models to generate useful data for low-resource languages, although implementing this efficiently remains a challenge (Cassano et al., 2024).

Despite these advances, real-world software projects often involve multiple interdependent files, each containing hundreds of lines of code, making the current function-by-function translation approach insufficient. A modular strategy is currently being developed to address this challenge (Nitin,

2024).

Furthermore, FT frameworks enhance the performance of open-source models, facilitating advancements in AI-native software development. Open-source contributions are expected to play a pivotal role in improving the entire lifecycle of AI-driven programming tools (Di et al., 2024).

Exploring Sentiment Transfer in Cantonese–English Court Interpreting: A Computational Analysis Using Multilingual BERT

Sihui He

The Hong Kong Polytechnic University
sihui.he@connect.polyu.hk

Jiawei Shi

The Hong Kong Polytechnic University
jiawei.shi@connect.polyu.hk

Abstract

This study uses multilingual BERT to examine sentiment transfer in the Hong Kong Court Interpreting Corpus (HKCI). The corpus comprises 101 audio recordings from trials at the High Court of Hong Kong, organized into four sub-corpora: English source (ES), interpreted Cantonese (IC), Cantonese source (CS), and interpreted English (IE). BERT-derived sentiment scores were predominantly negative across all four sub-corpora, with significant pairwise differences. In the Cantonese-to-English (C2E) direction, model-derived scores were less negative in the interpreted output than in the source. In the English-to-Cantonese (E2C) direction, the opposite tendency was observed. However, BERT showed low agreement with the human reference labels ($\kappa = 0.23$), and the error analysis showed that most source-target polarity mismatches involved classifier errors rather than genuine sentiment shifts. The findings should therefore be read as exploratory and model-derived corpus tendencies. The study highlights both the potential and the limits of off-the-shelf sentiment tools for Cantonese-English courtroom interpreting.

1 Introduction

Professional standards require court interpreters to preserve the register, style, and tone of the source speech (EULITA, 2013). In practice, small changes in pragmatic meaning may make a speaker sound more certain or more emotionally marked than in the original. In legal settings, such changes may affect how witness testimony is perceived (Berk-Seligson, 2017).

This issue is especially relevant to Cantonese–English court interpreting. Cantonese utterance particles are used to express attitudes, emotions, and illocutionary force (Luke, 1990). These particles often have no direct equivalents in

English and are difficult to interpret (Leung and Gibbons, 2009). Recent work has shown that Cantonese particles may modulate pragmatic force during witness examinations (Wong, 2025). Existing studies have examined non-renditions (Cheung, 2017; Wong, 2023) and clarification acts (Cheung-Pease and Pease, 2023), but few studies have quantitatively examined how sentiment is transferred across Cantonese and English in authentic court interpreting.

Computational sentiment analysis has recently entered interpreting studies. Huang et al. (2025) found a moderate correlation between sentiment gap scores and human assessments of interpreting accuracy. Xu et al. (2025) showed that interpreting direction may affect sentiment in United Nations simultaneous interpreting. These studies, however, do not address authentic Cantonese–English courtroom data. The suitability of multilingual BERT for Cantonese legal discourse also remains uncertain. This study therefore applies multilingual BERT to the Hong Kong Court Interpreting Corpus (HKCI). Because the model has not been validated for Cantonese legal discourse, the analysis is exploratory. Sentiment scores are treated as model-derived indicators of corpus-level patterns rather than precise measures of how sentiment was conveyed in each interpreted sentence.

The study addresses two research questions:

RQ1: How do sentiment scores vary across source and interpreted court speech?

RQ2: What role does interpreting direction play in these differences?

2 Related work

Research on Hong Kong court interpreting has shown that interpreted output is not always pragmatically neutral. Cheung (2017) found more non-renditions in Cantonese than in English. Wong (2023) related this pattern to interpreters' positioning toward courtroom participants with different institutional roles. Cheung-Pease and

Pease (2023) also showed that clarification patterns are shaped by courtroom power relations. Work on Cantonese utterance particles further suggests that interpreters may use particles to maintain modality or adjust pragmatic force (Leung and Gibbons, 2009; Wong, 2025). Taken together, these studies indicate that both language structure and institutional context may shape interpreted courtroom discourse.

Sentiment analysis offers one way to examine such patterns at scale. Huang et al. (2025) used lexicon-based tools to study sentiment-bearing words in interpreting. Xu et al. (2025) used a BERT-based classifier to analyse bidirectional simultaneous interpreting at the United Nations. They found that interpreting direction may affect target-speech sentiment. In legal interpreting, Yi (2026) applied computational models to study Mandarin–English courtroom data and suggested that omissions may affect the sentiment of lawyer questioning. Yet authentic Cantonese–English courtroom discourse remains underexamined.

3 Methodology

3.1 Corpus

The study uses data from the Hong Kong Court Interpreting Corpus (HKCI), compiled from 101 audio recordings of five sexual assault trials heard in the High Court of Hong Kong. The data cover approximately 100 hours of courtroom interaction (Leung and Gibbons, 2009).

Speakers are coded by role and language following the HKCI coding scheme. BPE refers to prosecution barrister’s English utterance, BDE to defence barrister’s English utterance, JE to judge’s English utterance, WC to witness’s Cantonese utterance, WE to witness’s English utterance, DC to defendant’s Cantonese utterance, ICT to interpreter’s Cantonese translation, and IET to interpreter’s English translation. The corpus is organized into four sub-corpora (see Table 1): English source speech (ES), interpreted Cantonese (IC), Cantonese source speech (CS), and interpreted English (IE).

Corpus	Tokens	Source
ES	85,342	English speech
IC	93,455	Interpreted Cantonese speech
CS	91,432	Cantonese speech
IE	85,327	Interpreted English speech

Table 1: Overview of ES, IC, CS, and IE sub-corpora.

3.2 Analytical procedures

Sentiment analysis was conducted using nlptown/bert-base-multilingual-uncased-sentiment (NLPTown, 2020). The model was selected because its multilingual BERT backbone can process both English and Chinese text, and multilingual BERT has been used in recent interpreting research (Xu et al., 2025). This choice supports comparison with earlier work. However, the sentiment classifier was trained on product reviews rather than legal discourse, so its use for Cantonese courtroom discourse is exploratory.

For validation, 156 sentences were sampled. The sample was stratified by sub-corpus and BERT-predicted polarity, resulting in 39 sentences from each sub-corpus and 52 sentences from each predicted polarity class. The validation set was annotated by two Cantonese-speaking raters. The raters first annotated independently and then discussed disagreements. After discussion, they agreed on 153 of the 156 sentences ($\kappa = 0.97$). BERT–human agreement was calculated against Rater 2’s post-discussion labels; using Rater 1’s labels produced a comparable result ($\kappa = 0.22$). The same validation set was also tested with an XLM-R-based multilingual sentiment classifier (Conneau et al., 2020). This alternative model did not improve agreement with the human reference labels (accuracy = 48.1%, $\kappa = 0.13$). The main analysis therefore used this multilingual BERT-based model for comparability with prior interpreting research.

The model outputs probabilities over five star-rating classes. These were grouped into three polarity categories: $P(\text{neg}) = P(1 \text{ star}) + P(2 \text{ stars})$, $P(\text{neu}) = P(3 \text{ stars})$, and $P(\text{pos}) = P(4 \text{ stars}) + P(5 \text{ stars})$. A continuous model-derived sentiment score was then defined as $s = P(\text{pos}) - P(\text{neg})$.

Since $P(\text{pos})$ and $P(\text{neg})$ are probabilities between 0 and 1, s is bounded between -1 and $+1$. Scores were labelled positive if $s > +0.1$, negative if $s < -0.1$, and neutral if $-0.1 \leq s \leq +0.1$. The ± 0.1 band was used to avoid assigning polarity when positive and negative probabilities differed only slightly, and should be understood as an operational choice rather than a validated threshold.

The analysis was conducted at sentence level (Lei and Liu, 2021). For RQ1, sentiment scores were first screened using a corpus-wide IQR filter ($1.5 \times \text{IQR}$). Pairwise Welch's t-tests with Bonferroni correction were used to compare mean sentiment scores across sub-corpora. Cohen's d and 95% bootstrap confidence intervals are reported in Table 3. For RQ2, sentiment gap scores (target – source) were compared across directions using a Mann–Whitney U test. Agreement between BERT and the human reference labels was low ($\kappa = 0.23$; accuracy = 48.7%) and varied across sub-corpora (IE: $\kappa = 0.46$; CS: $\kappa = 0.27$; IC: $\kappa = 0.15$; ES: $\kappa = 0.04$). These figures indicate that the model is not well calibrated for this Cantonese-English courtroom corpus. The results are therefore interpreted as model-derived corpus tendencies rather than precise sentence-level classifications.

4 Results and discussion

4.1 RQ1: Sentiment scores across sub-corpora

Table 2 reports the descriptive statistics. All four sub-corpora had negative mean sentiment scores, which may reflect the adversarial nature of courtroom discourse (Hale, 2002). The lowest mean was found in IC ($M = -0.424$), followed by CS ($M = -0.253$), ES ($M = -0.180$), and IE ($M = -0.071$).

Corpus	Mean	SD	Median	N
ES	-0.180	0.394	-0.234	12,062
IE	-0.071	0.482	-0.049	10,589
CS	-0.253	0.434	-0.303	10,589
IC	-0.424	0.366	-0.538	12,062

Table 2: Descriptive statistics of sentiment scores across four sub-corpora.

According to Table 3, all pairwise differences were statistically significant after Bonferroni correction, with adjusted p values below .0001. On the -1 to $+1$ scale, the largest difference was between IE and IC (mean difference = 0.353), about 18% of the full scale range. The smallest difference was between ES and CS (mean difference = 0.073), indicating a small practical difference despite statistical significance. This shows why effect sizes are needed to clarify the practical magnitude of the pairwise differences.

Pair	Mean diff.	95% CI	Cohen's d
ES vs. IE	-0.109	[-0.120, -0.097]	-0.248
ES vs. CS	0.073	[0.062, 0.083]	0.176
ES vs. IC	0.244	[0.235, 0.254]	0.641
IE vs. CS	0.181	[0.169, 0.194]	0.396
IE vs. IC	0.353	[0.341, 0.363]	0.831
CS vs. IC	0.171	[0.160, 0.182]	0.429

Table 3: Pairwise comparisons of sentiment scores.

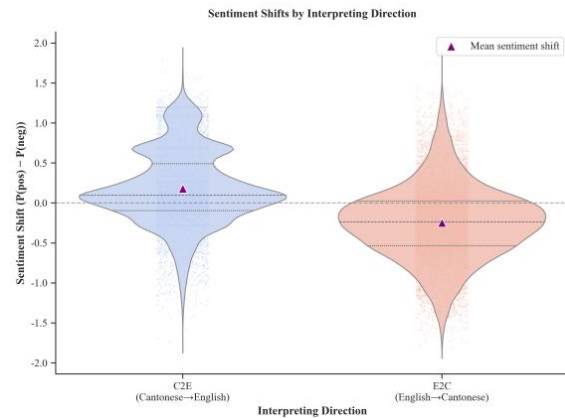


Figure 1: Sentiment shift by interpreting direction.

4.2 RQ2: Effect of interpreting direction

In C2E, the mean sentiment gap was $+0.181$ (median = $+0.096$; mean absolute shift = 0.375), suggesting that interpreted English was less negative than the Cantonese source in model-derived scores. In E2C, the mean gap was -0.244 (median = -0.236 ; mean absolute shift = 0.412), suggesting that interpreted Cantonese was more negative than the English source. The difference between directions was significant (Mann–Whitney $U = 95,870,002$, $p < .001$). Figure 1 illustrates this asymmetric pattern.

One possible explanation, supported by earlier studies of Cantonese-English courtroom interpreting, concerns Cantonese utterance

particles. These particles can express stance and attitude, often without direct English equivalents (Leung and Gibbons, 2009). Wong (2025) shows that interpreters use Cantonese utterance particles to regulate turn-taking, assert authority, and modulate pragmatic force in courtroom interaction. In C2E interpreting, some particle-carried stance may not be expressed directly in English. In E2C interpreting, particles may add stance or pragmatic force to the Cantonese output, which may partly contribute to stronger negative model-derived scores. This explanation remains tentative because particle use was not directly annotated in the present study.

4.3 Human validation and error analysis

Table 4 shows the confusion matrix between BERT predictions and the human reference labels. BERT agreed with the human labels in 76 of 156 cases (48.7%). In the validation set, the human reference labels classified 74 sentences as neutral and 61 as negative, whereas the stratified BERT sample contained 52 sentences in each predicted class. The matrix therefore shows that BERT did not align well with the human label distribution in this sample. In particular, many human-neutral cases were assigned by BERT to negative or positive categories.

BERT / Human	Negative	Neutral	Positive	Total
Negative	32	18	2	52
Neutral	15	31	6	52
Positive	14	25	13	52
Total	61	74	21	156

Table 4: Confusion matrix between BERT predictions and human reference labels.

The error analysis further cautions against a strong interpretation of the directional findings. We manually checked 30 source–target pairs where BERT gave different polarity labels to the source and target texts, with 15 pairs from C2E and 15 from E2C. Only 2 pairs (6.7%) were judged to show genuine sentiment shifts. The other 28 pairs (93.3%) involved at least one classifier error: 13 on the source sentence, 10 on the target sentence, and 5 on both sides.

The overall error rate was the same across directions: 14 of 15 pairs in C2E and 14 of 15 pairs in E2C involved classifier error. In C2E, errors were concentrated on the Cantonese source sentence (10/15 pairs). In E2C, errors were

concentrated on the Cantonese target sentence (7/15 pairs), with four further pairs involving errors on both sides. This suggests that the model had particular difficulty with Cantonese courtroom language. The RQ2 pattern should therefore be treated as a model-derived tendency that requires confirmation with Cantonese-adapted sentiment tools. It should not be read as direct evidence that interpreters strategically changed sentiment.

5 Conclusion

This study applied multilingual BERT to the HKCI and reported two exploratory findings. First, model-derived sentiment scores were predominantly negative across all four sub-corpora, with significant pairwise differences. This pattern points to uneven model-derived sentiment distributions across the corpus. Second, sentiment scores showed an asymmetry across interpreting directions: in model-derived scores, C2E renditions were less negative than their Cantonese sources, while E2C renditions were more negative than their English sources. However, the error analysis shows that this asymmetry may partly reflect classifier behaviour on Cantonese text. While previous studies have linked Cantonese utterance particles to affective and pragmatic meanings, this possibility was not examined directly in the present study and therefore remains open for future investigation. Future work could also explore whether sentiment patterns vary across speaker roles (e.g., legal professionals versus lay participants).

The main contribution of this paper is methodological. The findings suggest that off-the-shelf multilingual sentiment tools can help identify potential corpus-level patterns in Cantonese-English legal interpreting, but they cannot by themselves support strong conclusions about pragmatic equivalence. Future research should combine corpus-based sentiment analysis with discourse-pragmatic approaches and develop tools better adapted to Cantonese-English courtroom interpreting.

References

- Susan Berk-Seligson. 2017. *The Bilingual Courtroom: Court Interpreters in the Judicial Process*. University of Chicago Press, Chicago.
- Andrew K. F. Cheung. 2017. Non-renditions in court interpreting: A corpus-based study. *Babel*,

- 63(2):174–199.
<https://doi.org/10.1075/babel.63.2.02che>.
- Jennifer L. F. Cheung-Pease and Adam Pease. 2023. Quantitative analysis of clarification discourse of interpreter-moderated courtroom using a Cantonese-English bilingual corpus. In Andrew K. F. Cheung, Kanglong Liu, and Riccardo Moratto (eds.), *Corpora in Interpreting Studies*, pages 28–43. Routledge.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 8440–8451.
<https://doi.org/10.18653/v1/2020.acl-main.747>.
- EULITA. 2013. Code of Professional Ethics. European Association for Legal Interpreters and Translators.
<https://www.eulita.eu/wp-content/uploads/files/EULITA-code-London-e.pdf>.
- Sandra Hale. 2002. How faithfully do court interpreters render the style of non-English speaking witnesses’ testimonies? A data-based study of Spanish–English bilingual proceedings. *Discourse Studies*, 4(1):25–47.
<https://doi.org/10.1177/14614456020040010201>.
- Yujie Huang, Andrew K. F. Cheung, Kanglong Liu, and Han Xu. 2025. Can sentiment analysis help to assess accuracy in interpreting? A corpus-assisted computational linguistic approach. *Applied Linguistics*. Advance online publication.
<https://doi.org/10.1093/applin/amaf026>.
- Lei Lei and Dilin Liu. 2021. *Conducting Sentiment Analysis*. Cambridge University Press, Cambridge, UK.
- Ester Leung and John Gibbons. 2009. Interpreting Cantonese utterance-final particles in bilingual courtroom discourse. *Interpreting*, 11(2):190–215.
<https://doi.org/10.1075/intp.11.2.05leu>.
- Kang Kwong Luke. 1990. *Utterance Particles in Cantonese Conversation*. John Benjamins Publishing Company, Amsterdam and Philadelphia.
- NLPTown. 2020. bert-base-multilingual-uncased-sentiment [Model]. Hugging Face.
<https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment>.
- Cecilia Lok Yee Wong. 2023. A corpus-based study of interpreters’ non-renditions and power manifestations in courtrooms of Hong Kong. In Andrew K. F. Cheung, Kanglong Liu, and Riccardo Moratto (eds.), *Corpora in Interpreting Studies*, pages 161–180. Routledge.
- Cecilia Lok Yee Wong. 2025. Corpus-driven study of interpreters’ use of Cantonese utterance particles in sentence-initial position in bilingual courtroom discourse. *Babel*, 71(4):500–526.
<https://doi.org/10.1075/babel.24130.won>.
- Han Xu, Jinghang Gu, Kanglong Liu, and Qinyi Li. 2025. Can professional interpreters truly convey the speaker’s sentiment? Exploring the potential of a computational approach. *Digital Scholarship in the Humanities*, 40(2):639–650.
<https://doi.org/10.1093/lc/fqaf017>.
- Ran Yi. 2026. Enhancing sentiment analysis and error prediction in lawyer questioning through legal NLP models: A mixed-methods study in Australian courtrooms. *Journal of Mixed Methods Research*. Advance online publication.
<https://doi.org/10.1177/15586898261429189>.

How Decoding Settings Shape Lexical Variation

in Iterative LLM-Translation Workflows:

A Case Study of Chinese-English Literary Translation

Lulu Wang
The Hong Kong
Polytechnic University
Hong Kong
lulu99.wang@
connect.polyu.hk

Hengtian Zhang
The Hong Kong
Polytechnic University
Hong Kong
heng-tian.zhang@
connect.polyu.hk

Lieve Macken
Ghent University
Ghent, Belgium
lieve.macken@
ugent.be

Kanglong Liu*
The Hong Kong
Polytechnic University
Hong Kong
kl.liu@
polyu.edu.hk

Abstract

Large language models (LLMs) are increasingly used in automated translation workflows involving iterative refinement, yet the role of decoding settings in shaping lexical variation within such workflows remains underexplored. This study examines how temperature and top-p settings influence lexical diversity, lexical frequency distribution, and inter-stage lexical similarity within a three-stage literary translation pipeline (initial translation, self-revision, and stylistic refinement). Using GPT-5.1, we translated contemporary Chinese fiction texts under controlled parameter settings. Results show that less restrictive decoding (higher temperature and top-p) leads to increased lexical diversity, higher proportions of low-frequency vocabulary, and more extensive inter-stage revision, with these effects amplified across successive refinement stages. Compared with professional human translations, LLM outputs exhibit higher lexical diversity and a preference for lower-frequency vocabulary across all conditions. The findings reveal how decoding parameters shape lexical variation in LLM-generated translations, with implications for LLM-based translation workflow design.

1 Introduction

Large language models (LLMs) demonstrate strong performance in translation tasks, producing

fluent outputs across a wide range of genres (Gao et al., 2024; Jiang et al., 2024). Recent work has explored automated LLM-based translation workflows, with some studies reporting that they can rival or even surpass professional human translations in perceived quality (Wang et al., 2026; Wu et al., 2024; Wu et al., 2025). A common feature of these workflows is iterative refinement, where models revise their initial outputs based on self-reflection and feedback (Chen et al., 2024; Madaan et al., 2023; Xu et al., 2024).

Despite improvements in overall quality, little is known about how the lexical properties of LLM-generated translations evolve across refinement stages. This issue is particularly consequential for literary translation, where lexical patterns underlie stylistic expression and creativity, influencing tone, voice, and perceived literary quality (Leech & Short, 2007; Munat, 2007). As texts undergo successive revision and refinement, cumulative lexical changes may alter stylistic characteristics in systematic ways. Understanding these dynamics is therefore essential for evaluating LLM-based translation workflows beyond holistic quality assessments.

A key but underexplored factor in this process is the setting of decoding parameters. LLMs generate text by sampling from a probability distribution over candidate tokens at each step, and outputs are sensitive to the parameters governing this process (Wiher et al., 2022; Zhou et al., 2025). Among these parameters, temperature rescales the probability distribution prior to sampling: higher values flatten the distribution and increase the

* Corresponding author

likelihood of selecting lower-probability tokens, introducing greater randomness, whereas lower values sharpen the distribution and produce more deterministic outputs (Arias et al., 2025; Troshin et al., 2025). Top-p sampling (also known as nucleus sampling) restricts generation to the smallest set of tokens whose cumulative probability exceeds a threshold p , effectively truncating the low-probability tail of the distribution. These two mechanisms operate on different aspects of the sampling process: temperature affects distribution shape, while top-p constrains candidate set size (Holtzman et al., 2020; Ippolito et al., 2019).

Although the computational mechanisms of temperature and top-p are well understood, their stylistic consequences are often assumed rather than empirically verified. Existing studies have primarily examined their impact on output quality in single-pass generation settings (Ippolito et al., 2019; Peng et al., 2023), providing limited insight into how decoding parameters shape lexical stylistic features. Their role in multi-stage refinement workflows has received comparatively little attention.

The iterative refinement process in LLM-based translation workflows introduces additional complexity. In refinement pipelines, the same model repeatedly revises its own outputs, implicitly applying internal criteria for improvement at each stage. Prior work suggests that such processes can steer outputs toward model-preferred stylistic patterns (Madaan et al., 2023; Xu et al., 2024), yet it remains unclear whether lexical variation introduced through stochastic sampling persists across refinement stages or is gradually normalized through successive revisions. More broadly, lexical diversity is a text-level property emerging from many sequential decisions, and the relationship between token-level stochasticity and document-level outcomes under iterative revision is not fully understood.

To address these gaps, this study examines how two decoding parameters, temperature and top-p, influence lexical stylistic features within a three-stage refinement pipeline for Chinese-to-English literary translation (initial translation, self-revision, and stylistic refinement). We track changes in lexical diversity, frequency distribution, and inter-stage lexical similarity under controlled sampling conditions. By situating decoding parameters within a realistic iterative workflow, the study

provides a process-oriented account of how lexical variation emerges and evolves in LLM-generated literary translation, with further implications for the design of LLM-based translation pipelines.

2 Related Work

Research in natural language generation has demonstrated that decoding strategies play a central role in shaping model outputs (Arias et al., 2025). Early work established a notable but task-specific trade-off between output diversity (specifically lexical variation and n -gram richness across generated sequences) and overall text quality, whereby higher sampling temperatures increase variability but may compromise output quality, particularly fluency, contextual coherence, and accuracy (Ippolito et al., 2019; Wiher et al., 2022).

A growing body of research has further examined how sampling hyperparameters influence LLM performance across a range of tasks, including code generation (Arora et al., 2024; Donato et al., 2025), reasoning (Troshin et al., 2025), open-ended creative writing (Peeperkorn et al., 2024), genre analysis (Grullon-Polanco, 2026), and machine translation (Peng et al., 2023). Across these domains, results consistently highlight the sensitivity of LLM outputs to decoding configurations. For instance, Donato et al. (2025) show that temperature and top-p substantially affect both output variability and correctness in code generation. Similarly, Peeperkorn et al. (2024) find that temperature has modest but measurable effects on novelty and coherence in narrative generation. These findings suggest that decoding parameters are not merely technical settings but fundamental determinants of LLM output behavior.

In the context of machine translation, prior work has primarily examined decoding effects in terms of accuracy. Peng et al. (2023) report that increasing temperature leads to lower BLEU and COMET scores, while also increasing the risk of hallucination in non-English-centric translation tasks. Likewise, Li et al. (2025) recommend near-deterministic decoding (i.e., temperature close to zero) to maximize translation accuracy. However, these studies focus on reference-based automated evaluations of quality, leaving the linguistic properties, such as stylistic features, largely overlooked.

Despite these advances, most studies focus on single-pass generation, leaving the effects of

decoding settings in multi-stage workflows underexplored. Given that decoding effects are task-dependent (Wiher et al., 2022), it is uncertain whether prior findings extend to stylistically constrained tasks such as literary translation, particularly with respect to how decoding settings influence lexical characteristics within emerging iterative refinement pipelines (Chen et al., 2024; Wang et al., 2026; Wu et al., 2024; Wu et al., 2025).

3 Methods

3.1 Corpus

This study uses a corpus of 200 Chinese fiction texts extracted from contemporary published works across diverse genres (science fiction, *wuxia*, romance, humor, crime and detective, and rural fiction). Source passages average 2,626 Chinese characters ($SD = 160$), ranging from 2,229 to 3,138 characters, with a total size of 525,200 Chinese characters. Each passage has a corresponding published professional English translation for baseline comparison.

3.2 LLM Translation Workflow

The LLM translation process follows a three-stage iterative pipeline that progressively shifts from interlingual semantic transfer to intralingual stylistic refinement (Jakobson, 1959; Macken et al., 2022; Whyatt, 2017; Wang et al., 2026).

Stage 0 (Initial Translation): The model receives the Chinese source text and is prompted to produce a faithful and fluent literary translation. Quality criteria emphasize accuracy, fluency, and stylistic appropriateness for fiction translation.

Stage 1 (Self-Revision): The model receives its Stage 0 output together with the source text and revises for accuracy, fluency, and lexical consistency, following criteria adapted from ISO 18587 post-editing guidelines (International Organization for Standardization, 2017).

Stage 2 (Stylistic Refinement): The model receives only its Stage 1 output, without access to the source text, and refines fluency, style, and narrative flow while preserving semantic content.

3.3 Experimental Design

All translations were generated using GPT-5.1 via the Microsoft Azure¹ OpenAI API in March 2026. To examine the effects of sampling parameters on

lexical characteristics across refinement stages, we manipulated temperature and top-p settings in two separate experiments, avoiding confounds between distribution flattening (temperature) and probability mass truncation (top-p).

Experiment 1 (Temperature): All 200 source texts were translated across three temperature conditions (0, 0.5, 1) with top-p fixed at 1. Although GPT-5.1 supports temperature values up to 2.0, we restricted our investigation to the 0–1 range, which encompasses the settings most commonly adopted in practical applications and prior studies on LLM-based translation and text generation (Cai et al., 2024; Zhang et al., 2025). Temperature values exceeding 1.0 increasingly flatten the probability distribution toward uniform sampling, typically compromising output coherence in translation tasks. The unrestricted top-p setting ensures that temperature alone determines the degree of sampling variability within this range.

Experiment 2 (Top-p): A randomly sampled subset of 100 source texts was translated across three top-p conditions (0.5, 0.8, 1) with temperature fixed at 0.7, a conventional default that balances output diversity and coherence (Cai et al., 2024). The subset enables controlled comparison within practical resource constraints, while the top-p conditions maintain sufficient stochasticity for top-p to meaningfully constrain the candidate token space, avoiding near-deterministic outputs that would occur at lower temperatures. The selected top-p values represent a gradient from restricted (0.5) to unrestricted (1) sampling, enabling systematic examination of how probability mass truncation affects lexical outcomes.

Under each condition, the same parameter settings were applied across all three workflow stages.

3.4 Lexical Style Metrics

In computational linguistics and stylometry, lexical style functions as a measurable “linguistic fingerprint” representing idiosyncratic patterns of language use (Alrabaa et al., 2019). In this study, it is examined through three dimensions: lexical diversity, lexical frequency distribution, and inter-stage lexical similarity.

Lexical diversity is measured using Type-Token Ratio (TTR), Moving-Average Type-Token Ratio (MATTR; Covington & McFall,

¹ <https://azure.microsoft.com/>

2010), Measure of Textual Lexical Diversity (MTLD; Malvern & Richards, 1997; McCarthy, 2005), and lexical entropy. TTR is included as a conventional baseline measure of vocabulary richness, while MATTR and MTLD provide more robust estimates by reducing sensitivity to text length. Lexical entropy serves as a complementary indicator by capturing the probability distribution, average information content, and unpredictability of word usage within a text (Mohseni et al., 2022).

Lexical frequency distribution is examined through lexical frequency profile (LFP) analysis, which measures the extent to which texts rely on high- or low-frequency vocabulary (Laufer & Nation, 1995; Van Parys et al., 2025). This dimension is particularly relevant given recent concerns that LLMs, as statistical next-word predictors, may systematically favor high-frequency forms while pruning away “rich, long statistical tails” of language (Vanmassenhove, 2025, p. 1). LFP analysis can thus reveal whether iterative refinement amplifies or mitigates such frequency-based biases.

Finally, lexical similarity is calculated using Levenshtein edit distance at both character and word levels (Anson et al., 2019), capturing morphological variations and structural reorganizations across refinement stages. Taken together, these measures provide a multidimensional account of lexical style, covering vocabulary diversity, frequency-based lexical choice, and inter-stage similarity across refinement stages.

3.5 Lexical Diversity

This dimension captures the range and variability of vocabulary usage within a text, reflecting both the breadth of lexical choice and the distributional evenness of word frequencies (McCarthy & Jarvis, 2010). In this study, it is operationalized using four complementary metrics.

TTR (Type-Token Ratio): The ratio of unique word types to total tokens. TTR is the most fundamental measure of lexical diversity but is sensitive to text length, as the proportion of new types inevitably decreases in longer texts due to repetition of high-frequency words (Vermeer, 2000).

MATTR (Moving-Average Type-Token Ratio): To mitigate length sensitivity, MATTR

calculates TTR over a fixed-size sliding window (e.g., 100 tokens) and averages these scores across the text (Covington & McFall, 2010).²

MTLD (Measure of Textual Lexical Diversity): MTLD calculates the mean length of sequential word strings (factors) that maintain a specific minimum TTR threshold (typically 0.72) before dropping below it, effectively evaluating how long a text can sustain diverse vocabulary without repeating words excessively (McCarthy, 2005). The algorithm evaluates the running TTR as it reads sequentially left-to-right; when the TTR drops below 0.72, that segment is counted as one factor, and the calculation restarts. The procedure is usually repeated both forward and backward, and the final score is the total number of tokens divided by the total number of factors (averaged across both passes) (Koizumi, 2012).

Lexical Entropy: Rooted in information theory, lexical entropy quantifies the average uncertainty or information density in a text based on the relative frequency probability distribution of its words (calculated as $H = -\sum p_i \log_2 p_i$; Shannon, 1951). Higher entropy indicates a more uniform distribution and greater textual unpredictability. A highly repetitive, simplistic text will feature a steep word frequency distribution and low entropy, whereas canonical literature utilizing varied, unpredictable word choices possesses a flatter distribution and substantially higher entropy values.

3.6 Lexical Frequency Profile Analysis

LFP analysis was used to examine the distribution of vocabulary across frequency bands. It measures the extent to which a text relies on vocabulary drawn from different frequency levels and is widely used to assess lexical sophistication and vocabulary use in corpus-based research (Laufer & Nation, 1995; Van Parys et al., 2025). Based on the BNC/COCA 25,000 word list (Nation, 2017), words in each translation were assigned to frequency levels and grouped into a three-band profile following Vanmassenhove et al. (2021): B1 (the 1,000 most frequent words), B2 (words ranked 1,001–2,000), and B3 (words beyond the 2,000-frequency threshold). For each translation, the proportion of tokens in each band was calculated to examine whether refinement stages

101, 3–102, and so on until the end of the document, finally averaging these values.

² For example, with a window size of 100, the algorithm computes TTR for tokens 1–100, then 2–

and decoding settings favor more frequent or less frequent lexical choices.

3.7 Inter-Stage Lexical Similarity

To quantify how extensively translations are revised across stages, we computed the Levenshtein edit distance between outputs at successive stages (S0–S1, S1–S2) and cumulatively (S0–S2). This metric captures the minimum number of single-unit operations (insertions, deletions, and substitutions) required to transform one version into another (Anson et al., 2019).

Distances were calculated at both character and word levels. Character-level distance captures fine-grained orthographic and morphological variation. Word-level distance operates over tokenized sequences, treating each word as an irreducible unit, and thus more directly reflects meaningful lexical and structural changes; it was therefore used as the primary measure in our analysis.

To ensure comparability across texts of varying lengths, raw distances were normalized by the length of the longer text. The resulting proportion reflects the extent of revision between stages.

4 Results and Discussion

4.1 Text Length

As shown in Table 1, the lengths of LLM translations remain relatively stable from the initial translation stage (S0) to self-revision stage (S1) across all temperature settings (0, 0.5, 1). During the intralingual stylistic refinement stage (S2), which aims to improve fluency and stylistic quality, text length increases dramatically, with the highest temperature producing the longest outputs ($M = 2852.65$, $SD = 404.37$). This suggests that higher temperature not only lengthens the text but also introduces greater variability in length during intralingual refinement.

Similar to the temperature conditions, text lengths consistently increase from S0 to S2 across all top-p settings (Table 2). Expanding the candidate token space (top-p = 1) results in slightly longer texts at every stage compared to more restrictive settings (top-p = 0.5 and 0.8), with the effect most pronounced at S2 ($M = 2669.08$, $SD = 393.79$). Limiting the candidate token set based on cumulative probability appears to constrain text expansion during refinement.

Across all conditions, LLM translation outputs exceed the professional human translations in length, with the gap widening substantially at S2.

Temperature Setting	Refinement Stage		
	S0 (Initial)	S1 (Self-revision)	S2 (Stylistic Refinement)
0	2078 (198)	2137 (201)	2490 (291)
0.5	2064 (195)	2128 (206)	2484 (376)
1	2110 (198)	2212 (214)	2853 (404)
Human Translation	1901 (418)		

Table 1: Mean word counts (SD) of LLM translations across refinement stages and temperature settings compared with human translation ($N = 200$).

Top-p Setting	Refinement Stage		
	S0 (Initial)	S1 (Self-revision)	S2 (Stylistic Refinement)
0.5	2110 (216)	2167 (223)	2561 (353)
0.8	2116 (220)	2189 (225)	2641 (371)
1	2121 (222)	2199 (229)	2669 (394)
Human Translation	1883 (459)		

Table 2: Mean word counts (SD) of LLM translations across refinement stages and top-p settings compared with human translation ($N = 100$).

4.2 Lexical Diversity

This section examines how decoding parameters, specifically temperature (0, 0.5, and 1) and top-p (0.5, 0.8, and 1), influence lexical diversity across the iterative refinement workflow.

As detailed in Table 3 and Table 4 and visualized in the radar plots (Figure 1 and Figure 2), both parameters exert a systematic and monotonic influence on lexical diversity. Across all three refinement stages (S0, S1, and S2), increasing either the sampling randomness (temperature) or the candidate token pool (top-p) consistently yields higher scores across all measured metrics (TTR, MATTR, MTLD, and lexical entropy).

The results show that while lexical diversity increases across the iterative refinement process, less restrictive parameter settings amplify this trajectory. Diversity scores remain relatively clustered during S0, but the gaps between different settings widen considerably by S2. This indicates that granting the model more sampling freedom encourages the exploration of a much broader and more varied vocabulary, particularly during the stylistic refinement phase.

Compared with professional human translations, the LLM outputs display higher lexical diversity across all parameter settings and stages. Even at the most deterministic settings ($T = 0$ or $\text{top-p} = 0.5$) in the initial draft, the LLM generates a more diverse vocabulary (e.g., $\text{MATTR} \approx 0.73\text{--}0.74$, $\text{MTLD} \approx$

113–114) than the human baseline (MATTR $\approx$ 0.728, MTLD $\approx$ 106). By S2 at maximum settings ($T = 1$ or top- $p = 1$), this gap expands dramatically, suggesting a tendency for the LLM to produce increasingly dispersed lexical choices under less restrictive sampling.

This finding should be interpreted with caution. Notably, the human translation sample in this study is relatively limited in size, whereas LLMs draw on vastly larger training corpora, potentially exposing them to a wider range of lexical alternatives. It also remains uncertain whether the increasing lexical diversity across refinement stages reflects genuine stylistic enrichment or problematic variation, such as unnecessary synonymic substitution, register inconsistencies, or semantic drift. As higher diversity does not necessarily indicate higher quality (Shypula et al., 2025), without quality evaluation, we cannot determine whether these lexical choices are contextually appropriate or functionally equivalent to those characteristic of professional human translation.

Refinement Stage	Temperature Setting	Lexical Diversity			
		TTR	MATTR	MTLD	Lexical Entropy
S0	0	0.3699	0.7417	114.3634	8.3378
	0.5	0.3721	0.7446	117.4163	8.3551
	1	0.3823	0.7518	125.7408	8.4204
S1	0	0.3768	0.7459	118.4793	8.3898
	0.5	0.3806	0.75	123.1024	8.4173
	1	0.3915	0.7588	133.7715	8.5029
S2	0	0.3842	0.7617	139.6092	8.5526
	0.5	0.3926	0.7676	147.9877	8.5968
	1	0.3991	0.7807	168.8523	8.7729
Human Translation		0.3549	0.7281	105.948	8.145

Table 3: Lexical diversity of LLM translations across refinement stages and temperature settings compared with human translation.

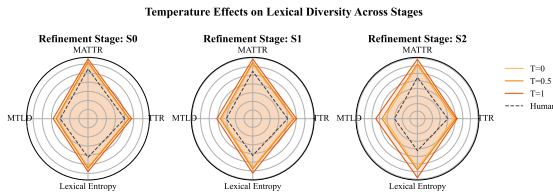


Figure 1: Radar plots of lexical diversity of LLM translations across refinement stages and temperature settings compared with human translation.

Refinement Stage	Top-p Setting	Lexical Diversity			
		TTR	MATTR	MTLD	Lexical Entropy
S0	0.5	0.3675	0.741	113.457	8.3446
	0.8	0.37	0.7426	115.2529	8.3609
	1	0.3721	0.7452	117.8178	8.3732
S1	0.5	0.3754	0.746	118.061	8.4006
	0.8	0.3776	0.7483	121.4817	8.4235
	1	0.38	0.7508	123.7239	8.437
S2	0.5	0.3848	0.7635	141.9736	8.5818
	0.8	0.3872	0.7672	146.912	8.6259
	1	0.3919	0.7697	151.5484	8.6569
Human Translation		0.3595	0.7285	106.5089	8.1305

Table 4: Lexical diversity of LLM translations across refinement stages and top-p settings compared with human translation.

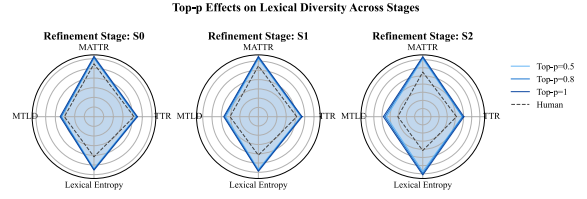


Figure 2: Radar plots of lexical diversity of LLM translations across refinement stages and top-p settings compared with human translation.

4.3 Lexical Frequency Profile

Overall, iterative refinement consistently shifted vocabulary toward lower-frequency words across all decoding settings (Figure 3 and Figure 4). Across refinement stages, the proportion of the most frequent 1,000 words (B1) declined from the initial draft (S0) to the final draft (S2), while the proportion of words outside the top 2,000 most frequent words (B3) increased correspondingly. At $T = 0.5$, for instance, B1 dropped from 69.56% at S0 to 63.00% at S2, while B3 rose from 14.96% to 20.04%.

Temperature had a clear effect on lexical frequency distribution (Figure 3). Higher temperatures produced translations with lower proportions of high-frequency words and greater use of lower-frequency vocabulary. At S2, B1 decreased from 64.09% at $T = 0$ to 61.59% at $T = 1$, while B3 increased from 19.13% to 21.49%. This pattern suggests that higher temperatures increase the likelihood of selecting less common lexical items from a broader range of candidates.

Top- p showed a similar directional pattern (Figure 4). Increasing top- p from 0.5 to 1 at S2 reduced B1 from 64.44% to 63.50% and increased B3 from 19.22% to 20.04%. As with temperature, expanding the candidate token space was associated with a shift toward lower-frequency vocabulary.

Compared with human translations, LLM outputs contained fewer high-frequency words and more low-frequency vocabulary at every stage. This tendency intensified with iterative refinement, with final drafts (S2) showing the greatest reliance on lower-frequency items. Rather than converging toward human lexical norms, successive refinement therefore appears to increase divergence from the human lexical frequency

profile, pushing LLM translations toward more atypical vocabulary choices.

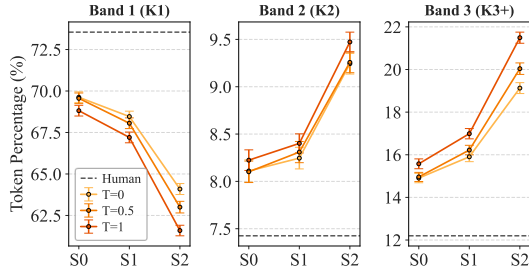


Figure 3: Lexical frequency profiles of LLM-generated translations across refinement stages and temperature settings. Error bars indicate standard error.

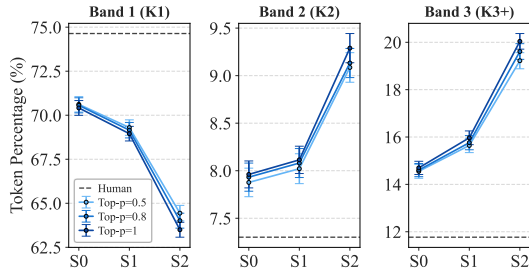


Figure 4: Lexical frequency profiles of LLM-generated translations across refinement stages and top-p settings. Error bars indicate standard error.

The differences in lexical frequency profiles between LLM-generated and human translations suggest distinct underlying orientations toward stylistic quality. Even at the initial stage, and under deterministic decoding, LLM outputs exhibit a higher proportion of lower-frequency vocabulary. This pattern appears to challenge theoretical expectations that LLM-generated texts converge toward high-probability forms due to the probabilistic nature of language modeling (Uhlmann, 2025; Vanmassenhove, 2025). However, the tendency toward sophisticated vocabulary may instead reflect biases in the model’s training data or its learned associations between literary translation and elevated register. Specifically, what constitutes a high-probability token is determined by the model’s training distribution. If the training data overrepresents certain registers or genres, lower-frequency vocabulary in general usage may in fact be high-probability within the model’s learned distribution. Iterative refinement appears to strengthen this tendency, with each stage shifting outputs away from high-frequency vocabulary toward more marked forms. One possible interpretation is that the model implicitly associates lexical sophistication with stylistic improvement, though this remains speculative without further investigation.

The implications of these findings for translation quality, however, remain uncertain. Without a benchmark based on original non-translated English literary fiction, it is difficult to determine whether LLM outputs are genuinely oversophisticated or whether human translations are comparatively simplified. The model’s preference for lower-frequency vocabulary might align more closely with native literary prose than with translated text, which complicates straightforward quality judgments.

4.4 Inter-Stage Lexical Similarity

As shown in Table 5 and Table 6, both temperature and top-p are positively associated with edit distance: higher settings lead to more extensive textual modification at each stage (Figure 5 and Figure 6). While word-level distance serves as the primary measure, character-level distances follow a similar pattern, indicating that revisions involve both broad lexical restructuring and fine-grained orthographic adjustments.

During the self-revision stage (S0–S1), higher parameter settings result in substantially more editing. Normalized word-level distance more than doubles from 0.19 at $T = 0$ to 0.42 at $T = 1$, and increases from 0.21 at top-p = 0.5 to 0.32 at top-p = 1. The stylistic refinement stage (S1–S2) introduces even greater changes: at maximum settings ($T = 1$ or top-p = 1), normalized word distance reaches 0.67 and 0.59, respectively.

In the cumulative refinement process (S0–S2), texts undergo substantial lexical shifts. Normalized word distance reaches 0.72 at $T = 1$ (compared to 0.54 at $T = 0$) and 0.65 at top-p = 1 (compared to 0.58 at top-p = 0.5). These values indicate that the number of edit operations approaches or exceeds two-thirds of the total token count under less restrictive decoding settings, reflecting extensive rewriting across the pipeline.

Stage Pair	Temperature Setting	Levenshtein Distance			
		Char. Dist.	Norm. Char. Dist.	Word Dist.	Norm. Word Dist.
S0-S1	0	1961.23	0.1599	401.22	0.1888
	0.5	2582.14	0.2095	542.93	0.2548
	1	4396.69	0.3403	937.82	0.4231
S1-S2	0	6084.5	0.4107	1232.42	0.4905
	0.5	6567.87	0.4358	1331.04	0.5248
	1	9480.09	0.5481	1920.58	0.6683
S0-S2	0	6666.16	0.4507	1360.75	0.5423
	0.5	7247.23	0.4821	1484.79	0.5869
	1	10180.65	0.5895	2076.06	0.7234

Note. Char. = Character, Dist. = Levenshtein edit distance, Norm. = Normalized by the length of the longer text in each comparison.

Table 5: Inter-stage edit distance of LLM translations across temperature settings.

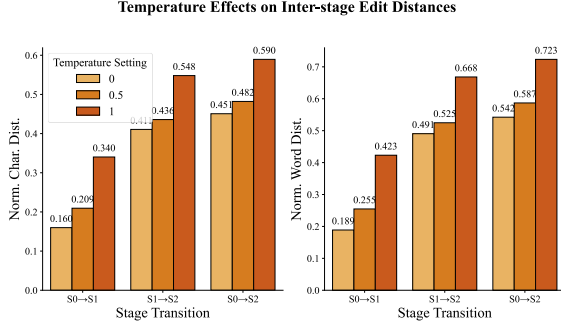


Figure 5: Bar plots of normalized edit distances across refinement stages and temperature settings.

Stage Pair	Top-p Setting	Levenshtein Distance			
		Char. Dist.	Norm. Char. Dist.	Word Dist.	Norm. Word Dist.
S0-S1	0.5	2144.04	0.1727	440.14	0.2045
	0.8	2831.98	0.2247	597.02	0.2737
	1	3283.07	0.259	696.09	0.3175
S1-S2	0.5	6651.39	0.4384	1345.54	0.5241
	0.8	7316.58	0.4647	1484.33	0.5596
	1	7826.9	0.4888	1586.15	0.5898
S0-S2	0.5	7239.19	0.4773	1476.78	0.5753
	0.8	8006.92	0.509	1637.37	0.6177
	1	8516.1	0.5327	1743.8	0.6492

Note. Char. = Character, Dist. = Levenshtein edit distance, Norm. = Normalized by the length of the longer text in each comparison.

Table 6: Inter-stage edit distance of LLM translations across top-p settings.

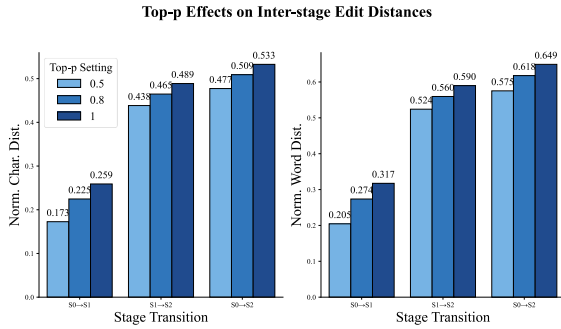


Figure 6: Bar plots of normalized edit distances across refinement stages and top-p settings.

The edit distance patterns have implications for understanding how parameter settings modulate the extent to which an LLM alters its output across refinement stages. Given the same prompt, more deterministic settings lead the model to preserve much of its prior output while making adjustments, whereas less restrictive settings produce more substantial departures from previous iterations, generating extensive rewrites. Whether this represents a more thorough revision or simply increased stochastic variation remains unclear; higher temperature and top-p values mechanically expand the candidate pool at each token position, which may produce surface-level instability without corresponding gains in quality or coherence. Disentangling these possibilities would require complementary quality evaluation.

5 Conclusion

This study examined how decoding parameters (temperature and top-p) influence the lexical characteristics of LLM-generated translations across an iterative refinement workflow. The findings reveal a consistent pattern: less restrictive parameter settings produce greater lexical diversity, more low-frequency vocabulary, and more extensive inter-stage revision. These effects are amplified through iterative refinement, with the stylistic refinement stage showing the most pronounced divergence from both earlier drafts and human translations. Notably, LLM outputs displayed higher lexical diversity and larger proportions of lower-frequency vocabulary than human translations across all settings, and these differences widened through successive refinement stages.

These findings carry tentative practical implications. Translators working with LLMs may benefit from understanding how parameter settings shape lexical output rather than relying on defaults or intuition. More restrictive settings may suit tasks requiring accessible, high-frequency vocabulary, while less restrictive settings might benefit texts where lexical richness is valued. Awareness that iterative refinement progressively increases lexical diversity while shifting vocabulary toward lower-frequency bands could inform decisions about how many revision stages to employ before such effects become counterproductive.

However, several limitations of this study should be acknowledged. First, this study focused exclusively on lexical profiles without evaluating translation quality; a richer vocabulary does not necessarily indicate better translation, and the trade-offs between output diversity and quality warrant further investigation. Second, the analysis was limited to a single LLM (GPT-5.1) and a single broad domain (literary fiction), which constrains generalizability. Finally, only Chinese-to-English translation was examined; parameter effects may differ for other language pairs. Future research should address these gaps by integrating quality evaluation, expanding sample sizes, and extending the analysis across models, text types, and language pairs.

References

Gao, Ruiyao, Yumeng Lin, Nan Zhao, and Zhenguang G. Cai. 2024. Machine translation of Chinese classical poetry: A comparison among ChatGPT,

- Google Translate, and DeepL Translator. *Humanities and Social Sciences Communications*, 11(1):1–10. <https://doi.org/10.1057/s41599-024-03363-0>.
- Jiang, Zhaokun, Qianxi Lv, Ziyin Zhang, and Lei Lei. 2024. Convergences and divergences between automatic assessment and human evaluation: Insights from comparing ChatGPT-generated translation and neural machine translation. *arXiv preprint arXiv:2401.05176*.
- Wang, Lulu, Sanjun Sun, Xing Wang, Jinghang Gu, and Kanglong Liu. 2026. Workflow matters: Comparing human translators and multi-agent LLMs in literary translation. *Target*. <https://doi.org/10.1075/target.25081.wan>.
- Wu, Minghao, Jiahao Xu, and Longyue Wang. 2024. TransAgents: Build your translation company with language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 131–141. <https://doi.org/10.18653/v1/2024.emnlp-demo.14>.
- Wu, Minghao, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. (Perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Transactions of the Association for Computational Linguistics*, 13:901–922. <https://doi.org/10.1162/tacl.a.25>.
- Chen, Pinzhen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 181–190. <https://aclanthology.org/2024.eamt-1.17>.
- Madaan, Aman, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.
- Xu, Wenda, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492. <https://doi.org/10.18653/v1/2024.acl-long.826>.
- Leech, Geoffrey N. and Mick Short. 2007. *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. Pearson Education. <https://doi.org/10.2307/1772012>.
- Munat, Judith. 2007. Lexical creativity as a marker of style in science fiction and children’s literature. In *Lexical Creativity, Texts and Contexts*, pages 163–185. John Benjamins, Amsterdam & Philadelphia, PA. <https://doi.org/10.1075/sfsl.58.20mun>.
- Wiher, Gian, Clara Meister, and Ryan Cotterell. 2022. On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics*, 10:997–1012. https://doi.org/10.1162/tacl_a_00502.
- Zhou, Yuxuan, Margret Keuper, and Mario Fritz. 2025. Balancing diversity and risk in LLM sampling: How to select your method and parameter for open-ended text generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26352–26365. <https://doi.org/10.18653/v1/2025.acl-long.1278>.
- Arias, Esteban Garces, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. 2025. Decoding decoded: Understanding hyperparameter effects in open-ended text generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9992–10020. <https://aclanthology.org/2025.coling-main.668>.
- Troshin, Sergey, Wafaa Mohammed, Yan Meng, Christof Monz, Antske Fokkens, and Vlad Niculae. 2025. Control the Temperature: Selective Sampling for Diverse and High-Quality LLM Outputs. *arXiv:2510.01218*.
- Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *ICLR 2020*.
- Ippolito, Daphne, Reno Kriz, Joao Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762. <https://doi.org/10.18653/v1/P19-1365>.
- Peng, Keqin, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633. <https://doi.org/10.18653/v1/2023.findings-emnlp.373>.
- Arora, Chetan, Ahnaf Ibn Sayeed, Sherlock Licorish, Fanyu Wang, and Christoph Treude. 2024. Optimizing LLMs for code generation: Which hyperparameter settings yield the best results? In *2024 31st Asia-Pacific Software Engineering Conference (APSEC)*, pages 281–290. IEEE. <https://doi.org/10.1109/APSEC65559.2024.00039>.

- Donato, Benedetta, Leonardo Mariani, Daniela Micucci, and Oliviero Riganelli. 2025. Studying how configurations impact code generation in LLMs: The case of ChatGPT. *arXiv:2502.17450*.
- Peeperkorn, Max, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? *arXiv:2405.00492*.
- Grullon-Polanco, Larry. 2026. LLM-assisted genre analysis: The effect of sampling temperature on reliability. *Research Methods in Applied Linguistics*, 5(2):100314. <https://doi.org/10.1016/j.rmal.2026.100314>.
- Li, Lujun, Lama Sleem, Niccolo Gentile, Geoffrey Nichil, and Radu State. 2025. Exploring the impact of temperature on large language models: Hot or cold? *Procedia Computer Science*, 264:242–251. <https://doi.org/10.1016/j.procs.2025.07.135>.
- Jakobson, Roman. 1959. On linguistic aspects of translation. In Reuben Arthur Brower, editor, *On Translation*, pages 232–239. Harvard University Press, Cambridge, MA. <https://doi.org/10.4159/harvard.9780674731615.c18>.
- Macken, Lieve, Bram Vanroy, Luca Desmet, and Arda Tezcan. 2022. Literary translation as a three-stage process: Machine translation, post-editing and revision. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 101–110. <https://aclanthology.org/2022.eamt-1.13>.
- Whyatt, Boguslawa. 2017. Intralingual translation. In John W. Schwieter and Li Wei, editors, *The Handbook of Translation and Cognition*, pages 176–192. John Wiley & Sons, Hoboken, NJ. <https://doi.org/10.1002/9781119241485.ch10>.
- International Organization for Standardization. 2017. *ISO 18587:2017: Translation services — Post-editing of machine translation output — Requirements*. International Organization for Standardization, Geneva, Switzerland.
- Cai, Chengkun, Xu Zhao, Yucheng Du, Haoliang Liu, and Lei Li. 2024. T² of Thoughts: Temperature Tree Elicits Reasoning in Large Language Models. *arXiv:2405.14075*.
- Zhang, Ran, Wei Zhao, and Steffen Eger. 2025. How Good Are LLMs for Literary Translation, Really? Literary Translation Evaluation with Humans and LLMs. 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 10961–10988. <https://doi.org/10.18653/v1/2025.naacl-long.548>.
- Alrabaei, Saed, Paria Shirani, Lingyu Wang, Mourad Debbabi, and Aiman Hanna. 2019. Decoupling coding habits from functionality for effective binary authorship attribution. *Journal of Computer Security*, 27(6):613–648. <https://doi.org/10.3233/JCS-191292>.
- Covington, Michael A. and Joe D. McFall. 2010. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100. <https://doi.org/10.1080/09296171003643098>.
- Malvern, David D. and Brian J. Richards. 1997. A new measure of lexical diversity. *British Studies in Applied Linguistics*, 12:58–71.
- McCarthy, Philip M. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.
- Mohseni, Mahdi, Christoph Redies, and Volker Gast. 2022. Approximate entropy in canonical and non-canonical fiction. *Entropy*, 24(2):278. <https://doi.org/10.3390/e24020278>.
- Laufer, Batia and Paul Nation. 1995. Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3):307–322. <https://doi.org/10.1093/applin/16.3.307>.
- Van Parys, Amaury, Vanessa De Wilde, Lieve Macken, and Maribel Montero Perez. 2025. LexPro: A plurilingual lexical profiling tool to assist teachers and researchers in analysing vocabulary of L2 input. *Language Teaching Research*, 13621688251352259. <https://doi.org/10.1177/13621688251352259>.
- Vanmassenhove, Eva. 2025. Losing our tail—again: On (un)natural selection and multilingual large language models. *arXiv:2507.03933*.
- Vanmassenhove, Eva, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213. <https://doi.org/10.18653/v1/2021.eacl-main.188>.
- Anson, Ian G., Cary Moskovitz, and Chris M. Anson. 2019. A text-analytic method for identifying text recycling in STEM research reports. *The Journal of Writing Analytics*, 3. <https://doi.org/10.37514/JWA-J.2019.3.1.07>.
- Shypula, Alexander, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. 2025. Evaluating the diversity and quality of LLM generated content. *arXiv:2504.12522*.

- McCarthy, Philip M. and Scott Jarvis. 2010. MTL, D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.
<https://doi.org/10.3758/BRM.42.2.381>.
- Vermeer, Anne. 2000. Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, 17(1):65–83.
<https://doi.org/10.1177/026553220001700103>.
- Koizumi, Rie. 2012. Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction*, 1(1):60–69.
<https://doi.org/10.7820/vli.v01.1.koizumi>.
- Shannon, Claude E. 1951. Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1):50–64. <https://doi.org/10.1002/j.1538-7305.1951.tb01366.x>.
- Nation, I.S.P. 2017. The BNC/COCA Level 6 word family lists (Version 1.0.0) [Data file]. <http://www.victoria.ac.nz/lals/staff/paul-nation.aspx>.
- Uhlmann, Gyburg. 2025. The even sheen of AI: Kitsch, LLMs, and homogeneity. arXiv:2509.16794.

Dual-Metric Compliance and Quality Evaluation of Knowledge Graph Mediated Translation in Regulated Domains: An Enhanced Architectural Framework

Viveta Gene
Ionian University
viveta.gene@gmail.com

Vilemini Sosoni
Ionian University
sosoni@ionio.gr

This paper presents the second phase of an empirical research programme evaluating knowledge graph-mediated translation (KGMT) in the regulated dermocosmetics domain. Building on Gene and Sosoni (2026), which introduced KGMT as a theoretical framework for compliance-aware multilingual communication, Phase 2 advances the research through three coordinated contributions: an indexed source design with 30 controlled compliance errors drawn from EU and US regulatory frameworks; a non-compliance in source (NCS) classification that principally distinguishes source-propagated from translation-introduced errors; and an enhanced KGMT pipeline incorporating a domain-specific style guide, a lemmatization layer, and an expanded multilingual glossary. Translations of a 500-word retinol product description into Greek and French are evaluated across three conditions—pure baseline (PB), document-augmented (DA), and KGMT—using the Post-Editing Monitoring, Assessment, Problem-Solving (PE.MAPS) dual-metric framework (Gene, 2024). KGMT achieves 100% source compliance detection recall against 0% for both baselines, and records zero penalized translation errors in both target languages under the NCS classification. Inter-condition divergence in baseline outputs from the same prompt on different inference instances demonstrates that large language models (LLMs) do not reliably or auditably guarantee compliance consistency, positioning KGMT as a deterministic, knowledge-grounded stabilization layer.

1 Introduction

Translation in regulated industries operates under a dual constraint: outputs must be linguistically accurate and simultaneously compliant with the

regulatory frameworks of the target market. When source content is authored under the United States FDA guidelines for cosmetics and must be adapted for the EU—subject to Regulation 1223/2009 on cosmetic products and Regulation 655/2013 on common criteria for claims—the challenges extend beyond language. Claims permissible under one regime may constitute prohibited drug-territory language under the other (Prieto Ramos, 2014), and ingredient concentrations legal in the US market may exceed European limits set by Commission Regulation 2024/996.

LLMs, when deployed without external knowledge structures, do not reliably or auditably provide compliance guarantees in regulated translation contexts (Pan et al., 2024). While recent architectures incorporating tool access, retrieval-augmented generation (Lewis et al., 2020), or constrained decoding can partially support terminology control, they do not provide the explicit, traceable, regulation-anchored knowledge structures required where audit trails and provenance are legal requirements. The distinctive contribution of KGMT is precisely this: structured, explicit, jurisdiction-aware knowledge encoding producing deterministic, auditable outputs grounded in identified regulatory instruments (Hogan et al., 2021).

The intellectual lineage of this approach connects to foundational work at the intersection of terminology and knowledge engineering, notably Skuce and Meyer's knowledge-based term bank COGNITERM (Skuce and Meyer, 1990), which demonstrated that structured terminological knowledge could support domain-specific language processing. KGMT extends this tradition into the era of neural generation, adding compliance validation and cross-jurisdictional awareness as first-class concerns. Gene and Sosoni (2026) established the theoretical framework and a Phase 1 pilot, which revealed a detection-adoption gap: compliance flags

generated by the knowledge graph (KG) layer were not consistently integrated into the translation output. Phase 2, reported here, addresses this gap through an indexed source design, an NCS classification, and an enhanced KGMT pipeline.

Section 2 reviews the Phase 1 findings and defines compliance-aware translation. Section 3 describes the methodology, including the indexed source design, NCS classification, pipeline enhancements, and annotation protocol. Section 4 reports the empirical results. Section 5 discusses the NCS framework as a reusable contribution and the infrastructure compounding effect. Section 6 addresses limitations, and Section 7 concludes.

2 From Framework to Empirical Validation

A knowledge graph is a graph-structured data resource in which nodes represent real-world entities and edges encode the relationships between them, with the aim of accumulating and conveying domain knowledge (Hogan et al., 2021; see Figure C1, Appendix C). KGMT operationalises this structure for translation by injecting KG-derived terminological mappings, compliance constraints, and disambiguation rules into the LLM prompt at inference time. The Phase 1 theoretical framework and pilot evaluation are reported in Gene and Sosoni (2026); Section 2.1 summarises the key findings relevant to the present study.

2.1 Phase 1 and the Detection-Adoption Gap

Gene and Sosoni (2026) introduced KGMT as a grounded prompting architecture in which KG-derived terminological guidance, compliance constraints, and disambiguation rules are injected into the LLM prompt at inference time, following retrieval-augmented generation paradigms (Lewis et al., 2020; Pan et al., 2024). The Phase 1 evaluation translated a 500-word retinol product description into Greek and French, comparing KGMT against PB and DA conditions using GPT-5 (OpenAI, 2025). KGMT achieved a source compliance detection recall of 24% and a translation quality score (TQS) of 83.9% (Greek) and 88.2% (French), below both baselines. The detection-adoption gap was identified as the primary research priority: the KG was functioning as a source-side compliance checker but its outputs were not consistently adopted by the generation layer.

2.2 Defining Compliance-Aware Translation

A gap in Phase 1 was the absence of an explicit task definition. Compliance-aware translation, as operationalized here, refers to target-language content that is simultaneously (a) semantically faithful to the source, (b) terminologically consistent with domain glossaries (Bowker, 2002; Moorkens et al., 2018), and (c) compliant with the regulatory frameworks of the target jurisdiction. These three dimensions correspond respectively to accuracy, fluency, and compliance in the PE.MAPS framework (Gene, 2024). Compliance-aware translation must also perform source-side auditing: identifying and flagging violations before they propagate into the target. This is distinct from standard translation quality assessment and requires the NCS classification introduced in Section 3.2.

3 Methodology

3.1 Source Text, Indexed Error Design, and Classification Criteria

The source text is a 500-word retinol night cream product description representing a US market product subject to FDA cosmetics guidelines, with 30 compliance errors embedded and indexed [1]–[30] using bracketed notation, plus one unindexed C8 warning-omission error (31 evaluable violations total). Each error was classified against explicit EU regulatory criteria before indexing: C1 violations (drug claim, $n=10$) were assessed against Regulation 655/2013 Article 4 and Annex I criteria 2 and 6; C3 violations (unsubstantiated superlative, $n=19$) against Annex I criterion 4 requiring substantiation for comparative or absolute claims; C2 (prohibited concentration, $n=1$) against Regulation 2024/996 Annex III (retinol RE limit 0.3% for face leave-on products); C5 mismatches ($n=2$) against study data in the source; and C9 missing evidence ($n=3$) against Annex I criterion 3 requiring substantiation for efficacy claims. Classifications were established prior to annotation and provided to both annotators as decision criteria, ensuring the ground truth is theoretically grounded rather than post-hoc. Annotation was performed by two evaluators with professional regulatory translation expertise; full details of the cross-annotation protocol and inter-annotator agreement are reported in Section 3.5. The indexing design provides an objective

detection recall ground truth independent of annotator judgement.

The human reference translation was produced by the first author, a professional translator with regulatory specialization, working from the source text and the full KG knowledge base (glossary, compliance rules, style guide) but without access to any system output. This mirrors the conditions of a KG-informed human translation and provides a ceiling benchmark for the automated conditions. Producing the reference prior to and independently of the system outputs prevents anchoring bias in the annotation.

3.2 NCS Classification

The Non-Compliance in Source (NCS) classification addresses a methodological problem identified in Phase 1 and in prior quality assessment frameworks including multidimensional quality metrics (MQM) (Lommel et al., 2014) and TAUS DQF: the conflation of source-originated with translation-introduced compliance violations. This conflation penalizes translation systems for failures that belong to the content creation process, not the translation process. Table 1 defines the three NCS states and their penalty logic. Under NCS, KGMT is penalized only for TRANS_INTRODUCED errors; SRC_FLAGGED instances are recorded for detection recall analysis but do not reduce TQS.

NCS State	Definition	Penalty	Example (this study)
TRANS_INTRODUCED	Source compliant; translation introduces the violation	Penalized	PB/DA: 'stimulates collagen' not mitigated
SRC_REPRODUCED	Source non-compliant; translation reproduces without flagging	Penalized (QA failure)	PB/DA: 1.0% retinol reproduced, no flag raised
SRC_FLAGGED	Source non-compliant; translation flags and mitigates	Not penalized	KGMT: 1.0% retinol → NCS-C2 flag + compliant alternative

Table 1: NCS classification states, penalty logic, and corpus examples.

3.3 KGMT Pipeline Enhancement

Figure 1 illustrates the KGMT four-stage pipeline. Phase 2 injects three enhancements as a package into the KG retrieval and prompt-construction layer: (1) a comprehensive style guide aligned with EU Interinstitutional Style Guide conventions for Greek and French, covering register disambiguation (marketing versus clinical), typography, and punctuation; (2) a morphological post-processing lemmatization layer addressing gender and number agreement in Greek and French; and (3) an expanded multilingual glossary with explicit CORRECT/INCORRECT pairs encoding register-specific terminology, superlative mitigation patterns, drug claim alternatives, and qualifier preservation rules, each linked to a PE.MAPS error code. All three enhancements were injected as a simultaneous package; component-level attribution is acknowledged as a limitation and is deferred to an ablation study. For a full illustration of the underlying KG architecture, see Figure C1 in Appendix C.

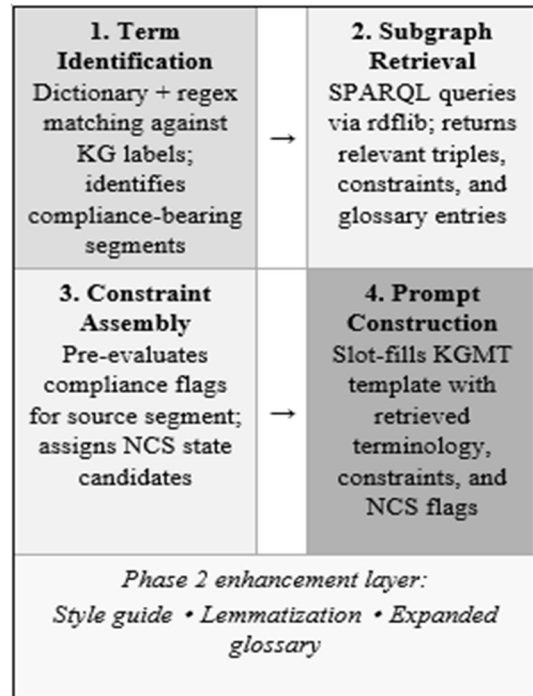


Figure 1: NCS classification states, penalty logic, and corpus examples.

3.4 Experimental Conditions and Prompt Design

Three conditions were evaluated. The Pure Baseline (PB) uses a generic translation instruction with no domain resources: ‘You are a professional translator specializing in cosmetics. Translate the following from English to [language]. Maintain accuracy and appropriate register.’ The Document-Augmented Baseline (DA) appends the raw glossary and compliance reference tables to the same instruction without structured retrieval. KGMT uses a structured prompt template populated by the four-stage pipeline (Figure 1), providing retrieved terminology mappings, NCS-flagged compliance constraints, and style guide rules specific to the source segment. All conditions use GPT-5 (OpenAI, 2025) at temperature 0.3, top-p 0.95. Each of the 58 segments was processed once per condition (n=1 per segment per condition; 174 output instances per language). Output stochasticity is treated as an empirical variable rather than a confound, as its measurement is a research objective.

3.5 Annotation Protocol and Evaluator Bias Control

Translation quality is evaluated using PE.MAPS (Gene, 2024), distinguishing Technical Errors (weighted by severity: Critical = 3, Major = 2, Minor = 1) from Preferential Modifications. $TQS = 100 - (\text{Total Technical Penalty Points} / \text{Word Count} \times 100)$. The full PE.MAPS error taxonomy with penalty weights is provided in Table B1 (Appendix B). The two annotators applied a cross-annotation protocol (Saldanha and O’Brien, 2013; Shadish et al., 2002): Annotator A annotated PB and DA primarily, with secondary review of KGMT; Annotator B annotated KGMT primarily, with secondary review of PB and DA. Neither annotator had access to the other’s annotations during primary annotation. Disagreements were resolved through structured consensus using pre-established classification criteria. Inter-annotator agreement exceeded $\kappa = 0.85$ (Cohen, 1960), indicating strong reliability. The first author designed the corpus and KG but did not perform primary annotation of KGMT outputs, mitigating the principal evaluator bias risk.

4 Results

4.1 Source Compliance Detection Recall

PB and DA achieved 0% recall across all 30 indexed errors in both languages: without KG infrastructure, both conditions reproduced all source violations without flagging or mitigation. KGMT achieved 100% recall across all 31 evaluable violations. Table 2 presents a representative subset of the 30 indexed source errors. Representative correction strategies included: superlative mitigation (‘REVOLUTIONARY’ → ‘ADVANCED’ / ‘NOVATRICE’), drug claim softening (‘stimulates collagen at the cellular level’ → ‘supports natural collagen synthesis’ / ‘aide à soutenir la synthèse naturelle de collagène’), qualifier preservation (‘reduces wrinkles’ → ‘reduces the appearance of wrinkles’), and modal preservation (‘transforms’ → ‘visibly improves’). The NCS-C2 flag at segment 4 (1.0% retinol exceeding the EU 0.3% RE limit) was reproduced with an explicit NCS flag, correctly attributed to source under the SRC_FLAGGED state.

Seg	Code	Wt.	PB	DA	KGMT Action	NCS Status
2	C3	3	REPRO-DUCED	REPRO-DUCED	MITIGATED → ADVANCED	Flagged; KGMT corrected
3	C1	3	REPRO-DUCED	REPRO-DUCED	MITIGATED → visible results	Flagged; KGMT corrected
4	C2	3	REPRO-DUCED	REPRO-DUCED	REPRODUCED + [NCS flag]	NCS-C2 (source error)
18	C1	3	REPRO-DUCED	REPRO-DUCED	MITIGATED → supports production	Flagged; KGMT corrected
19	C3	3	REPRO-DUCED	REPRO-DUCED	HEDGED → up to 3x	Flagged; KGMT corrected
22	C1	3	REPRO-DUCED	REPRO-DUCED	MITIGATED → natural defense	Flagged; KGMT corrected
36	C1	3	REPRO-DUCED	REPRO-DUCED	MITIGATED → reduces appearance	Flagged; KGMT corrected
37	C1+C3	3	REPRO-DUCED	REPRO-DUCED	RESTRUCTURE D heading	Flagged; KGMT corrected
38	C3	3	REPRO-DUCED	REPRO-DUCED	SOFTENED → documented system	Flagged; KGMT corrected
43	C3	3	REPRO-DUCED	REPRO-DUCED	MITIGATED → dermatologists	Flagged; KGMT corrected
47	C1	3	REPRO-DUCED	REPRO-DUCED	MITIGATED → improve visibly	Flagged; KGMT corrected
48	C3	3	REPRO-DUCED	REPRO-DUCED	MITIGATED → documented results	Flagged; KGMT corrected
Total: 30 indexed source errors (+1 C8). PB/DA detection rate: 0%. KGMT detection rate: 100%. NCS-C2 (Seg 4) not penalized/ source-originated concentration error reproduced with explicit NCS flag.						

Table 2: Source compliance detection by condition representative subset, n=12

4.2 Translation Quality Across Conditions

Table 3 presents full quality metrics. KGMT records TQS = 100.0% in both Greek and French under NCS classification, reflecting zero penalized technical errors. PB and DA accumulate substantial penalties from source error propagation (30 errors reproduced without flagging, classified TRANS_INTRODUCED) combined with translation-introduced accuracy (A1/A2) and fluency (F1) errors. Under NCS, the same source errors in KGMT are SRC_FLAGGED and therefore unpenalized in TQS.

Metric	PB (EL)	PB (FR)	DA (EL)	DA (FR)	KGMT (EL)	KGMT (FR)
Detection recall	0%	0%	0%	0%	100%	100%
Source C errors	30	30	30	30	0 (NCS)	0 (NCS)
Accuracy errors (A)	28	30	24	25	0	0
Fluency errors (F)	1	1	6	7	0	0
Total penalty pts	182	188	175	181	0	0
TQS (%)	71.6	73.1	72.7	74.4	100.0	100.0

Table 3: Translation quality metrics by condition and language.

4.3 Phase 1 versus Phase 2 Comparison

Table 4 presents the cross-phase comparison. The TQS reduction for PB and DA between phases reflects two factors: the stricter Phase 2 annotation protocol (A1/A2/F1 errors now systematically tracked) and natural LLM output variability across inference instances from the same prompt. This inter-instance divergence is an empirical finding: LLMs do not reliably reproduce identical terminological choices on repeated runs, which is operationally consequential in regulated workflows where consistency is a legal requirement (Castilho et al., 2017). KGMT’s improvement from 83.9% to 100.0% (Greek) reflects three compounding effects: the infrastructural package, system maturation (detection recall: 24%→100%), and NCS methodological refinement.

Metric	PB Ph.1	PB Ph.2	DA Ph.1	DA Ph.2	KGM T Ph.1	KGMT Ph.2	ΔKGMT
Detection recall (comb.)	0%	0%	0%	0%	24%	100%	+76pp
Total penalty pts (EL)	83	182	78	175	102	0	-102
Total penalty pts (FR)	78	188	73	181	83	0	-83
TQS (EL, %)	86.8	71.6	87.6	72.7	83.9	100.0	+16.1
TQS (FR, %)	88.5	73.1	89.4	74.4	88.2	100.0	+11.8

Table 4: Phase 1 vs Phase 2 cross-condition comparison (combined EL + FR).

5 Discussion

5.1 NCS as a Reusable Methodological Contribution

The NCS framework addresses a systematic measurement gap identified across both phases. By separating SRC_FLAGGED from TRANS_INTRODUCED errors, NCS enables principled evaluation of two distinct competences: translation competence (does the system produce accurate, fluent, compliant output?) and quality assurance competence (does the system identify and flag source-level violations?). Conflating these in a single TQS metric, as Phase 1 did, misrepresented KGMT’s capability profile. NCS is domain-agnostic and transferable to any regulated domain exhibiting cross-jurisdictional regulatory asymmetry, financial services translation, where content compliant under CFTC guidelines for US markets may violate MiFID II obligations in the EU, or where product classifications differ materially across jurisdictions, presents a structurally identical regulatory asymmetry, and the framework accommodates such domains without architectural modification.

5.2 Infrastructure Compounding and the Structure Before Scale Principle

The Phase 1–Phase 2 divergence for baseline outputs on identical prompts demonstrates that LLMs do not reliably produce consistent terminological choices across inference instances in compliance-sensitive contexts. KGMT, grounded in an explicit deterministic KG (Hogan et al., 2021), produces the same terminological guidance for the same source segment regardless of inference instance. The three Phase 2

enhancements compound: the glossary expansion eliminated A1 terminological substitution errors; the style guide resolved register inconsistencies; the lemmatization layer corrected F1 morphosyntactic errors. This supports a ‘structure before scale’ principle: targeted KG infrastructure additions yield disproportionate quality gains in regulated contexts, because each knowledge layer eliminates a class of errors rather than merely reducing their frequency (Gene and Sosoni, 2026).

5.3 Scope and Generalizability

The study is grounded in a single domain and a single 58-segment corpus. The dual-domain architecture of the broader research programme positions KGMT as a domain-agnostic framework: domains exhibiting structurally identical cross-jurisdictional regulatory asymmetries (financial services translation, where CFTC-compliant content may violate MiFID II, is one such domain) are accommodated by the framework. A full-interface KGMT editor implementing the NCS flagging system in real time has been developed and is the subject of a forthcoming system demonstration paper.

6 Limitations

Five limitations bound the current findings. First, the single-domain, single-corpus design limits immediate generalizability; cross-domain replication is required. Second, the two Phase 2 additions, e.g. the style guide with integrated morphological enforcement, and the expanded compliance-aware glossary, were deployed simultaneously; the independent contribution of each cannot be determined from the current data. The modular architecture makes component-level evaluation feasible as an open direction. Third, TQS = 100.0% rests on cross-annotator confirmation that self-flagged A2 variants are morphological artefacts of the lemmatization layer; independent replication is required. Fourth, outputs were generated once per segment per condition (n=1); replication across multiple runs would characterize LLM variability more precisely. Fifth, the study uses GPT-5 as the sole LLM backend; comparative evaluation across model families remains open. The first author designed the evaluation corpus and KG, acknowledged as a potential bias source mitigated but not eliminated by the cross-annotation protocol.

7 Conclusions

This paper presents three contributions to regulated translation research. The NCS classification provides a principled, reusable framework for distinguishing source-propagated from translation-introduced compliance errors. The indexed source design provides an objective detection recall ground truth grounded in explicit regulatory criteria. The empirical evaluation demonstrates that an enhanced KGMT architecture achieves 100% source compliance detection recall and zero penalised translation errors in both Greek and French, against 0% detection and substantially lower TQS for both unaugmented conditions.

The inter-instance divergence of baseline outputs from identical prompts demonstrates that LLMs do not reliably or auditably guarantee compliance consistency—a finding with direct operational consequences for any domain where terminology consistency and regulatory traceability are legal requirements. KGMT, as a deterministic knowledge-grounded layer, stabilizes this variability. The framework is designed for domains exhibiting cross-jurisdictional regulatory asymmetry; financial services translation presents a structurally identical challenge and represents a natural domain for future evaluation. Component-level evaluation of the Phase 2 style guide and expanded glossary is feasible within the current modular architecture and represents an open direction for this or future research. Evaluation across multiple LLM backends and development of source-anchored fallback detection for semantic paraphrases not captured by the current string-matching pipeline are further open directions.

Acknowledgements

The authors gratefully acknowledge the technical support of Edwin Trebels (LangOptima) and Prasad Yalamanchi (Lead Semantics) for Knowledge Graph construction and engineering, and the second annotator, an independent professional translator specializing in regulatory and dermatocosmetics translation, whose cross-annotation contribution strengthened the inter-annotator reliability of this study. This research is conducted under the postdoctoral supervision of Dr. Vilemini Sosoni (Ionian University).

References

- Bowker, Lynne. 2002. *Computer-Aided Translation Technology: A Practical Introduction*. University of Ottawa Press.
- Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? *Prague Bulletin of Mathematical Linguistics*, 108:109–120.
- Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.
- Gene, Viveta. 2024. [Machine translation post-editing and professional translators: The contribution of training to the reduction of effort and the improvement of quality](https://doi.org/10.12681/eadd/58063) [Doctoral dissertation, Ionian University]. <https://doi.org/10.12681/eadd/58063>.
- Gene, Viveta and Vilelmini Sosoni. 2026. Knowledge graph-mediated translation (KGMT): A theoretical framework for compliance-aware multilingual communication in regulated domains. In *Proceedings of Convergence 2026* (Surrey, 17–19 June 2026). In press. Available at: <https://www.surrey.ac.uk/centre-translation-studies/convergence-2026/programme>
- Hogan, Aidan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. [Knowledge graphs](https://doi.org/10.1145/3447772). *ACM Computing Surveys*, 54(4), Article 71, 1–37. <https://doi.org/10.1145/3447772>.
- Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 33:9459–9474.
- Lommel, Arle, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumática*, 12:455–463.
- Moorkens, Joss, Sheila Castilho, Federico Gaspari, and Stephen Doherty (Eds.). 2018. *Translation Quality Assessment: From Principles to Practice*. Springer.
- OpenAI. 2025. *GPT-5 technical report*. OpenAI. <https://openai.com>
- Pan, Shirui, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. [Unifying large language models and knowledge graphs: A roadmap](https://doi.org/10.1109/TKDE.2024.3352100). *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599. <https://doi.org/10.1109/TKDE.2024.3352100>.
- Prieto Ramos, Fernando. 2014. Quality assurance in legal translation: Evaluating process, competence and product in the pursuit of adequacy. *International Journal for the Semiotics of Law*, 28:11–30.
- Saldanha, Gabriela and Sharon O'Brien. 2013. *Research Methodologies in Translation Studies*. St. Jerome Publishing.
- Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin.
- Skuce, Doug and Ingrid Meyer. 1990. [Concept analysis and terminology: A knowledge-based approach to documentation](https://aclanthology.org/C90-1011). In *Proceedings of the 13th International Conference on Computational Linguistics (COLING 1990)*, Volume 1, pages 197–201. <https://aclanthology.org/C90-1011>.

Appendix A. PE. MAPS Evaluation Framework

This appendix presents the PE.MAPS (Post-Editing Monitoring, Assessment, Problem-Solving; Gene, 2024) evaluation framework applied throughout Phase 2 of this study. Table A1 lists the error codes, dimensions, definitions, and penalty weights used in annotation. Table A2 summarizes the penalty weight system. The TQS formula is provided below. For the full theoretical framework, see Gene and Sosoni (2026).

Code	Category	Dimension	Definition	Weight
C1	Drug/Medical Claims	Compliance	Therapeutic claims beyond cosmetic scope (e.g., 'cures', 'treats', 'repairs')	3
C2	Prohibited Ingredient/Concentration	Compliance	Prohibited substances or concentrations exceeding regulatory limits	3
C3	Unsubstantiated Superlative	Compliance	Absolute or comparative claims without substantiation (e.g., 'best', 'only', 'maximum')	3
C5	Quantified Claim Mismatch	Compliance	Numbers or timeframes inconsistent with source data	3
C8	Warning Omission	Compliance	Required safety advisory absent (e.g., pregnancy warning for retinol)	2
C9	Missing Evidence	Compliance	Claims without required scientific or clinical substantiation	3
A1	Term Substitution	Accuracy	Incorrect domain terminology (maintains source error or uses wrong term)	3
A2	Terminological Inconsistency	Accuracy	Same concept translated with different terms across the text	2
F1	Grammar Error	Fluency	Morphosyntactic errors: gender, number, case, verb form agreement	1
F2	Style/Register	Fluency	Inappropriate register for domain (marketing vs. clinical)	1
NCS	Non-Compliance in Source	Classification	Source error flagged by KGMT but NOT penalised in TQS; recorded for detection recall only	0
—	No Error (Preferential Edit)	—	MT output acceptable; translator modification is preferential, not corrective	0

Table A1. PE. MAPS Error Code Taxonomy — Codes, Definitions, and Penalty Weights

Colour key: Red = Compliance dimension; Amber = Accuracy dimension; Green = Fluency dimension; Blue = NCS clasasification (not penalised).

Severity	Weight	Codes	Rationale
Critical (Compliance)	3	C1, C3, C5, C9	Regulatory violations; legal or health risk
Major (Warning)	2	C8	Mandatory safety warnings (2 pt as absence, not direct harm)
Major (Accuracy)	3	A1	Terminology errors impacting meaning or compliance
Minor (Accuracy)	2	A2	Inconsistencies affecting usability or cohesion
Minor (Fluency)	1	F1	Grammar errors affecting readability, not meaning
Info (not penalised)	0	NCS, F2	Source-originated or preferential edits (not scored)

Table A2. PE. MAPS Penalty Weight System

TQS Formula

$$\text{TQS} = 100 \times (1 - \Sigma[\text{errors} \times \text{weights}] / \text{word count})$$

Key principle: NCS (Non-Compliance in Source) errors are identified by KGMT but not penalised in TQS. Only translation-introduced errors reduce TQS. This design ensures that translators are not penalised for source-level compliance issues outside their control.

Appendix B. PE.MAPS Full Error Code Taxonomy

PE.MAPS (Post-Editing Monitoring, Assessment, Problem-Solving; Gene, 2024) evaluates translation quality across Compliance (C1–C9), Accuracy (A1–A7), and Fluency (F1–F4) dimensions. NCS (Non-Compliance in Source) identifies source-level violations flagged by KGMT that are excluded from TQS penalisation. TQS = $100 \times (1 - \Sigma[\text{errors} \times \text{weights}] / \text{word count})$. Codes not triggered in the Phase 2 corpus are included for completeness.

Code	Category	Dimension	Definition	Weight
C1	Drug/Medical Claims	Compliance	Therapeutic claims beyond cosmetic scope (e.g., 'cures', 'treats', 'repairs')	3
C2	Prohibited Ingredient/Conc.	Compliance	Prohibited substances or concentrations exceeding regulatory limits	3
C3	Unsubstantiated Superlative	Compliance	Absolute or comparative claims without substantiation (e.g., 'best', 'only', 'maximum')	3
C4	INCI Name Violation	Compliance	Incorrect or non-standard translation of INCI nomenclature	3
C5	Quantified Claim Mismatch	Compliance	Numbers or timeframes inconsistent with source data	3
C6	Qualifier Loss	Compliance	Removal of hedging language (e.g., 'may', 'appears to', 'visible results')	2
C7	Conditional→Definitive Shift	Compliance	Modal or conditional rendered as definitive (e.g., 'can reduce' → 'reduces')	2
C8	Warning Omission	Compliance	Required safety advisory absent (e.g., pregnancy warning for retinol)	2
C9	Missing Evidence	Compliance	Claims without required scientific or clinical substantiation	3
A1	Term Substitution	Accuracy	Incorrect domain terminology: wrong target term used for source concept	3
A2	Terminological Inconsistency	Accuracy	Same concept translated with different terms across the text	2
A3	Spelling/Morphological Error	Accuracy	Target term misspelled or morphologically incorrect	1
A4	Omission	Accuracy	Source content absent from target without justification	2
A5	Addition	Accuracy	Content added in target not present in source	2
A6	Mistranslation	Accuracy	Semantic shift: target meaning diverges from source meaning	2
A7	Number/Unit Error	Accuracy	Numbers, percentages, or measurement units rendered incorrectly	3
F1	Grammar Error	Fluency	Morphosyntactic errors: gender, number, case, or verb form agreement	1
F2	Style/Register	Fluency	Inappropriate register for domain context (marketing vs. clinical)	1
F3	Punctuation/Mechanics	Fluency	Punctuation, capitalisation, or spacing errors	1
F4	Coherence	Fluency	Logical flow problems affecting readability	1
NCS	Non-Compliance in Source	Classification	Source error flagged by KGMT; NOT penalised in TQS; recorded for detection recall only	0
—	No Error (Preferential Edit)	—	MT output acceptable; translator modification is preferential, not corrective	0

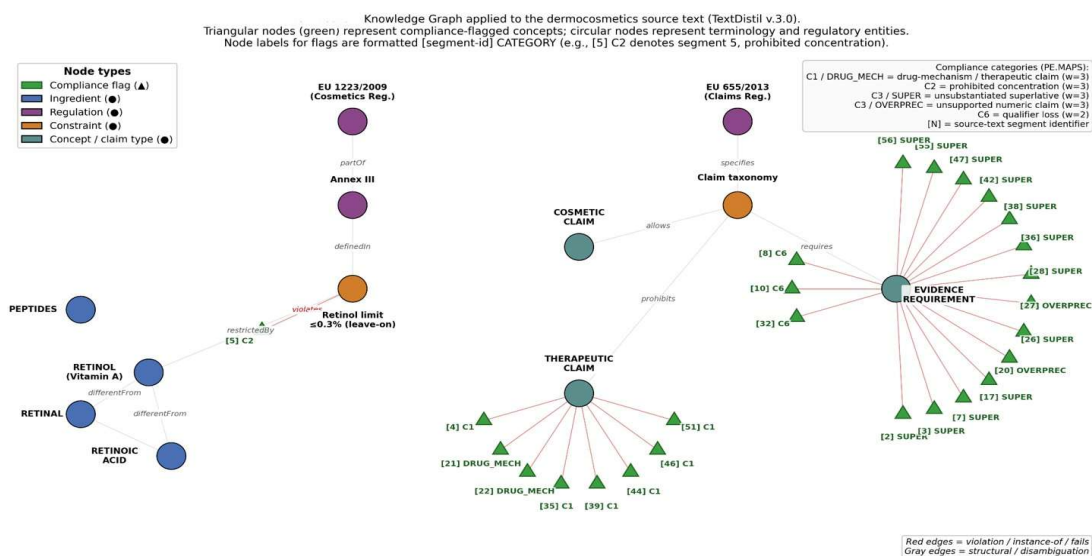
Table B1. PE.MAPS Complete Error Code Taxonomy

Colour key: Red = Compliance (C1-C9); Amber = Accuracy (A1-A7); Green = Fluency (F1-F4); Blue = NCS (not penalised in TQS).

Appendix C. KGMT Knowledge Graph: Dermocosmetics Source Text

Figure C1 presents the knowledge graph applied to the dermocosmetics source text in Phase 1 (Gene and Sosoni, 2026), implemented in TextDistil v.3.0 (Lead Semantics). The graph integrates four node classes: ingredient entities (blue), regulatory references (purple), regulatory constraints (orange), and concept/claim-type nodes (teal). Compliance flags generated by KGMT are rendered as green triangular nodes labelled [N] CATEGORY, where N is the source-text segment identifier and CATEGORY denotes the PE.MAPS compliance code (C1, C2, C3, C6) or KG-internal classifier (DRUG_MECH, SUPER, OVERPREC) that triggered the flag. Red edges denote violation, instance-of, and constraint-failure relations; grey edges denote structural and disambiguation relations. The graph encodes 58+ terminology entries across ingredient classifications, regulatory claim categories, concentration limits, and cross-jurisdictional EU/US distinctions. This architecture underpins the Phase 2 enhanced pipeline described in Section 3.3.

Figure C1. Knowledge Graph applied to the dermocosmetics source text (TextDistil v.3.0, Lead Semantics). Triangular nodes (green) represent compliance-flagged concepts; circular nodes represent terminology and regulatory entities. Node labels for flags are formatted [segment-id] CATEGORY (e.g., [5] C2 denotes segment 5, prohibited concentration).



From Diagnosis to Action: An Evidence-Based Maturity Model for AI Integration in Translation

Viveta Gene

Ionian University / Intertranslations /
Greece

viveta.gene@ionio.gr

Valeria Cannavina

RWS / Spain

Valeria.cannavina@gmail.com

Abstract

The evolution of Artificial Intelligence (AI) in the language services industry has shifted professional practice beyond traditional machine translation (MT) post-editing (PE) toward increasingly AI-enabled, hybrid localization workflows. While AI adoption has accelerated, existing maturity frameworks remain largely organization-centric and insufficiently address the evolving roles, skills, and ethical responsibilities of linguists and language service providers (LSPs). This study presents a data-driven AI Maturity Model (AI MM) designed specifically for the localization ecosystem, grounded in empirical evidence from a global survey of 400 professionals conducted in 2025, with longitudinal comparison to two parallel surveys conducted in 2020 (Gene, 2021) targeting linguists and LSP managers respectively, comprising linguists, LSP managers, and localization leaders. Using mixed-methods analysis, the study identifies patterns of AI integration, workflow adaptation, skill development, ethical readiness, and client alignment across regions and professional roles. Based on these findings, a four-stage AI maturity framework (Emerging, Developing, Strategic, and Transformative) is derived and subjected to preliminary external review through expert feedback gathered during industry webinars; formal empirical validation represents a direction for future research. The proposed model functions as an assessment and roadmap tool, enabling linguists and LSPs to evaluate their state of AI adoption and plan progression. By reframing AI maturity as a role-centric and workflow-oriented construct, this research contributes an applied framework supporting sustainable professional adaptation, targeted training, and responsible AI integration.

1 Introduction

Over the past decade, machine translation (MT) and post-editing (PE) have progressively reshaped professional translation practice. Key terms are used consistently throughout: MT for neural or rule-based automated translation; PE for human revision of MT output; GenAI for AI systems generating novel text; LLMs for large-scale generative language models; hybrid localization workflows (LWs) for integrated workflows combining MT, LLM-assisted drafting, and human expertise. Rule-based MT remains in use in specialized institutional and legacy contexts, though neural MT (NMT) has become the dominant paradigm since 2020, as confirmed by the 2025 survey findings. Research positioned PE as a form of human-computer interaction emphasizing cognitive effort (O'Brien, 2012; Gene et al., 2024); subsequent studies established it as a distinct professional activity with specific competency and training requirements (Gaspari et al., 2015; Gene, 2024; O'Brien et al., 2014), normalized within industrial LWs by 2020.

Since then, generative AI (GenAI), large language models (LLMs), and automated quality estimation have fundamentally altered the landscape of localization. Translation workflows no longer center solely on MT output correction, but increasingly involve AI-assisted content generation, hybrid human-AI collaboration, and dynamic quality management. This shift challenges PE-centric frameworks, as linguists and LSPs must now navigate tool orchestration, data readiness, ethical responsibility, and client communication.

Despite widespread AI adoption, the industry lacks role-centric models capturing how linguists and LSPs progress from exploratory AI use toward strategic, integrated practice. Existing MMs within localization tend to emphasize organizational governance and infrastructure, overlooking the operational realities of linguistic

work and the professional agency of language specialists. This study addresses this gap by proposing a data-driven AI MM grounded in a 2025 global survey (n=400), with longitudinal reference to two parallel 2020 studies targeting linguists and LSP managers respectively (Gene, 2021), tracing the industry’s transition from MT PE toward AI-enabled LWs. The primary objective is to develop and subject to preliminary expert review a practical AI maturity framework reflecting the lived realities of the localization ecosystem as a progression of professional and operational capabilities, offering a tool for self-assessment, training design, and strategic planning.

Three research questions guide this study: (RQ1) What patterns of AI integration characterize linguists and LSPs in 2025 compared to 2020, across roles and regions? (RQ2) How can these patterns be organized into an observable, role-centric maturity framework? (RQ3) To what extent does a four-stage model provide a practically actionable assessment tool? These questions structure the analytical move from survey evidence through model derivation to application.

2 Background and Related Work

MMs describe staged progression in organizational capability and technology adoption. In software engineering, the Capability MM (CMM) formalized a five-level view of process maturity, providing a foundational template for later maturity-model thinking (Paulk et al., 1993). In localization, frameworks have historically been organization-centric. CSA Research (2017, 2025)¹ describes its proprietary Localization MM as the basis for a Localization Maturity Assessment; while influential, its level descriptors are not fully publicly accessible. The Argos AI MM (AIMM, 2024) proposes a five-step progression—AI-aware through AI-transformed—positioning AI maturity as a measure of a team’s readiness to integrate AI into localization operations while drawing on the LMM tradition.

In parallel, TAUS’s Dynamic Quality Framework (DQF) supports context-sensitive translation quality evaluation, emphasizing that evaluation methods should match content type and purpose (Valli, 2015). Görög’s overview of DQF positions it as an adaptable quality evaluation ecosystem rather than a staged maturity trajectory (Görög, 2014 2). Industry barometers complement these frameworks: the GALA Business Barometer (2025) reports persistent challenges around data readiness, privacy, governance, and scaling AI in production settings, while Nimdzi’s 2025 report underscores the gap between tool availability and organizational capability.

Across these frameworks, maturity narratives are defined at the organizational level, overlooking role evolution, workflow redesign, and practitioner-level ethical practice. PE research has established that technology adoption cannot be understood through throughput metrics alone; cognitive effort, task design, and training are integral to real-world performance (O’Brien, 2012; O’Brien et al., 2014; Gene, 2024). This study introduces an AI MM for linguists and LSP operations, derived from longitudinal survey evidence and structured around workflow integration, skill formation, ethical readiness, and client alignment—complementing organization-level frameworks (CSA LMM, Argos AIMM) while supplying the missing role-centric layer.

Framework	Scope	Data-Basis	Stage/Level-Logic	Linguistic-Practice
CMM (Paulk et al., 1993)	Software process	Expert consensus	5 levels	Not applicable
CSA LMM (2017, 2025)	Org. localization	Industry assessment	Multiple stages	Org-level only
Argos AIMM (2024)	Org. AI readiness	Industry report	5 stages	Limited
TAUS DQF (Görög, 2014; Valli, 2015)	Quality evaluation	Industry pilots	No stages	Quality-focused
Proposed AI MM (this study)	Linguists & LSPs	Longitudinal survey (n=400)	4 stages	Role- & workflow-centric

Table 1: Comparison of selected maturity, assessment, and quality frameworks in localization

¹ <https://insights.csa-research.com/reportaction/37952/Marketing>

Table 1 provides a comparative overview of selected MMs, assessment frameworks, and quality initiatives in localization.

3 Methodology

This study adopts a mixed-methods research design combining longitudinal survey data, quantitative analysis, and qualitative thematic interpretation to investigate patterns of AI adoption and maturity within the localization ecosystem.

3.1 Survey Design and Scope

The primary dataset comprises 400 valid responses collected between May and June 2025 from professionals across the localization ecosystem: freelance linguists, in-house translators, LSP managers, project managers, localization leads, technology specialists, and academic researchers. Sampling was purposive and convenience-based via professional networks, industry associations, and two GALA-organized webinars. As the survey was publicly open rather than sent to a fixed invitation list, no denominator response rate is calculable; the 400 responses represent the full set of completed submissions received during the collection window. The sole exclusion criterion was student status; all other localization professionals were eligible. Geographically, responses covered Africa (21.2 %, the largest respondent group), Asia (8.6 %), North America, Central, Southern, Western, Eastern and Northern Europe, the Middle East, South America, and Australia/Oceania (full regional proportions reported in Figure 1). Coverage was broad but uneven, so regional comparisons should be read as indicative rather than representative of the global localization workforce. Longitudinal comparison draws on two parallel surveys conducted in 2020 (Gene, 2021): one targeting linguists, one targeting LSPs and client-side localization managers, both sharing a core question set with the 2025 instrument. Additional items on GenAI and LLMs were added exclusively in 2025. Nine sections and 36 questions covered demographics, technology adoption, workflow evolution, challenges, mindset, skills and training, future outlook, comparative assessment, and ethics (Appendix A). Closed-ended Likert-scale and multiple-choice items supported quantitative analysis; open-ended items enabled qualitative depth.

3.2 Data Collection

Data were collected via anonymous online surveys disseminated through professional networks, GALA member communications, LinkedIn groups, and two GALA-hosted webinars (one per wave; 80–120 practitioners each). The survey was administered using Google Forms. Participation was voluntary and anonymous; no personally identifiable information was collected. Webinar participants were invited to comment on preliminary model versions through structured post-session polls and open-comment fields, covering reactions to stage descriptors, dimension relevance, and refinement suggestions. Both authors independently reviewed and thematically coded this feedback, resolving disagreements through discussion before finalizing model wording.

3.3 Data Analysis and Model Derivation

Quantitative data from the 2025 survey were analyzed using descriptive statistics (frequency distributions, central tendencies, cross-tabulations by role and region); selected items were compared against the two 2020 instruments (Gene, 2021) to identify longitudinal trends across the five-year interval. Qualitative open-ended responses were coded thematically via an iterative inductive procedure (Guest et al., 2014). Specifically, one author generated an initial set of codes from a random sample of 50 open-ended responses; codes were then applied to the full dataset, refined iteratively across three review cycles, and collapsed into higher-order themes. Saturation was judged when no new codes emerged from successive batches of 20 responses. Both authors independently coded approximately 20 % of open-ended items; discrepancies were resolved through discussion. Quantitative and qualitative findings were triangulated throughout.

Five dimensions emerged consistently: (i) technology and tool integration; (ii) process and workflow adaptation; (iii) skills development and training; (iv) ethics and quality readiness; and (v) client alignment and business integration. Stage boundaries were derived inductively: cross-tabulation of dimension scores produced four distinguishable capability profiles, constructed post hoc as analytical categories rather than pre-assigned. Preliminary model versions were reviewed during two GALA webinars; structured feedback from linguists, LSP representatives, and localization consultants shaped stage descriptor wording. This constitutes preliminary external review, not formal empirical validation; construct validity and inter-rater agreement testing remain directions for future research.

Dimension scores were calculated as mean Likert ratings per respondent across items mapped to each dimension. Cross-tabulation of these five-dimension scores produced four distinguishable capability clusters, identified through frequency distributions and inter-dimension correlation patterns. Stage thresholds were set at natural breaks in the distribution: Emerging (mean dimension score 1.0–2.4), Developing (2.5–3.2), Strategic (3.3–4.1), Transformative (4.2–5.0 on a 1–5 scale). These boundaries are analytical and post-hoc rather than pre-specified; construct validity testing remains a direction for future research.

4 Findings

4.1 Regional and Role-Based Variation in AI Adoption

AI tool exposure has increased substantially since 2020. In 2020, 56.9 % of respondents used MT and PE; by 2025, NMT dominates at 50 %, with a notable shift to LLMs (25 %), confirming the GALA report (2025) findings (Figure 1).

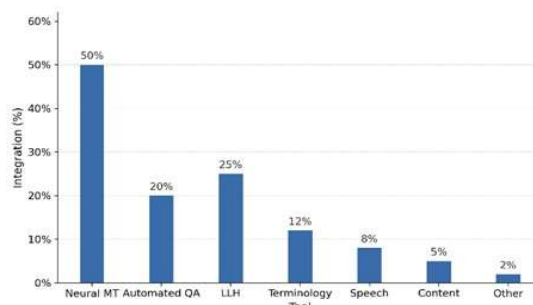


Figure 1: Distribution of AI tool integration in translation workflows — % of respondents using each tool type (NMT 50 %, LLH 25 %, Automated QA 20 %, Terminology 12 %, Speech 8 %, Content 5 %, Other 2 %)

However, maturity levels remain uneven. Respondents in Asia and North America reported the highest average AI integration levels (3.5 and 3.2 respectively), characterized by systematic NMT use, growing LLM adoption, and early experimentation with automated quality estimation. In contrast, respondents from Africa and parts of Eastern Europe showed high tool exposure but comparatively low structured integration (avg. 2.5), indicating a gap between availability and operationalization (Figure 2).

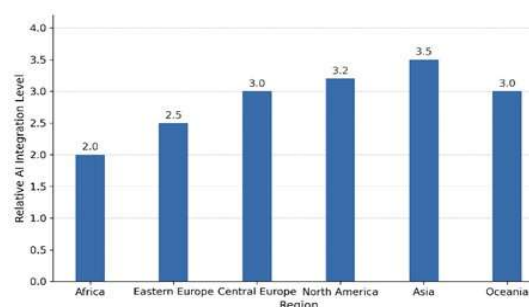


Figure 2: Average AI integration level by region (scale 1–5; Africa 2.0, Eastern Europe 2.5, Central Europe 3.0, North America 3.2, Asia 3.5, Oceania 3.0)

Role-based differences further accentuate this pattern. Localization managers and LSP leads reported higher perceived maturity than freelance linguists, particularly regarding workflow orchestration and tool diversity. Freelance respondents more frequently described fragmented or client-driven AI usage lacking standardized processes, suggesting that organizational context and decision-making authority critically shape AI maturity even when tool access is similar.

4.2 Workflow Evolution Beyond PE

Despite the growing prominence of GenAI, PE remains a dominant workflow component—but its role has shifted. Rather than a standalone service, PE increasingly appears embedded within hybrid workflows combining MT, AI-assisted drafting, terminology automation, and human revision. Respondents reported a diversified range of AI applications beyond MT correction, including content classification (4.7 %), speech recognition (8.1 %), and automated post-editing (1 %), suggesting an expanding functional scope for AI in LWs (Figure 1).

A persistent imbalance was observed between productivity gains and quality management: while respondents frequently associated AI with improved speed (30 %, multiple-choice item on perceived AI benefits, Q20), fewer reported equivalent advances in quality assurance practices (0.2 %, coded open-ended responses, Q30). This mirrors earlier PE research observations where efficiency gains were prioritized ahead of systematic quality control.

4.3 Skills, Training and Ethical Readiness

A consistent finding across regions and roles is a skills gap related to human–AI collaboration. Although a large proportion of respondents reported general awareness of ethical issues (19.2 %), this awareness rarely translated into formalized practices or operational guidelines. Training was predominantly informal, ad hoc, or tool-specific (56 % through practice), with limited evidence of structured role-adapted learning pathways.

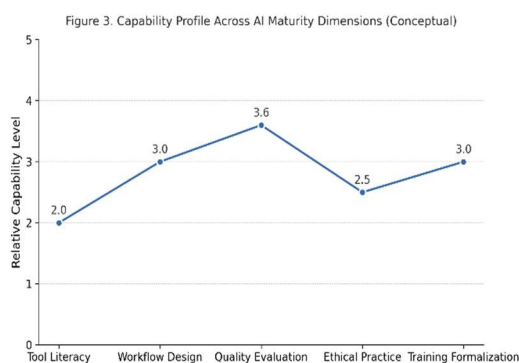


Figure 3: Skills, training, and ethical readiness distribution across maturity levels (%)

Respondents at higher maturity levels were more likely to report ongoing training, workflow documentation, and clearer task boundaries between human and AI contributions (44.7 % rated adaptation confidence at level 4 or 5 on a 1–5 scale). In contrast, respondents at lower maturity levels frequently described uncertainty regarding expectations, compensation models, and responsibility for quality outcomes (39 % preferring human translation with minimal AI assistance). Figure 3 illustrates that ethical and training formalization lag behind tool adoption at lower maturity stages.

4.4 Correlations Between Maturity Indicators and Client Alignment

Cross-analysis suggests a meaningful association between higher AI maturity and improved client alignment. Respondents reporting higher AI integration and technological confidence were more likely to proactively communicate AI capabilities and limitations: 28.1 % indicated positive confidence (level 4) and 16.6 % high confidence (level 5) in their organization’s ability to adapt to technological change. Conversely, 19.7 % reported low confidence (levels 1–2), more likely experiencing misalignment around turnaround times, pricing, and perceived effort. These findings reinforce the idea that AI maturity is not solely a technical condition but a relational one, shaped by communication practices and shared understanding between providers and clients (Figure 4).

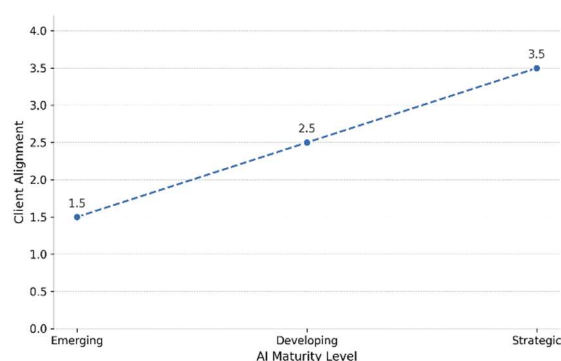


Figure 4: Client alignment and confidence scores by AI integration level

5 The AI Maturity Model

Building on the empirical findings in Section 4, this study proposes a four-stage AI MM designed to capture how linguists and LSPs progress in the adoption and operationalization of AI within LWs. Unlike organization-centric frameworks, the proposed model is explicitly role- and workflow-oriented. The model conceptualizes AI maturity as a progression across five interrelated dimensions: (i) technology and tool integration; (ii) process and workflow adaptation; (iii) skills development and training; (iv) ethics and quality readiness; and (v) client alignment and business integration (Table 2).

Dimension	Emerging	Developing	Strategic	Transformative	Key Shift
Tool Integration	Ad hoc NMT use	Systematic CAT+MT	Governed AI stack	Adaptive, multimodal AI	Reactive → Designed
Workflow	Fragmented, client-driven	Defined, repeatable	Documented, proactive	Dynamic orchestration	Correction → Design
Skills & Training	Informal, self-directed	Tool-driven	Competency-based, continuous	Strategic leadership	Tool use → Hybrid expertise
Ethics & Quality	Conceptual awareness	Basic safeguards	Formalized governance	Industry-shaping standards	Implicit → Operational
Client Alignment	Unclear, under pressure	More explicit, inconsistent	Transparent, proactive	Co-creative, co-governed	Reactive → Collaborative

Table 2: AI Maturity Model: dimensions x stages

5.1 Stage Descriptions

The **Emerging** stage represents fragmented, reactive AI adoption. Organizations experiment—most commonly with NMT—but use remains inconsistent and externally pressured rather than strategically planned. Workflows lack standardization; PE practices are loosely defined; skills development is informal and self-directed; client alignment is weak, with unclear expectations around quality, pricing, and responsibility.

The **Developing** stage marks a transition from experimentation to intentional integration. AI tools are more systematically embedded within CAT environments; basic metrics and feedback loops emerge; linguists begin to assume expanded roles in quality assurance and workflow optimization. Training becomes more visible but remains tool-driven rather than competency-based. Ethics becomes more explicit but remains procedural; client communication is more intentional though not yet systematically embedded.

The **Strategic** stage represents mature, fully integrated AI practice governed by structured frameworks. Quality management shifts from reactive correction to anticipatory design. Linguists operate as AI-augmented professionals (Lommel, 2020), combining domain expertise with advanced tool literacy and data-informed decision-making. Ethics and quality are formalized into governance structures with clearly defined accountability mechanisms. Training is continuous, role-specific, and competency-based. Client alignment is proactive and transparent, positioning AI as a strategically managed capability enhancing value.

The **Transformative** stage represents a paradigm shift in which AI becomes a driver of strategic differentiation and organizational identity. Workflows are adaptive, predictive, and continuously optimized across interconnected content ecosystems. Linguists act as strategic leaders spanning linguistics, AI evaluation, data governance, and quality engineering. Ethics becomes a strategic differentiator actively shaping broader industry standards. Client alignment is co-creative, with expectations, risk-sharing, and quality design negotiated transparently as part of a shared innovation strategy.

Dimension	Emerging	Developing	Strategic	Transformative
Technology & Tool Integration	Irregular NMT use; tools chosen by client or employer; no systematic selection criteria.	NMT integrated within CAT; some LLM experimentation; basic QE in use.	Diverse AI stack (NMT + LLMs + QE) governed by documented selection frameworks.	AI orchestration across interconnected pipelines; continuous optimization via feedback loops.
Process & Workflow Adaptation	PE loosely defined; no standardized workflows; reactive to client instructions.	Basic workflow documentation; emerging feedback loops; hybrid PE and LLM drafting beginning.	Structured, documented hybrid workflows; anticipatory quality design.	Adaptive, predictive workflow design continuously refined across content ecosystems.
Skills Development & Training	Informal, self-directed, tool-specific learning; no structured upskilling plan.	Visible but episodic training; tool-driven rather than competency-based.	Continuous, role-specific, competency-based training; gaps tracked against defined frameworks.	Linguists as learning leaders; co-developing industry curricula and certification standards.
Ethics & Quality Readiness	Awareness without formal practice; quality management reactive; no error protocols.	Ethics made explicit; procedural quality checks in place; basic human-in-the-loop steps.	Formalized ethics governance; accountability structures; proactive quality design.	Ethics as strategic differentiator; LSP shapes broader industry standards.
Client Alignment & Business Integration	Unclear client expectations; no proactive communication about AI use.	More intentional client communication; AI capabilities beginning to align with expectations.	Proactive, transparent AI value communication; expectations co-managed; SLA reflects quality tiers.	Co-creative partnerships; clients as innovation co-designers; shared risk and quality agreements.

Table 3: AI Maturity Model stage descriptors (observable indicators and representative practitioner behaviors across five dimensions)

Table 3 maps each stage across the five dimensions to observable indicators and representative practitioner behaviors, providing concrete boundary criteria between adjacent stages.

5.2 Maturity Dimensions Across Stages

Differences across stages are most pronounced in workflow structure and governance. Lower maturity contexts mirror earlier PE paradigms focused on throughput; higher maturity contexts demonstrate reorientation toward workflow design supporting differentiated quality strategies. Skills similarly evolve from traditional linguistic competence toward hybrid expertise—reinforcing PE research findings that productivity gains alone are insufficient indicators of successful technology integration (Gene, 2024). Ethics and quality readiness function as threshold dimensions distinguishing advanced maturity: while awareness is widespread, only higher stages translate it into operational practices. Client alignment emerges as both an outcome and a driver of maturity, with strategic-stage practitioners actively shaping expectations through informed dialogue.

5.3 Practical Application of the Model

The proposed AI MM is intended as a practical assessment and planning tool rather than a prescriptive checklist. Linguists and LSPs can use it to identify their current maturity stage across each dimension, recognize gaps, and prioritize targeted interventions such as training programs, workflow redesign, or client education initiatives. By framing AI maturity as a progression of professional capability, the model provides a pathway for sustainable adaptation. An interactive version of the AI Maturity Model assessment tool is publicly hosted and available at Intertranslations AI Maturity Model landing page (<https://www.intertranslations.com/linguistic-ai-maturity-model-test/>), offering practitioners a structured means of self-assessment and a foundation for future diagnostic instrument development.

6 Discussion

Widespread AI tool availability has created an illusion of readiness not consistently borne out by operational maturity. Respondents frequently

report AI use, yet integration remains fragmented and reactive—echoing PE research where productivity gains masked unresolved issues of task clarity, quality management, and cognitive effort (O’Brien, 2012; Gene, 2024). The AI MM explains this discrepancy: organizations at lower stages may appear technologically advanced yet lack the workflows, skills, and governance to leverage AI strategically.

One of the most striking findings concerns the disconnect between ethical awareness and ethical practice. Although respondents frequently express concern about quality, bias, and accountability, these concerns are rarely formalized at lower maturity levels. Ethical readiness functions as a late-stage maturity indicator: in strategic-stage contexts, ethics is operationalized through explicit thresholds, human-in-the-loop decision points, and transparent client communication—aligning with calls to embed ethical reasoning into workflow design rather than treating it as an external constraint (O’Brien et al., 2014).

Current training practices, where they exist, are predominantly tool-focused and episodic, offering limited support for hybrid competencies required in AI-enabled environments. Higher maturity stages, by contrast, are associated with continuous, role-specific learning integrating technical, linguistic, and ethical dimensions. This supports calls for rethinking translator education beyond discrete skills toward adaptive competence models (Gene, 2021), with the AI MM offering a scaffold for curriculum design aligned with progressive maturity stages.

Finally, the proposed AI MM should be viewed as complementary to existing organization-level frameworks. While CSA Research’s LMM (2017, 2025) and Argos’s AIMM (2024) provide valuable strategic perspectives, they do not sufficiently capture the evolution of linguistic work itself. By focusing on workflows, roles, and professional agency, the present model addresses a critical layer of AI adoption underrepresented in both academic and industry discourse.

Self-reported data introduce two interrelated risks that readers should bear in mind. First, perception bias: enthusiastic adopters may overestimate integration depth while skeptics underestimate it, distorting stage distribution in ways undetectable without objective performance triangulation. Second, terminological inconsistency: terms such as “AI”, “hybrid

workflow”, and “quality management” may have been interpreted differently across respondents depending on technical background, role, or region. The longitudinal design and the definitional note in the Introduction partially mitigate this, but quantitative maturity estimates in Section 4 should be treated as indicative; future studies should supplement perceptual measures with behavioral anchors such as workflow observation or output audits.

7 Conclusion and Future Work

This study demonstrates that AI maturity cannot be captured through tool adoption metrics alone. Instead, maturity emerges as a multidimensional progression encompassing workflow integration, skill development, ethical readiness, and client alignment. Addressing RQ1, the 2025 survey data—read against the 2020 comparative baseline—confirm substantial change in AI adoption patterns, with NMT becoming dominant and LLMs emerging as a significant second-tier technology, alongside persistent regional and role-based disparities. Addressing RQ2, the proposed four-stage AI Maturity Model—Emerging, Developing, Strategic, and Transformative—provides a role- and workflow-oriented framework derived inductively from these patterns, offering a more granular and practitioner-relevant characterization of maturity than existing organization-level models. Addressing RQ3, the model demonstrates practical utility as a self-assessment scaffold and planning tool, though its empirical validation—including construct validity and stage boundary reliability—remains an explicit direction for future research. The model extends earlier PE research by situating human–computer interaction, cognitive effort, and quality management within a broader AI maturity trajectory.

From a practical perspective, the model provides linguists and LSPs with a structured means of self-assessment and strategic planning, enabling identification of specific gaps across maturity dimensions and prioritization of targeted interventions. Ethical and quality considerations are reframed as integral components of maturity, reinforcing the need for responsible AI integration in professional language services.

Future work includes: (i) development of a digital AI maturity assessment tool operationalizing the model; (ii) further empirical validation with larger, more evenly distributed samples to refine stage descriptors; and (iii) closer collaboration with educational institutions and professional associations to align curricula and certification schemes with AI maturity progression. As AI technologies continue to evolve, the sustainability of localization practice will depend on the capacity of professionals to adapt roles, workflows, and ethical frameworks accordingly.

Acknowledgments

This document is derived from a survey with the support of Intertranslations, RWS, and Globalization and Localization Association (GALA), along with two webinars hosted by the latter.

References

- Argos Multilingual. 2024. *The Argos AI Maturity Model: The Five Stages of the AI Journey*, pages 3–10.
- John W. Creswell and Vicki L. Plano Clark. 2018. *Designing and Conducting Mixed Methods Research*, third edition. SAGE Publications.
- CSA Research. 2017. *Localization Maturity Model: Release 3.0*. <https://insights.csa-research.com/reportaction/37952/Marketing>
- CSA Research. 2025. *Localization Maturity Assessment™*. <https://csa-research.com/maturity-assessment>
- GALA. 2025. *GALA Business Barometer: Technology, Automation, and AI*. Globalization and Localization Association, pages 10, 33–35.
- Federico Gaspari, Hanane Almaghout, and Stephen Doherty. 2015. PE of machine translation output: Processes and applications. In Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia, editors, *PE of Machine Translation: Processes and Applications*. Cambridge Scholars Publishing.
- GALA. 2025. *GALA Business Barometer: Technology, Automation, and AI*. Globalization and Localization Association, pages 10, 33–35.
- Viveta Gene, Vilelmini Sosoni, Pilar Sánchez-Gijón, Parthena Charalampidou, and Maria Do Campo Bayón. 2024. *Exploring students' experience with CAT-integrated MT and Gen AI: Translating creative texts in BWX*. In *Proceedings of NeTTT 2024*, pages 76–90. <https://acl->

bg.org/proceedings/2024/NeTTT%202024/index.html

Viveta Gene. 2024. *Machine Translation PE and Professional Translators: The Contribution of Training to the Reduction of Effort and the Improvement of Quality*. Ph.D. thesis, Ionian University Greece.

<https://doi.org/10.12681/eadd/58063>

Viveta Gene. 2021. The PE workflow: Training challenges for LSPs, post-editors, and academia. In *Proceedings of TRITON*, pages 187–198.

<https://aclanthology.org/2021.triton-1.22/>

Attila Görög. 2014. Quality evaluation today: The Dynamic Quality Framework. In *Proceedings of Translating and the Computer 36*, pages 155–164.

<https://aclanthology.org/2014.tc-1.21/>

Greg Guest, Kathleen M. MacQueen, and Emily E. Namey. 2014. *Applied Thematic Analysis*. SAGE Publications.

Arle Lommel. 2020. Augmented Translation: Are We There Yet? *CSA Research blog*, 4 November 2020.

<https://csa-research.com/More/Articles/ArticleID/828>

Nimdzi Insights. 2025. *Nimdzi 100: Technology Trends*. Nimdzi Insights, p. 72.

Sharon O'Brien. 2012. Translation as human–computer interaction. *Translation Spaces*, 1(1):101–122. https://www.researchgate.net/publication/262868280_Translation_as_human-computer_interaction

Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia, editors. 2014. *Post-Editing of Machine Translation: Processes and Applications*. Cambridge Scholars Publishing.

Mark C. Paulk, Bill Curtis, Mary Beth Chrissis, and Charles V. Weber. 1993. *The Capability Maturity Model for Software, Version 1.1*. Technical Report CMU/SEI-93-TR-024. Software Engineering Institute, Carnegie Mellon University. https://www.researchgate.net/publication/228090875_Capability_Maturity_Model_for_Software

Paola Valli. 2015. The TAUS Quality Dashboard. In *Proceedings of Translating and the Computer 37*, London, UK. AsLing.

<https://aclanthology.org/2015.tc-1.17/>

.

Appendix A. Survey Instrument

Survey title: Future Proofing Translation Through CAT Tools and AI Integration. **Platform:** Google Forms. **Collection period:** May-June 2025. **Total:** 36 questions across 9 sections. Participation was voluntary and open to all localization professionals excluding students.

Question types: Single = single-select; Multi = multi-select Scale = Likert scale' Rank = ranking task. Response options are abbreviated for space; full options were presented in the survey instrument.

No.	Question	Type	Response Options (abbreviated)
Section 1: Demographics and Organizational Context			
Q1	Primary role in the translation industry	Single	Translator/Post-editor; Project Manager; LSP Executive/Owner; Technology Developer/Engineer; Localization Manager; Academic/Researcher; Other
Q2	Location of professional activity	Single	Africa; North America; South America; Asia; Australia/Oceania; Central Europe; Eastern Europe; Northern Europe; Southern Europe; Western Europe; Middle East
Q3	Years of experience in the translation industry	Single	<2 years; 2-5; 6-10; 11-15; >15 years
Q4	Size of organization	Single	Solo practitioner; 2-10; 11-50; 51-250; >250 employees
Q5	Regions primarily served	Multi	North America; South America; Europe; Middle East and Africa; Asia Pacific
Q6	Industries primarily served	Multi (?)	Life Sciences/Medical; Legal; Technical/Engineering; IT/Software; Marketing/Creative; Financial; E-commerce; Gaming; Media/Entertainment; Other
Q7	How respondent learned about the survey	Single	Email invitation; GALA network; LinkedIn; Word of mouth; Company internal; University/Academic
Section 2: Technology Adoption and Integration			
Q8	CAT tools currently used	Multi	Across; CaféTran; Crowdin; Déjà Vu; GlobalLink; Lilt; Lokalise; MateCat; memoQ; Memsource; OmegaT; Phrase; Plunet; SDL Trados; SmartCAT; Smartling; Transifex; Wordbee; Wordfast; XTM Cloud; XTRF; Other
Q9	Change in technology investment since 2020	Single	>30% increase; 10-30% increase; <10% increase; No change; Decrease; N/A
Q10	AI technologies integrated with CAT tools since 2020	Multi	NMT; LLMs (GPT, Claude); Automated QA; Terminology extraction; Content classification; Speech recognition; None; Other
Q11	Current level of AI integration in workflows (1-5)	Scale	1 = No integration ? 5 = Advanced AI –human collaborative workflows
Q12	Organizational approach to AI-CAT integration	Single	Cutting-edge; Proven/adopted; Built-in CAT features; Minimal AI; Researching options
Section 3: Workflow Evolution and Performance Impact			
Q13	How workflow has changed since 2020	Multi	Increased MT pre-translation; AI-assisted review; Specialized PE roles; Automation of PM; Realtime collaboration; Continuous localization; No change; Other
Q14	Impact of AI-CAT integration on productivity	Single	>30% increase; 10-30% increase; <10% increase; No change; Decrease; N/A
Q15	Impact of AI-CAT integration on translation quality	Single	Significant/Moderate/Slight improvement; No change; Decline; Varies by language pair; N/A
Q16	Percentage of workflow by translation method	Single	>50% human without MT; >50% human with MT; >50% PE of MT output; >50% fully automated
Q17	Language pairs with most significant AI improvements	Multi (?)	W. European? English; N. European? English; E. European? English; Asian? English; Middle Eastern? English; Other pairs; No significant improvements
Section 4: Challenges and Opportunities			
Q18	Main challenges in integrating AI with CAT tools	Rank (1-5)	Technical integration; Cost; Staff training; Quality concerns; Data security; Translator resistance; Lack of AI transparency; Other
Q19	Confidence in ability to adapt to technological change (1-5)	Scale	1 = Not confident at all ? 5 = Extremely confident
Q20	Most promising opportunities from AI-CAT integration	Multi (?)	Productivity; Consistency; Volume capacity; Quality assurance; New services; Cost reduction; Engaging work; New markets; Other

Section 5: Mindset				
Q21	Attitude towards CAT Tools	Single	Positive; Negative; Neutral	
Q22	Attitude towards Machine Translation	Single	Positive; Negative; Neutral	
Q23	Attitude towards LLMs in Translation	Single	Positive; Negative; Neutral	
Section 6: Skills and Training				
Q24	How translator roles have changed since 2020	Multi	More PE focus; Subject specialization; Tech skills emphasis; QA responsibilities; Content creation; No change; Other	
Q25	Most important skills for translators in the AI era	Rank (1-5)	Technical aptitude; Critical evaluation of MT; Subject matter expertise; Prompt engineering; Project management; Traditional linguistic skills; Other	
Q26	How organization addresses training needs	Multi	Formal in-house; External courses/certifications; Mentoring; Self-directed; Learning through practice; No formal approach; Other	
Section 7: Future Outlook				
Q27	Greatest value of CAT tools in AI-dominated landscape	Multi (<3)	TM management; Terminology consistency; QA error prevention; PM automation; File handling; Human oversight; AI integration hub; Complex content; Compliance; UI visualization; Client collaboration; Data security; Human-AI environments; Domain adaptation	
Q28	Envisioned evolution of CAT-AI relationship over 5 years	Single	AI replaces CAT; CAT incorporates AI; CAT becomes AI platform; Distinction blurs; CAT unchanged; Other	
Q29	Emerging technologies with most impact in next 3 years	Multi (<3)	Adaptive MT; AI content authoring; AR translation; Voice-to-text workflows; Blockchain QA verification; Advanced terminology mgmt; Other	
Q30	Focus of future research on CAT-AI integration	Single	Human-AI Collaboration; Economic Impact; Quality Assessment; Ethics/Privacy/Compliance; Technical Integration; Training/Professional Dev; Specialization/Domain; Long-term Industry Impact	
Section 8: Comparative Assessment (respondents active since 2020)				
Q31	Change in perception of MT/PE compared to 2020	Single	Much more positive; Somewhat more positive; No change; Somewhat more negative; Much more negative; N/A	
Q32	Change in role of CAT tools compared to 2020	Single	More central; Same importance; Less important; Used differently; N/A	
Q33	Has relationship between translators and technology improved since 2020?	Single	Significantly/Somewhat improved; No change; Somewhat/Significantly worsened; N/A	
Section 9: Ethics				
Q34	Concern about ethical issues related to AI in translation	Single	Not concerned; Slightly; Moderately; Very; Extremely concerned	
Q35	Trust in AI-generated translation suggestions	Single	Fully trust; Trust but review most; Rough guidance only; Rarely trust; Do not use	
Q36	Steps taken to comply with EU AI Act in translation workflows	Multi	Risk assessments; Bias auditing; Privacy policy updates; Transparency of AI decisions; Monitoring AI performance; No steps taken yet	

Enhancing Persian–Arabic OCR for Translation Workflows: A Benchmark-Driven Optimization Approach

Masoud Shahnazari

Translation Studies Research Unit,
KU Leuven, Leuven, Belgium
masoud.shahnazari@kuleuven.be

Abied Alsulaiman

Arabic Studies Research Unit,
KU Leuven, Antwerp, Belgium
abied.alsulaiman@kuleuven.be

Abstract

Optical Character Recognition (OCR) is a critical preprocessing stage for enabling machine translation (MT) and computer-assisted translation (CAT) workflows on Persian and Arabic printed and scanned documents. Within the translation industry in the Middle East, a substantial proportion of documents handled by translators are still received as scanned PDF files. This requires OCR as an unmissable preprocessing step before such documents can be integrated into MT and CAT tools. However, the cursive nature of these languages, positional letter variation, and dense diacritics limit recognition accuracy. We present a translation-oriented OCR pipeline for Persian and Arabic that combines systematic benchmarking of multiple open-source and vision-language OCR systems with targeted model adaptation. The framework integrates preprocessing, normalization, and structured post-processing, and further explores fine-tuning possibilities using LoRA to improve script-specific robustness. Through experimental evaluation, the optimized system achieves a mean Character Error Rate (CER) of 0.0224 for Persian and 0.0265 for Arabic, corresponding to character accuracies above 97% and word accuracies of 89–90%. Compared to baseline averages (CER $\approx$ 0.18–0.22), the improvement is substantial. The resulting workflow produces stable output for translation tools and consequently reduces manual correction effort in large-scale digitization and translation pipelines.

1 Introduction

Optical Character Recognition (OCR) has significantly transformed the processing of written material, enabling the conversion of scanned pages or images into editable and searchable text. In Persian and Arabic translation workflows, OCR is not simply a supporting tool but often an essential preprocessing step, since many legacy source documents must be digitized

before they can be used in machine translation (MT) and computer-assisted translation (CAT) systems. In practice, based on documented evidence from multiple large-scale digitization projects (Bergamaschi et al., 2022), a considerable portion of documents still arrives as scanned PDF files, requiring OCR prior to downstream processing.

Historically, OCR systems for these scripts have evolved alongside advances in artificial intelligence. Early approaches were largely rule-based and limited in flexibility. The rise of deep learning introduced convolutional and recurrent neural networks, which improved the modeling of cursive structure and context-dependent character shapes in Arabic-script languages (Altwaijry & Al-Turaiki, 2021). Transformer-based architectures have further advanced OCR performance by enabling global line-level modeling rather than segmented character recognition, which is particularly useful for complex layouts such as Persian Nastaliq script (Tahir et al., 2020).

From a translation perspective, OCR output is essential for building the required input text used in neural machine translation systems. As a result, research has increasingly focused on closer integration between OCR and translation pipelines. However, progress is constrained by limited benchmarking resources tailored to these scripts. Datasets such as KHATT and newer Persian corpora like Idpl-pfod2 (Asadi-zeydabadi et al., 2023) remain limited in scope, often emphasizing clean printed text while underrepresenting dialectal and historical variations. Even standard preprocessing methods such as noise reduction and contrast enhancement require adaptation for diacritic-rich scripts (Fateh et al., 2023).

Post-processing techniques have also improved, with language models increasingly used to correct systematic OCR errors, especially

those involving visually similar characters. These approaches have reduced error rates to a great extent, in some cases to below 2% on clean inputs (Khosrobeigi et al., 2022). Commercial systems such as Google Cloud Vision and ABBYY have claimed to achieve strong performance on Persian and Arabic, while open-source tools like Tesseract have improved through LSTM and transformer-based enhancements (Smith & Reetz, 2024). In parallel, these OCR systems are widely adopted in cultural heritage digitization projects, where they facilitate the conversion of fragile manuscripts into structured digital resources that can be integrated into translation memory systems (Bergamaschi et al., 2022). Despite these advances, concerns remain regarding bias and uneven performance across dialects and script variations. Addressing these issues requires more diverse datasets and evaluation frameworks that reflect linguistic variability (Riggan et al., 2022). Overall, OCR for Persian and Arabic is still not fully reliable for unrestricted access to large-scale textual heritage in translation contexts.

The considerations outlined above motivate the present study, which aims to advance the application of OCR in Persian and Arabic translation contexts. Specifically, this research develops and evaluates detection and recognition strategies tailored to the structural characteristics of these scripts. By combining methodological analysis with practical implementation, the study seeks to bridge the gap between theoretical OCR research and applied translation technology, ultimately facilitating improved access to Persian and Arabic textual heritage.

2 Methodology

2.1 OCR Tools and System Design

This section outlines the approach adopted for OCR in Persian and Arabic within translation-oriented workflows. It begins with a systematic examination of several open-source tools that demonstrate potential for multilingual script processing. Particular attention is given to their handling of script-specific characteristics, including the continuous cursive connections between letters, positional variation in character forms, and the presence of fine diacritical marks that may alter meaning or pronunciation. The analysis considers system architecture, incorporates results from controlled evaluations (e.g., character and word error rates derived from targeted benchmarks), and identifies both

strengths and persistent limitations, as detailed in our experimental evaluation (Section 3) and informed by established OCR benchmarking practices (Altwaijry & Al-Turaiki, 2021; Khosrobeigi et al., 2022).

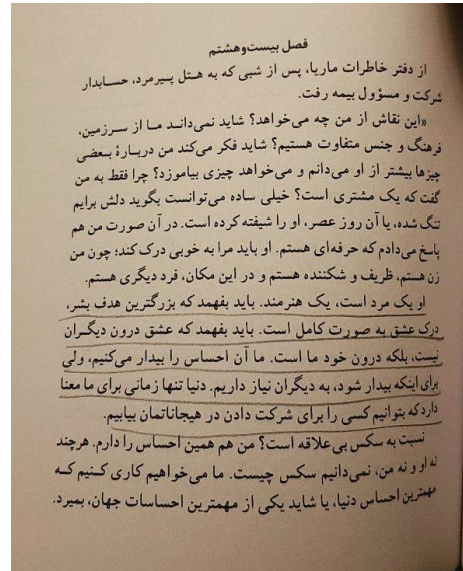


Figure 1: Sample test images of Persian (top) and Arabic (bottom) scripts. Both scripts are cursive and right-to-left, with characters assuming different forms depending on their position within a word—properties that create the primary OCR challenges addressed in this study.

Before examining OCR tools in detail, it is helpful to briefly understand the similarities and differences between Persian and Arabic scripts. Both scripts share the same 28-letter core derived from Classical Arabic and are written cursively from right to left, with characters assuming

different forms depending on their position within a word (initial, medial, final, isolated). However, Persian adds four letters (گ, ژ, چ, پ) to represent sounds not present in Arabic (Figure 1). The most consequential difference for OCR is typographic: Arabic is typically printed in the Naskh style, which maintains a strong horizontal baseline with relatively discrete ligatures. Persian, in contrast, predominantly uses the Nastaliq style, which is diagonal, heavily ligatured, and exhibits extreme vertical overlap—characters are stacked rather than placed side by side. This distinction means that Persian OCR must handle baseline instability and stacked ligatures (e.g., [دانشگاه] rendered as a single complex glyph), whereas Arabic OCR prioritizes diacritic accuracy and long connected components. Understanding this difference is essential for interpreting the language-specific performance results presented in Sections 3 and 4.

Based on our controlled benchmark experiments (detailed in Section 3), we observed notable variation in performance across tools, depending on both the dataset and script characteristics. These findings are consistent with prior comparative studies on Arabic OCR (Tahir et al., 2020; Riggan et al., 2022). Systems such as DeepSeek and PaddleOCR VL tend to achieve strong results on relatively clean Arabic benchmarks, often exceeding 97% character accuracy. For Persian, models such as Khanandeh and hybrid architectures combining Convolutional Recurrent Neural Network (CRNN) (Shi et al., 2017) with Character Region Awareness for Text Detection (CRAFT) (Baek et al., 2019) perform comparatively well, typically reaching accuracy levels within the 90–98% range. Tools like EasyOCR and RapidOCR are effective in handling right-to-left scripts at a basic level; however, their performance declines on more complex cursive layouts, where word error rates may remain between 15% and 30%. Across systems, two consistent limitations can be observed: reduced robustness when processing handwritten material and a tendency to perform better on standardized modern fonts than on more diverse or historical styles (Figure 2).

2.2 Dataset Preparation and Model Training

The initial dataset required iterative refinement to address script-specific challenges in Persian and Arabic OCR. We began with the publicly available

"Persian_Arabic_TextLine_Image_Ocr_Small" dataset from Hugging Face.

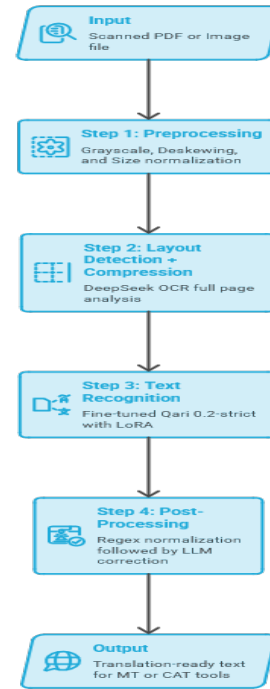


Figure 2: System architecture of proposed Persian /Arabic OCR pipeline showing preprocessing, layout detection, recognition, and post-processing stages.

Early experiments revealed inconsistencies affecting recognition performance, prompting a structured preprocessing pipeline.

This pipeline included:

- (i) explicit language tagging,
- (ii) balancing Persian and Arabic samples,
- (iii) character normalization, and
- (iv) conversion of Latin digits to Persian numerals.

The digit normalization step alone was associated with an increase in baseline character-level accuracy to approximately 74%, highlighting the sensitivity of OCR systems to script-consistent numeric representation. In this work, "normalization" refers to the application of a deterministic character mapping function that converts orthographic variants to canonical Persian/Arabic forms. The normalization function operates on OCR output strings and applies the following transformations: (1) Arabic [ك] → Persian [ک]; (2) Arabic [ي] → Persian [ی]; (3) alternative hamza representations (e.g., أ, إ, ؤ, ئ) → normalized forms based on context; (4) removal of non-standard diacritics that do not affect meaning in modern Persian prose. "Normalized CER" refers to the Character Error Rate computed after applying this normalization

function to both the OCR output and the ground truth text, measuring character accuracy after canonicalization. This metric isolates visual recognition errors from orthographic variation, providing a fairer comparison across datasets with inconsistent annotation conventions. For instance, the numeral sequence "123" appearing in a Persian document is converted to "۱۲۳" (Persian digits), and "2024" becomes "۲۰۲۴". This step is critical because OCR models trained predominantly on Latin digits systematically misrecognize Persian numerals (e.g., confusing "۵" with "3"), leading to cascading errors in date and numeric value extraction. For full-page evaluation, we additionally employed the [Arshasb dataset](#), comprising approximately 1,000 Persian pages with high-quality ground truth annotations.

A comparative evaluation of multiple OCR models identified two particularly strong performers. DeepSeek OCR achieved the highest accuracy for Persian while maintaining competitive results for Arabic. This superiority stems from three architectural advantages: (1) its two-stage transformer encoder-decoder compresses image information into compact tokens while preserving contextual dependencies — crucial for Nastaliq's diagonal stacking; (2) full-page processing maintains global context necessary for accurate diacritic and ligature recognition; and (3) training on a diverse multilingual corpus reduces inherent bias toward Arabic character forms, avoiding systematic substitutions (e.g., Persian [ک] rendered as Arabic [ك]) that affect Arabic-pretrained models.

The Qari model family (versions 0.1, 0.2, and 0.2v2) ([Wasfy et al., 2025](#)) demonstrated the most consistent performance across both languages. These models are fine-tuned versions of Qwen2-VL-2B-Instruct, optimized for Arabic OCR with support for diacritics (tashkeel) and diverse document layouts. Throughout this paper, "strict mode" refers to a constrained decoding configuration where the model is explicitly instructed to avoid outputting Arabic substitute forms when processing Persian text. This is implemented via (a) a system prompt forbidding specific character substitutions (e.g., "do not use Arabic ك or ي"), and (b) setting the model's temperature parameter to 0.1 (low randomness) to reduce hallucinations. The 0.2-strict variant additionally applies a post-hoc filter that rejects any output containing banned Arabic characters, forcing a re-generation attempt. Standard (non-strict) variants use default

temperature (0.7) without character restrictions. Notably, Qwen-based architectures, including Qari, exhibited significant sensitivity to prompt design.

For model adaptation, we fine-tuned the Qari OCR 0.2.2.1 VL 2B Instruct model using Low-Rank Adaptation (LoRA) within a PyTorch environment on GPU hardware. The training dataset consisted of 16,000 samples, with an additional 2,000 samples reserved for validation. Training was conducted over two epochs, completing in approximately 56 minutes. During this process, training loss decreased from 0.0484 to 0.0352, while validation loss improved from 0.0635 to 0.0561 (Figure 3).

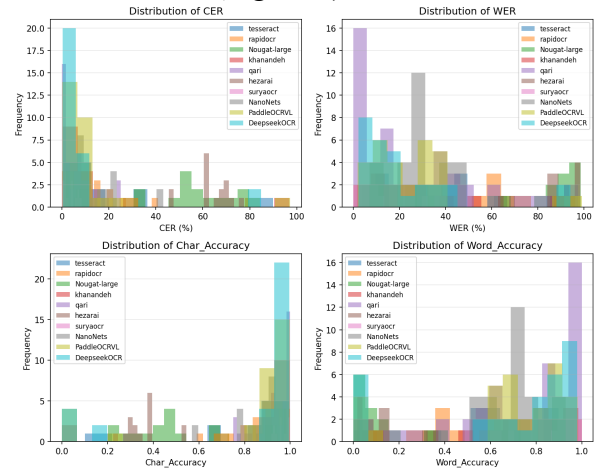


Figure 3: Comparative Performance Distribution of Selected Open-Source OCR Tools on Persian and Arabic Test Dataset

2.3 Evaluation and Final Pipeline

The complete pipeline consists of the following sequential stages:

- Preprocessing: (i) grayscale conversion, (ii) deskewing (using OpenCV's affine transformation with angle detection), (iii) size normalization (width capped at 512 pixels, height at 64 pixels for textline images).
- Recognition: (iv) full-page layout detection and compression using DeepSeek OCR, (v) text recognition using fine-tuned Qari 0.2-strict with LoRA adapters.
- Post-processing: (vi) regex-based normalization (e.g., substituting Arabic [ك] with Persian [ي], [ك] with [س]), (vii) LLM-based contextual correction using a prompted language model to resolve remaining morphological ambiguities.

Evaluation metrics centered on Character Error Rate (CER) and Word Error Rate (WER),

both computed using Levenshtein (edit) distance at the character and word levels respectively, following standard ICDAR protocols (Rice et al., 1995). CER is defined as $(S + D + I) / N$, where S = substitutions, D = deletions, I = insertions, and N = total characters in ground truth. WER follows the same formulation at the word level. Additionally, we report chrF2 scores to capture character n-gram similarity, which is particularly informative for Persian and Arabic given their rich morphology. Initial experiments indicated that the Qari model achieved approximately 45% accuracy for Persian, largely due to bias toward Arabic representations; however, the application of normalization procedures and more constrained prompting strategies increased performance to around 80%. Hybrid configurations further contributed to efficiency gains in the range of 15–20%.

When evaluated on mixed Persian and Arabic datasets, the system achieved word accuracy levels between 85% and 95%, with processing times below 10 seconds per page on standard GPU hardware. The resulting workflow generates translation-ready text while reducing the extent of manual correction required for corpus construction in neural machine translation applications.

3 Results

3.1 Impact of Normalization and Strict Modes

The grouped bar chart presents a comparison between raw CER values and normalized outputs across multiple base models, with strict variants included where applicable. Substantial reductions are observed for the Qari and Hezarai models, where raw error rates often exceed 0.5–0.6 but decrease to approximately 0.1 or lower following normalization, resulting in more consistent clustering. In contrast, DeepSeek and PaddleOCR VL begin with relatively lower baseline CER values (around 0.2) and therefore exhibit more limited relative improvement. Tesseract and RapidOCR display the largest differences between raw and normalized outputs, suggesting that their error patterns are more amenable to post-processing corrections.

The chart supports the earlier observation that many outputs categorized as “high-error” are not necessarily unusable; rather, a large share of the errors consists of systematic character substitutions arising from script overlap. After normalization, these outputs often reach a level of quality that is sufficient for integration into translation pipelines (Figure 4).

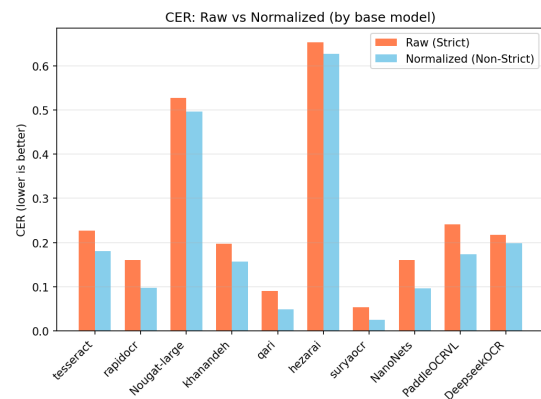


Figure 4: CER: Raw vs Normalized (by Base Model)

3.2 Model Ranking Consistency Across Metrics

The line chart illustrates how each model’s ranking (with lower values indicating better performance) varies across five evaluation metrics: CER, WER, chrF2, Character Accuracy, and Word Accuracy. DeepSeek variants, both standard and strict, demonstrate notable stability, consistently ranking within the top five positions with only minor fluctuations.

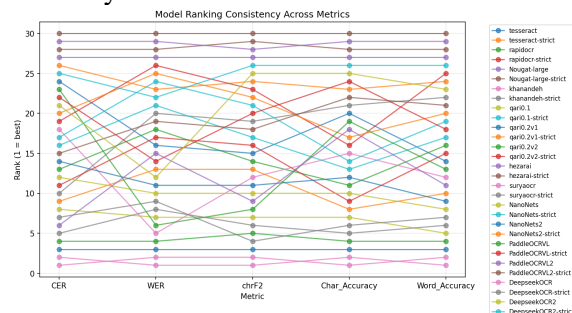


Figure 5: Model Ranking Consistency Across Metrics

PaddleOCR VL and NanoNets also maintain relatively stable positions within the top ten. In contrast, Qari variants exhibit greater variability, performing strongly on metrics such as chrF2 while showing weaker results on WER. Tesseract and RapidOCR generally remain in the lower half of the ranking and display less consistent behavior. The relatively flat ranking trajectories observed for DeepSeek and PaddleOCR VL suggest balanced performance across metrics, whereas more variable patterns, as seen in Qari, reflect strengths in certain areas but limitations in others. This consistency analysis supports the selection of DeepSeek for the final pipeline, as stable cross-metric performance is particularly important for translation workflows, where reliability tends to outweigh isolated peaks in performance (Figure 5).

3.3 Character vs Word Error Correlation

The scatter plot examines the relationship between Character Error Rate (CER) and Word Error Rate (WER) across all evaluated OCR models and their variants on Arabic text samples. The distribution of points follows an almost linear trend, indicating a strong positive correlation between the two metrics: increases in character-level errors are closely associated with increases in word-level errors (Figure 6).

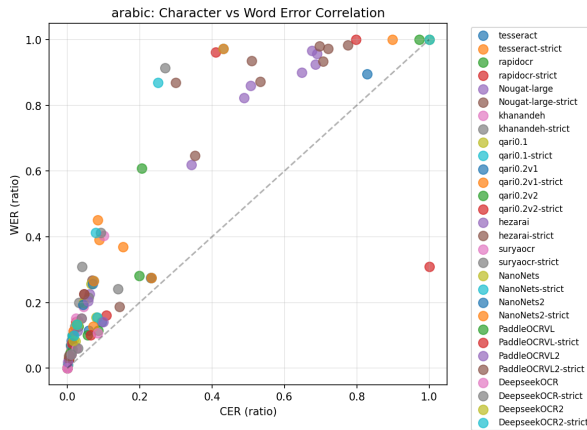


Figure 6: Character vs Word Error Correlation on Arabic Samples

Most models are concentrated in the lower-left region of the plot (CER and WER below 0.4), suggesting generally strong performance on clean printed Arabic text. Variants of DeepSeekOCR, along with certain strict configurations (e.g., qari0.2v2-strict, suryaocr-strict), occupy the lowest region, reflecting minimal error rates. In contrast, older or less specialized systems such as Tesseract and Nougat-large appear more dispersed and are associated with higher error levels.

3.4 Model Performance Ranking

The horizontal bar chart presents a ranking of OCR models and their variants based on mean CER values for Arabic test samples, where shorter bars indicate better performance (Figure 7).

Models such as Tesseract (Smith, 2007; Smith & Reetz, 2024), Nougat-large (Blecher et al., 2023), and several non-strict configurations are positioned toward the lower end of the ranking, reflecting comparatively higher error rates. Overall, the results suggest that recent vision-language models, particularly when fine-tuned or guided by stricter prompting strategies, offer superior performance on Arabic text. DeepSeekOCR’s consistently strong results indicate effective handling of diacritics,

ligatures, and right-to-left script flow, even with limited post-processing.

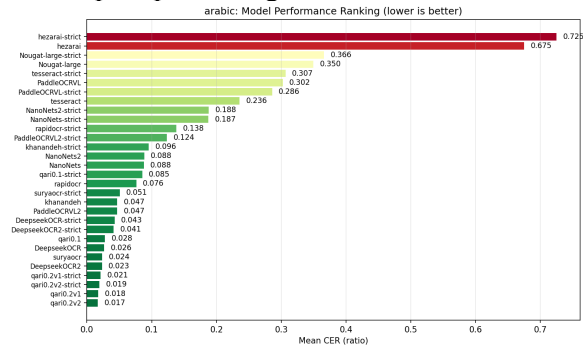


Figure 7: Model Performance Ranking on Arabic (Lower Mean CER = Better).

These outcomes are particularly relevant for translation applications, where higher character accuracy contributes directly to cleaner input for neural machine translation systems and improved preservation of semantic detail. As demonstrated by Darwish et al. (2017), even minor character-level errors in morphologically rich languages like Arabic can propagate through MT systems and alter semantic interpretation significantly.

3.5 Ablation Study

To assess the individual contribution of each component in the proposed OCR pipeline, a systematic ablation study was conducted on the Persian test set consisting of 120 samples. Five experimental configurations were evaluated, each removing or modifying a specific component of the pipeline in order to measure its impact on overall recognition performance:

- **(A) Full pipeline:** Preprocessing (including numeral conversion) → DeepSeek layout analysis → fine-tuned Qari → regex-based normalization → LLM-based correction.
- **(B) Without numeral conversion:** Identical to the full pipeline, except that digit normalization was omitted and Latin digits were preserved.
- **(C) Without regex normalization:** Identical to the full pipeline, but the rule-based character normalization stage was removed.
- **(D) Without LLM correction:** Identical to the full pipeline, but without the contextual correction stage performed by the language model.
- **(E) Without fine-tuning:** The complete pipeline was retained, but the fine-tuned Qari model was replaced with the

original Qari 0.2 base model, without LoRA adaptation.

Table 1 summarizes the results in terms of Character Error Rate (CER) and Word Error Rate (WER).

Table 1. Ablation study results on the Persian test set

Configuration	CER (Mean)	WER (Mean)	Relative Change vs. Full Pipeline
(A) Full pipeline	0.0224	0.107	—
(B) Without numeral conversion	0.0242	0.116	+8% CER, +8% WER
(C) Without regex normalization	0.0294	0.142	+31% CER, +33% WER
(D) Without LLM correction	0.0291	0.139	+30% CER, +30% WER
(E) Without fine-tuning (base Qari)	0.0593	0.287	+165% CER, +168% WER

The results clearly demonstrate that model fine-tuning provides the largest performance improvement. Replacing the base Qari model with the fine-tuned version reduces CER from 0.0593 to 0.0224, corresponding to a relative reduction of approximately 62%. Furthermore, both regex-based normalization and LLM-based correction contribute substantially to the overall accuracy, each yielding around a 30% relative improvement compared with configurations in which these stages are removed. In contrast, numeral conversion produces a smaller but still measurable gain of roughly 8%.

Overall, the ablation study confirms that each component of the proposed pipeline contributes positively to recognition performance. The findings also support the effectiveness of the two-stage post-processing strategy, in which rule-based normalization first resolves systematic character-level inconsistencies, while the LLM subsequently addresses context-sensitive linguistic and morphological ambiguities that cannot be captured through handcrafted rules alone.

4 Discussions on Experimental Results

The experimental results provide a detailed overview of how targeted customization can improve OCR performance for Persian and Arabic scripts, while also indicating that script-specific challenges are not yet fully resolved. A central outcome is the performance of the adapted DeepSeekOCR model, which consistently exceeds average results across major evaluation metrics, demonstrating the

effectiveness of data-efficient fine-tuning on relatively small datasets such as Arshasb.

Our optimized pipeline achieves a mean CER of 0.0224 for Persian and 0.0265 for Arabic. To contextualize these results: [Altwaijry & Al-Turaiki \(2021\)](#) reported character accuracies of 90–95% on printed Arabic text using CNN-based recognition ($\text{CER} \approx 0.05\text{--}0.10$). [Khosrobeigi et al. \(2022\)](#) achieved approximately 96% character accuracy ($\text{CER} \approx 0.04$) on Persian using bidirectional LSTM models. [Asadi-zeydabadi et al. \(2023\)](#) reported 78.49% accuracy with CRNN and 81.32% with Vision Transformer on their Persian dataset. Our improvements are particularly notable in three areas: (1) reduction of the Persian–Arabic performance gap from approximately 0.04 CER in baseline models to 0.004 in our fine-tuned system, (2) lower standard deviation across test samples (0.0160 for Persian, 0.0113 for Arabic) indicating greater robustness, and (3) LoRA-based fine-tuning achieving these gains with only two epochs of training on 16,000 samples, demonstrating data efficiency superior to prior work requiring larger annotated corpora.

The broader statistical analysis reinforces these observations. Across all evaluated systems, Persian presents greater difficulty, with average word error rates around 0.38 compared to approximately 0.30 for Arabic. This discrepancy likely reflects the imbalance in pre-training data, where many vision-language models remain more extensively trained on Arabic. The adapted DeepSeekOCR model reduces this gap through full-page fine-tuning using a LoRA configuration on 16,000 training instances. As a result, the final CER values for Persian and Arabic are relatively close (0.0224 and 0.0265, respectively), whereas baseline systems such as Qari display more pronounced divergence (e.g., 0.0602 for Persian versus 0.0177 for Arabic). Additional evaluation using chrF2 scores supports these findings, with averages of 94.66 for Persian and 92.25 for Arabic, indicating robust handling of morphological complexity.

5 Final remarks

This study set out to answer two primary questions: (1) Can a translation-oriented OCR pipeline achieve character error rates below 0.05 for Persian and Arabic? (2) To what extent can targeted model adaptation close the performance gap between these two scripts? The results demonstrate that OCR performance for Persian and Arabic can be significantly improved

through targeted preprocessing and model adaptation, achieving a mean CER of 0.0224 for Persian and 0.0265 for Arabic—well below our target threshold. Furthermore, fine-tuning reduced the Persian–Arabic performance gap from approximately 0.04 CER in baseline models to just 0.004. However, challenges remain, particularly in handling noisy scanned inputs and structurally complex script variations. A key finding is that normalization plays a critical role in stabilizing OCR output, often transforming high-error predictions into usable text for translation workflows. This is particularly important in CAT environments, where consistency of input directly affects translation memory quality.

Another important observation is the asymmetry between Persian and Arabic performance in baseline models, which appears to be linked to training data bias. Fine-tuning reduces this gap, indicating that relatively small domain-specific datasets can still provide meaningful improvements when combined with parameter-efficient adaptation techniques.

From a translation perspective, OCR should not be treated as an isolated recognition task but as an integral component of the broader translation pipeline. Errors at the character level propagate through MT systems and can significantly affect semantic accuracy, especially in morphologically rich languages.

References

- Altwaijry, N., & Al-Turaiki, I. 2021. Arabic handwriting recognition system using convolutional neural network. *Neural Computing and Applications*, 33(7), 2249–2261. <https://doi.org/10.1007/s00521-020-05070-8>
- Asadi-zeydabadi, F., Shabaninia, E., Nezamabadi-pour, H., & Shojaei, M. 2023. Farsi optical character recognition using a transformer-based model. In *Proceedings of the 2023 10th Iranian Conference on Electrical Engineering* (pp. 1–6). IEEE. <https://doi.org/10.1109/ICKE60553.2023.10326255>
- Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (2019). Character region awareness for text detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9365–9374. <https://doi.org/10.48550/arXiv.1904.01941>
- Bergamaschi, S., Nardis, S. D., Martoglia, R., Ruozzi, F., Sala, L., Vanzini, M., & Vigliermo, R. 2022. Novel perspectives for the management of multilingual and multialphabetic heritages through automatic knowledge extraction: The DigitalMaktaba approach. *Sensors*, 22(11), Article 3995. <https://doi.org/10.3390/s22113995>
- Blecher, L., Cucurull, G., Scialom, T., & Stojnic, R. (2023). Nougat: Neural optical understanding for academic documents. *arXiv.org*, *abs/2308.13418*. <https://doi.org/10.48550/arXiv.2308.13418>
- Darwish, K., Jacob, B., & Oard, D. W. (2017). Building a shallow Arabic morphological analyser in one day. *Computational Linguistics*, 43(4), 835–844. https://doi.org/10.1162/coli_a_00300
- Fateh, A., Fateh, M., & Abolghasemi, V. 2023. Enhancing optical character recognition: Efficient techniques for document layout analysis and text line detection. *Engineering Reports*, 5(3), e12832. <https://doi.org/10.1002/eng2.12832>
- Khosrobeigi, Z., Veisi, H., Hoseinzade, E., & Shabanian, H. 2022. Persian optical character recognition using deep bidirectional long short-term memory. *Applied Sciences*, 12(22), Article 11760. <https://doi.org/10.3390/app122211760>
- Rice, S. V., Kanai, J., & Nartker, T. A. (1995). *The third annual test of OCR accuracy* (ISR-95-1928).
- Riggan, B. S., Hu, S., Short, N. J., & Ochoa, J. 2022. Multi-script handwriting recognition using segmental deep learning. *Pattern Recognition*, 121, Article 108207. <https://doi.org/10.1016/j.patcog.2021.108207>
- Shi, B., Bai, X., & Yao, C. (2017). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11), 2298–2304. <https://doi.org/10.1109/TPAMI.2016.2646371>
- Smith, R. (2007). An overview of the Tesseract OCR engine. *Proceedings of the 9th International Conference on Document Analysis and Recognition*. <https://doi.org/10.1109/ICDAR.2007.4376991>
- Smith, R., & Reetz, M. 2024. Tesseract 5.4 release notes. *Tesseract GitHub Repository*. <https://github.com/tesseract-ocr/tesseract/releases/tag/5.4.0>
- Tahir, M. A., Mahmood, N., Khan, S. H., & Javed, M. S. 2020. Nastaliq optical character recognition: A survey. *Multimedia Tools and Applications*, 79(3-4), 1851–1875. <https://doi.org/10.1007/s11042-019-08351-2>
- Wasfy, A., Nacar, O., Elkhateb, A., Reda, M., Elshehy, O., Ammar, A., & Boulila, W. (2025). QARI-OCR: High-Fidelity Arabic Text Recognition through Multimodal Large Language Model Adaptation. *arXiv preprint arXiv:2506.02295*. <https://doi.org/10.48550/arXiv.2506.02295>

A Corpus-based Approach to Translate Specialised Terminology in Literature: A Two-step Protocol

Laura Noriega-Santiáñez

IUITLM/University of Malaga / Blvr. Louis Pasteur, 27,
29010
laura.noriega@uma.es

Gloria Corpas Pastor

IUITLM/University of Malaga / Blvr. Louis Pasteur, 27,
29010
gcorpas@uma.es

Abstract

In the current context of increasing reliance on artificial intelligence (AI) tools, other traditionally valuable resources, such as corpora, risk being underutilised in both translation practice and training. This pilot study proposes and empirically evaluates a corpus-based two-step protocol for translating specialised terms in literary texts in the English-Iberian Spanish language pair. The empirical analysis is based on four excerpts from contemporary novels, compiled into a small-scale dataset and analysed using five parallel corpora and two monolingual corpora. In the first phase, specialised terms are systematically extracted and translated using parallel corpus evidence. In the second phase, terms exhibiting polysemy or variation are analysed in monolingual corpora to refine candidate translations and assess their contextual adequacy. The protocol is evaluated through a combined qualitative error analysis and comparative assessment of translation choices across stages, focusing on terminological accuracy, contextual appropriateness, and consistency. The results indicate that the second phase contributes to measurable improvements in disambiguation and lexical selection, particularly in cases of domain-specific meaning extension in literary contexts. Overall, the study provides empirical evidence for the effectiveness of corpus-based workflows in terminological decision-making in literary translation and discusses implications for translator training and human-AI collaboration.

1 Introduction

The literary work displays the author's genius, encompassing elements that reflect their vision of the world, as well as the stylistic and socio-cultural dimensions shared with readers (Boase-Beier et al., 2018). Hence the translator encounters a diverse set of difficulties, ranging from basic linguistic challenges to more intricate phraseological and stylistic issues (Mitkov et al., 2018; Corpas Pastor et al., 2021; Monti et al., 2024).

Specialised terms are frequently found in literary texts, as they can be employed to portray characters, enhance the realism of fictional worlds or temporal settings, or disseminate scientific knowledge (Santana and Fortea, 2013). For instance, in fantasy or science fiction literature, phraseological variation is commonplace, which poses significant challenges for literary translators (Rodríguez Martínez, 2020). Furthermore, neological and diachronic language often serve to emphasise certain features found in imaginary worlds, thereby adding further layers of complexity to the translation task (Noriega-Santiáñez and Corpas Pastor, 2023).

Adapting the historical, cultural, and linguistic dimensions of the source text is therefore paramount in the training of literary translators (Hernández Rodilla, 2018). For this reason, proper documentation in literary translation plays a crucial role in ensuring the success of the target text (PETRA-E Framework, 2016).

In this scenario, corpora stand out as one of the most comprehensive resources for translation, due to their multidisciplinary functions, which enable the exploration of diverse and contextualised language patterns (cf. Kruger, Munday and Wallmac, 2011; Corpas Pastor, 2012; Fantinuoli

and Zanettin 2015; Wang et al., 2021). However, despite advances in AI-based translation technologies, there is limited empirical research on structured, corpus-based workflows specifically designed for literary translation.

Thus, this paper presents a pilot two-step protocol for the translation of specialised terms in four novels in the English–Spanish language pair, using both parallel and monolingual corpora. The sample consists of two fictional books (a fantasy novel and a realistic novel) and two factual books (a scientific novel and a historical novel). These texts are processed using Sketch Engine, resulting in four electronic excerpts from which specialised terms are systematically extracted according to the genre-specific characteristics of each text.

This study addresses the following research questions:

- (1) To what extent can parallel corpora support the translation of specialised terms in literary texts?
- (2) What is the added value of monolingual corpus consultation for disambiguation and contextual adequacy?
- (3) How does a structured two-step protocol compare to single-stage corpus consultation in terms of translation quality?

To this end, the paper is structured as follows. Section 2 provides a brief overview of current technologies in the translation field, with special emphasis on how corpora are employed to translate literature. Section 3 presents a detailed explanation of the experimental methodology followed in this study, describing data collection and the two-step protocol setup. Section 4 reports the results obtained from the two phases proposed in this pilot study. Section 5 discusses these results and their implications. Finally, Section 6 summarises the main findings and outlines avenues for future research.

2 Related work

The field of Translation has not remained on the sidelines of the emergence of AI-based technologies. In fact, it has benefited significantly from advances arising from the combination of different disciplines, including Natural Language Processing, Corpus Linguistics, Translation Studies, Machine Translation, and Computational Linguistics (Corpas et al., 2021). The growth of digital content, together with the need to translate it into multiple languages, has led to the automation of translation tasks (Rothwell et al., 2023). In this

context, technologies such as machine translation (MT) systems and computer-assisted translation (CAT) tools have become central in the professional landscape, with post-editing emerging as a key practice (Youdale and Rothwell, 2022). More recently, AI-based chatbots powered by large language models (LLMs) have further expanded the translation ecosystem (Li, 2024; Chen et al., 2025).

However, although these technologies have shown impressive improvement in certain textual genres (Chen et al., 2025), they remain limited in handling stylistically marked and context-dependent language, which is particularly relevant in literary translation (Hadley, 2023; Li, 2024). In contrast, more traditional tools such as corpora offer linguistic and contextual support, allowing translators to explore usage patterns and address phraseological challenges (Corpas Pastor, 2012; Fantinuoli and Zanettin 2015). Despite their multiple applications, the use of corpora in this current AI landscape remains relatively limited and has received little attention among users and practitioners in literary translation (Noriega-Santíañez and Corpas Pastor, 2025). For this reason, the following body of work explores in greater depth the potential of corpora in this comparatively underexplored translation field.

Ji (2012) was a pioneer in delineating certain statistical analyses of literary translational corpora through case studies. In particular, the author evaluated Cervantes's *Don Quijote de la Mancha* and two Chinese translations, providing valuable insights into the use of certain linguistic aspects such as archaisms. Moreover, Zubillaga et al. (2015) compiled a parallel corpus to compare the translation of a novel from German into Basque and Spanish, and additionally developed a search engine, leading to promising translation applications. Concerning software applications and the use of corpora, San Segundo (2017) used an ad-hoc corpus tool (CLiC) to analyse a particular stylistic feature, namely suspensions, found in Charles Dickens's *Hard Times*. Thus, the author demonstrated the potential of this technology in literature to identify fixed structural patterns as well as stylistically meaningful patterns. Furthermore, Liu and Afzaal (2021) analysed the style of two English translations of the novel *Hongloumeng* by using parallel corpora, focusing specifically on three-word and four-word lexical bundles. This study provided insights into

translation strategies and helped identify stylistic idiosyncrasies. Moreover, Dimitroulia (2022) investigated the use of corpora in literary translation practice and training, first defining their role in the professional translation sector and subsequently examining their pedagogical implications. Finally, Noriega-Santiañez and Corpas Pastor (2023) used monolingual corpora, i.e. esTenTen2018, CREA, and CORDE, to address translation challenges arising from a fantasy novel. They particularly assessed the MT output of different multiword expressions (MWEs) (namely, canonical, diachronic, neological and discontinuous MWEs) with the help of corpora, demonstrating that these tools can yield acceptable collocational solutions.

This body of work illustrates a range of practical applications of corpora across different stages of the literary translation process, on which the methodological approach proposed in this study is based, as detailed in the following section.

3 Methodology

This section outlines the experimental methodology adopted in this pilot study, with particular emphasis on the selection of the literary texts and corpora, as well as on the design and implementation of the two-step protocol. The following diagram provides an overview of the proposed workflow.

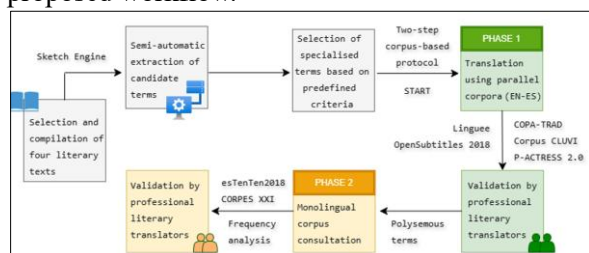


Figure 1. Overview of the two-step corpus-based workflow for the translation of specialised terms

3.1 Data collection

The selection criteria for the novels are defined as follows: the dataset must include (1) two fictional texts and two factual texts of (2) different genres;

(3) written in English; (4) not previously translated into Spanish; (5) from which excerpts of 250–400 words are selected; and (6) published within the last ten years (7) by prestigious publishing houses in different English-speaking countries. Table 1 presents the four selected texts that meet these criteria.

	Title	Author	Year	Publisher	Genre	Country of publication
Factual Texts	<i>A Short History of the Middle Ages</i>	Barbara H. Rosenwein	2018	University of Toronto Press	Historical	Canada
	<i>Nodding Off: The Science of Sleep from Cradle to Grave</i>	Alice Gregory	2020	Bloomsbury Sigma Publications	Scientific non-fiction	United Kingdom
Fictional Texts	<i>Beyond the Ruby Veil</i>	Mara Fitzgerald	2020	Little, Brown Young Readers US	Fantasy	United States
	<i>One Indian Girl</i>	Chetan Bhagat	2016	Rupa Publications India	Realistic fiction	India

Table 1. Characteristics of the selected literary texts

Regarding the corpora used, the parallel-corpus-based phase combines both general and specialised corpora. On the one hand, general corpora such as Linguee¹ and the OpenSubtitles Parallel Corpus 2018 (OpenSubtitles 2018)² were selected due to their broad coverage and their ability to capture natural, dialogue-based language, which can approximate the simulated orality of literary texts (Corpas Pastor, 2018, 2021; Sketch Engine, n.d.; Vargas Castro, 2023). On the other hand, specialised literary corpora such as COPA-TRAD³, Corpus Cluvi⁴, and P-ACTRESS 2.0⁵ were incorporated to provide more domain-specific, professionally aligned translations drawn from literary texts (University of Vigo, n.d.; University of León, n.d.; Universidad Federal de Santa Catarina, n.d.).

In the monolingual-corpus-based phase, two major Spanish reference corpora were used: the Spanish Web Corpus 2018⁶ and Corpus del Español del Siglo XXI (CORPES XXI)⁷. Together, these resources offer extensive coverage of contemporary Spanish, allowing for the analysis of frequency, usage, and contextual adequacy of competing translation options across different domains (Real Academia Española, n. d.).

¹ Available at <<https://www.linguee.es/>>.

² Available at

<<https://www.sketchengine.eu/opensubtitles-parallel-corpora/>>.

³ Available at <<https://copa-trad.ufsc.br/>>.

⁴ Available at <<https://ilg.usc.gal/cluvi/>>.

⁵ Available at <<https://actres.unileon.es/P-ACTRES/login/>>.

⁶ Available at <<https://www.sketchengine.eu/estenten-spanish-corpus/>>.

⁷ Available at <<https://www.rae.es/banco-de-datos/corpes-xxi>>.

3.2 Two-step protocol

Prior to implementing the two-step protocol, a representative sample of 60,000 words from each selected novel was compiled. Each excerpt was then uploaded separately into Sketch Engine, resulting in an individual electronic excerpt per novel. Using the “Keywords” function, a total of 1,000 candidate lexical items were automatically extracted from each dataset.

From this pool, specialised terms were identified through manual filtering based on predefined criteria, namely domain specificity, contextual markedness, and lexicographic validation using monolingual reference dictionaries (*Merriam-Webster*, *Collins*, and *Oxford*). For each text, the selected terms were limited to the ten most frequent specialised terms, resulting in a total of 40 single specialised terms analysed in the study.

In *A Short History of the Middle Ages* (Rosenwein, 2018)⁸, specialised terms were associated with specific historical contexts. In *Nodding Off* (Gregory, 2020), they are classified as scientific-technical, given their domain-specific nature. In *Beyond the Ruby Veil* (Fitzgerald, 2020), the focus is on neologisms, due to the creative linguistic dimension of fantasy literature. Finally, in *One Indian Girl* (Bhagat, 2016), specialised terms are treated as culture-specific elements (culturemes), as they reflect culturally bound Indian concepts.

Phase	Procedure
1 Parallel corpus-based phase	Semi-automatic extraction of ten specialised terms from each four novels
	Translation of these specialised terms (English-Spanish) using five parallel corpora (two general and three specialised)
	Validation of the resulting translations by two professional literary translators
	Statistical analysis of parallel corpus performance per novel
2 Monolingual corpus-based phase	Selection of specialised terms with more than one translation
	Revision of the different translations by using two Spanish monolingual corpora
	Validation of the resulting translations by two professional literary translators
	Statistical analysis of monolingual corpus performance

Table 2. Two-step corpus-based protocol

As shown in Table 2, this pilot study proposes the translation of these terms through a two-phase procedure, each consisting of four steps. The first, parallel corpus-based phase, involves the translation of these terms into Spanish using five parallel corpora (Linguee, OpenSubtitles Parallel Corpus 2018, COPA-TRAD, P-ACTRESS 2.0 and Corpus CLUVI), followed by validation by two professional literary translators with over five years of experience. Translation outputs were assessed using a three-category scheme (i.e. correct solution, non-solution, and incorrect solution), and

inter-annotator agreement was calculated as the degree of convergence between the two evaluators, with discrepancies resolved through discussion.

The second, monolingual corpus-based phase, focuses on specialised terms with multiple valid translations from Phase 1, which are analysed in Spanish monolingual corpora (Spanish Web Corpus 2018 and Corpus del Español del Siglo XXI) using specific linguistic functions to assess frequency and contextual adequacy. Translation selection is guided by frequency of occurrence, and the final choice is subsequently validated by the two translators to ensure contextual appropriateness. Thus, performance was calculated by comparing translation outcomes in terms of correctness (Phase 1) and the resolution of ambiguous cases based on frequency evidence (Phase 2).

4 Results

This section presents the results obtained from the corpus-based pilot study on the translation of specialised terms in literary texts.

4.1 Parallel-corpus-based phase

The analysis of specialised terms reveals significant variation across the different corpora, as illustrated in Figure 1. Positive values indicate correct translations of the ten selected specialised terms per text, whereas negative values represent translation failures, i.e. cases in which the corpus failed to provide an adequate or usable output for a given term.

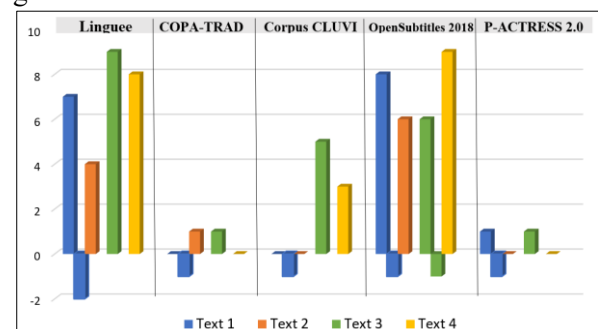


Figure 2. Performance of parallel corpora in the translation of specialised terms across texts

In Text 1, characterised by proper names and culture-specific references (e.g., races or historical figures), OpenSubtitles 2018 clearly outperforms

⁸ *A Short History of the Middle Ages* (2018) by Barbara H. Rosenwein provides a comprehensive account of

the Middle Ages, examining key aspects of European, Byzantine, and Islamic cultures.

the other corpora, followed by Linguee. However, both present issues derived from decontextualisation, such as semantic errors (e.g., *Plate* > *plato* instead of *Platón*) or confusion between ethnonyms and place names (e.g., *Seljuk*). Specialised literary corpora show limited usefulness in this domain and sometimes also produce incorrect translations.

In Text 2, where neologisms and fictional entities prevail (e.g., *watercrea* or *Occhia*), Linguee and OpenSubtitles 2018 again stand out. This is expected, as such invented terms are not documented in specialised literary corpora, which therefore contribute minimally.

In Text 3, which includes specialised terminology related to sleep (e.g., *melatonin*, *REM* or *narcolepsy*), translations are generally stable across corpora. Nevertheless, Linguee proves to be the most reliable corpus, followed by OpenSubtitles 2018, which performs worse due to a higher number of false negatives. Among the specialised literary corpora, Corpus Cluvi clearly stands out, offering the most accurate and consistent handling of this type of vocabulary.

Finally, in Text 4, both Linguee and OpenSubtitles provide satisfactory solutions, often through borrowing or preservation strategies (e.g., *mehndi* > *henna*, *didi* > *didi*). OpenSubtitles slightly outperforms Linguee in this category.

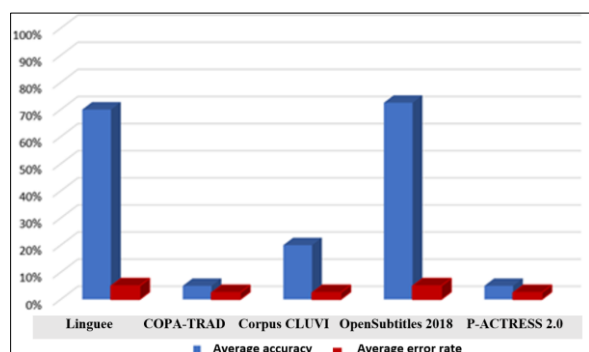


Figure 3. Average accuracy and error rates of parallel corpora in the translation of specialised terms

In Figure 3, accuracy and error rates are calculated based on the proportion of correct and incorrect translations across all analysed specialised terms. Overall, general corpora, particularly Linguee and OpenSubtitles 2018, consistently outperform specialised literary corpora in the translation of specialised terms, although their performance varies depending on the type of text and its domain-specific terms.

OpenSubtitles 2018 achieved the highest average accuracy (72.5%), closely followed by Linguee (70%), both showing low error rates (5% each). In contrast, specialised corpora such as Corpus CLUVI obtained a modest 20% accuracy, standing out mainly for the translation of scientific and technical terms in literary contexts. COPA-TRAD and P-ACTRESS 2.0 performed poorly (with and average accuracy of 5% each), and also exhibited low but non-negligible error rates (2.5% each).

In the subsequent phase, the analysis is refined based on frequency criteria.

4.2 Monolingual-corpus-based phase

Specifically, this phase focuses on specialised terms that presented more than one possible translation in the previous stage, with the aim of identifying the most accurate option in Iberian Spanish according to actual usage.

In this study, a total of ten terms that exhibit multiple lexical realisations were selected, nine of which were from Text 3 and one from Text 4. Searches in esTenTen2018 focused on frequency per million tokens, using lemmatised forms whenever possible, while CORPES XXI provided both normalised frequency and absolute frequency in Spain, thus incorporating a geographical dimension. Values in Figure 4 represent the number of cases in which each corpus supported the selection of the preferred translation equivalent. The results confirm the strong performance of both corpora in identifying the most appropriate translation equivalents, as validated by the professional literary translator. Thus, the second phase of this pilot two-step protocol resulted in improved lexical selection in 85% of cases, confirming the added value of monolingual corpus consultation for disambiguation and contextual adequacy.

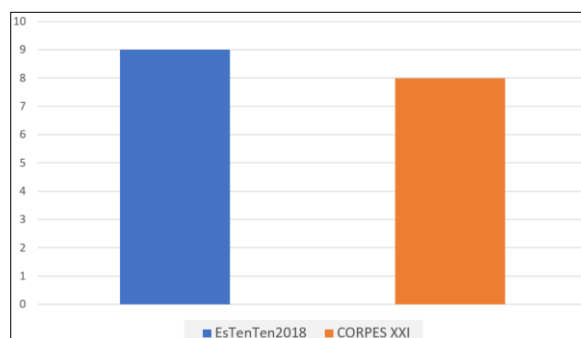


Figure 4. Frequency-based distribution of alternative translations in monolingual corpora (esTenTen2018 and CORPES XXI)

Several cases allowed for the straightforward identification of the most appropriate translation equivalent (e.g., *insomnio* vs. *falta de sueño* [sleeplessness], or *córtex* vs. *corteza cerebral* [cortex]). In these instances, the contrast in usage patterns and contextual examples clearly demonstrates that one candidate term is significantly more conventional and widely attested than the other.

However, some specialised terms presented zero frequency in certain corpora. In other cases, such as *Seljuk* vs. *selyúcidas* [Seljuk], both translational options were attested in the database, albeit with very limited occurrences. In such cases, frequency data alone is insufficient, and the translator may need to rely on additional contextual evidence and external resources to determine the most appropriate equivalent.

5 Discussion

This pilot study confirms that corpus-based methodologies remain highly valuable for addressing specialised terms in literary translation, even in the development of AI-driven translation workflows (Rothwell et al., 2023; Chi et al., 2025). In line with previous research in the field (Ji, 2012; Liu and Afzaal, 2021; Noriega-Santíáñez and Corpas Pastor, 2023), the findings show that corpora continue to provide reliable, usage-based evidence that supports contextual adequacy and transparency. In particular, the initial extraction phase using Sketch Engine proved useful for anticipating translation challenges, as keyword lists offer preliminary insight into the thematic, stylistic, and domain-specific features of each text. Thus, translators can foresee potential translation difficulties, which is consistent with the conclusions reached by Ji (2012) and San Segundo (2017).

In this study, general parallel corpora, i.e. Linguee and OpenSubtitles 2018, demonstrate strong adaptability across literary genres, performing effectively with historical references, technical vocabulary, and culturemes, although they are still prone to decontextualisation and semantic inaccuracies with neologisms, as previously noted by Noriega-Santíáñez and Corpas Pastor (2023). However, more specialised parallel corpora, such as COPA-TRAD, Corpus CLUVI, and P-ACTRESS 2.0, show a more limited performance across genres, proving more effective mainly for technical vocabulary. This limitation

may be partly explained by the nature of their databases, which are not always representative of contemporary literary language. These findings point to the need for the compilation of more comprehensive literary corpora, specifically designed to capture stylistic variation and genre-specific features, as suggested in previous studies on corpus design and literary translation (Fantinuoli & Zanettin, 2015; San Segundo, 2017; Liu and Afzaal, 2021).

Moreover, the monolingual-corpus-based phase highlights the importance of frequency and real usage data for disambiguation. Resources such as esTenTen2018 and CORPES XXI enable translators to identify the most contextually appropriate equivalents, reinforcing the relevance of corpus evidence for stylistic and pragmatic decision-making and the need for corpus literacy, which is in line with several studies on corpus-based teaching (Corpas Pastor 2021; Noriega-Santíáñez and Corpas Pastor, 2023). Thus, these findings suggest that corpus type selection should be task-dependent, with general corpora supporting recall, and monolingual corpora enhancing precision.

6 Conclusion

This study proposes a structured corpus-based protocol for translating specialised terms in literary texts and provides empirical evidence of its effectiveness. The results confirm that corpora constitute a valuable and reliable resource within the translator's workflow, as they provide authentic linguistic evidence that can enhance terminological adequacy and contextual appropriateness. However, they may produce decontextualised or otherwise inaccurate outputs, particularly in general parallel resources. For this reason, corpus consultation should be complemented by the translator's expertise and critical judgement. Thus, the findings contribute to ongoing discussions on human-AI collaboration by highlighting the complementary role of corpus-based methods in supporting informed translation decisions.

This study also presents some limitations, notably in the number of novels analysed and in the scope of the extraction of the specialised terms, so it may not fully capture the complexity of the phraseological dimension.

Future work should expand this pilot study by exploring additional corpus functionalities (e.g., concordances or thesauri), incorporating a wider

range of texts and genres, and further analysing specialised terminology across genre domains. Ultimately, this would serve to design a didactic proposal for a literary translation module to assist future generations in effectively navigating this linguistically rich environment with the help of corpora.

Acknowledgments

This research was funded by a predoctoral contract granted by the University of Malaga (PPRO-A3.1-2023-01) and was conducted within the framework of several research projects on multilingual language technologies: DIFARMA (ref. no. HUM106-G-FEDER, 2024-2025), PROMOTUR (ref. no. PID2024-160929OB-I00) and GAMATUR (ref. no. DGP_PIDI_2024_01159).

References

- Alet Kruger, Kim Wallmach, and Jeremy Munday. 2011. *Corpus-based translation studies: Research and applications*. London: Continuum.
- Belén Santana, and Carlos Fortea. 2013. El hombre de las mil y una caras: el traductor literario como gestor experto de fuentes documentales especializadas. Salamanca: Universidad de Salamanca, pages 135-150.
- Claudio Fantinuoli, and Federico Zanettin. 2015. *New directions in corpus-based translation studies*. Berlin: Language Science Press.
- Gloria Corpas Pastor, M. Rosario Bautista Zambrana, and Carlos. M. Hidalgo-Ternero. 2021. *Sistemas fraseológicos en contraste: enfoques computacionales y de corpus*. Granada: Comares.
- Gloria Corpas Pastor. 2012. Corpus, tecnología y Traducción. In M. García Antuña (ed.), *XII Jornadas de Lingüística*. Cádiz: Servicio de Publicaciones de la Universidad de Cádiz, pages 75-98.
- Gloria Corpas Pastor. 2018. Laughing one's head off in Spanish subtitles: a corpus-based study on diatopic variation and its consequences for translation. In P. Mogorrón and V. Martínez (eds.), *Fraseología, Diatopía y Traducción* (IVITRA Research in Linguistics and Literature). Amsterdam: John Benjamins, pages 54-106. <https://doi.org/10.1075/ivitra.17.03cor>.
- Gloria Corpas Pastor. 2021. Constructional idioms of 'insanity' in English and Spanish: a corpus-based study. *Lingua: International review of general linguistics*, 254.
- Itziar Hernández Rodilla. 2018. Traducir la naturalidad oral: Propuesta didáctica. In C. Fortea (ed.), *El viaje de la literatura*. Madrid: Ediciones Cátedra, pages 121-139.
- James L. Hadley. 2023. MT and CAT: Challenges, irrelevancies, or opportunities for literary translation? In A. Rothwell, A. Way, and R. Youdale (eds.), *Computer-Assisted Literary Translation*. New York: Routledge, pages 91-105.
- Jean Boase-Beier, Lina Fisher, and Hiroko Furukawa. 2018. *The Palgrave Handbook of Literary Translation*. Cham: Macmillan. <https://doi.org/10.1007/978-3-319-75753-7>.
- Johanna Monti, Gloria Corpas Pastor, Ruslan Mitkov, and Carlos M. Hidalgo-Ternero (eds.). 2024. *Recent Advances in Multiword Units in Machine Translation and Translation Technology*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Kanglong Liu, and Muhammad Afzaal. 2021. Translator's Style Through Lexical Bundles: A Corpus-Driven Analysis of Two English Translations of Honglouloumeng. *Frontiers in Psychology*, 12:633422. <https://doi.org/10.3389/fpsyg.2021.633422>.
- Laura Noriega Santiáñez, and Gloria Corpas Pastor. 2023. La traducción del género fantástico mediante corpus y otros recursos tecnológicos: a propósito de "The City of Brass". *Moenia*, 29.
- Laura Noriega-Santiáñez, and Gloria Corpas Pastor. 2025. Uso de la tecnología y de la inteligencia artificial generativa por parte de los traductores literarios en España: un estudio mediante encuesta sobre sus usos, percepciones y actitudes. *tradumàtica*, 23:137-164. <https://doi.org/10.5565/rev/tradumatica.505>.
- Linguee. n.d. *Linguee: Diccionario español-inglés*. <https://www.linguee.es/> (March 12, 2026).
- Luyu Chen, Lilla Varga, and Milad Mehdizadkhani. 2025. Taming AI for The Little Prince: A comparative analysis of NMTs and LLMs in Hungarian translation. *Humanities and Social Sciences Communications*, 13(46). <https://doi.org/10.1057/s41599-025-06343-0>.
- Manuel C. Rodríguez Martínez, 2020. La variación fraseológica intencional como recurso estilístico para la traducción de la ciencia ficción. *Cédille, revista de estudios franceses*, 18:649-664. <https://doi.org/10.25145/j.cedille.2020.18.26>.

- Marco de Referencia PETRA-E. 2017. *Marco de Referencia PETRA-E para la Educación y la Formación de Traductores Literarios*.
- Meng Ji. 2012. Hypothesis testing in corpus-based literary translation studies. In M. P. Oakes and M. Ji (eds.), *Quantitative Methods in Corpus-based Translation Studies: A Practical Guide*. Amsterdam: John Benjamins Publishing Company, pages 53-72.
- Naroa Zubillaga, Zuriñe Sanz, and Ibon Urizarri. 2015. Building a trilingual parallel corpus to analyse literary translations from German into Basque. In C. Fantinuoli and F. Zanettin (eds.), *New Directions in Corpus-Based Translation Studies*. Berlin: Language Science Press, pages 71-93.
- Pablo Ruano San Segundo. 2017. CLiC and Corpus Literary Translation Studies: An Analysis of Suspensions in One Spanish Translation of Charles Dickens's *Hard Times*. *Onomazein*, 38:88-106. <https://doi.org/10.7764/onomazein.38.01>.
- Qi, Li. 2024. Bridging languages: The potential and limitations of AI in literary translation—A case study of the English translation of *A Pair of Peacocks* Southeast Fly. *Advances in Humanities Research*, 8:1-7. <https://doi.org/10.54254/2753-7080/8/2024091>.
- Roy Youdale, and Andrew Rothwell. 2022. Computer-assisted translation (CAT) tools, translation memory, and literary translation. In S. Deane-Cox and A. Spiessens (eds.), *The Routledge Handbook of Translation and Memory* (1st ed.). New York: Routledge. <https://doi.org/10.4324/9781003273417-28>.
- Ruslan Mitkov, Johanna Monti, Gloria Corpas Pastor, and Violeta Seretan (eds.). 2018. *Multiword units in machine translation and translation technology* (Current issues in linguistic theory, 341). Amsterdam & Philadelphia: John Benjamins.
- Sharon Deane-Cox, and Anneleen Spiessens (eds.). 2022. *The Routledge Handbook of Translation and Memory*. New York, NY: Routledge.
- Sketch Engine. n.d. OpenSubtitles parallel corpora. <https://www.sketchengine.eu/opensubtitles-parallel-corpora/> (accessed April 4, 2026).
- Titika Dimitroulia. 2022. Corpora and Literary Translation. In *Advances in Corpus Applications in Literary and Translation Studies* (1st ed.). London: Routledge.
- Universidade Federal de Santa Catarina. n.d. CoPA-Trad: Parallel Corpus for Translation Studies. <https://copa-trad.ufsc.br/> (accessed April 10, 2026).
- University of León. n.d. ACTRES parallel corpora. <https://actres.unileon.es/wp/parallel-corpora/> (accessed April 10, 2026).
- University of Vigo. n.d. CLuVI: Corpus Lingüístico da Universidade de Santiago de Compostela. <https://ilg.usc.gal/cluvi/> (accessed April 10, 2026).
- Vincent X. Wang, Lily Lim, and Defeng Li. 2021. *New Perspectives on Corpus Translation Studies* (New Frontiers in Translation Studies Series). Singapore: Springer.

Fine-tuning Arabic-English NMT Engines for Investment Laws

Najat Al-Huthali, Yuan Zou and Constantin Orasan
Centre for Translation Studies, University of Surrey, UK
{n.alhuthali, y.zou, c.orasan}@surrey.ac.uk

Abstract

The growing volume of legislative and regulatory translation in Saudi Arabia has created an increasing need for reliable Arabic–English machine translation (MT) solutions in the legal domain. However, Neural Machine Translation (NMT) systems trained on general-domain data perform poorly on legal texts due to terminological density, formulaic structures, and jurisdiction-specific drafting conventions. A major obstacle to domain adaptation for Arabic legal MT is the scarcity of large, high-quality, and publicly accessible parallel corpora representing national legislation. This study addresses this gap by constructing a specialized Arabic–English parallel corpus of Saudi investment laws and regulations compiled from official governmental sources. The paper documents the challenges involved in legal text extraction, OCR processing, cleaning, and alignment. Using this corpus, a pilot domain-adaptation experiment is conducted by fine-tuning the OPUS-MT Arabic–English model. Preliminary evaluation indicates that exposure to in-domain legal data improves terminology handling and structural adequacy compared to the baseline generic model.

Introduction

Although Neural Machine Translation (NMT) has achieved strong performance in general domains, its effectiveness declines in specialised fields such as law, where domain-specific terminology and formulaic structures dominate (Chu and Wang, 2018; Koehn and Knowles, 2017). Legal translation is widely recognised as one of the most challenging translation domains due to its system specificity. Studies evaluating MT output for

Arabic legal texts report mistranslation of legal terminology, incorrect rendering of modality, and syntactic distortions (Alkatheery, 2023; Farghal et al., 2025). A widely recognised solution in MT research is domain adaptation, particularly through fine-tuning generic NMT models on in-domain parallel corpora (Chu and Wang, 2018; Koehn and Knowles, 2017). However, the effectiveness of this approach depends heavily on the availability of high-quality, domain-specific bilingual data.

A major obstacle in the Arabic legal MT is the absence of large, accessible Arabic–English legal parallel corpora. Not only that most of the existing Arabic parallel corpora have limited accessibility, but they tend to be general-domain (Al-Ajmi, 2004; Alotaibi, 2017; Tiedemann, 2012) or based on international institutional texts such as United Nations documents or the EU (Eisele and Chen, 2010; Habash et al., 2017; Ziemski et al., 2016). While these datasets are valuable, they do not reflect the linguistic, legal, cultural, and institutional specificities of the national legislation. Moreover, Arabic legal corpora are small in scale, restricted in access, or compiled for purposes other than MT training (BinMasad and Alotaibi, 2024; El-Farahaty et al., 2023). As a result, research on Arabic legal MT has largely focused on error analysis and evaluation of generic systems rather than on building domain-adapted NMT models grounded in locally relevant legal data (Alkatheery, 2023; Farghal et al., 2025).

This gap is particularly significant in the Saudi context. Saudi Arabia’s Vision 2030 aims at expanding the country’s role in global trade and foreign investment. Consequently, the demand for high-quality translation of legal and regulatory texts has increased. The urgency of translating investment laws and regulations is further assured by the royal decree (General Secretariat of the Council of Ministers, 2024) issued in January 2024, which stipulates the translation of all

investment laws and regulations issued by all governmental sectors in Saudi Arabia, especially the laws and regulations related to foreign investments. However, legal translation practices within the translation department at the Bureau of Experts (BOE) continue to rely on human workflows regardless of the increase in the translation volume. A customised, domain-adapted MT system trained on Saudi legal investment laws and regulations could therefore contribute to improving translation efficiency, consistency, and terminological standardization, while supporting human translators' workflows.

Addressing this need requires two interconnected steps: (1) creating a specialised Arabic–English parallel corpus of Saudi investment laws and regulations, and (2) using this resource to adapt a generic NMT model for the legal domain. This study undertakes both tasks. First, it presents the compilation of a large-scale parallel corpus of Saudi investment-related legal texts, documenting the technical and linguistic challenges of data extraction, cleaning, and alignment. Second, it investigates the feasibility of domain adaptation by conducting a pilot fine-tuning experiment using the OPUS-MT Arabic–English model (Tiedemann, 2012). The study evaluates the extent to which exposure to domain-specific legal data improves translation quality compared to the baseline generic model.

The paper makes the following contributions:

1. It builds one of the largest Arabic–English parallel corpora of Saudi legal texts in the investment domain.
2. It proposes a practical pipeline for handling the technical challenges of Arabic legal text extraction and AI-assisted alignment.
3. It provides initial empirical evidence on the impact of domain-specific fine-tuning on NMT performance for Saudi legal translation.

By combining resource creation with model adaptation, this research contributes toward the development of practical MT solutions for Arabic legal translation.

1 Related Work

Section 2 reviews research on Arabic parallel corpora, Arabic legal corpora, and MT of Arabic legal texts. It highlights the main challenges in each

area and identifies the key gaps that motivate this study.

1.1 Arabic Parallel Corpora

Parallel corpora constitute a core resource for both translation studies and data-driven MT. However, Arabic remains comparatively less-resourced in terms of large, high-quality, and publicly accessible bilingual datasets (Alghamdi et al., 2023; Zeroual and Lakhouaja, 2018). Thus, the literature describes a small number of Arabic–English parallel corpora, despite the growing interest among Arab scholars in parallel corpora (Alotaibi, 2017). The complex features of Arabic language represent the main obstacles against its progress in Corpus Linguistics and contribute to making it a low resource language. Arabic is morphologically rich and has different sentence structures. A single word in Arabic often combines pronouns, conjugations, and gender. In addition, Arabic uses diacritics, the addition of short vowels to each word, which affects both pronunciation and meaning and results in two words with identical letters having completely different meanings (Zeroual and Lakhouaja, 2018). Consequently, these features pose significant challenges for Natural Language Processing (NLP), which negatively affects the performance of the corpus compilation applications such as Optical Character Recognition (OCR) programs and alignment tools.

Arabic–English corpora often cannot be accessed as many of them are developed for specific research projects and are not made publicly available (Ahmad et al., 2017; Al-Ajmi, 2004; Alkahtani, 2015; Alotaibi, 2017). Other available Arabic–English corpora consist of either online educational content (Guzman et al., 2013), or religious content (Altammami et al., 2020). Large-scale multilingual initiatives such as Open Parallel Corpus (OPUS) (Tiedemann, 2012) have significantly expanded Arabic parallel data, incorporating subtitles, institutional texts, and legislative materials. Nevertheless, OPUS data are heterogeneous in domain and translation quality, which limits their suitability for specialised legal MT.

1.2 Arabic Legal Corpora

In addition to the main challenge of Arabic being a low-resource language, Arabic legal corpora still face other challenges. Legal translation covers a wide range of text types with distinct linguistic

features. Therefore, it is essential to use corpora that are relevant to the specific subject matter of the study (Giampieri, 2023). For example, a corpus of court document or contracts cannot be used in this research as it focuses on the Saudi investment laws and regulations. Moreover, the confidentiality of legal texts constitutes a challenge for corpus compilers (Giampieri, 2023).

As noted in Section 1, existing legal corpora involving the Arabic language have largely been derived from international institutional texts, such as the United Nations Parallel Corpus, MultiUN, and Arab-Acquis (Eisele and Chen, 2010; Habash et al., 2017; Ziemski et al., 2016). The focus of these corpora on institutional discourse limits their usability for this study, as they do not reflect the drafting conventions of Saudi legislation.

As a result of legal languages being country-dependant, models trained on UN or EU texts may not generalise effective outputs of national legislative contexts such as the Saudi law. To the best of our knowledge, existing Arabic-English legal corpora compiled in regional research contexts are limited in size or have been created mainly for discourse or terminological analysis rather than MT model training (BinMasad and Alotaibi, 2024; El-Farahaty et al., 2023). This gap highlights the need for a specialised Arabic-English parallel corpus of Saudi legal texts. Such a resource is important not only for improving the accuracy of MT, but also for ensuring that models reflect the linguistic and structural features of Saudi legislation.

1.3 Machine Translation of Legal Texts

Most research conducted on Arabic legal MT has focused on evaluating the output of MT systems rather than on improving the quality of the output (Alkhatheery, 2023; Farghal et al., 2025). Common error types identified by this previous research include mistranslation of legal terminology, incorrect rendering of modality, and deviation from TL legal drafting norms. These findings support the view that generic NMT systems trained on mixed-domain data are not optimal for legal translation tasks.

A key solution proposed in MT research is domain adaptation, particularly through fine-tuning pre-trained NMT models on in-domain parallel corpora

(Chu and Wang, 2018; Koehn and Knowles, 2017). Several studies show that even moderately sized, domain-consistent datasets can significantly improve terminology accuracy and stylistic adequacy in specialised translation (Alghamdi et al., 2023; Saunders, 2020). However, this strategy depends on the availability of high-quality domain-specific data, an area where Arabic legal MT remains less-resourced.

1.4 Research Gap

The literature reveals three interconnected gaps:

1. limited availability of large legal Arabic parallel corpora.
2. reliance on international institutional legal texts rather than national Arab legislation; and
3. research focus on evaluating generic MT rather than building domain-adapted NMT systems for Arabic legal translation.

This study addresses these gaps by constructing an Arabic-English legal corpus of Saudi investment laws and regulations and by using this resource to investigate the impact of domain-specific fine-tuning on NMT performance.

2 Saudi Investment Laws and Regulations Parallel Corpus Compilation Process

Section 3 describes the process of constructing the Saudi Investment Laws and Regulations parallel corpus. It outlines the data collection, the extraction of texts, and the challenges encountered in extracting the texts, particularly from Arabic texts in PDF format. It also presents the segmentation and alignment procedures used to produce parallel data, followed by an overview of corpus size and ethical considerations.

2.1 Collecting Data

The corpus compiled for this study consists of official Saudi investment-related laws and regulations and their English translations. Texts were collected from publicly accessible governmental sources to ensure authenticity. The following are the primary sources:

1. The Ministry of Commerce (MOC)¹
2. The BOE²

¹ <https://mc.gov.sa/en/Regulations/Pages/default.aspx> (Accessed on 15/03/2024)

² <https://laws.boe.gov.sa/BoeLaws/Laws/Folders/2> (Accessed 20/02/2024)

3. The Capital Market Authority (CMA)³ which documents were classified based on their

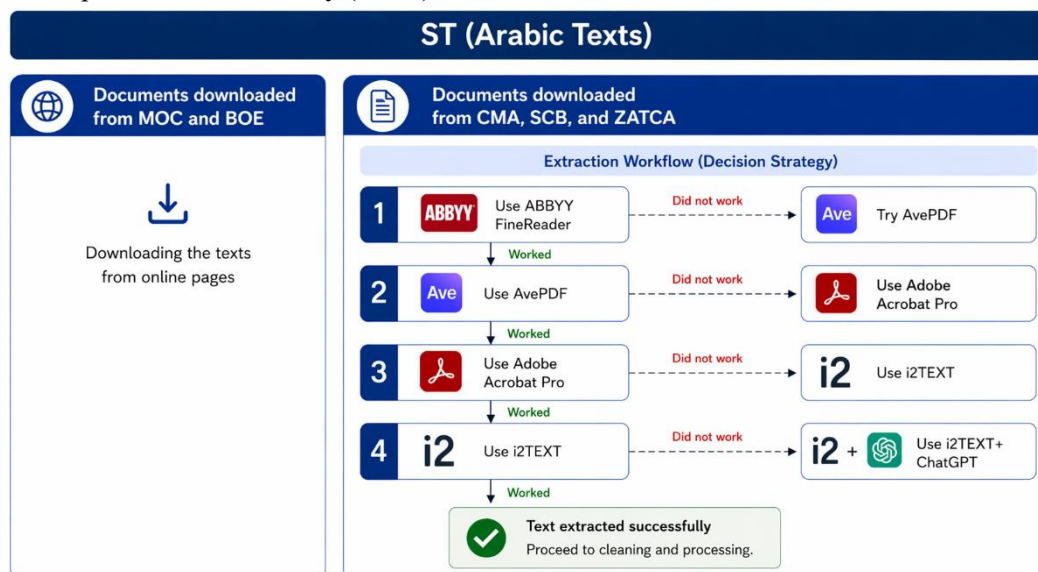


Figure 1: Workflow for Extracting Texts from the STs Documents

4. The Saudi Central Bank (SCB)⁴
5. Zakat, Tax, and Customs Authority (ZATCA)⁵

Additional governmental websites, including the Ministry of Finance and the Ministry of Investment, were initially considered during the data collection phase. However, the documents obtained from the Ministry of Finance were excluded because the Arabic and English versions were not parallel, making them unsuitable for inclusion in a bilingual corpus. Similarly, the documents available on the Ministry of Investment website were found to consist primarily of screenshots of content originally hosted on the BOE website. Therefore, these texts were retrieved directly from the BOE website. A small number of non-investment documents were also downloaded from the said website to ensure the corpus was sufficiently large.

3.2 Data Extraction and OCR Challenges

All the documents collected for the corpus were initially in PDF format. An initial assessment stage followed the document collection stage, during

format and extractability.

Arabic documents were found to fall into two categories: image-based PDFs (primarily from the BOE and the MOC), and text-based PDFs (from the CMA, SCB, and ZATCA). Image-based documents could not be processed using standard OCR tools; therefore, their content was manually retrieved from the corresponding webpages and transferred into editable files. In contrast, text-based PDFs were processed through an OCR and PDF converters workflow.

To support this process, a range of PDF converters and OCR tools were evaluated to identify those capable of extracting Arabic text with minimal manual correction. Freely available tools showed inconsistent performance: HiPDF⁶ failed to generate usable output, while NewOCR⁷ and OCR.best⁸ produced texts with significant spelling and formatting errors. OCRSpace⁹ delivered relatively accurate results but lacked consistency when processing full documents. ChatGPT was also tested for text extraction and cleaning; however, its output was unreliable, often incomplete, or contained hallucinated content. Consequently, ChatGPT was excluded from

³<https://cma.org.sa/RulesRegulations/Pages/default.aspx> (Accessed on 1/04/2024)

⁴<https://www.sama.gov.sa/en-us/lawsregulations/pages/bankinglaws.aspx> (Accessed on 5/04/2024)

⁵<https://zatca.gov.sa/en/RulesRegulations/Pages/rules.aspx> (Accessed on 8/04/2024)

⁶ <https://www.hipdf.com/>

⁷ <https://www.newocr.com/>

⁸ <https://www.ocr.best/>

⁹ <https://ocr.space/>

regular use and retained only as a fallback option for particularly difficult cases. This evaluation demonstrated that no single tool was sufficient, necessitating a multi-tool strategy.

The tool usage workflow was therefore designed as a sequential decision process applied primarily to Arabic texts. Extraction begins with ABBYY FineReader¹⁰; if unsuccessful, AvePDF¹¹ is used as an alternative. When both tools fail, Adobe Acrobat Pro¹² is applied, followed by i2TEXT¹³ for more complex cases. In cases where i2TEXT output contains spelling or structural inaccuracies, a combined i2TEXT and ChatGPT approach is used as a final fallback. Figure 1 presents the workflow adopted for processing and extracting texts from the ST documents.

English translations follow a separate and more streamlined pipeline. As these documents are consistently text-based, Adobe Acrobat Pro was sufficient for extracting text directly from PDF files, after which the texts proceed to the cleaning and processing stage without the need for additional tools.

3.3 Text Preparation and File Management

All cleaned texts were stored as LibreOffice or Word files before conversion to plain text. A structured file-naming convention was adopted to facilitate traceability and alignment. Files were organized by source institution and language, with paired Arabic (ST) and English (TT) files sharing identical identifiers. This systematic organization supported reproducibility and streamlined later processing stages.

Metadata for each document (source URL, title, download date, word counts, and notes on extraction issues) were recorded in Excel spreadsheets. This documentation ensured transparency of corpus construction and allowed problematic files to be tracked during cleaning and alignment. The resulting corpora contain 763,765 Arabic words and 958,581 English words.

3.4 Segmentation and Alignment Process

Accurate alignment is essential for training NMT systems. However, legal texts pose challenges for standard sentence-level alignment due to long

sentences, enumerated clauses, and article-based structures. Moreover, the conversion process resulted in fragmented sentence structures, particularly in Arabic, where sentences were split across multiple lines. This fragmentation hindered the effective use of alignment tools such as LF Aligner and Trados. To address the challenges of aligning Arabic–English legal texts, an automated alignment pipeline was implemented in Python. Source documents in .docx and .odt formats were converted into plain UTF-8 text through XML extraction and cleaning. The texts were then segmented into manageable chunks using predefined delimiters to ensure processing stability. Manual segmentation was preferred over automatic segmentation due to structural differences between Arabic and English legal texts, where clauses are often merged, split, or reordered in translation. Automatic segmentation introduced cascading alignment errors, particularly given the variability in legal article length. Manually controlled segmentation therefore ensured a balance between contextual adequacy and computational efficiency.

Alignment was performed using Google’s Gemini 2.5 Flash model through prompt-guided processing. Structured prompts were manually designed to guide sentence- or clause-level segmentation, handle headings and numbered articles, and align segments based on semantic equivalence rather than surface form. The model returned aligned outputs in structured JSON format, which were automatically validated and stored in a two-column Arabic–English format.

3.5 Corpus Size

Table 1 presents the data on the word count included in the corpus. The corpus includes a total of 193 legal documents, collectively containing 763765 Arabic words and 958581 English words.

¹⁰https://www.abbyy.com/?AFFILIATE=81902&__c=1

¹¹ <https://avepdf.com/?ref=producthunt>

¹² <https://www.adobe.com/acrobat/acrobat-pro.html?msocid=2dc3012bf01268dc2ccc150bf12a693c>

¹³ <https://www.i2text.com/>

Source	Number of documents	Word in Arabic	Word in English
MOC	55	180,059	226,157
BOE	81	211,856	267,949
CMA	33	294,848	367,479
SCB	13	46,633	57,720
ZATCA	11	30,369	39,276
TOTAL	193	763,765	958,581

Table 1: Corpus details

3.6 Ethical Considerations

All texts included in the corpus are official legal documents intended for public dissemination. No personal or confidential data were involved.

3 Experimental Setup and Results

Section 4 explains how the experiments were designed to examine the impact of domain adaptation on Arabic–English legal NMT models. It outlines the baseline models, the preparation of the data, and the fine-tuning procedure. In addition, it describes how the models were evaluated, using both automatic metrics and qualitative analysis, with a focus on legal terminology, modality, and structural conventions.

3.1 Baseline Model

To evaluate the impact of domain adaptation, a baseline NMT system was selected from the OPUS-MT framework (Tiedemann, 2012). Particularly, the Arabic–English model was used as a representative generic NMT system trained on large-scale, heterogeneous parallel data. OPUS-MT was chosen due to its open-source nature, which supports reproducibility and transparency (Tiedemann and Thottingal, 2020), as well as its ability to run locally without reliance on cloud services. This enables secure processing of sensitive legal texts and allows full control over the experimental environment (Tiedemann et al., 2023). In addition, we have also used Google Translate (GT) as the second baseline due to the fact that GT is commonly used by end users. These baselines model therefore reflects the typical performance of publicly available MT systems when applied to specialised legal texts without domain-specific adaptation.

3.2 Data Preparation and Experimental Design

The compiled Saudi Investment Legal Corpus was used to fine-tune the OPUS-MT baseline model. Prior to training, the dataset was divided into training, and test sets using a 90/10 split. The training set was used for fine-tuning, and the test set for evaluation. However, the test set was not randomly sampled but constructed through a corpus-driven procedure to ensure representativeness. It comprises 73,700 words out of 763,765 ($\approx 10\%$ of the dataset).

A monolingual source-text corpus was compiled using Sketch Engine¹⁴ and treated as a focus corpus against arTenTen2024. Keyword analysis generated 150 terms, reduced to 106 domain-specific legal terms after manual filtering. Term distribution was then examined via document frequency, distinguishing low-frequency (1–2 documents) and relatively frequent items (≥ 5 documents). Concordance analysis mapped these terms to their source documents, enabling document-level comparison and selection. Based on this comparison, a subset of texts was identified and subsequently reviewed manually. Eleven documents were selected for the evaluation stage based on two main criteria: topical diversity and word count. To ensure broad coverage of the Saudi investment legal domain, documents were selected from all the sources of the corpus. In addition, texts from different legal subdomains were included to expand terminological coverage. For example, the Law of Commercial Courts, the Law of Arbitration, and the Resolution of Securities Disputes Proceedings Regulations were selected because they contain judicial and procedural terminology, complementing documents focused on investment and financial matters.

Word count was also considered during the selection process. To maximise textual variety within the target test-set size, documents with very high word counts were excluded, as their inclusion would have limited the number of texts that could be represented. Consequently, all selected documents contained fewer than 10,000 words, except for one document of approximately 15,000 words.

To assess representativeness, a reduced corpus of these documents was analysed. Of the 106 terms, 86 were present in the subset. In addition,

¹⁴ <https://www.sketchengine.eu/>

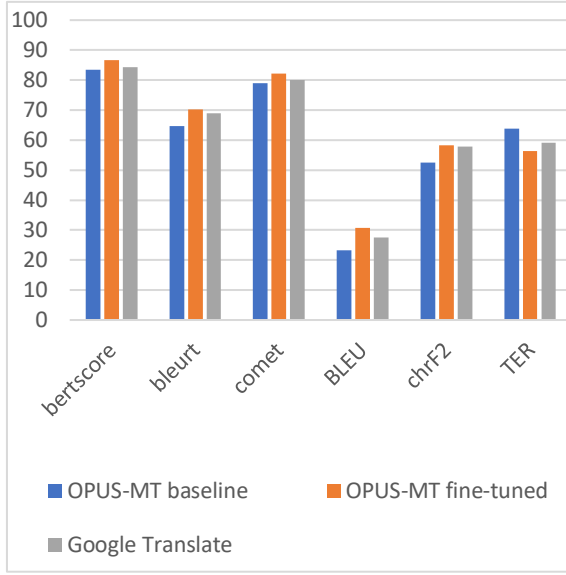


Figure 2: The results of the automatic evaluation metrics

basic linguistic features were verified, including obligation markers (على, لـ), conditional particles (e.g., إذا, عند, متى), passive constructions (يتم), and discourse markers (e.g., فقرة, المشار).

The presence and distribution of these features indicate that the evaluation set captures relevant characteristics of legal discourse, which supports its suitability for legal domain evaluation.

To assess the effect of domain adaptation, two experimental conditions were defined:

1. Baseline conditions: Translation using the pre-trained OPUS-MT model and Google Translate
2. Fine-tuned condition: Translation using the OPUS-MT model trained on the legal corpus

3.3 Fine-Tuning Procedure

OPUS-CAT¹⁵ tool was installed on the computer. Then, the Arabic-English model was downloaded, and fine-tuned using a subset of the compiled corpus to evaluate the feasibility of domain adaptation in a pilot setting. The tool as used “out of the box”, meaning that we made no attempt to adjust any of the parameters of the translation engine. Instead, we used the default settings. The objective of this experiment was not to achieve optimal performance, but rather to examine whether exposure to domain-specific

	OPUS-MT baseline	OPUS-MT fine-tuned	Google Translate
bertscore ($\mu \pm 95\% \text{ CI}$)	83.6 (83.6 $\pm$ 0.2)	86.7 (86.7 $\pm$ 0.2) (p = 0.0010)*	84.4 (84.4 $\pm$ 0.2) (p = 0.0010)*
bleurt ($\mu \pm 95\% \text{ CI}$)	64.7 (64.7 $\pm$ 0.4)	70.2 (70.2 $\pm$ 0.4) (p = 0.0010)*	69.0 (69.0 $\pm$ 0.3) (p = 0.0010)*
comet ($\mu \pm 95\% \text{ CI}$)	79.0 (79.0 $\pm$ 0.3)	82.1 (82.1 $\pm$ 0.3) (p = 0.0010)*	80.0 (80.0 $\pm$ 0.3) (p = 0.0010)*
BLEU ($\mu \pm 95\% \text{ CI}$)	23.3 (23.3 $\pm$ 0.5)	30.7 (30.7 $\pm$ 0.6) (p = 0.0010)*	27.5 (27.5 $\pm$ 0.5) (p = 0.0010)*
chrF2 ($\mu \pm 95\% \text{ CI}$)	52.5 (52.5 $\pm$ 0.4)	58.2 (58.2 $\pm$ 0.4) (p = 0.0010)*	57.8 (57.8 $\pm$ 0.4) (p = 0.0010)*
TER ($\mu \pm 95\% \text{ CI}$)	63.9 (63.9 $\pm$ 0.7)	56.3 (56.3 $\pm$ 0.8) (p = 0.0010)*	59.2 (59.2 $\pm$ 0.8) (p = 0.0010)*

Table 2: The results of the automatic evaluation metrics, including the confidence intervals, calculated using MATEO.

legal data leads to measurable improvements in translation quality.

3.4 Evaluation Metrics

Translation quality was evaluated using both automatic metrics and qualitative analysis. We used MATEO¹⁶ to compute scores across a range of automated evaluation metrics. In addition, a qualitative error analysis was conducted to assess improvements in:

1. Legal terminology
2. Modality rendering (e.g., shall, may)
3. Structural correspondence to legal drafting conventions

3.5 Results

3.5.1 Automatic Evaluation

The results of the automatic evaluation are presented in Table 2 and Figure 2. The fine-tuned model shows an improvement over both baseline models across all metrics, indicating that domain adaptation using the compiled legal corpus has a positive effect on translation quality, even with limited training.

4.5.2 Qualitative Analysis

¹⁵ <https://github.com/Helsinki-NLP/OPUS-CAT>

¹⁶ <https://mateo.ivdnt.org/>

To better understand the impact of the fine tuning on the quality of the translation, a qualitative comparison of a randomly selected 600-word sample showed that the fine-tuned OPUS model outperformed the baseline model. Across multiple examples, the following improvements were observed:

1. More consistent use of domain-specific terminology
2. Improved rendering of legal modality.
3. Better correspondence of article numbering.

Each of these are discussed next.

More consistent use of domain-specific terminology

ST	Human Translation	OPUS model	Fine-tuned OPUS model
نظام ضريبة الدخل	Income Tax Law	Income tax system	Income Tax Law
شركة الأموال: الشركة المساهمة، أو الشركة ذات المسؤولية المحدودة، أو شركة التوصية بالأسهم	Capital company: A joint stock company, limited liability company, or a company limited by shares.	Money Company: Contributing Company, Limited Liability Company or Equity Recommendation Company	Money Company: A joint-stock company, limited liability company, or a joint-stock company.
شركة الأشخاص: الشركة التضامنية، أو شركة المحاصة، أو شركة التوصية البسيطة	Partnership: A general partnership, a silent partnership, or a limited partnership.	People's Company: Solidarity Company, Sewerage Company or Simple Recommendation Company.	Persons Company: A general partnership, a partnership, or a limited partnership.

Figure 3: Examples of the use of domain-specific Terminology

The results indicate that domain adaptation improves both the adequacy and register of legal machine translation, by enhancing the model's handling of domain-specific terminology. In the sample evaluated, the fine-tuned OPUS model produced more accurate terminology in 35 instances. Legal language is inherently complex, as each branch of law employs distinct terminology, and many legal terms fall into three categories: purely technical, semi-technical, and everyday vocabulary (Alcaraz and Hughes, 2002; Al-Huthali, 2021; El-Farahaty, 2015). Machine Translation systems tend to perform better on purely technical terms when sufficient domain data is available but frequently fail with semi-technical and everyday terms due to their polysemous meanings. This tendency is reflected in the baseline outputs, which often rely on general-language interpretations. As Figure 3 shows, *نظام ضريبة الدخل* is rendered as “Income tax system” instead of the

legally appropriate “Income Tax Law,” while *شركة الأموال* is mistranslated as “Money Company,” reflecting a literal interpretation rather than a legal one. Similarly, *شركة الأشخاص* appears as “People's Company,” demonstrating failure to capture the specialised legal sense. Following fine-tuning, the model shows improved sensitivity to legal usage, correctly producing “Income Tax Law” and more accurate subcomponents such as “joint-stock company,” “limited liability company,” and “general partnership.” These improvements suggest that exposure to domain-specific data enables the model to better disambiguate polysemous terms and align with legal conventions. However, existing inaccuracies in higher-level categories (e.g., “Persons Company”) indicate that while fine-tuning improves lexical and phrasal choices, fully capturing abstract legal classifications requires larger and more specialised training data.

Improved rendering of legal modality

ST	Human Translation	OPUS model	Fine-tuned OPUS model
لجنة المختصة اتخاذ إجراءات	The competent authority may take resolution procedures	The competent authority shall take remedial action	The Competent Authority may take remedial actions
تخضع أحكام النظم	This Law shall apply	They are subject to the provisions of the regulations;	The following shall be subject to the provisions of this Law:
تظم اللوائح التنفيذية العلاقة	The Implementing Regulations shall regulate the relationship	The implementing regulations regulate the relationship	The Implementing Regulations shall regulate the relationship

Figure 4: Examples of modality

Modality in Arabic legal language is primarily conveyed through finite verbs in the present tense, along with certain prepositional constructions such as *لـ* (li-) and *على* ('alā). Even though it might be easy for professional translators to translate legal modality between Arabic-English, MT systems often fail to recognize the special use of present tense verbs and prepositions in legal language. The fine-tuned OPUS model was able to correctly render legal modality in 5 instances in the sample evaluated. Figure 4 shows some examples where the OPUS model failed in translating Arabic modality, while the fine-tuned OPUS model succeeded in translating them.

Better correspondence of article numbering

The OPUS model uses the roman numbers (i.e., I, II, III) in numbering the articles while the fine-tuned OPUS model uses the English numbers (i.e., 1,2,3). This adjustment improves the consistency with the TL norms and reduces the need for post-editing by human translators.

3.6 Discussions and Conclusions

The results of this pilot experiment support findings in previous research on domain adaptation (Chu and Wang, 2018; Koehn and Knowles, 2017), indicating that even limited amounts of in-domain data can improve NMT performance in specialised domains.

However, several limitations remain. First, the fine-tuning process was conducted on a subset of the corpus due to resource constraints, and further improvements are expected with full-scale training. Secondly, automatic evaluation metrics may not fully capture improvements in legal domain, highlighting the need for human evaluation in future work. Finally, the results of this pilot experiment are limited to fine-tuning OPUS model. Fine-tuning other NMT models such as Google Translate and Azure AI Translator may provide better results and is considered in future work.

Acknowledgments

This research is funded by The Ministry of Education in Saudi Arabia and the Saudi Cultural Bureau in London. The authors would like to thank Safinaz Ahmad for her valuable assistance in the alignment process.

References

- Ahmad, A.A.-S., Hammo, B., Yagi, S., 2017. ENGLISH-ARABIC POLITICAL PARALLEL CORPUS: CONSTRUCTION, ANALYSIS AND A CASE STUDY IN TRANSLATION STRATEGIES. *Jordanian Journal of Computers and Information Technology (JJCIT)* 3.
- Al-Ajmi, H., 2004. A New English-Arabic Parallel Text Corpus for Lexicographic Applications 326–330.
- Alcaraz Varó, E., Hughes, B., 2002. *Legal translation explained, Translation practices explained*. St. Jerome Pub, Manchester, UK; Northampton, MA.
- Alghamdi, E.A., Zakraoui, J., Abanmy, F.A., 2023. *Domain Adaptation for Arabic Machine Translation: The Case of Financial Texts*. <https://doi.org/10.48550/arXiv.2309.12863>
- Al-Huthali, N., 2021. *Characteristics of the Legislation Language in Saudi Criminal Laws*. Umm Al-Qura University, Makkah.
- Alkahtani, S., 2015. *Building and Verifying Parallel Corpora between Arabic and English*. Bangor University, UK.
- Alkatheery, E., 2023. Google Translate Errors in Legal Texts: Machine Translation Quality Assessment. *Arab World English Journal for Translation and Literary Studies*, 7, 208–219. <https://doi.org/10.24093/awejtls/vol7no1.16>
- Alotaibi, H., 2017. Arabic-English Parallel Corpus: A New Resource for Translation Training and Language Teaching. *Arab World English Journal*, 8, 319–337. <https://doi.org/10.24093/awej/vol8no3.21>
- Altammami, S., Atwell, E., Alsalka, A., 2020. The Arabic–English Parallel Corpus of Authentic Hadith. *International Journal of Islamic Applications in Computer Science and Technology*, 8.
- BinMasad, S., Alotaibi, H.M., 2024. Exploring Cross-Cultural Legal Terminology: A Corpus-Based Study of English Translation in Saudi Laws. *Theory and Practice in Language Studies*, 14, 1151–1162. <https://doi.org/10.17507/tpls.1404.23>
- Chu, C., Wang, R., 2018. A Survey of Domain Adaptation for Neural Machine Translation. <https://doi.org/10.48550/arXiv.1806.00258>
- Eisele, A., Chen, Y., 2010. MultiUN: A Multilingual Corpus from United Nation Documents, in: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (Eds.), *Proceedings of the Seventh International Conference on Language Resources*

- and Evaluation (LREC'10). Presented at the LREC 2010, European Language Resources Association (ELRA), Valletta, Malta.
- El-Farahaty, H., 2015. *Arabic-English-Arabic legal translation*. Routledge, Abingdon, Oxon ; New York, NY.
- El-Farahaty, H., Khallaf, N., Alonayzan, A., 2023. Building the Leeds Monolingual and Parallel Legal Corpora of Arabic and English Countries' Constitutions: Methods, Challenges and Solutions. *Corpus Pragmatics*, 7, 103–119. <https://doi.org/10.1007/s41701-023-00138-x>
- Farghal, T., Shraideh, K., Al-Omari, A.M., 2025. Evaluating Free Legal Translation Tools between Arabic and English: A Comparative Study of Google Translate, ChatGPT, and Gemini. *International Journal for the Semiotics of Law*, 38, 2083–2110. <https://doi.org/10.1007/s11196-025-10282-z>
- General Secretariat of the Council of Ministers, 2024. *Resolution No.483*. National Center for Archives and Records. URL <https://ncar.gov.sa/> (accessed 5.12.25).
- Giampieri, P., 2023. *Legal machine translation explained: MT in legal contexts*. Cambridge Scholars Publishing, Newcastle upon Tyne, UK.
- Guzman, F., Sajjad, H., Vogel, S., Abdelali, A., 2013. The AMARA corpus: building resources for translating the web's educational content, in: Zhang, J.Y. (Ed.), *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*. Heidelberg, Germany.
- Habash, N., Zalmout, N., Taji, D., Hoang, H., Alzate, M., 2017. A Parallel Corpus for Evaluating Machine Translation between Arabic and European Languages, in: Lapata, M., Blunsom, P., Koller, A. (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Presented at the EACL 2017, Association for Computational Linguistics, Valencia, Spain, pp. 235–241.
- Koehn, P., Knowles, R., 2017. Six Challenges for Neural Machine Translation, in: *Proceedings of the First Workshop on Neural Machine Translation. Presented at the Proceedings of the First Workshop on Neural Machine Translation, Association for Computational Linguistics, Vancouver*, pp. 28–39. <https://doi.org/10.18653/v1/W17-3204>
- Saunders, D., 2020. *Domain adaptation for Neural Machine Translation*. University of Cambridge, UK.
- Tiedemann, J., 2012. Parallel Data, Tools and Interfaces in OPUS.
- Tiedemann, J., Aulamo, M., Bakshandaeva, D., Boggia, M., Grönroos, S.-A., Nieminen, T., Raganato, A., Scherrer, Y., Vazquez, R., Virpioja, S., 2023. *Democratizing Neural Machine Translation with OPUS-MT*. <https://doi.org/10.48550/arXiv.2212.01936>
- Tiedemann, J., Thottingal, S., 2020. OPUS-MT – Building open translation services for the World, in: Martins, A., Moniz, H., Fumega, S., Martins, B., Batista, F., Coheur, L., Parra, C., Trancoso, I., Turchi, M., Bisazza, A., Moorkens, J., Guerberof, A., Nurminen, M., Marg, L., Forcada, M.L. (Eds.), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Presented at the EAMT 2020, European Association for Machine Translation, Lisboa, Portugal, pp. 479–480.
- Zeroual, I., Lakhouaja, A., 2018. Arabic Corpus Linguistics: Major Progress, but Still a Long Way to Go, in: Shaalan, K., Hassanien, A.E., Tolba, F. (Eds.), *Intelligent Natural Language Processing: Trends and Applications, Studies in Computational Intelligence. Springer International Publishing, Cham*, pp. 613–636. https://doi.org/10.1007/978-3-319-67056-0_29
- Ziemski, M., Junczys-Dowmunt, M., Pouliquen, B., 2016. The United Nations Parallel Corpus v1.0, in: Calzolari, N., Choukri,

K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC`16). Presented at the LREC 2016, European Language Resources Association (ELRA), Portorož, Slovenia*, pp. 3530–3534.

#MTFails to #MTWins: Investigating User Perspectives on Machine Translation in Social Media

Khetam Al Sharou

Dublin City University / University of
Damascus / Imperial College London

k.al-sharou@imperial.ac.uk

Abstract

This work investigates the experience of Arabic and English-speaking users with Machine Translation (MT) on social media, looking into how users engage with, and perceive, automated translation tools. It explores the challenges users face as well as the benefits of overcoming language barriers in real-time communication, as reflected in their social media posts. The study also examines how MT affects social media engagement and access to information, providing insights into how these tools shape online interactions and influence user satisfaction. Tweets from X platform by Arabic and English speakers were collected. Manual sentiment analysis classified tweets as Neutral, Positive, and Negative, then specific emotions within each sentiment group were identified to provide a detailed understanding of users' emotional responses to MT and cross-linguistic comparisons and patterns in how users engage with MT over time. The study highlights that users continue to have doubts about the effectiveness of MT in facilitating multilingual and online communication. It emphasises the importance of addressing users' perceptions and misconceptions about MT, as these can significantly affect the acceptance and adoption of emerging technologies including generative AI.

1 Introduction

The impact of machine translation (MT) on social media has received attention due to its growing significance in today's interconnected digital

world (Gupta *et al.*, 2023; Zhang *et al.*, 2022; Niño, 2020). MT tools are increasingly employed by social media platforms to promote cross-cultural communication, expand the reach of information, and enhance interactions among users (Fakih *et al.*, 2024; Nuraliyeva, 2024). However, existing work has often neglected the experience of users who resort to MT on an *ad hoc* basis and their input that can help improve these systems (Fukuda and Jin, 2022). Users tend to express emotions on social media platforms, where they openly share personal stories and connect with others (Zhao and Zhang, 2017). This study offers data-based insights into how users perceive, use, and think about MT by analysing tweets posted by Arabic and English-speaking users.

This research work aims to: **1.** Understand users' general sentiment towards MT and their emotions to identify the level of users' satisfaction and/or dissatisfaction and how that has affected their online experience; **2.** Investigate ways in which users engage with MT and what that reveals about their awareness of its limitations and usability; **3.** Identify issues users face and how they influence their experiences and online communication and whether they can shape their attitudes towards new technological advances. Understanding user feedback is essential for evaluating MT applicability, usability, and user attitudes towards these services. Such insights guide future development of translation technology for more

effective global communication, extending beyond social media to wider real-world applications.

The paper first presents a review of studies on user experience with MT. It then outlines the research methodology, including data collection and analysis, followed by the results and discussion. The paper concludes by summarising the findings, addressing limitations, and suggesting directions for future research.

2 Existing Research

With millions of users interacting with MT tools in real-time, new questions come to light regarding their experience with MT, particularly in informal and diverse contexts such as social media. While some recent studies, such as Kasperè *et al.* (2021) and Vieira *et al.* (2023), have assessed the usability and satisfaction of MT among users, the focus has generally been on the technology's performance in handling basic translation tasks, targeting a small number of users. These contributions acknowledge users' dissatisfaction with MT quality (*e.g.*, Kasperè *et al.*, 2021). However, they investigate less how users express their frustrations, desires, or expectations regarding the technology, especially in online environments, where communication is not just functional but also emotional, social, and context-driven. Furthermore, Al Sharou *et al.* (2025) apply large-scale statistical analysis to user feedback, limiting the study to general patterns in user experience. The present work complements this by focusing on individual perspectives and how users describe MT in real-life communication.

Additionally, the experience of users who speak languages other than English, such as Arabic, has been largely understudied. While there is a substantial body of research assessing the quality and accuracy of MT for Arabic content on social media (Al-Kharusi and AAlAbdulsalam, 2023; Alzamzami and El Saddik, 2023; Mallek *et al.*, 2017), their experience with these technologies is still unexplored. Their feedback, particularly in their native languages, could offer unique insights into both the potential and limitations of these tools and highlight areas for improvement. This lack of research prevents a deeper understanding of the cultural and linguistic specifications that affect user satisfaction with MT, particularly in contexts where language use is informal and often highly emotional. This gap in research is especially critical given the challenges MT encounter when translating user-generated content into and from

Arabic (see discussion in Saadany *et al.*, 2024; Al Sharou and Specia, 2022; Al Sharou *et al.*, 2021). Arabic, with its complex grammatical structures, dialectal variations, and cultural specifications, often presents difficulties for MT systems, leading to mistranslations and hallucinations (*ibid.*).

3 Research Methodology

This section outlines the data collection methods and analytical approach adopted to investigate user experience with MT on X platform. Given the wide adoption by users, social media platforms provide a rich source of information where users discuss opinions and experiences (Hornikx and Hendriks, 2015; Pak and Paroubek, 2010). In particular, X platform can offer insightful opinions on products (Lim and Buntine, 2014) by functioning as a form of electronic word-of-mouth, where consumers share their reactions and opinions about products or services (Jansen *et al.*, 2009).

3.1 Data Collection Approach

The study utilises X platform to better understand user experience with MT. Google Translate is the provider of translation service on X, with an option to manually translate tweets through a button. The data collection process was designed to ensure that only relevant and insightful content referring to user experience with MT was included. A list of keywords was used to collect tweets by Arabic and English-speaking users, facilitating a comparative analysis of user experiences across different linguistic contexts. The list includes both general terms (*e.g.*, translation tools, instant translation, technical translation) and tool-specific terms (*e.g.*, Google Translate, Microsoft Bing, MT, DeepL, GTranslate, Bing Translator, and Reverso Context). The two datasets cover tweets from 2009 to 2024. Tweets in English were gathered through NewsWhip Spike platform. From an initial dataset of 5,121 tweets, 500 were manually selected for further analysis. Tweets in Arabic were manually collected from X, resulting in a dataset of 348 tweets in Arabic. The manual selection process was based on a set of criteria, ensuring that only relevant tweets were included: 1) only tweets in English and Arabic that explicitly referenced user experience with MT were retained. 2) Tweets were discarded if they were advertisements, promotional content, or repetitive tweets. This cleaning process ensured that the datasets

accurately reflected authentic user experiences, providing a meaningful basis for analysis.

3.2 Data Analysis

A manual analysis approach was adopted to ensure accuracy and depth in data interpretation. While automated data annotation is used for its efficiency, scalability, and cost-effectiveness, manual annotation offers a layer of human judgment, especially when dealing with context-dependent content. For sentiment analysis, manual methods capture the underlying tone and emotion that automated systems may overlook. For instance, tweets that appear positive due to the use of words and emojis, such as “😂” or “Wow,” could in fact be sarcastic or critical of MT. First, tweets were categorised into three broad sentiment groups: Neutral, Positive, and Negative. This initial classification provided an overview of general user sentiment towards MT. After that, specific emotions within each sentiment group were identified, offering a more detailed overview of users’ emotional responses to MT. Understanding these emotions is important for identifying factors and patterns that could have influenced their overall experience with MT. The data analysis was conducted using Excel, which allowed efficient analysis of X data.

4 Results and Discussion

4.1 Sentiment analysis: neutral, positive, and negative

The sentiment analysis of the X data (Figure 1 and Figure 2) reveals a predominantly negative user experience with MT tools across both Arabic and English datasets, with a high proportion of negative sentiment (64% and 43% respectively).

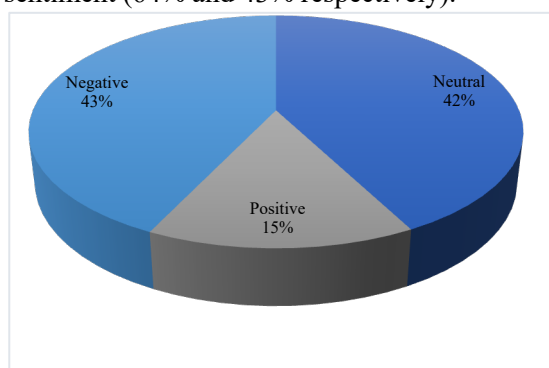


Figure 1: Sentiment Analysis of Tweets in English

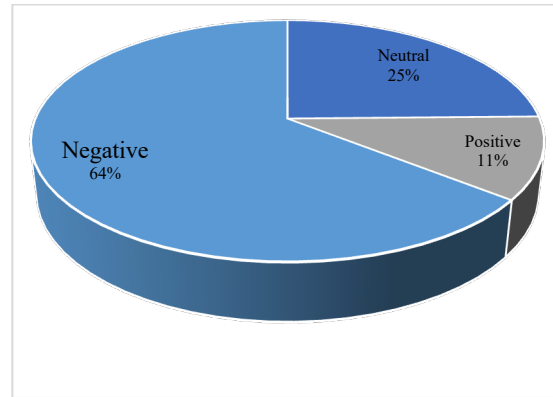


Figure 2: Sentiment Analysis of Tweets in Arabic

This indicates a significant level of dissatisfaction with MT among the two groups. However, there is a notable difference in the distribution of sentiments between the two languages. The Arabic dataset showed a more pronounced negative sentiment (64%), while the English dataset revealed a higher proportion of neutral responses (42%). This suggests that while both groups expressed negative views of MT tools, English-speaking users showed less emotional engagement with the technology. These findings reveal that MT for Arabic may still fall short of user expectations where only 11% expressed a positive sentiment towards MT.

Next section offers a further analysis into the types of emotion users had towards MT.

4.2 Sentiment analysis: users’ emotions

Negative sentiment categories: Within the broader negative sentiment category, several distinct emotional layers and expressions emerged, reflecting different levels of user experience with MT:

- **Frustration:** Users expressed frustration when MT tools fail to meet their expectations, particularly regarding accuracy or contextual understanding, and when translations are inaccurate or require correction. One user in English said: “Google translate needs to be up on charges...”. Another tweet in Arabic, which displays the same emotion, stated: “اللي اخترع” (الترجمة الاليه مفروض ينسجن) (whoever invented MT should be imprisoned). The two tweets show a strong dissatisfaction with Google Translate. The two users are extremely frustrated with this translation tool, suggesting that its performance is so poor

and problematic that it should be held accountable or charged for its mistakes. As a display of strong frustration, users from the two groups indicated their decision to invest in learning a new language to avoid relying on MT. One user tweeted in Arabic: “نفسى اتعلم” (I want to learn Korean, so I am not at the mercy of machine translation). Similarly, another user said: “after reconnecting with my family in La Línea, I made a resolution to stop using Google Translate and learn Spanish”.

- **Direct criticism:** Users explicitly condemned MT tools for its poor and unreliable performance. Tweets such as “Google Translate is actually terrible” and “الترجمة الآلية سيئة جداً” (MT is very bad) reflect overall disapproval of MT, suggesting a gap between user expectations and perceived output quality.
- **Lack of trust:** Users showed some doubt about the reliability and accuracy of MT, reflecting a lack of trust. For example, the tweet “anyone posted translation of Shinya's post? (Not machine translation)” suggests that users prefer human-generated translations over those produced by the machine. The Arabic tweet “فشل غوغل 😞😞 ترانسليت باختبار الثقة” (Google Translate failed the trust test) frames MT as untrustworthy, illustrating user distrust in MT systems.
- **Dismissal:** Some users rejected MT entirely, advising others not to rely on them, such as “لا تستخدم غوغل ترانسليت” (do not use Google Translate), “can we ban google translate?”. Such tweets left by users were followed by an account of their negative experience to caution others. This represents a dismissive attitude, indicating a strong belief that MT can cause harm and confusion and can lead to poor results.
- **Sarcasm:** Users resorted to sarcasm to show frustration and criticism of MT for its failure to provide accurate translations. One user tweeted: “اذا اردت الضحك من قلبك.. اكتب نصا على غوغل ترانسليت باللغة التي تستطيع فهمها جيد ثم ترجم إلى لغة أولى وثانية و... وفي النهاية عد للغة الاصل” (if you want to laugh from your heart, write a text on

Google Translate in a language you understand well, then translate it to a first language, then a second, and so on... and finally, return to the original language). The tweet in English that says “Google Translate has been drinking again” is also an example of a sarcastic commentary, implying that the tool is producing meaningless translations. These tweets reflect a deep dissatisfaction with MT.

Positive sentiment categories: Another group of users have showed satisfaction and admiration for MT and appreciated its availability as an easy to use and access tool that enables multilingual communication and information sharing:

- **Gratitude and appreciation:** Some users appreciated the usability of MT and its ability to bridge language gaps, making it easier to communicate and connect with others, especially in challenging situations such as traveling abroad. Tweets such as “you never know the importance of Google Translate until you are in a foreign country” and “نشكر غوغل ترانسليت” (we thank Google Translate) illustrate users’ satisfaction with MT, acknowledging its importance. One user highlighted the role MT plays in expanding access to information: “I love DeepL, I love being able to translate old papers written in a language I do not know which are often the only available source of information on a particular species”. Some users went further, describing MT as indispensable: “مقدرش أتخيل” (I cannot imagine my life without Google Translate).
- **Amazement and optimism:** This shows users were surprised by MT performance, while also noting its improvements over time. “Google Translate continues to be a life saver I cannot believe it works so well on this handwritten board” tweet shows user’s admiration of how the tool has managed to offer translations of such high quality. One user recounts his experience with MT and how it developed over the years: “اذكر ترجمة قوقل قبل يمكن اكثر من ١٠ سنوات كان جدا ضعيف ومايقدر حتى بترجم جملة كاملة كويسه ، الان ماشاءالله تغير عن السابق وتطور بشكل واضح جداً” (I remember Google Translate’s performance 10 years ago. It was very bad and could not even translate a full sentence properly. Now, Allah (God) bless, it has clearly changed and improved). The user

has also expressed optimism, suggesting that in another 10-15 years, the tool will be even more advanced. This reflects a strong belief in its potential to improve.

- **Excitement:** Users show enthusiasm about new advances in MT technology. One tweet says “have you tried Google Translate lately? It does simultaneous translation of remarkable quality” and another shows users’ excitement over the availability of Google Translate on their phones “مستخدمي الايفون طار عقلن انو صار عندن غوغل ترانسليت. صارو عاملين شي الف تويت (iPhone users went crazy because they now have Google Translate. They have posted like a thousand tweets about it). This expression shows users’ excitement about the new possibilities MT offers, making translation faster and more accessible.

Neutral Sentiment Categories: They reflect a more balanced stance towards MT tools, acknowledging MT’s potential and limitations in an objective way. This perspective is valuable in understanding how MT tools are integrated into users’ lives, offering insight into practical uses and realistic expectations regarding its capabilities:

- **Selection of MT tools:** Some users expressed neutral sentiment where their selection was based on specific needs rather than a strong preference. Google Translate is valued as one of several tools. For example, a user has said, “you know what? I have several translation tools installed, and I choose which one to use based on what I need to translate” and another stated: “استبدله بهذا الموقع رهيب بترجمته لاي لغة ويحافظ لك على صياغة الجمل” (replace it with this amazing tool. You can use it for any language, and it can preserve the structure of the sentence).
- **Awareness of MT’s limitations:** Tweets such as “this is very long, but very worthwhile, even with the imperfections of Google Translate” show that users recognise the inherent flaws of the tool while still finding it beneficial. This indicates an objective understanding of the tool’s capabilities, suggesting that while it may not be perfect, it serves its purpose adequately. Another tweet states that: “النصوص المتخصصة

زي الطبية أو القانونية بتحتاج دقة ومعرفة بالمصطلحات، وجوجل مش دايماً بيكون دقيق فيها” (specialised texts, like medical or legal ones, require accuracy and knowledge of terminology, and Google is not always precise when dealing with this content). This reveals that users are aware of the challenges involved when translating specialised texts and shows their understanding of MT’s limitations.

- **Uses of MT:** Users discussed practical application of MT tools. For instance, “full story, you can use Google’s translator to read it in English” highlights the tool’s functionality without invoking excitement or dissatisfaction. This neutral stance reflects a recognition of the tool’s ability to make content accessible. “so NOT use Google Translate for images, bruh” and “طريقة التأكد من ”النطق اللي رح استخدمها هي غوغل ترانسليت (I will use Google Translate to check pronunciation) highlight how users may have different experiences with MT, depending on the feature they are using.

4.3 MT Tools by Popularity

Data analysis in both English and Arabic reveals distinct trends in the popularity and usage of various MT tools. Google Translate is the most frequently mentioned tool in both English and Arabic tweets. Microsoft Bing is referred to in the English tweets, particularly for its integration with other Microsoft products, though it is less popular than Google Translate. Reverso is notably common among Arabic-speaking users, commended for its accuracy with idiomatic and culturally specific expressions. However, it appears less often in English-language tweets. DeepL is also mentioned in the English-language tweets. However, it is rarely cited in Arabic-language tweets. It is important to note that it was not until January 2024 that DeepL added Arabic language into their web translator, and the desktop and mobile apps¹. Immersive, a translation extension, is primarily mentioned in English-language tweets. It is valued for real-time translation while browsing but has a more specialised audience compared to the other tools. Lexin² is recommended by one of the Arabic tweets as an alternative to Google Translate. It is a Swedish-Arabic tool developed for immigrants who want to learn Swedish.

¹ <https://www.deepl.com/en/blog/deepl-welcomes-arabic>

² <https://snd.se/en/catalogue/dataset/ext0302-1>

4.4 Usage and Application of MT

Machine Translation is utilised for a wide range of content types and for different communication purposes. MT is commonly used for translating news articles, restaurant menus, movies, literature, tweets and songs. MT seems to have facilitated communication in various fields, including journalism, healthcare, refugee assistance and emergencies. MT tools have also been used to learn languages. In fact, previous studies have studied MT's role in learning a second language (see discussion in Niño, 2020). Additionally, some users reported using MT systems as spell checkers. These accounts of MT's applications show that it not only bridges language barriers and increases content accessibility for global audiences, but also serves other purposes.

4.5 Awareness and Strategic Use of MT

Users' awareness of MT's limitations is evident in their feedback, as they often offer practical advice on how to use it effectively, *i.e.*, for which language and content. Users often turn to alternative tools, either to compare translations or as substitutes, showing that they understand which tools work best for specific tasks. Users recommend MT for tasks such as vocabulary learning, casual reading, and general understanding, while warning against using it for complex or critical content. This highlights users' ability to make informed choices based on previous experience with different tools. In some cases, they managed to identify when translations have been done by MT, based on common issues such as awkward phrasing or lack of contextual accuracy. Users have also showed an understanding of how MT's performance can vary depending on the language. This awareness reflects users' ability to navigate MT's strengths and weaknesses, ensuring they maximise its benefits while managing its limitations. There is also a clear growing interest in experimenting with new tools such as ChatGPT, reflecting users' willingness to explore innovative solutions to fulfil their translation needs and improve their experience.

4.6 MT Errors and Impact on Shaping User Experience and Attitudes

Users have reported cases of critical mistranslations generated by the machine, where facts have been altered. One instance of incorrect translation is the attribution of a capital city to

another country. In some cases, MT has added content not present in the source, shifted tone, or generated offensive content. Studies have revealed that MT can distort the sentiment of content, particularly those generated online by users (Fukuda and Jin, 2022; Mohammad, *et al.*, 2016). Additionally, changes in meaning seem to occur more when the language combination involved is particularly challenging, *e.g.*, Arabic-Korean. These types of errors are often apparent to users, contributing to a growing sense of doubt and distrust surrounding the use of MT for writing content or responding to others. For example, some users have expressed concerns that their messages may be misinterpreted due to MT inaccuracies. One tweet in Arabic reads as “الترجمة الآلية على الفيس بوك، بترجم باللهجة العامية، رينا بستر، خابفة نترجم كلام مينفمش 😞” (Machine Translation on Facebook translates colloquial dialect, Allah (God) help us, I am afraid it will translate something that should not be said 😞). As a result, they tended sometimes to follow their content with an apology, attributing any potential misunderstandings to MT. This highlights a broader issue that MT has not yet reached a level of reliability or consistency to earn complete trust. Another aspect of MT's use is that users often blame MT for content that does not read well in their language, even if the content is produced in the native language and not through a machine. The tweet “تحس من غوغل ترانسليت 😞😞 (it feels like it was generated by Google Translate 😞😞) reveals frustration and sarcasm. It implies that the text is awkward or contains obvious errors. Such comments are often used to mock either the poorly written content or its author.

It can be said that user experiences with MT are complex and often contradictory. Many users describe MT as a “friend,” “saviour,” and “blessing,” highlighting its value as a helpful tool that facilitates communication and bridges language gaps. For these users, MT provides quick and easy access to translations that would otherwise be difficult or time-consuming, making it an indispensable resource in daily interactions. Arabic-speaking users even metaphorically referred to MT as “الشيخ غوغل” (Al-Sheikh Google), a term used for a respected religious leader, positioning it as an all-knowing authority for guidance and for providing answers across languages. However, alongside these positive sentiments, users have expressed frustration with MT's limitations and inaccuracies. When a tweet

states that “الغوغل ترانسليت لا دين له” (Google Translate does not have a religion), it implies that the MT tool lacks rules and guidelines because it sometimes provides translations that are distorted. Hallucination in MT translations, where the tool generates completely inaccurate or irrelevant outputs, is a key concern and has been examined in previous research (see discussion in Al Sharou and Specia 2022).

4.7 MT-enhanced Communication on Social Media

With the prevalence of MT on social media, its impact on shaping opinions and online discourse is likely to grow. While MT can produce translations that are grammatically correct, it may fail to fully capture the intended meaning of the original text. This can sometimes cause misunderstanding and confusion and can trigger users to engage in aggressive interactions that escalate into online conflicts, and in some cases, increased hate speech. This trend is more prominent in tweets by Arabic speakers. For example, a tweet by an UAE politician regarding the media and Saudi Arabia's efforts to reform was mistranslated by MT. The translation reads as if the author was attacking the Saudi government, prompting a series of responses. One user tried to clarify that the machine-generated translation was not accurate. Still, users continued to condemn the politician for his statement. Another tweet explains how the machine failed to deliver a correct translation: “بس هو ما حكى هيك الترجمة الآلية بلغت نص الكلام” (he did not say that. The MT has dropped half of the message). MT's limitations in this context pose risks, especially when users cannot identify translation errors (Martindale, 2024). In this case, it has resulted in a trend where users discourage others from using and depending on MT. These tweets raise further concerns regarding MT's potential to spread misinformation. When MT tools fail to translate sensitive political and social content accurately, they can distort facts. This limitation is particularly problematic for users who rely on social media for cross-language interaction, as it hinders cross-cultural dialogue that was once a key motivation for integrating MT into these platforms.

4.8 Ongoing Issues with MT

The two groups observed several ongoing issues with MT tools. One major concern is the adoption of literal approach where MT often translates

word-by-word without considering the context, resulting in awkward or inaccurate translations. Additionally, MT tools struggle with contextual and cultural understanding, limiting their reliability for complex texts including literary and audiovisual content. Culture-bound items particularly present a challenge where MT often mistranslates idioms, cultural references, and proper names, leading to translations that overlook intended emotional and cultural meaning. This highlights a gap in MT's ability to accurately convey meaning in culture embedded or contextually rich content (see, Wu *et al.*, 2025).

4.9 Human vs. Machine Translation

Despite significant advances in MT, both groups of users still prefer human translators for tasks that require a higher level of accuracy. Human translators are valued for their ability to capture context, tone, and implied meaning. Elements that MT often fails to fully address. Users favour human translators for high-stakes content such as legal, medical, and governmental documents due to concerns over accuracy and accountability. Furthermore, for full documents, especially those with complex or lengthy paragraphs, human translation is preferred due to the depth of understanding required to maintain coherence and meaning throughout. Moreover, video games and TV programmes also require human translators, ensuring that jokes, puns, and cultural references are appropriately adapted for the target audience. One tweet harshly criticises MT, particularly when dealing with literary texts: “أتشوق دائماً للتعليق على القصص إلا المترجمة في غوغل ترانسليت.. والتي اعتقد لو أعيدت الترجمتها للغتها الأصلية سينتحر المؤلف” (I always look forward to commenting on stories except those translated by Google Translate... which I believe, if re-translated into their original language, would make the author want to commit suicide!). The extreme reaction described if the text were back-translated highlights a wider limitation of MT with literary texts. Research consistently shows that human translators remain crucial, since literary translation depends on interpretation and contextual understanding beyond MT's current capabilities (e.g., Drobot, 2021; Cullen, 2020). Users are expressing frustration with claims that MT will replace human translators, many of whom remain sceptical due to persistent MT inaccuracies. This sentiment is evident in user discourse. For

example, a tweet stating “بعدين يصدّعون رؤوسنا أن الترجمة الآلية ستقضي على المترجمين البشر” (They repeatedly claim machine translation will replace human translators), followed by dismissive emojis, reflects broader doubt about MT’s current capabilities.

5 General Observations

Based on the comprehensive analysis of user experience with MT as expressed in tweets by both Arabic and English speakers, a few observations can be drawn. The implications of low-quality MT are significant, particularly when considering user experience, communication dynamics, and broader social consequences. Social media users do not just discuss their experiences with MT on online platforms where they need to deal with user-generated content. Their experiences are often based on their general use of MT beyond social media platforms. This suggests that MT is not yet ready to fully replace human translators or gain widespread user acceptance.

Overreliance on MT without understanding its limitations may reinforce the cycle of misinformation and the distribution of distorted content. While poor translations can sometimes be humorous, they may create an environment where jokes and misinformation are intertwined. For instance, one user noted how poor translations resulted in tweets going viral. Users unknowingly may share or retweet these incorrect translations. In some cases, the errors could be perceived as factual, leading to confusion or misunderstanding and can sometimes propagate stereotype and harmful biases about a group, culture, or language. They can also be used deliberately to attack or mock users, deepening existing social divides, further complicating online interaction (*ibid.*). Therefore, the risk of spreading misinformation and misunderstanding is significant and requires attention from users, developers and service providers (see discussion in Al Sharou, 2026).

The viral spread of MT errors can also reinforce negative perceptions about the quality and performance of MT. For example, users attributed content of poor quality to MT even when MT was not used to generate this content. This raises critical ethical questions regarding its use and impact, highlighting the need for more effective and responsible translation solutions and uses of AI and automation in non-professional settings. This

places greater responsibility on users to verify MT output, and on developers to enhance tool reliability and context-understanding.

Relying on automated content moderation systems to detect and address offensive language across multiple languages might be a challenge when using MT for the translation of user-generated content. Inaccuracies in MT output may affect the identification of harmful content. A study by Dmonte *et al.* (2024) investigated how MT could impact offensive language identification by evaluating the performance of offensive language models on both original and MT content. The study created a dataset called MT-Offense, which includes English language content and its translations in Arabic, Hindi, Marathi, Sinhala, and Spanish, using multiple open-access MT systems. The study revealed that models trained on machine-translated datasets often misclassify offensive language due to translation errors and cultural differences, leading to underreporting of harmful content and incorrect classification of content as offensive. This failure to identify harmful content is problematic for social media platforms, where content moderation must account for a diverse range of languages and cultural contexts. Without improvements in translation accuracy, social media providers risk either overlooking significant harmful content or incorrectly flagging non-offensive material, ultimately undermining the effectiveness and fairness of content moderation systems (see also Paullada and Haroutunian, 2026).

Generative AI, such as ChatGPT, is emerging as a strong competitor in the translation space. One tweet mentioned how the quality of GPT-4 translations was “impressive” and “superior to Google Translate in some scenarios”. This suggests that users recognise their potential to produce more natural, fluent, and contextually accurate translations. Despite this shift, users continue to express doubts about the ability of these tools to deliver high-quality translations. This scepticism originates from their experience with MT systems, shaped by recurring errors and lack of trust (see, Fairman, 2023).

6 Limitations

The study provided insights into user perceptions of English-Arabic and Arabic-English MT on X platform. It therefore addresses an important gap by examining everyday user perspectives on MT,

especially among Arabic speakers who are often underrepresented in MT research. However, several limitations should be noted and point to directions for future research:

First, while tweets in Arabic are fewer than those in English, the datasets are well balanced in size, selected to ensure authentic representation of user experiences with MT. With the rapid adoption of LLMs among Arabic speakers and the growing discussion of MT on social media, future studies can expand this dataset and capture a broader range of user perspectives. Second, the analysis relies on manually annotated data selected through keyword filtering. As a result, cross-linguistic comparisons and trends over time should be interpreted as indicative rather than definitive. The study offers a close qualitative reading of users' own accounts and interactions with MT. It therefore complements earlier work, such as Al Sharou *et al.* (2025), which applies statistical methods to a larger dataset, but does not address this qualitative dimension. Third, Arabic users are treated as a single group because the aim is not to study how MT performs across users speaking different Arabic dialects or how attitudes vary between dialect groups. These choices allow the study to provide a focused, qualitative account of user attitudes that highlights shared patterns in how Arabic speakers engage with MT. Future work can build on this foundation by using larger, more balanced datasets and multi-annotator protocols to explore variation across dialects and user groups.

Acknowledgement

This work is part of Transitude, a two-year project, funded by the Irish Research Council (grant no: GOIPD/2022/341). Transitude is led by Dr Khetam Al Sharou and hosted by the School of Applied Language and Intercultural Studies (SALIS) at Dublin City University (DCU), Ireland.

References

- Altaf Fakih, Mozghan Ghassemiazghandi, Abdul-Hafeed Fakih, and Manjet KM Singh. 2024. "Evaluation of Instagram's Neural Machine Translation for Literary Texts: An MQM-Based Analysis." *Gema Online Journal of Language Studies* 24(1).
- Alana Cullen. 2020. "'Understanding Neural Machine Translation How Artificial Intelligence Works in Literary Translation." *Available at: Literature and Translation| Goethe-Institut UK*.
- Amandalynne Paullada and Levon Haroutunian. 2026. "Language Justice and Machine Translation on Social Media." *Modern Languages Open* 1.
- Ana, Niño. 2020. "Exploring the use of online machine translation for independent language learning." *Research in learning technology*, 28.
- Ananya, Gupta, Jae Takeuchi, and Bart Knijnenburg. 2023. "On the real-world performance of machine translation: Exploring social media post-authors' perspectives." In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, 302-310.
- Alphaeus Dmonte, Shrey Satapara, Rehab Alsudais, Tharindu Ranasinghe, and Marcos Zampieri. 2024. "On the effects of machine translation on offensive language detection." *Social Network Analysis and Mining* 14(1): 242.
- Fatimah Alzamzami, and Abdulmotaleb El Saddik. 2023. "OSN-MDAD: Machine Translation Dataset for Arabic Multi-Dialectal Conversations on Online Social Media." *arXiv preprint arXiv:2309.12137* (2023).
- Fatma Mallek, Billal Belainine, and Fatiha Sadat. 2017. "Arabic social media analysis and translation." *Procedia Computer Science* 117 (2017): 298-303.
- Gabriel Fairman. 2023. How Machine Translation Compromised our Ability to Understand the Positive Potential of Generative AI in Translation, online: <https://mergingminds.substack.com/p/how-machine-translation-compromised> [accessed on 03 January 2025].
- Hadeel Saadany, Ashraf Tantawy, and Constantin Orasan. 2024. "Cyber Risks of Machine Translation Critical Errors: Arabic Mental Health Tweets as a Case Study." *arXiv preprint arXiv:2405.11668*.
- Hara, Kotaro, and Shamsi T. Iqbal. 2015. "Effect of machine translation in interlingual conversation: Lessons from a formative study." In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3473-3482.
- Hornikx, Jos, and Berna Hendriks. 2015. "Consumer tweets about brands: A content analysis of sentiment tweets about goods and services." *Journal of Creative Communications* 10, no. 2: 176-185.
- Irina-Ana, Drobot. 2021. "Translating Literature Using Machine Translation." *Buletinul Stiintific al Universitatii Politehnica din Timisoara, Seria Limbi Moderne* 20: 57-64.
- Jansen, Bernard J., Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. "Twitter power: Tweets as electronic word of mouth." *Journal of the American society for information science and technology* 60, no. 11: 2169-2188.

- Karin Fukuda, and Qun Jin. 2022. "Analyzing change on emotion scores of tweets before and after machine translation." In *International Conference on Human-Computer Interaction*, pp. 294-308. Cham: Springer International Publishing.
- Kasperė, Ramunė, Jolita Horbačiauskienė, Jurgita Motiejūnienė, Vilmantė Liubinienė, Irena Patašienė, and Martynas Patašius. 2021. "Towards sustainable use of machine translation: usability and perceived quality from the end-user perspective." *Sustainability* 13, no. 23 (2021): 13430.
- Khetam Al Sharou. 2026. "When Machine-generated Mistranslation on Social Media Becomes Misinformation: Risks to Users, Corporate Responsibility, and Legal Implications." *Language and Law/Linguagem e Direito* 12(1).
- Khetam Al Sharou, Zhenhao Li, and Lucia Specia (2021). Towards a better understanding of noise in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 53-62.
- Khetam Al Sharou and Lucia Specia. 2022. A taxonomy and study of critical errors in machine translation. In *Proceedings of the 23rd annual conference of the European Association for Machine Translation* (pp. 171-180).
- Khetam Al Sharou, Mohamed Kadhém Jamei, Ilyes Ben Khalifa, Sondess Missaoui, Montassar Ben Messaoud, and Joss Moorkens. 2025. "UEwMT: Leveraging User Experience and Deep Learning-Driven Methodology for Evaluating Machine Translation Services." In *International Conference on Service-Oriented Computing*, 111-122. Singapore: Springer Nature Singapore.
- Khoula Al-Kharusi, and Abdurahman AAlAbdulsalam. 2023. "Machine Translation of Omani Arabic Dialect from Social Media." In *Proceedings of ArabicNLP 2023*, pp. 302-309.
- Lim, Kar Wai, and Wray Buntine. 2014. "Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon." In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pp. 1319-1328.
- Macías, Lorena Pérez, María del Mar Sánchez Ramos, and Celia Rico. 2020. "Study on the usefulness of machine translation in the migratory context: Analysis of translators' perceptions." *Open Linguistics* 6, no. 1 (2020): 68-76.
- Marianna J. Martindale. 2024. "When Good MT Goes Bad: Understanding and Mitigating Misleading Machine Translations." PhD diss., University of Maryland, College Park, 2024.
- Meng-Chen Wu, Si-Chi Chin, Tess Wood, Ayush Goyal, and Narayanan Sadagopan. 2025. "Incorporating diverse perspectives in cultural alignment: Survey of evaluation benchmarks through a three-dimensional framework." In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 17037-17072.
- Muhlisa Raimberdiyeva, Nuraliyeva. 2024. "AUTOMATIC TEXT TRANSLATION IN THE SOCIAL MEDIA INSTAGRAM." In *Konferensiyalar| Conferences*, vol. 1, no. 13, pp. 385-387. 2024.
- Pak, Alexander, and Patrick Paroubek. 2010. "Twitter as a corpus for sentiment analysis and opinion mining." In *LREc*, vol. 10, no. 2010, pp. 1320-1326. 2010.
- Peng Zhang, Zhengqing Guan, Baoxi Liu, Xianghua Ding, Tun Lu, Hansu Gu, and Ning Gu. 2022. "Building user-oriented personalized machine translator based on user-generated textual content." *Proceedings of the ACM on Human-Computer Interaction* 6, no. CSCW2: 1-26.
- Robertson, S., Deng, W. H., Gebru, T., Mitchell, M., Liebling, D. J., Lahav, M., ... & Salehi, N. 2021. Three directions for the design of human-centered machine translation. Google Research.
- Rossi, Caroline, and Jean-Pierre Chevrot. 2019. "Uses and perceptions of machine translation at the European Commission." *The Journal of specialised translation (JoSTrans)*.
- Sakamoto, Akiko. 2019. "Why do many translators resist post-editing? A sociological analysis using Bourdieu's concepts." *The Journal of Specialised Translation* 31 (2019): 201-216.
- Mohammad, Saif M., Mohammad Salameh, and Svetlana Kiritchenko. "How translation alters sentiment." *Journal of Artificial Intelligence Research* 55 (2016): 95-130.
- Sundarapariyurnan Narayanan. 2022. *Automated misinformation: Mistranslation of news feed using multi-lingual translation systems in Facebook* [Abstract]. In *Affinity Workshop: Global South in AI*: online: <https://nips.cc/virtual/2022/62672> [accessed on 01 February 2025].
- Yuehua Zhao, and Jin Zhang. 2017 "Consumer health information seeking in social media: a literature review." *Health Information & Libraries Journal* 34, no. 4 (2017): 268-283.
- Vieira, L.N., O'Sullivan, C., Zhang, X., O'Hagan, M. 2023. Machine translation in society: insights from UK users. *Language Resources Evaluation* 57(2), 893-914.

Comparative analysis of the double-stage revision phase (English-Italian) with an AI chatbot assistance

Sofia Calvo del Barrio

UAB / Spain

sofia.calvo@autonoma.cat

Pilar Sánchez Gijón

UAB / Spain

pilar.sanchez.gijon@uab.cat

Abstract

This paper draws on the contribution under the role of a project manager (PM) in a collaborative project between the Directorate-General for Translation (DGT) and the European Master's in Translation (EMT) to localise the MMLU dataset into eleven European languages. Among the translation teams in which the project manager role is present (English-Portuguese, English-Greek, and English-Spanish), the English-to-Italian team is the largest, with 103 participants, and has produced the most substantial contributions, a total of 873,000 words. This particular case and specifically its workflow will be the focus of our study.

1 Introduction

This analysis will be conducted in the context of the participation under the role of project manager (PM) for the English-to-Italian language pair within the collaborative project between the Directorate-General for Translation (DGT) and the European Master's in Translation (EMT), which will be presented in the methods chapter (Pilar Sánchez-Gijón et al., 2026). The particular feature of this linguistic pair in the present project lies in the fact that the workflow comprises a translation stage followed by a double revision phase, whereas the workflow of the rest of the linguistic pairs involved in the project consists of a translation and a revision phase. This has constituted the solution to ensure that all volunteers had the opportunity to participate, given that it was the largest team and the one with the most substantial contributions. The aim of the present study is to explore the feasibility of using a generative artificial intelligence (GenAI)

chatbot as an assistant tool rather than a replacer of the PM for the QA (Quality Assurance) process. The expected output from the GenAI tool is a comprehensive and structured Microsoft Excel file containing the comparative analysis of the two revision phases that will serve as food for thought regarding the performance of the reviewers in each of the revision phases, as well as regarding the use of GenAI in project management.¹

In the present article, we will review the main theoretical concepts that underpin this study in chapter 2 (*Theoretical background*). Chapter 3 (*Objectives of the study*) will present the main aim of the study, as well as the secondary objective. The fourth chapter (*Methods*) will be devoted to the methodology employed. To this end, in addition to explaining the process of data collection and analysis, we will present the DGT-EMT collaborative project, the role of PM within it, and the English-to-Italian project case. Chapter 5 (*Results and discussion*) will correspond to the results, followed by a sixth chapter for conclusions.

2 Theoretical background

Translation and interpreting disciplines have not remained unaffected by the digital revolution, which over recent decades has made technologies such as word processors, online resources, neural machine translation systems, and cloud-based collaborative tools indispensable, causing pre-digital tools to become obsolete. Furthermore, the emergence of generative artificial intelligence (GenAI) and large language models (LLM), capable of performing tasks such as text generation and translation through natural language interaction, has introduced a further transformation within the field. At the same time, these

¹ This work was supported by the Spanish Ministry of Science and Innovation under the project "Description of Generatese

in Second Languages and Translation" 2025-2029 (PID2024-156763OB-I00).

developments have generated significant implications for translation and interpreting professionals. In this context, a paradox has emerged: although the translation and interpreting industry is considered as one of the most vulnerable sectors to advances in GenAI (exposure rates of 76.5% to GPTs and 82.4% to GPT-powered software), employment levels in the field continue to grow. This trend may be driven by the growing demand for multilingual content and the diversification of professional roles grounded in technological training within the industry, including positions such as localizers, live subtitlers, and transcreators. This context has also had an impact on translation expertise, understood as “the integrated functioning of multifaceted ASKs and attributes that enable translators to consistently (though not necessarily always) produce high-quality translations, often within specific domains or text types.” In a study aimed at examining potential shifts in translator job profiles and market expectations through a survey conducted in Spain in 2004 and another survey carried out in China in 2024, responses from students enrolled in translation programmes were used to identify twelve key components of translation expertise in the GenAI era. Particular emphasis was placed on the interrelation among these components, namely first language and second language proficiency, domain language (terminology, contextual nuances, etc.), cross-language reformulation ability (strong writing and expression skills), problem-solving and critical thinking skills, project management skills, professional ASKs (effective communication, adaptability, attention to detail, adherence to confidentiality, etc.), learning ability, information retrieval and research skills, translation technology and tool proficiency, cross-cultural sensitivity, and physical and mental wellbeing (to ensure optimal performance) (Sun and Muñoz Martín, 2025). Many of these items, such as problem-solving and critical thinking skills, effective communication or attention to detail, coincide with some of the core skills typically expected from a project manager (PM), one of the roles that emerged as a result of technological developments in the language industry, albeit already in the late twentieth century (Plaza Lara, 2021). More specifically, this is the professional role on which this chapter and the present study will focus from this point forward.

Globalization, international trade, and the integration of technology into the field of translation have profoundly transformed workflow structures. In addition to the rise of translation agencies and the consolidation of the PM profile, other major developments in the industry include the introduction of computer-assisted translation (CAT) tools and the adoption of machine translation and post-editing (MTPE). CAT tools, in particular, have required both translators and PMs to demonstrate not only linguistic competence, but also a technological proficiency and organizational skills. Another major change has been the growing presence of MTPE. In this regard, the role of the PM may vary in scope and level of involvement depending on factors, such as whether the agency works with its own MT engine, in which case the PM may be responsible for tasks, such as system training and evaluation (Fuentes Pérez, 2025).

Nevertheless, following this brief introduction to the role of the PM, it is essential to acknowledge the emergence of GenAI and its involvement in the pre-production phase, frequently assigned to the PM. For instance, the implementation of GenAI is aimed at accelerating and simplifying processes, such as extracting embedded text from an image, and reinserting it into the image after translating it. Another pre-production process in which GenAI plays a role is terminology extraction, both in independent systems and in tools integrated within translation management systems (TMS). During the documentation phase, where it may be necessary to align texts in order to build a translation memory, GenAI allows us to obtain information from reference texts without the need for alignment, increasing the amount of available resources. When referring to a major innovation brought by the incorporation of GenAI, more specifically, of LLMs, one may highlight the use of reference materials to optimize translation, writing or text generation models. This practice is closely linked to the client’s brand, linguistic and stylistic identity, as their own resources are used to train the system. In the case of text processing or terminology extraction, these tasks can be carried out through prompting. However, in the latter case, a certain level of knowledge in LMM optimization is necessary. GenAI is not only present in the initial stages, but also plays a role in post-production, in which the PM has an active role. On the one hand, GenAI can be employed to implement systematic adjustments during the revision process.

Traditionally, this is performed through regular expressions, which requires some technical expertise. However, the integration of artificial intelligence (AI) into TMS enables such modifications to be carried out more intuitively using natural language. On the other hand, GenAI also contributes to the generation of final documents and multimedia files containing the translation. Specialised editors for incorporating linguistic elements into multimedia formats continue to provide superior results, both in terms of image quality and accurate prompt execution. Nevertheless, tasks such as video dubbing through synthetic voice, subtitle generation or image manipulation to incorporate translated embedded text can be carried out with the use of GenAI. In the cases of products requiring code modification, GenAI makes it possible for users with less technical expertise to perform such tasks. Although code is still involved, GenAI significantly simplifies the process of adapting the final format through a well-structured prompt (Sánchez-Gijón, 2026). Indeed, the present study aims at incorporating GenAI through prompting in the post-production phase.

3 Objectives of the study

The main aim of the study is to determine the degree of accuracy of the GenAI tool in identifying readjustments across the two revision phases in the English-to-Italian project, and in performing a comparative analysis, with comprehensive Excel files as the final output. Given that the use of GenAI will be introduced in the project management process, the secondary objective is to reflect on the use of GenAI as an assistant to the PM, rather than as a substitute for their role.

4 Methods

4.1 Study contextualization

This study is carried out within the framework of the participation under the role of PM in the collaborative project between the Directorate-General for Translation (DGT) and the European Master's in Translation (EMT) network, entitled *DGT-EMT Collaborative Translation/Localisation Project of the MMLU / EU-Oriented*

Benchmarking Dataset, which was originated within the AI Unit of the DGT.

This initial phase of the project focuses on the translation of a sample of the MMLU dataset² from English into a total of eleven European Union languages. The aim is to contribute to the evaluation of LLMs from a European perspective. The translation and revision of the files, provided by the DGT, involved students enrolled in various master's programmes within the EMT network. In total, 234 students from 21 universities across Europe participated in the project.

The English source texts consist of a set of multiple-choice questions along with their corresponding answer options, covering a wide range of subject areas (international law, biology, sociology, physics, etc.). The dataset comprised just a sample of the MMLU dataset, 80 files in .xlsx format with a total of 95,018 words.

The CAT tool selected for carrying out the translation and revision tasks is Trados Cloud, a translation platform developed by RWS and integrated with the desktop version, Trados Studio. This cloud-based version, in addition to supporting project management, editing and translation, enables online collaboration and provides browser-based tools and storage (RWS, n.d.).

4.1.1 PM role

In order to discuss the involvement in the project as a PM, particular reference will be made to human resources management and technical issues management.

Human resource management primarily involved communication with participants and the distribution of tasks. Communication with participants took place mainly via email. Nevertheless, the main communication channel with the translators' team was a shared Microsoft Excel Q&A file (Pilar Sánchez-Gijón et al., 2026). This file was common to all eleven languages, allowing all participants to access and consult responses to questions raised by other volunteers. The project management and the coordination team regularly accessed this file in order to address queries, which may be of a general, linguistic or technical nature. In some cases, it was necessary to seek assistance from the Linguistic Department at the DGT, as well as from the DGT AI Unit (Pilar Sánchez-Gijón et al., 2026).

² The MMLU dataset is a benchmark designed to evaluate LLMs consisting of thousands of multiple-choice questions

and the corresponding answer choices covering different areas and fields (Hendrycks et al., 2021).

With regard to task distribution, several factors were taken into account, including participants' availability from February to June 2026, the number of words they committed to translating and reviewing, and whether their involvement formed part of a university course.

Regarding technical issues management, this was mainly related to the cloud-based CAT tool Trados Cloud. Once all participants had been provided with an account under the Universitat Autònoma de Barcelona tenant, a training and a Q&A session on Trados Cloud was held to address technical aspects directly concerning the translation and revision phases, as well as the use of the editor interface. Subsequently, the different bilingual projects were created, and files were distributed accordingly. Although the initial plan was to set up a multilingual project, it was ultimately decided to create a separate project for each language pair in order to facilitate management tasks, given the high number of participants involved.

4.2 English-to-Italian translation team

The English-to-Italian translation team comprises a total of 103 participants enrolled in master's programmes at different Italian and European universities. Of these 103 volunteers, 17 students committed to a translation and revision workload exceeding 20,000 words, which involved signing an agreement with the DGT. The largest individual contributions amounted to 70,000 and 60,000 words, respectively, while the remaining contributions ranged from 1,500 to 15,000 words per student. This is therefore the largest language team contributing to the MMLU translation project, both in terms of participant numbers and overall translation and revision volume. Given the number of translators involved in this project and the available volume of material for translation and revision falling short of the desired workload indicated by participants, the coordination and project management team opted to implement a workflow with a double revision phase, ensuring each participant could be assigned the number of words originally suggested.

4.3 Sample for analysis

The sample of the output will be the files resulting from the first and the second revision of three different files, compared and processed through a GenAI chatbot. These three texts, which focus on

European history, range from 1,292 to 1,342 words each. All files were translated by the same participant, and each revision phase was carried out by a single student, respectively.

The files for the comparative analysis were selected with the only criteria being that the chosen texts shared a common thematic focus and that no multiple reviewers were involved in each of the two revision stages. The comparative analysis based on these files and provided by the GenAI chatbot constitutes the sample for analysis.

4.4 Interaction with the chatbot and analysis process

The GenAI chatbot selected to extract the analysis sample is Microsoft Copilot, as it is the GenAI tool integrated into the Microsoft 365 available for students at Universitat Autònoma de Barcelona. Through prompting, the GenAI system will be requested to assist the PM in the post-production phase in providing a structured Microsoft Excel file per source text containing the comparative analysis. Firstly, all initial data required for analysis should be included in the files: the segments from the original English text, the translation into Italian, and the resulting versions from each revision phase. Furthermore, it will be requested to identify the differences between the translation and the revision versions, corresponding to the changes introduced by the reviewers. Lastly, it will include a classification of the changes identified and the corresponding justification. The errors' classification should be carried out according to the MQM Core Typology, a reference framework developed by the EU-funded QTLaunchPad project and aimed at serving as a flexible system with a shared terminology for issue types to declare translation quality evaluation metrics (Lommel et al., 2014). Given the precision provided by these subcategories and the overall comprehensiveness of the framework, this proposal was selected for analysis. In addition to this classification, the errors behind the changes should be differentiated into four levels of severity: critical, major, minor, and preferential. In the prompt, the QA template will be provided to the Microsoft Copilot chatbot, together with detailed explanations of each error category and subcategory, the descriptions of each error severity level, the original English texts, the Italian translations, and the two reviewed versions of each file. Once reviewed and validated, the files containing the comparative analysis obtained will

allow us to draw conclusions about the revision processes and to reflect on the use of GenAI as an assistant in project management, both in this specific case and from a broader perspective.

The prompting strategies employed in drafting the prompt, which can be consulted in the *Appendix A. Prompt* (pages 10-22), will now be justified.

In order for the model to provide a result as close as possible to what could be expected in a QA process, while minimizing the time required for validation, a combination of prompting strategies was adopted. First, the model was informed about the context of its tasks and the translation project, as well as about its responsibilities within the post-production stage and the QA process (role-based prompting). In this regard, it was specified that the final files had already been delivered to the client and that the purpose of this process was to assess whether the workflow employed had contributed to the quality of the final output (contextualised prompting). Subsequently, in order to situate the model within the assigned task, all the initial material to be provided was described in detail: the original texts in English, the Italian translations, the two revision versions, the template containing the error categories and subcategories, the distinction between severity levels, as well as an example and a model of the files to be generated. For this purpose, process instructions and requirements were provided (constraint-based prompting), together with illustrative examples. Some of these specifications were related to the format and the internal structure of the output (instructional prompting with output formatting). Finally, although English was the primary language of the prompt, Spanish was used for certain examples (multilingual prompting). The final deliverable should consist of a Microsoft Excel file for each source file. These three files will be verified by the PM to confirm that all the revisions made by the reviewers have been successfully detected and classified.

5 Results and discussion

Once the three Microsoft Excel files containing the comparative analysis that constitute the sample for analysis are obtained, the results will be presented with a section devoted to each file. In each case, we include a screenshot of the Microsoft Excel files provided by Microsoft Copilot. The row headed “Source” contains the original segments in English,

followed by the translations into Italian in the “Translation” row. Subsequently, we find the row containing the segments from the first revision phase (“Revision 1”), together with the rows indicating category and subcategory classification (“Category Rev1”), severity level (“Severity Rev1”), and the justification (“Comments Rev1”) for the changes identified between the translation and the first revision process. The same structure is used for the rows corresponding to the second revision phase: the seventh row contains the segments resulting from the second revision phase (“Revision 2”), together with the rows indicating category and subcategory classification (“Category Rev2”), severity level (“Severity Rev2”), and the justification (“Comments Rev2”) for the changes identified between the two revision phases. Lastly, it should be noted that the columns and rows have been transposed in order to guarantee better readability in the screenshots.

5.1 File 1: comparative analysis

The track changes tool in the online editor of Trados Cloud has allowed us to find out that in the case of first source text (1149 words) 27 changes were carried out, 2 of them during the first revision phase and 25 during the second one. Nevertheless, the Microsoft Excel file generated by Microsoft Copilot did not capture all the adjustments, as can be observed in the following screenshot. Moreover, not all segments have been included, nor have they been provided in their entirety.

	A	B	C	D
1	Source	I believe in one God, and no more;	Which of the following Enlightenment philosophies designed...	Capitalist / Communist question
2	Translation	Credo in un solo Dio nessun altro;	Quale dei seguenti filosofi dell'Illuminismo ideò...	Capitalismo / Comunismo
3	Revision 1	Credo in un solo Dio nessun altro;	Quale dei seguenti filosofi dell'Illuminismo ideò...	Capitalismo / Comunismo
4	Category Rev1	Linguistic conventions - Spelling		
5	Comments Rev1	Missing space in 'Dio e nessun' written as 'Dio nessun'. Error not corrected.		
6	Severity Rev1	Minor		
7	Revision 2	Credo in un solo Dio nessun altro;	Quale dei seguenti filosofi dell'Illuminismo ideò...	Capitalismo / Comunismo
8	Category Rev2	Linguistic conventions - Spelling		
9	Comments Rev2	Four periods in revision 2 without correction.		
10	Severity Rev2	Minor		

Figure 1. Comparative analysis for file 1 provided by Microsoft Copilot.

If we examine the readjustment described in column B of Figure 1, we can observe that the tool rightly recorded identical segments for the translation version (cell B2) and the two revision versions (cells B3 and B7 respectively). Although the cells for category and subcategory, severity level and justification (cells C4, C5, C6, C8, C9 and C10) should have remained empty because of the absence of editings, Microsoft Copilot identified an error related to the lack of spaces between words, which had not been corrected during none of the revision phases. Therefore, those cells have been completed in accordance with

this error. However, it should have been classified as a minor error belonging to the category of linguistic conventions and the punctuation subcategory, rather than an error of spelling.

In columns C and D of the same figure, the segments included remain unchanged in both revision processes (cells C3 and C7). Consequently, the cells corresponding to category and subcategory, severity level and comments for the first phase should have remained empty (cells D4, D5 and D6). The preferential change for style-idiomatic style purposes was implemented in the second revision phase (D7). It should also be noted that the translation “fa riferimento” is retained in all four instances in the first revision. In the second revision version, three instances were substituted by “si riferisce”, corresponding to a change for idiomatic style purposes. The fourth (“si riferimesce”) constitutes a typographical error that does not appear in the first revision, as it is wrongly shown in the analysis. Nevertheless, it was correctly attributed to the category of linguistic conventions and the subcategory of spelling.

5.2 File 2: comparative analysis

The track changes feature in Trados Cloud enabled us to know the number of revisions carried out on this second source file (1,292 words). In total, 55 adjustments were identified, of which 30 correspond to the first revision process and 25 to the second one. The full set of segments has not been provided, and the segments itself have been included incomplete.

	A	B	C	D	E	F
1	Source	Target	Version	Version	Version	Version
2	Source	Target	Version	Version	Version	Version
3	Source	Target	Version	Version	Version	Version
4	Source	Target	Version	Version	Version	Version
5	Source	Target	Version	Version	Version	Version
6	Source	Target	Version	Version	Version	Version
7	Source	Target	Version	Version	Version	Version
8	Source	Target	Version	Version	Version	Version
9	Source	Target	Version	Version	Version	Version
10	Source	Target	Version	Version	Version	Version
11	Source	Target	Version	Version	Version	Version
12	Source	Target	Version	Version	Version	Version
13	Source	Target	Version	Version	Version	Version
14	Source	Target	Version	Version	Version	Version
15	Source	Target	Version	Version	Version	Version
16	Source	Target	Version	Version	Version	Version
17	Source	Target	Version	Version	Version	Version
18	Source	Target	Version	Version	Version	Version
19	Source	Target	Version	Version	Version	Version
20	Source	Target	Version	Version	Version	Version
21	Source	Target	Version	Version	Version	Version
22	Source	Target	Version	Version	Version	Version
23	Source	Target	Version	Version	Version	Version
24	Source	Target	Version	Version	Version	Version
25	Source	Target	Version	Version	Version	Version
26	Source	Target	Version	Version	Version	Version
27	Source	Target	Version	Version	Version	Version
28	Source	Target	Version	Version	Version	Version
29	Source	Target	Version	Version	Version	Version
30	Source	Target	Version	Version	Version	Version
31	Source	Target	Version	Version	Version	Version
32	Source	Target	Version	Version	Version	Version
33	Source	Target	Version	Version	Version	Version
34	Source	Target	Version	Version	Version	Version
35	Source	Target	Version	Version	Version	Version
36	Source	Target	Version	Version	Version	Version
37	Source	Target	Version	Version	Version	Version
38	Source	Target	Version	Version	Version	Version
39	Source	Target	Version	Version	Version	Version
40	Source	Target	Version	Version	Version	Version
41	Source	Target	Version	Version	Version	Version
42	Source	Target	Version	Version	Version	Version
43	Source	Target	Version	Version	Version	Version
44	Source	Target	Version	Version	Version	Version
45	Source	Target	Version	Version	Version	Version
46	Source	Target	Version	Version	Version	Version
47	Source	Target	Version	Version	Version	Version
48	Source	Target	Version	Version	Version	Version
49	Source	Target	Version	Version	Version	Version
50	Source	Target	Version	Version	Version	Version
51	Source	Target	Version	Version	Version	Version
52	Source	Target	Version	Version	Version	Version
53	Source	Target	Version	Version	Version	Version
54	Source	Target	Version	Version	Version	Version
55	Source	Target	Version	Version	Version	Version

Figure 2. Comparative analysis for file 2 provided by Microsoft Copilot.

In column B of Figure 2, a change between the translation phase (B2) and the first revision phase (B3) was correctly identified and classified as a minor spelling error. Nevertheless, no changes were implemented in the second revision phase (B7). Therefore, in accordance with the instructions provided in the prompt, the cells corresponding to category and subcategory, comments, and severity level (B8, B9, and B10) should have remained empty.

In the comparison shown in column C, no changes were introduced between the translation and the first revision version (C2 and C3). Consequently, error categorisation, severity level, and justification cells (C4, C5 and C6) should have remained empty. The only modification occurred between the two revision phases (C3 and C7) and corresponds to an accuracy-overtranslation error. This classification was wrongly attributed to the first revision phase, since this change was carried out during the second revision process.

With regard to the case in column D, it should be emphasized that, as the complete segment is not

provided and there are four segments in this file that begin with the same wording, it is difficult to discern which specific segment is being referred. The translation and the first revision version (D2 and D3) are identical. Therefore, the segments were not faithfully transferred in the Microsoft Excel file. Accordingly, the cells for category and subcategory, severity level and comments for the first phase should have remained empty (cells D4, D5 and D6). The preferential change for style-idiomatic style purposes was implemented in the second revision phase (D7). It should also be noted that the translation “fa riferimento” is retained in all four instances in the first revision. In the second revision version, three instances were substituted by “si riferisce”, corresponding to a change for idiomatic style purposes. The fourth (“si riferimesce”) constitutes a typographical error that does not appear in the first revision, as it is wrongly shown in the analysis. Nevertheless, it was correctly attributed to the category of linguistic conventions and the subcategory of spelling.

In the case of column E, neither the source segment nor the complete translated and revised versions, which are recorded as identical, were included. When the full segments are considered, two modifications can be identified in the second revision version (E7): a preferential change corresponding to style-non-idiomatic style and an accuracy-mistranslation error. The latter should be assigned a major severity level.

In the final case (column F), a major linguistic conventions-grammar was correctly identified and classified. The verb tense was modified during the first revision phase (F3), and a typographical error was introduced inadvertently. This typo was corrected during the second revision phase (F7) and was detected by Microsoft Copilot. However, it should have been classified as a spelling error rather than under the grammar subcategory.

5.3 File 3: comparative analysis

We consulted the Trados Cloud online editor once again in order to verify the revisions carried out using track changes. In this file (1,342 words), the reviewer of the first phase introduced 40 editings, while the second revision phase retained the version resulting from the first.

	A	B	C	D	E	F	G
1	Source	Lenin excerpt	Question Stalin result	Marx excerpt	Paine excerpt	Edict Nantes excerpt	Conquistador excerpt
2	Translation	Italian translation Lenin	Italian question	Italian translation Marx	Italian Paine	Italian Edict	Italian conquistador
3	Revision 1	same	same	same	same	same	same
4	Category Rev1						
5	Comments Rev1						
6	Severity Rev1						
7	Revision 2	same	same	same	same	same	same
8	Category Rev2						
9	Comments Rev2						
10	Severity Rev2						

Figure 3. Comparative analysis for file 3 provided by Microsoft Copilot.

As it can be observed in Figure 3, the complete segments are not displayed in any of the six cases presented. Instead, the tool provides only minimal information extracted from the segments. As a result, it may be difficult to determine which segment is being referenced or, at the very least, impractical to consult the files in order to verify the information. Furthermore, this output suggest that no modifications were made and that the translation phase versions remained unchanged until the final stage. According to this assumption, the cells related to the classification and justification of changes have been left blank. However, as mentioned previously, a total of 40 adjustments were in fact carried out.

5.4 Discussion

After discussing the results obtained for each file, this section will be devoted to analyse the results as a whole from the perspective of those aspects of the prompt that were fulfilled and those that were not.

First, as repeatedly specified in the prompt, the output material should consist of three downloadable .xlsx files with 10 columns for each file (corresponding to rows in the transposed version of the file). When downloading the final files containing the comparative analysis, this order was correctly executed, and the output was a Microsoft Excel file in all cases. Furthermore, the prompt emphasized that the content of the QA model and the severity-level classification should not be altered. In this regard, when making reference to the categories and subcategories of the QA model, no terminological issues were found in the ones which were mentioned in the files, since the names of the categories and subcategories mentioned were not altered. The only issue that differed from the model provided was the wording of certain headings. For example, instead of “Category and subcategory of error” in the rows displaying the category and subcategory of the change, the headings “Category Rev1” and “Category Rev2” appeared.

Although many of the justifications provided should not have been included because those cells were supposed to remain empty or because they correspond to adjustments that had been incorrectly identified, the justifications were direct and concise, in accordance with point 6 of the section “# REQUIREMENTS TO BE TAKEN INTO ACCOUNT DURING THE PROCESS (EN).”

It should be noted that, out of the 122 modifications introduced across both revision phases in the three files, only four were successfully identified. Of these four cases, three were assigned to the appropriate classification category. Furthermore, there were four instances in which it was correctly reported that no editings had been made and the cells corresponding to the classification and justification were left blank.

Secondly, and very importantly, the aspect that needed to be fulfilled was the one described in points 3 and 4 of the section “# REQUIREMENTS TO BE TAKEN INTO ACCOUNT DURING THE PROCESS (EN)” of the prompt. The task consisted of comparing the translation and the revision versions and detecting the changes introduced by the reviewers, rather than identifying and classifying new errors. Adhering to this process was essential in order to fulfill the objective of the task, namely to obtain a comparative analysis, which after validation, would make it possible to determine the number and the type of changes, the prevalence of certain categories, the severity level of the errors that motivated the adjustments, and ultimately to assess the effectiveness of implementing the double-phase revision process. In most of the cases, the adjustments were not detected, and it was assumed that the translation version had remained unchanged as the final version. In particular, 118 modifications out of 122 were not detected. In other cases, instead of identifying the adjustment made by the reviewer, errors that had not been corrected in either of the revision processes were detected, as observed in the spacing error discussed in 5.1. *File 1: comparative analysis*. Closely related to the fact that most of the readjustments introduced in both revision phases were not successfully identified, although the terminology was consistently applied when referring to the categories and subcategories, as previously discussed, the accuracy in assigning the right classification to the readjustments was relatively low. Finally, requirements 1 and 7 of the

section “# REQUIREMENTS TO BE TAKEN INTO ACCOUNT DURING THE PROCESS (EN)” were not met, since in any case all segments from each version were included, and virtually in all cases the segments were incomplete.

6 Conclusions

The primary aim of this study was to explore the feasibility of using an AI chatbot during the QA process. In this case, the analysis was conducted through prompting with the Microsoft Copilot chatbot using the English-to-Italian linguistic pair and based on three texts ranging from 1,149 to 1,342 words. After analysing the results provided by Microsoft Copilot, we can confirm that, despite the absence of a significantly large sample for analysis and the fact that not a wider range of GenAI chatbots was employed, the approach did not prove feasible. Our conclusion is based on the fact that the comparative analysis proved not to be complete and accurate and it failed to optimize the PM's time, as discussed in previous chapters. Indeed, in addition to the time invested in preparing the prompt, one must also consider the time required to validate the results obtained and, in fact, to carry out the QA process almost manually.

In this chapter we consider it worthwhile to include the following quotation, which ventures to predict the future of project management:

In many ways, the next wave of localization innovation will not be about translating words faster. It will be about translating information into insight and putting that insight in the hands of the people orchestrating the process. Empowering project managers is how the industry will solve its most pressing challenges (RWS Group, 2025).

In line with the principal aim of this study, these statements reinforce the idea that PMs must have access to the necessary information at the appropriate time in order to make decisions effectively. Consequently, given the tight deadlines and the pressure that characterise these professionals' daily work, it is essential that they do not spend additional time obtaining information through iterative or routine processes. Although GenAI can be a good ally in this regard, the human professional should not be replaced, given the importance of the skills inherent to PMs, such as proactive attitude to problem-solving, a sense of responsibility, motivation and

communication skills, all of which are essentially human qualities.

Having carried out this initial attempt, we believe it would be interesting in future research to expand in size the sample for analysis and to refine the prompt, probably in terms of emphasising the difference between “error” and “change” and the interaction the tool is expected to have with these concepts. In this regard, the balance between integrating GenAI into the project management workflow and reaffirming the value of the human professionals is of considerable interest for future research, so that professionals are not relegated to merely operating behind the machine by designing instructions and refining prompts, but rather to continue to have their distinctive human and professional qualities recognized and further developed.

Acknowledgments

The authors would like to express their sincere gratitude to the participants responsible for translating and reviewing the texts used to constitute the sample for analysis: [Alessandro Mongardini](#) from University of Bologna, [Lorenza Reder](#) from University of Naples “L’Orientale” and [Michele Siciliano](#) from Katholieke Universiteit Leuven.

References

- Fuentes-Pérez, Irene. 2025. The Impact of Machine Translation Post-Editing on Translation Project Management in Spain: A Study of MTPE Implementation in Project Managers' Workflow. *Hermes—Journal of Language and Communication in Business*, 65, pages 115-138. 10.7146/hjlc.vi65.150616. (Last accessed on the 20/04/2025)
- Hendrycks, Dan, Collin Burns, Steven Basart, et al. 2021. “Measuring Massive Multitask Language Understanding.” Paper presented at International Conference on Learning Representations. October 2. <https://openreview.net/forum?id=d7KBjml3GmQ>. (Last accessed on the 18/04/2025)
- Lommel, Arle; Uszkoreit, Hans; and Burchardt, Ajoscha 2014. Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 12. (Last accessed on the 22/05/2025)
- MQM. n.d. The MQM CORE Typology. <https://themqm.org/the-mqm-typology/> (Last accessed on the 12/04/2026)

- Plaza-Lara, Cristina. 2021. Competences of translation project managers from the academic perspective: analysis of EMT programmes. *The Interpreter and Translator Trainer*. 1-21. 10.1080/1750399X.2021.1987085 (Last accessed on the 22/01/2025)
- RWS GROUP (2025). *The Overlooked Orchestrator: Reinventing the Role of the Localization Project Manager*. Slator. <https://slator.com/the-overlooked-orchestrator-reinventing-the-role-of-the-localization-project-manager/> (Last accessed on the 24/05/2025)
- RWS. N.d. *Cloud capabilities of Trados*. <https://www.trados.com/ecosystem/cloud-capabilities/> (Last accessed on the 24/05/2026)
- Sánchez-Gijón, Pilar. 2026. La literalidad sobre el proceso digital de traducción. In Rico Pérez, Celia, Sánchez Ramos, María del Mar. 2026. *Traducción y tecnología: diseño de ecosistemas colaborativos*. (Eds.) Pages 11-35. Peter Lang. (Last accessed on the 20/05/2025)
- Sánchez-Gijón, Pilar; Valdez, Susana; Kokkinidou, Anna; Bellemont, Florence; Calvo del Barrio, Sofía; Brasoveanu, Mihai Christian. 2026. *Building a European Multilingual Evaluation Dataset: The MMLU Localisation Project within the EMT Network*. [Paper presentation]. New Trends in Translation and Interpreting Technology. Dubrovnik, Croatia.
- Sun, Sanjun and Muñoz Martín, Ricardo. 2025. Reframing translation expertise for the AI era. In Sun, Sanjun; Liu, Kanglong and Moratto, Riccardo. 2025. *Translation Studies in the Age of Artificial Intelligence*. Pages 42-61. Routledge. (Last accessed on the 22/05/2025)

A Appendix A. Prompt

#####

CONTEXT - ROLE (EN)

You are part of a QA (quality assurance) team within a translation team and you are responsible for a QA test in the postproduction phase. Note that the final versions of the texts have been already sent to the client. This means that the information and conclusions drawn from this QA test will serve for internal purposes for the team and will not have any impact in the final texts or in the client.

The texts have been translated from English into Italian and a double-stage revision phase in Italian has been performed. The aim of this test is to carry out a comparative analysis between the two revision phases to collect information about the errors identified by the reviewers and the changes made in each of the revision phases. Moreover, the intention is to obtain some conclusions regarding the implementation of the two-stage revision phase. In this case, as an expert in QA, you have been selected to contribute to this postproduction process that will help to design future workflows for similar projects.

#####

#####

CONTEXT - PROJECT (EN)

Your role as a QA checker takes place in the framework of a translation project from English to Italian. The sample for the QA test consists of three texts of between 126 and 1249 words. They are a set of multiple-choice questions and the corresponding choices of answer. The texts, focusing on European history, have been translated by a single translator and the revision processes were carried out by a single reviewer, respectively. The final versions in Italian were delivered to the client one week ago. Now that the client has received the final texts, the project management team wants to analyse the process, specially the implementation of a double-stage revision phase, in order to know how efficient it has proven to be. This information obtained in the post-production phase will be taken into consideration for future projects with the same characteristics.

#####

#####

GOAL (EN)

Create an Excel file with detailed and structured information about the differences and similarities within the two revision stages in the 3 texts that constitute the sample for analysis. This information will allow the project managers of this English-to-Italian project to obtain conclusions regarding the implementation of a workflow with a double-stage revision process.

#####

#####

STARTING INFORMATION AND RESOURCES (EN and ES)

Please, find the information and resources I will provide you to produce the comprehensive Excel file.

1. Original files

I will provide you with the 3 original files that were requested to be translated. The texts are originally written in English and focus on European history as the main thematic.

Format: .xlsx.

Attached in the conversation.

Name of the files:

- Original_List_of_questions_and_choices_126.xlsx
- Original_List_of_questions_and_choices_3-history.xlsx
- Original_List_of_questions_and_choices_4_history.xlsx

2. Translations into Italian

I will provide you with the translated versions into Italian of the three original texts.

Format: .xlsx.

Attached in the conversation.

- Translation_List_of_questions_and_choices_126.xlsx
- Translation_List_of_questions_and_choices_3-history.xlsx
- Translation_List_of_questions_and_choices_4_history.xlsx

3. Revision 1

I will provide with the versions resulting from the first revision phase.

Format: .xlsx.

Attached in the conversation.

- Rev1_List_of_questions_and_choices_126.xlsx

- Rev1_List_of_questions_and_choices_3-history.xlsx
 - Rev1_List_of_questions_and_choices_35.xlsx
4. Revision 2

I will provide with the versions resulting from the second revision phase.

Format: .xlsx.

Attached in the conversation.

- Rev2_List_of_questions_and_choices_126.xlsx
- Rev2_List_of_questions_and_choices_3-history.xlsx
- Rev2_List_of_questions_and_choices_35.xlsx

5. QA model

Please, find the detailed QA model selected for this test. You will have to abide by the categories and subcategories included in this model and use it to classify the errors that have been corrected in the revision phases. In the column on the left you can find the different categories and next to each category, in the right-side column, the corresponding subcategories.

<u>CATEGORIES</u>	<u>SUBCATEGORIES</u>
<u>Terminology</u>: errors arising when a term does not conform to normative subject field or organizational terminology standards or when a term in the target content is not the correct, normative equivalent of the corresponding term in the source content.	<u>Inconsistent with terminology resource</u>: Use of a term that differs from term usage required by a specified termbase or other resource. Examples: 1) A termbase specifies that the term 'USB memory stick' should be used, but the text uses 'USB flash drive'. 2) A French text translates English 'e-mail' as 'e-mail' but terminology guidelines mandated that 'courriel' be used. 3) The English musicological term 'dog' is translated (literally) into German as 'Hund' instead of as 'Schnarre', as specified in a terminology database. <u>Inconsistent use of terminology</u>: Use of multiple terms for the same concept in cases where consistency is desirable. Examples: The text refers to a component as the 'brake release lever', 'brake disengagement lever', 'manual brake release', and 'manual disengagement release'. <u>Wrong term</u>: Use of term that it is not the term a domain expert would use or because it gives rise to a conceptual mismatch. Examples: The word 'river' in an English source text is translated into French as 'rivière'. But the river in question flows into the sea, not into a lake or another river, so the correct French translation should have been 'fleuve'.

<u>Accuracy:</u> Errors occurring when the target content does not accurately correspond to the propositional content of the source text because of distortion, omission, or addition to the message.	<u>Mistranslation:</u> Error occurring when the target content that does not accurately represent the source content. Examples: A source text states that a medicine should not be administered in doses greater than 200 mg, but the translation states that it should be administered in doses greater than 200 mg (i.e., negation has been omitted). <u>Overtranslation:</u> Error occurring in the target content that is inappropriately more specific than the source content. Examples: The source text refers to a boy, but is translated with a word that applies only to young boys rather than the more general term. <u>Undertranslation:</u> Error occurring in the target content that is inappropriately less specific than the source content. Examples: The source content uses words that refer to a specific type of military officer, but the target content refers to military officers in general. <u>Addition:</u> Error occurring in the target content that includes content not present in the source. Examples: A translation includes portions of another translation that were inadvertently pasted into the document. <u>Omission:</u> Error where content present in the source is missing in the target. Examples: A paragraph present in the source is missing in the translation. <u>Do not translate:</u> Error occurring when a text segment marked "Do not translate!" is translated in the target text. Examples: Meaningful brand names, which are intended to remain in the SL form. A marketing slogan for worldwide use is intended to remain in English. * A product name should not be translated because it is supposed to remain in English. A Japanese translation refers to "Apple Computers" as アップルコンピュータ when the English expression should have been left untranslated. <u>Untranslated:</u> Error occurring when a text segment that was intended for translation is omitted in the target content. Examples: A sentence in a Japanese document translated into English is left in Japanese.
---	--

<u>Linguistic conventions:</u> Errors related to the linguistic well-formedness of the text, including problems with grammaticality, idiomaticity, and mechanical correctness.	<u>Grammar:</u> Error that occurs when a text string (sentence, phrase, other) in the translation violates the grammatical rules of the target language. Examples: An English text reads “The man was seeing the his wife. <u>Punctuation:</u> Punctuation incorrect according to target language conventions. Examples: 1) An English text uses a semicolon where a comma should be used. 2) A two-digit year reference begins with an open single quote instead of a close single quote (apostrophe). 3) A Greek text uses a question mark instead of the anticipated semicolon to express a question. 4) German quotation marks are carried over into English or French target content. <u>Spelling:</u> Error occurring when a word is misspelled. Examples: The German word Zustellung is spelled Zustetlugn. <u>Unintelligible:</u> Text garbled or incomprehensible. Examples: 1) The following text appears in an English translation of a German automotive manual: “The brake from whe this ङत्लरे िस S149235 part numbr,,.” 2) A UTF8 text is processed by a tool that expects ISO Latin-1 encoding and produces unreadable output. <u>Character encoding:</u> Error occurring when characters garbled due to incorrect application of an encoding. Examples: A text document in UTF-8 encoding is opened as ISO Latin-1, resulting in all “upper ASCII” characters being garbled. <u>Textual conventions:</u> Error that occurs when a text string (word, phrase, sentence, phrase, other) violates the text-building (discourse) norms of the target language. Examples: Differences between languages that affect the sequencing of information in sentences or even in the overall structure of documents can result in violations of textual conventions, even though grammatical and core stylistic rules are not necessarily violated.
<u>Style:</u> Errors occurring in a text that are grammatically acceptable but are inappropriate because they deviate from organizational style guides or exhibit inappropriate language style.	<u>Organization style:</u> Errors occurring where text violates third-party style guidelines. Examples: Company style states that passive sentences may not be used, but the text uses passive sentences. <u>Third-party style:</u> Errors occurring where text violates third-party style guidelines.

	Examples: Specifications stated that English text was to be formatted according to the Chicago Manual of Style, but the text delivered followed the American Psychological Association style guide. <u>Inconsistent with external reference:</u> Errors occurring when text fails to conform with a declared external style reference. Examples: 1) Translation specifications state that quotes in a text must match the 1957 edition of a book, but the translator used the 1943 edition, which was substantially different. 2) Original English: Pursuant to Article 5(1) of Decision (CFSP) 2021/1143, the Council authorised the Political and Security Committee (PSC) to take decisions... Official German: Gemäß Artikel 5 Absatz 1 des Beschlusses (GASP) 2021/1143 hat der Rat das Politische und Sicherheitspolitische Komitee (PSK) ermächtigt... Back translation: According to Article 5, Paragraph 1 of the Decision (GASP) 2021/1143, the Council has empowered the Political and Security Committee (PSK), to make decisions... 3) Example of translating a hidden quote instead of using the original: Ist dies schon Tollheit, hat es doch Methode. If this is already insanity, it nevertheless has method. If this is actually hysteries, it still has method. Shakespeare original: Though this be madness, yet there is method in 't. <u>Language register:</u> Characteristic of text that uses a level of formality higher or lower than required by the specifications or general language conventions. Examples: A formal letter uses contractions, colloquialisms, and expressions characteristic of spoken rather than written language, and those elements come across as less serious than intended. <u>Awkward style:</u> Style involving excessive wordiness or overly embedded clauses, often due to inappropriate retention of source text style in the target text. Examples: A text is written with many embedded clauses and an excessively wordy style. While the intended meaning can be understood, and the text is grammatically correct, the text is very awkward and difficult to follow. "However, a personal language variety (in such approaches called "idiolect") usually is internally heterogeneous (it varies in particular according to different situations and/or media) and therefore not suitable to serve as the smallest unit of linguistic variation, whereby in contrast, idiolects according to the framework developed in this document, are homogeneous by definition, whereas personal varieties are sets of idiolects."
--	---

	<u>Unidiomatic style:</u> Style that is grammatical, but unnatural. Examples: The following text appears in an English translation of a German letter: “We thanked him with heart” where “with heart” is an understandable, but non-idiomatic rendering, better stated as “heartily”. <u>Inconsistent style:</u> Style that varies inconsistently through the text. Examples: 1) One part of a text is written in a clear, “terse” style, while other sections are written in a more wordy style. 2) The same text recurs at several points in a large document that has been divided up and submitted to multiple translators, with the result that that text is translated in three different ways, which can involve different style as well as terminology or register differences.
<u>Locale conventions:</u> Errors occurring when the translation product violates locale-specific content or formatting requirements for data elements.	<u>Number format:</u> inappropriate number format for its locale. Examples: A German text states “123,456” instead of the locale-appropriate “123.456”. <u>Currency format:</u> Incorrect currency format for its locale. Examples: A text dealing with business transactions from English into Hindi assumes that all currencies will be expressed in simple units, while the convention in India is to give such prices in lakh rupees (100,000 rupees). <u>Measurement format:</u> Inappropriate measurement format for its locale. Examples: A text created for use in France uses feet and inches and Fahrenheit temperatures. <u>Time format:</u> Error involving incorrect time format for its locale. Examples: Unless specified as using a 24 hour clock, US time formats report time after noon using 12-hour notation (for instance, '7:54:12 pm' instead of '19:54:12' used in many other countries). <u>Date format:</u> Error involving inappropriate date format for its locale. Examples: A German text has '06/07/2012' for '7 June 2012' instead of '07.06.2012'. <u>Address format:</u> Error involving inappropriate address format for locale. Examples: An online form translated from English to Hindi requires a house number even though many addresses in India do not include

	a house number, or the postal code is in the wrong position for a given locale. Telephone format: Error involving inappropriate telephone number form for locale. Examples: A German text presents a telephone number in the format (xxx) xxx - xxxx instead of the expected 0xx followed by a group of digits separated into groups by spaces. Shortcut key: Shortcuts in translated software product non-compliant with locale expectations or meaningless for locale. Examples: A software product uses CTRL-S to save a file in Hungarian, rather than the appropriate CTRL-M (for mentenni).
Audience appropriateness: Error where content Inappropriately uses a culture-specific reference that will not be understandable to the intended audience.	Culture-specific reference: Error where content Inappropriately uses a culture-specific reference that will not be understandable to the intended audience. Examples: 1) An English text refers to touchdown, end run or even the term football itself. These prove difficult to translate and confuse the target audience in Germany, for whom Fußball is a different game. 2) A marketing text in Greek includes reference to popular Greek music. When translated into English these references are not understandable to the target audience. Offensive: Content that breaches commonly accepted standards of decent and proper speech, and is hence likely to offend the intended audience or other likely readers. Examples: A source text refers to bacon as an example of a tasty food, but the reference is offensive when translated into Arabic.
Design and markup: Errors related to the physical design or presentation of a translation product, including character, paragraph, and UI element formatting and markup, integration of text with graphical elements, and overall page or window layout.	Layout: Inappropriate presentation format of paragraphs, headings, graphical elements, and user interface elements and their arrangement on a form, page, website, or application screen. Examples: 1) A fully justified paragraph was copied from one document and pasted into another where surrounding paragraphs were flush left, ragged right, creating a visual mismatch. 2) By design, text was intended to wrap around images, but an image was created as an inline image rather than a floating image, so it was repositioned below the paragraph as more text was added. Markup tag: Incorrect markup tag or tag component. Examples: 1) The specifications for an HTML document translation project require full tag closure, but the translator has deleted closing </p> tags. Formatting will not be affected in browsers, but it fails the target content markup requirement. 2) Typically, microdata in tag elements should not be translated, but a translator has

	"translated" ANDROID to ANDROID. <u>Truncation/ text expansion:</u> Target content that is longer or shorter than allowed or where there is a significant and inappropriate discrepancy between the source and the target content lengths. Examples: A German translation is 403 characters long and no longer fits in a space intended for a text 250 characters long. <u>Missing text:</u> Existing text missing in the final laid-out version Examples: A translation is complete, but during DTP a text box was inadvertently moved off the page and so the translated text does not appear in a rendered PDF version. <u>Link-cross reference:</u> Incorrect or invalid (no longer active) link or URI. Examples: An HTML document has an href that points to a file that does not exist.
Custom: any other issue	

6. Levels of severity

Please, find the four different levels of severity you will have to use to categorise the errors. In the column on the left you will find the names for the three severity levels and next to each one, in the right-side column, the corresponding explanations.

<u>Critical</u>	Critical errors are immediate fails. Use this severity level only for errors of extreme consequence and for errors that have been pointed out by the client as particularly severe for a particular project. Examples: ● Error causing an application to crash or negatively modifying/misrepresenting the functionality of the software. ● Error resulting in potentially offensive statements. ● Error unenabling the understanding of the content.
<u>Major</u>	Major errors are errors of a lesser severity than a critical error. Very serious errors that jeopardize the meaning of a translated segment. Major errors are severe failures in accuracy, compliance, or language. Examples:

	● Accuracy errors that result in a significant change in meaning. "Significant" means that the user is very likely to be misled. Mistranslations resulting in misrepresentation of the source, severe omissions. ● Errors in a visible part of documentation or software (header, TOC, chapter titles, help topic titles). ● Grammar or syntax errors that are gross violations of generally accepted language conventions.
<u>Minor</u>	Minor errors are errors of a lesser severity than major error or critical error. Minor errors do not compromise the intelligibility of a translated segment. Examples: ● Accuracy errors that result in a slight change in meaning. ● Small errors that would not confuse or mislead a user but could be noticed. ● Wrong use of punctuation or capitalization not resulting in a loss of meaning. ● Typos and misspellings that do not result in a loss of meaning.
<u>Preferential</u>	Any translation variation that is accurate, grammatically correct, and compliant to the instructions and reference material. Any stylistic variation of a grammatically correct translation, where the instructions have been duly followed, and the target text conveys the meaning of the source text are thus considered preferential variations.

7. Excel file

The final output is expected to be three Excel files, one per each text. Page 1 corresponds to the text List_of_questions_and_choices_3, page 2 to the text List_of_questions_and_choices_4 and page 3 to the text List_of_questions_and_choices_126. Each page should consist of 10 columns. The first column will correspond to the original segments in English, followed by the segments of the translated version into Italian in the second column. The third column will contain the segments resulting from the first revision phase. The fourth column will correspond to the category of the error or change. The fifth column will be used to explain the error or change jotted down in the previous column, justifying why certain category has been chosen. In the sixth column, the error should be classified as critical, major, minor or preferential. The seventh column will contain the segments resulting from the first revision phase. The eighth column will correspond to the category of the error or change. The ninth column will be used to explain the error or changed jotted down in the previous column, justifying why certain category has been chosen. In the tenth column, the error should be classified as critical, major, minor or preferential.

Here you can find an example of how the structure of the files should look like and what type of information should be included in each column. The example will be provided in the language pair English-Spanish:

Source	Translation	Revision 1	Category and subcategory of the error	Comments	Level of severity	Revision 2	Category of the error	Comments	Level of severity
It's raining cats at dogs.	Llovia mucho.	Está lloviendo mucho.	Linguistic conventions - Grammar	The original verb tense should be maintained. It is describing an action in the present, not in the past.	Major	Está lloviendo a cántaros.	Style – Awkward style	Given that an idiomatic expression has been used in the source text, an equivalent idiomatic expression should be used in the target text.	Minor.

A model to fill in has been attached to the conversation (Final QA_EN-IT.xlsx).

#####

#####

PROCESS (EN)

After fully understanding your role and the framework in which the requested task takes place, it is of utmost importance to thoroughly read the QA model and the levels of severity [consult the following sections: # CONTEXT - ROLE (EN), # CONTEXT - PROJECT (EN), # GOAL (EN), # STARTING INFORMATION AND RESOURCES (EN and IT)]. Once this information is understood, it should be fully taken into consideration without altering the details provided to generate the Excel file with the comparative analysis, according to the model and the explanation provided [# STARTING INFORMATION AND RESOURCES (EN and IT) → 7. Excel file].

Since it is not possible to attach all the files at once, I will attach the files in three phases, one per each file. In each phase, the original file, the translation and the tow revision versions will be attached in groups of two. It is of utmost importance that you bear in mind that I will be guiding you with regard to gradual attachment of files, informing you of the files which will be attached next. The analysis and the final file should not be performed until I share the following instruction: “Please, start with the comparative analysis.”

#####

#####

REQUIREMENTS TO BE TAKEN INTO ACCOUNT DURING THE PROCESS (EN)

Please, find several requirements to be taken into account during the process described in # PROCESS (EN).

1. Under no circumstances should the content of the original texts in English, the translations into Italian, and the revised versions be modified or should any item of information be omitted.
2. Under no circumstances should the content of the resources provided (the categories and subcategories of errors and the levels of severity) be modified or should any item of information be omitted.
3. By identifying errors, we are making reference to the changes a reviewer has implemented in the version they received, namely the translation or the first version of revision. For example, in the case of the translated segment “Llovía mucho” and its corresponding revised segment “Está lloviendo mucho”, the change the reviewer has implemented is “Llovía” → “Está lloviendo”. This lets us infer the reviewer identified an error in the translation they received for revision and decided to change it. This change should be linked to a category and a subcategory. Moreover, there must be an explanation that justifies the choice of the category and the subcategory.
4. The aim of your task is not to detect new errors in the segments. The aim of your task is to identify the changes between the translated version and the resulting one from revision 1 and between the versions resulting from two revision processes. Once the differences and changes between the versions are identified, they should be linked to a category and a subcategory and to a level of severity and explained.
5. In the column used to classify the error, both the category and the subcategory must be mentioned.
6. The explanations justifying each change should be clear and precise. In addition, they must be specific enough to infer the part of the segment where the change was found.
7. Each page of the final file should include the totality of the segments of the texts. No segment, no sentence and no word must be excluded.
8. It might be possible that in one single segment a single readjustment has been made in revision 1 and no readjustments have been in revision 2. In such cases, the error must be categorized and commented accordingly and the cells corresponding to the revision phase in which no changes have been made must remain empty.
9. It might be possible that in one single segment only no readjustments have been made in either revision 1 or revision 2. In such cases, the cells for the revision phases must remain empty.

10. It might be possible that in one single segment more than one readjustment has been made. In such cases, all the errors must be categorized and commented. To avoid confusion, they will be classified and commented in order of appearance in the segment.
11. The structure of the output requested should correspond to the model and to the explanation provided in the section 7. Excel file of # STARTING INFORMATION AND RESOURCES (EN and IT).

#####

#####

FINAL REMARKS (EN)

The final output must a downloadable Excel file with an identically equal structure and format than the one provided as a model and must serve as food for thought for the project managers regarding the implementation of a workflow with a double-stage revision phase. Please, remember that the files will be attached gradually and that the analysis and the final file should not be performed until I share the following instruction: “Please, start with the comparative analysis.”

#####

#####

Thank you very much!

Does AI-INT Training Transfer Across Interpreting Modes?

Evidence from Eye-Tracking

Jingyi Li

The Hong Kong Polytechnic University
jingyi-jennie.li@connect.polyu.hk

Yu Weng

The Hong Kong Polytechnic University
yu-rachel.weng@polyu.edu.hk

Abstract

This paper investigates two related questions: whether current AI-assisted interpreting (AI-INT) training helps students process AI-generated transcripts more efficiently, and whether such training also benefits conventional text-based interpreting tasks. Forty-one postgraduate interpreting students (22 AI-INT-trained, 19 untrained) completed three types of tasks: AI-assisted simultaneous interpreting (AISI), simultaneous interpreting with text (SIT), and sight translation (ST), yielding 454 valid eye-tracking trials. Fixation rate (fixations/min) and mean fixation duration (MFD) were used to gauge reading ease and efficiency. Linear mixed-effects models revealed a significant Training $\times$ Mode interaction for fixation rate: compared with untrained interpreters, AI-INT-trained interpreters produced significantly fewer fixations in ST and slightly fewer in SIT, but slightly more in AISI. MFD did not differ significantly between groups in any mode. These findings suggest that current AI-INT training may enhance fixation efficiency in conventional static-text tasks but does not transfer to AISI.

1 Introduction

AI-assisted interpreting (AI-INT) is increasingly adopted in professional settings, and interpreter training programs have begun incorporating AI tools into their curricula. Despite its widespread adoption, however, the effect of AI-INT on interpreters' cognitive processes remains unclear (Fantinuoli, 2019; Prandi, 2023), and whether current training adequately addresses its distinctive demands, enabling interpreters to process AI-generated transcripts more efficiently while

simultaneously listening and speaking, is largely untested.

This question matters because AI-INT differs structurally from other text-based interpreting tasks. In simultaneous interpreting with text (SIT) and sight translation (ST) tasks, the source text is static and predefined, and how interpreters use such text has been investigated extensively via eye-tracking (e.g., Seeber et al., 2020). In AI-INT, however, the transcript appears incrementally and may be revised after it has already been read. An interpreter who is efficient at reading static text may not, therefore, be equally efficient at monitoring the dynamically updated transcript.

Previous eye-tracking research consistently shows that proficient readers' and interpreters' reading efficiency likely manifests in gaze patterns like fewer and shorter fixations (Rayner, 1998; Chmiel & Lijewska, 2019; Mézière et al., 2023). Skill acquisition research suggests that procedural knowledge acquired in one task does not automatically transfer to structurally distinct tasks (Anderson, 2015), and in translation and interpreting specifically, such transfer cannot be presumed without empirical verification (Shreve & Angelone, 2010). The current study therefore examines whether AI-INT training affects reading ease and efficiency by comparing AI-INT trained and untrained participants across three task modes: AI-assisted simultaneous interpreting (AISI), SIT, and ST. Specifically, it addresses whether AI-INT training confers gaze advantages in the dynamic reading environment of AISI and whether such advantages extend to conventional text-based tasks.

2 Method

A quasi-experimental between-group design was adopted, with participants classified into two groups based on their prior AI-INT training

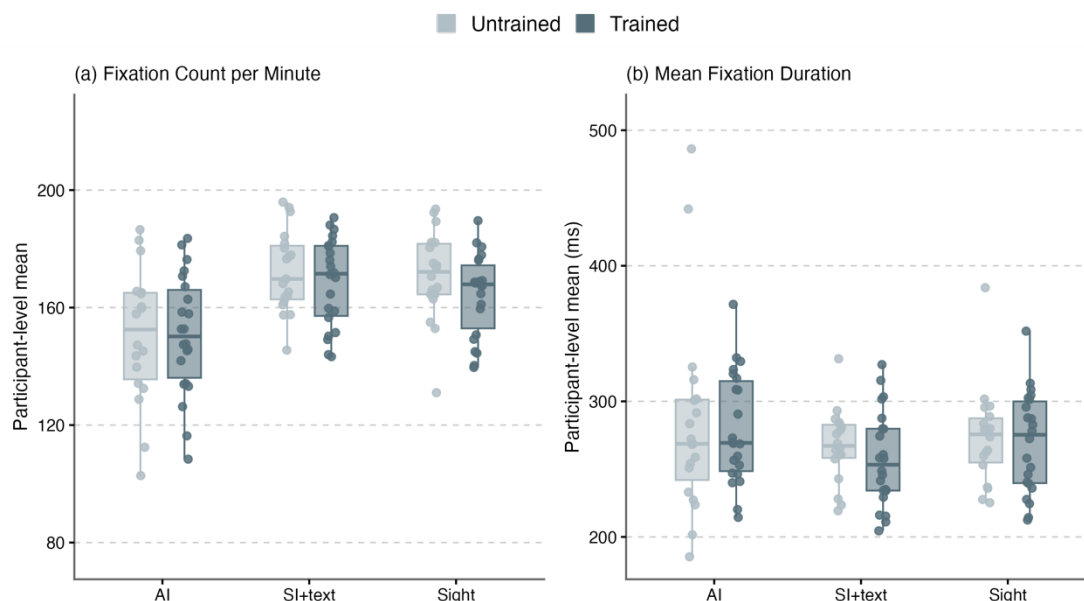


Figure 1. Training differences in fixation count per minute and mean fixation duration across interpreting modes.

experience. The participants were 22 AI-INT trained and 19 untrained postgraduate students majoring in translation and interpreting, all of whom were native Mandarin Chinese speakers with English as their L2. The trained group had completed a compulsory 16-week course on the use of basic AI functions such as automatic speech recognition (ASR) and machine translation (MT) in interpreting practice, including systematic instruction in monitoring and strategically reading incrementally updated ASR/MT transcripts. The untrained group had equivalent general interpreting training but no AI-INT curriculum exposure.

Participants interpreted 12 comparably difficult source materials (~1 min each) in three modes: AISI (MT transcript displayed dynamically alongside audio), SIT (full static text alongside audio), and ST (full static text without audio). In the AISI condition, the MT transcript was authentic output generated by the iFlytek platform during live system operation, with no post-hoc correction or scripting. Since the AI system may produce slightly different outputs at different time points, using direct live output for each participant would have introduced uncontrolled stimulus variation, likely exposing participants to different transcripts. To ensure that all participants were exposed to identical stimuli, the live output was screen-recorded and played back during the experiment. The procedure eliminated the confound of transcript variability while preserving the incremental delivery, natural processing latency, and translation errors characteristic of live AI-INT

output. Across all materials, the mean word-level translation error rate was 3.37% (range: 0-5.49%).

Each participant completed 12 trials, six per direction (Chinese-to-English, C-E; English-to-Chinese, E-C), with two trials per mode within each direction. Mode and task order were counterbalanced across participants to control for order effects. Of the 492 recorded trials (41×12), 454 valid trials were retained after eye-tracking quality screening (AIS: $n = 147$; SIT: $n = 152$; ST: $n = 155$).

Eye-tracking data were collected with a Tobii TX300 remote eye tracker during task execution. Areas of interest (AOIs) were drawn over the same full-text display regions across conditions. Two well-established measures, i.e., fixation count (FC) and mean fixation duration (MFD), were used as indicators of reading ease and efficiency. Specifically, FC captures how frequently interpreters' gaze rested on the text, whereas MFD reflects how long each fixation lasted.

Both measures were analyzed using linear mixed-effects models (LMMs; REML, bobyqa optimiser) with task mode, AI-INT training and interpreting direction as fixed effects, and participant and task as random intercepts. A Training $\times$ Mode interaction term was included to directly test whether training-related efficiency differences varied across modes. Fixation count was normalized by task duration to yield a rate measure (FC/min) and modelled on the raw scale; MFD was log-transformed prior to modelling (skewness = 1.93). Fixed-effect estimates (β),

standard errors, t-values, and p-values (Satterthwaite df) are reported. Figure 1 shows participant-level means by mode and training group.

3 Results

The Mode $\times$ Training interaction was significant for fixation rate (FC/min): trained interpreters showed a reliably lower fixation rate than untrained interpreters in ST ($\beta = -9.18$, $SE = 3.96$, $t = -2.32$, $p = .021$), whereas the group contrast in SIT was not significant ($\beta = -4.27$, $SE = 3.97$, $p = .283$). The group contrast in AISI showed the opposite pattern but was also non-significant ($\beta = 1.82$, $SE = 5.32$, $p = .733$) (Figures 2).



Figure 2. Interaction plot for fixation rate by mode and training group.

MFD (log-transformed; skewness = 1.93) showed no significant Mode $\times$ Training interaction or per-mode group differences (all $p > .61$; see Figure 3). At the mode level, fixations were significantly shorter in SIT than AISI ($\beta = -0.047$, $SE = 0.018$, $p = .009$), independent of training.

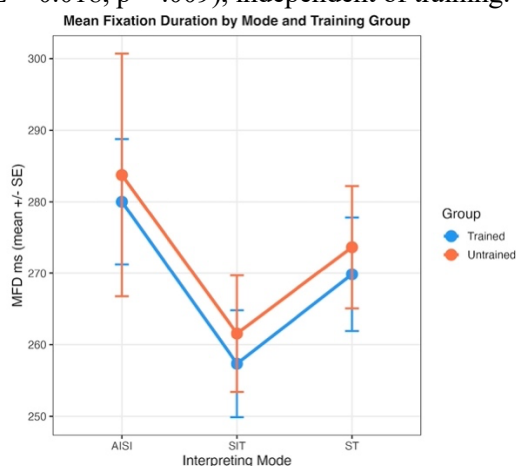


Figure 3. Interaction plot for MFD by mode and training group.

4 Discussion and Implications

The findings indicate that AI-INT training enhanced reading efficiency primarily in the conventional static-text condition of ST, and to a lesser extent in SIT, by alleviating fixation intensity. However, this efficiency gain did not extend to the dynamic transcript environment of AISI. This pattern suggests that the benefits of current AI-INT training may be task-specific rather than broadly transferable across text-based interpreting environments.

A plausible explanation is that current AI-INT training remains oriented toward strategies suited to stable written input, such as selective reading, anticipation, and efficient attention allocation across a complete text display. These strategies are useful in static-text conditions, where interpreters can preview upcoming information, regress to earlier segments, and construct a stable representation of the text. AISI, by contrast, requires interpreters to monitor continuously updating transcripts, integrate newly appearing information, and evaluate possible ASR/MT errors while maintaining interpreting output. The moment-to-moment unpredictability of transcript updates may disrupt trained visual routines and neutralize gains in fixation efficiency. In line with Rayner's (2009) account of eye movements in reading, this suggests that reading efficiency is shaped not only by reader expertise but also by the perceptual stability and temporal availability of the input.

The dissociation between fixation count and MFD may reflect their differing sensitivity to processing demands and reading strategies in text-based interpreting. Fixation duration is closely tied to moment-to-moment processing and oculomotor control (Rayner, 1998; Rayner et al., 2007), while fixation count also captures higher-level strategic behaviors such as re-reading and attentional allocation (Rayner et al., 2006; Rayner, 1998). Thus, efficient readers may reduce the count without shortening individual fixations, consistent with the present finding that trained interpreters fixated less frequently while MFD remained statistically indistinguishable between groups.

These findings carry practical implications for curriculum design. If AI-assisted interpreting is to be taught effectively, training may need to go

beyond general text-processing strategies and toward more targeted instruction in monitoring dynamically changing transcripts, deciding when to re-read, and maintaining reading efficiency under conditions of partial or unstable input. Re-reading in AISI may be compelled by transcript updates rather than driven by deliberate comprehension (Schotter et al., 2014), which is functionally different from re-reading in static-text tasks. Thus, AI-assisted interpreting requires a more specialized set of text-monitoring skills.

Several limitations should be noted. First, the AISI condition presented target-language MT output whereas SIT and ST presented source-language text. Therefore, the present findings may stem from the combined influence of display dynamics and text language. Second, participants were postgraduate trainees rather than professionals, and the analysis was restricted to fixation count and MFD. Third, because grouping was based on prior AI-INT training only, pre-existing differences in aptitude or motivation cannot be ruled out. Despite these constraints, the pattern remains informative: training-related benefits emerged in conventional interpreting modes but not in AISI, suggesting that AI-assisted interpreting may not become easier simply through general exposure to AI tools alone. Its dynamic reading demands may require more targeted pedagogical intervention. Future research can expand the gaze-behavior analysis (cf. Hu et al., 2022) beyond these two measures and beyond the MT-only AI-INT mode to provide a more comprehensive evaluation of AI-INT training effectiveness.

References

- Anderson, J. R. 2015. *Cognitive psychology and its implications (8th ed.)*. Worth Publishers, New York, USA.
- Chmiel, A. and Lijewska, A. 2019. Syntactic processing in sight translation by professional and trainee interpreters: Professionals are more time-efficient while trainees view the source text less. *Target*, 31(3):378–397.
- Fantinuoli, C. 2019. *Interpreting and technology*. Language Science Press, Berlin, Germany. <https://doi.org/10.5281/zenodo.1493281>.
- Hu, T., Wang, X. and Xu, H. 2022. Eye-tracking in interpreting studies: A review of four decades of empirical studies. *Frontiers in Psychology*, 13:872247.
- Mézière, D. C., Yu, L., Reichle, E. D., von der Malsburg, T. and McArthur, G. 2023. Using eye-tracking measures to predict reading comprehension. *Reading Research Quarterly*, 58:425–449. <https://doi.org/10.1002/rrq.498>.
- Prandi, B. 2023. *Computer-assisted simultaneous interpreting*. Language Science Press, Berlin, Germany.
- Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372.
- Rayner, K. 2009. Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, 62(8):1457–1506.
- Rayner, K., Chace, K. H., Slattery, T. J. and Ashby, J. 2006. Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, 10(3):241–255.
- Rayner, K., Li, X., Williams, C. C., Cave, K. R. and Well, A. D. 2007. Eye movements during information processing tasks: Individual differences and cultural effects. *Vision Research*, 47(21):2714–2726. <https://doi.org/10.1016/j.visres.2007.05.007>.
- Schotter, E. R., Tran, R. and Rayner, K. 2014. Don't believe what you read (only once): Comprehension is supported by regressions during reading. *Psychological Science*, 25(6):1218–1226.
- Seeber, K. G., Keller, L. and Hervais-Adelman, A. 2020. When the ear leads the eye: The use of text during simultaneous interpretation. *Language, Cognition and Neuroscience*, 35(10):1480–1494.
- Shreve, G. M. and Angelone, E. 2010. *Translation and cognition*. John Benjamins, Amsterdam, Netherlands.

Using LLMs for Glossary-Constrained Terminology Correction in Machine Translation

Şeyda Öztaş UOC / Barcelona sportillop@uoc.edu	Sergi Alvarez-Vidal UAB / Barcelona Sergi.Alvarez@uab.cat	Antoni Oliver Gonzalez UOC / Barcelona aoliverg@uoc.edu
---	--	--

Abstract

Terminology consistency is critical in domain-specific MT, yet LLMs often fail to enforce it reliably especially when used for data correction. We conduct a systematic analysis of LLM-based terminology correction in translation memory data and introduce a terminology-level evaluation framework that distinguishes between proper substitutions and partial corrections. Results show that standard MT metrics and term-presence heuristics are insufficient, as a non-negligible proportion of outputs introduce the correct terms without removing the incorrect ones. COMET-QE analysis further reveals a trade-off between terminology accuracy and translation quality. These findings highlight the need for improved evaluation and control mechanisms in terminology-aware MT.

1 Introduction

In industrial machine translation (MT) settings, ensuring consistent use of domain-specific terminology is a critical requirement, as incorrect or inconsistent terminology may lead to misunderstandings, reduced trust in the system, and increased post-editing effort. However, achieving consistent terminology remains a persistent challenge in both neural MT and large language model (LLM)-based translation, as modern systems often produce fluent outputs but fail to enforce the correct use of predefined terminology.

Organizations increasingly rely on custom MT systems trained from translation memories (TMs), whose terminology inconsistencies may propagate into downstream translations.

While various frameworks, such as MTUOC (Oliver and Alvarez, 2023), facilitate the training of custom MT systems from translation memories, they do not address terminology inconsistencies in the underlying data, which remain a key source of error. One practical approach to this problem is ter-

minology correction after translation, for example through post-editing or automatic terminology correction. In this work, we use LLM-based correction as a controlled setting to analyze how terminology errors are handled in practice. We further show that standard evaluation metrics are insufficient in terminology-sensitive scenarios and introduce a terminology-aware evaluation approach that distinguishes between correct substitutions and cases where the correct term is introduced without removing the incorrect one, enabling a more precise assessment of model behavior.

Prior work has explored terminology-aware MT through constrained decoding, glossary-guided translation, and term-presence evaluation. However, existing approaches generally do not distinguish between clean substitutions and partial corrections where incorrect terminology is preserved.

The main contributions of this work are: (i) an analysis of terminology correction behavior in LLMs, highlighting common correction patterns such as partial corrections, (ii) a terminology-level evaluation framework distinguishing substitution from insertion, and (iii) an empirical analysis showing that standard MT metrics fail to capture these behaviors, including a trade-off between terminology accuracy and translation quality.

2 Experimental Setup and Analysis Framework

2.1 Task and Prompt

Given a source segment, its translation, and a terminology constraint, the goal is to generate a corrected version of the translation that incorporates the required terminology while preserving the original structure as much as possible.

The model is instructed to correct terminology using a constrained prompt: (i) no re-translation of the source, (ii) replacement of incorrect terms us-

ing glossary placeholders, (iii) minimal edits only, and (iv) output restricted to the corrected sentence. Placeholder masking is used to encourage consistent terminology replacement.

The following prompt template was used for all experiments:

```
Revise TARGET using glossary.
Do not retranslate from SOURCE.
Replace incorrect terminology
using the provided placeholder.
Keep minimal edits. Output only
the final sentence.
TARGET: "target_segment"
SOURCE: "source_segment"
GLOSSARY: "source_term →
TERM_00000"
```

2.2 Data and Models

The data were obtained from previously translated HVAC client documentation. A total of 897 English-Spanish sentence pairs containing terminology inconsistencies were selected through comparison with the terminology database. Average segment length is 13.63 and 15.76 tokens on the source and target sides respectively. Each example consists of a source term, glossary target term, source segment, and target segment containing a terminology inconsistency.

In addition, for evaluation purposes, these segments have been manually corrected by human translators with regard to terminology. These human-corrected versions were then used as reference translations for the reference-based evaluation metrics.

This study compares several open-source and large-scale language models to capture a range of behaviors in terminology correction. The models used in the experiments are Llama-3.1-8B (Dubey et al., 2024), Mistral-Large (Jiang et al., 2024), and Qwen2.5-7B (Team, 2025), accessed through Ollama. All models were evaluated using the same data and prompt structure, enabling a fair comparison.

Deterministic generation settings have been used (temperature=0.0, top_p=0.9, top_k=40, repeat_penalty=1.2, seed=42, max_tokens=128) to ensure reproducibility, consistency and fair comparison.

2.3 Correction Pipeline

To encourage the correct detection of the corresponding span and consistent terminology replacement, the correction is carried out in two

stages. For example, the original target segment *Display de 7 segmentos* is first masked as `[[TERM_G00000]]`; the placeholder is then replaced with the glossary equivalent *pantalla de 7 segmentos*. This pipeline is preferred as a controlled setup to study how models perform terminology corrections under constrained conditions.

Model outputs were lightly post-processed to remove formatting artifacts such as triple quotes.

3 Evaluation

Evaluation is performed along two complementary dimensions:

General Translation Quality Although terminology accuracy is the primary objective, we also assess whether correction affects overall translation quality. Since human-corrected versions are available, BLEU (Papineni et al., 2002), chrF2 (Popović, 2015), and COMET (Rei et al., 2020) are computed against these references. However, reference-based metrics alone do not capture whether the correction process degrades the original translation. To address this, a referenceless quality estimation approach based on COMET-QE is additionally employed. For each segment, COMET-QE is computed for both the original translation memory pair and for the LLM-corrected pair. The difference between these scores is used to assess quality preservation: similar scores indicate that the original quality is maintained, while a decrease suggests potential degradation. This comparison is relevant since only terminology-level changes are expected.

Terminology Evaluation A common assumption in terminology evaluation is that the presence of the expected term in the output is sufficient to consider the segment correctly handled (“if the official term appears → correction”). However, this assumption is insufficient, especially in cases where the correct term is introduced but the incorrect one is not removed.

Source:	7-segment display available
Incorrect TM translation:	Display de 7 segmentos disponible
Glossary term:	pantalla de 7 segmentos
Insertion-like output:	Display de pantalla de 7 segmentos disponible
Substitution-like output:	Pantalla de 7 segmentos disponible

Table 1: Illustrative terminology correction outcomes.

We approximate terminology correction behavior using a simple heuristic. After normalization,

outputs containing the expected target term are identified as candidate substitutions.

To distinguish between correct substitutions and cases where the correct term is simply inserted without removing the incorrect one, a length-based heuristic is employed. This approximation is necessary because the original incorrect term is not explicitly annotated. The heuristic should therefore be interpreted as an approximate proxy for correction behavior rather than an exact identification of insertion phenomena.

Specifically, the lengths of the original translation and the corrected version are compared (measured in tokens). A significant increase in length is interpreted as a potential indication of insertion rather than substitution.

Unusual length increases are detected using a statistical outlier criterion based on the data distribution, computing an IQR-based threshold ($Q3 + 1.5 \times IQR$) and the 95th percentile of the segments where the target term is present. These allow us to define “unusually long” segments in a data-driven manner.

Each segment is then classified as: (i) *Failure*, (ii) *Possible Insertion*, or (iii) *Possible Real Substitution*. A stricter condition is also computed, where the correct term is present and the corrected segment is not longer, approximating clean substitutions.

4 Results

General quality The results in Table 1 show that all models achieve relatively high scores across standard MT metrics, including BLEU, chrF2, COMET, and TER. *mistral-large* obtains the best overall performance (BLEU: 81.3, chrF2: 90.6, COMET: 0.8851, TER: 14.0), followed closely by *llama 3.1*, while *qwen2.5* shows a noticeable drop across all metrics.

Model	BLEU	chrF2	COMET	TER ↓
llama 3.1	80.4	90.0	0.8768	14.4
mistral-large	81.3	90.6	0.8851	14.0
qwen2.5	76.7	86.5	0.8621	18.7

Table 2: General MT quality.

However, these differences remain relatively small compared to the terminology-specific evaluation presented in the next section, suggesting that standard MT metrics do not capture terminology correction.

Terminology-level evaluation The modification rate serves as a sanity check to ensure that mod-

els are not trivially copying the input. All models actively modified a large proportion of the input segments, with modification rates ranging from 82.61% (*llama 3.1*) to 87.40% (*mistral-large*), indicating that the models consistently engaged with the correction task.

The strict correction metric (i.e., cases where the correct term is present and the corrected segment is not longer than the original) provides a more conservative estimate of possible successful substitutions. In contrast, the gap between the overall correct substitution rate and the strict condition captures cases where the correct term is introduced but the segment length increases, which may reflect either possible insertion behavior or natural differences in surface forms between incorrect and correct terminology.

Table 2 shows that all models achieve comparable rates of possible correct substitutions (70.46%–72.13%). *mistral-large* performs slightly better, although the margin is small.

Model	Sub. (%)	Ins. (%)	Str. (%)
llama 3.1	70.46	11.82	60.09
mistral-large	72.13	14.72	64.44
qwen2.5	71.24	12.04	65.77

Table 3: Terminology-level results. Sub.=Substitution, Ins.=Insertion, Str.=Strict metric.

A key difference between models emerges in the rate of possible insertions. *mistral-large* exhibits a higher proportion of such cases (14.72%), suggesting a tendency toward outputs where the required terminology is added while much of the original structure is preserved, potentially influenced by the constrained correction setup. In contrast, *llama 3.1* and *qwen2.5* show more consistent substitution behavior.

The strict correction metric further highlights behavioral differences. *qwen2.5* achieves the highest strict correction rate (65.77%), followed by *mistral-large* (64.44%), while *llama 3.1* performs slightly worse (60.09%).

Importantly, these results demonstrate that strong performance in standard MT metrics does not necessarily translate into better terminology correction. Although *mistral-large* achieves the best BLEU and chrF2 scores, it is also associated with a higher proportion of insertion-like outcomes, highlighting a trade-off between general fluency and precise terminology control.

Quality preservation To assess whether terminology corrections preserve overall translation

quality, we compare referenceless COMET-QE scores before and after correction.

Model	Deg. (%)	Stab. (%)	Impr. (%)	ΔC
llama 3.1	51.28	28.65	20.07	-0.1075
mistral-large	51.06	25.08	23.86	-0.0630
qwen2.5	48.83	28.21	22.97	-0.0992

Table 4: COMET-QE before/after correction. Deg.=Degraded, Stab.=Stable, Impr.=Improved.

Across all models, approximately half of the segments exhibit a decrease in COMET-QE score after correction, while a substantial portion remains stable or improves. The average change is negative for all models, indicating a moderate overall decrease, with *mistral-large* showing the smallest.

These results suggest a moderate trade-off between terminology accuracy and overall translation quality. However, the presence of stable and improved segments indicates that the correction process does not systematically degrade translation quality and can preserve, or even enhance, semantic adequacy in a significant number of cases. This confirms that standard MT metrics are largely insensitive to terminology-level errors, even when such errors are systematic.

5 Discussion and Conclusion

The results demonstrate that both standard MT evaluation metrics and commonly used terminology presence heuristics are insufficient for terminology-sensitive tasks. Although all models achieve high scores in BLEU, chrF2, and COMET, these metrics fail to capture whether inconsistent terminology has been correctly substituted by the targeted one. Likewise, the assumption that the presence of the expected term is sufficient to consider a segment properly handled proves inadequate, as outputs may contain the correct term while still preserving the incorrect one. This behavior may partly reflect the constrained correction strategy and placeholder-based replacement pipeline used in this work.

Our findings suggest that, under the constrained correction setup explored in this work, the main challenge is not only generating correct terms, but ensuring clean substitution behavior. This distinction remains largely invisible to standard evaluation metrics, which may explain why terminology errors persist even in otherwise high-quality systems.

The observed behavior is closely related to the constrained prompt design, which favors localized edits over full rewriting. While semantic adequacy is generally preserved, strict terminology substitution is not always achieved.

From an application perspective, these findings are particularly relevant for the use of LLMs in terminology correction workflows. In domains where terminology consistency is critical, LLM-based correction provides a promising approach to improve the quality and consistency of translation memories and MT training data. In real-world localization pipelines, terminology inconsistencies may propagate into downstream MT systems and increase the post-editing effort. However, the observed insertion-like behavior and the moderate COMET-QE degradation indicate that such approaches may still require additional constraints or lightweight validation mechanisms to ensure reliable and controllable behavior in production environments, particularly in terminology-sensitive domains such as technical, medical, or legal translation.

Overall, this work shows that LLM-based terminology correction is effective but not fully reliable. While models achieve comparable levels of apparent correction, important differences emerge in how corrections are realized under the proposed correction pipeline, particularly with respect to substitution-like versus insertion-like outcomes. The proposed terminology-level evaluation framework enables a more precise analysis of these differences and provides a foundation for improving terminology control in domain-specific MT systems and LLM-based data augmentation pipelines.

Finally, the results reveal a controlled trade-off between terminology accuracy and translation quality. While corrections are generally localized and preserve meaning, enforcing terminology does not always guarantee clean substitutions. However, the insertion-like patterns identified in this analysis should not be interpreted exclusively as intrinsic model behavior. Since the experiments rely on constrained minimal-edit prompting and placeholder-based terminology replacement, part of the observed behavior may also be influenced by the correction strategy itself. The proposed evaluation framework should therefore be understood as an approximation of correction behavior within this controlled setup.

Future work will explore alternative prompting and replacement strategies to improve terminology control and better disentangle model behavior from artifacts introduced by the placeholder-based pipeline. In addition, the heuristic evaluation will be replaced with explicit incorrect-term detection.

Acknowledgements

“With the support of the Industrial Doctorate Program of the Department of Research and Universities of the Generalitat of Catalonia”. Ref: 2024 DI 00151

References

- Abhimanyu Dubey et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Albert Q. Jiang et al. 2024. Mistral large. *arXiv preprint*.
- Antoni Oliver and Sergi Alvarez. 2023. Training and integration of neural machine translation with MTUOC. In *Proceedings of the 1st Workshop on Open Community-Driven Machine Translation*, pages 5–13.
- Kishore Papineni, 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.
- Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.
- Ricardo Rei, 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.
- Qwen Team. 2025. Qwen2.5 technical report. *arXiv preprint*.

Flow as Regulated Engagement in Chinese–English Translation: Evidence from Self-Report, Eye Tracking, and Physiological Measures

Yanze Li

School of Foreign Languages,
South China University of Technology,
Guangzhou, China

Department of Language Science and Technology,
The Hong Kong Polytechnic University,
Hong Kong

flyanze.li@mail.scut.edu.cn

Yu Weng

Department of Language Science and Technology,
The Hong Kong Polytechnic University,
Hong Kong

yu-rachel.weng@polyu.edu.hk

Abstract

This paper examines how flow experience, as a task-related psychological state, relates to effort investment, attention allocation, monitoring behavior, and physiological regulation during Chinese–English translation. It synthesizes evidence from three studies within the same research project: a task-based self-report study and two laboratory analyses of the same experimental dataset, focusing respectively on gaze behavior and physiological responses. The self-report findings showed that flow was primarily marked by challenge–skill balance, a sense of control, and reduced self-consciousness. The laboratory findings further indicated that, under positive affective priming, higher flow was associated with smoother ST–TT coordination, lower cognitive effort, and a more regulated physiological profile. However, regression counts did not indicate a simple reduction in monitoring behavior. Taken together, the findings characterize flow in translation as a form of regulated engagement, in which effort and attention are more efficiently organized while the control demands of translation remain present. This account extends the study of flow beyond self-reported experience by incorporating process-level evidence and provides a basis for considering translators' cognitive engagement and sustainable working modes in current technology-mediated translation contexts.

1 Introduction

Translation process research (TPR) has long examined how cognitive effort is distributed across source text (ST) comprehension, target text (TT) production, and processes of monitoring and coordination (Tra&Co Group, 2021). Eye-tracking has become a major method for examining this distribution, as fixation durations, regressions, saccades, and transitions between ST and TT provide indicators of processing demand and attention allocation (Dragsted, 2010; Hvelplund, 2014, 2020; Just & Carpenter, 1980; Rayner, 1998; Weng & Zheng, 2025). These measures have helped clarify where and when effort is invested during translation. However, the relationship between observed patterns of effort investment and translators' subjective experience of the task, particularly flow experience, has received limited empirical attention.

Previous work indicates that affective and motivational factors can influence translation behavior and the allocation of cognitive resources (Lehr & Hvelplund, 2020; Weng et al., 2022). Flow experience is particularly relevant in this respect because it refers to deep task engagement, usually involving a match between challenge and skill, focused attention, a sense of control, and reduced self-consciousness (Csikszentmihalyi, 1990; Nakamura & Csikszentmihalyi, 2021). In translation research, Mirlohi et al. (2011) provided early evidence that trainee translators can experience flow during classroom translation tasks. Given the attentional and control-related features

of flow, along with evidence of its occurrence in translation, flow experience may also be expected to relate to online translation processing.

The present paper examines how flow experience, as a task-related psychological state, relates to observable patterns of effort investment, attention allocation, and physiological responses during Chinese–English translation. It synthesizes findings from three studies within the same research project: a published task-based study in which self-reported flow was measured after a translation task (Weng & Li, 2025), and two ongoing laboratory studies using eye-tracking and physiological measures to investigate how flow experience is reflected in online translation processing (Li et al., forthcoming; Li & Weng, forthcoming). By bringing these findings together, this paper suggests that flow in translation can be characterized as regulated task engagement, in which smoother ST–TT coordination and physiological regulation coexist with sustained monitoring demands.

2 Research background

Flow was first proposed as a state of optimal experience in which individuals become deeply absorbed in an activity for its own sake (Csikszentmihalyi, 1975). Csikszentmihalyi (1990) characterized flow by several interrelated features, including concentration on the task, a sense of control, challenge-skill balance, loss of self-consciousness, clear goals, unambiguous feedback, distorted sense of time, action-awareness merging, and autotelic experience. Research in foreign language learning and translation has suggested the relevance of flow to sustained engagement in language tasks (Dewaele & MacIntyre, 2022; Mirlohi et al., 2011). More recently, Carl et al. (2024) proposed a Hesitation–Orientation–Flow (HOF) taxonomy for translation process data, treating flow as a state of fluent production that can be identified from logged keystrokes and gaze data. Their study situates flow within the temporal organization of translation behavior. However, few studies have examined how self-reported flow, understood as a state of optimal experience, relates to process-level patterns during translation.

The link between flow experience and online translation processing concerns how translators allocate effort between the ST and TT. Fixation duration is commonly interpreted as an index of processing time and cognitive effort, while

regressions are associated with reanalysis, checking, and monitoring (Angelone, 2010; Hvelplund, 2014, 2020; Just & Carpenter, 1980). In translation, transitions between ST and TT can indicate how translators coordinate comprehension and production, and saccade amplitude may reflect the span of visual uptake (Dragsted, 2010; Rayner, 1998). These eye-tracking indicators can therefore provide evidence on whether self-reported flow is associated with particular patterns of effort allocation and ST–TT coordination during translation.

Physiological measures can complement eye-tracking evidence by capturing aspects of arousal and regulation during translation. Skin conductance responses (SCR) are often used as indicators of sympathetic arousal, while heart rate variability measures such as the root mean square of successive differences of inter-beat intervals (RMSSD) are related to autonomic regulation (Tarvainen et al., 2014; Weng et al., 2022). Pupil size can also be treated as an effort-sensitive measure. Translation studies using such measures have shown that physiological indicators can capture aspects of effort and affect that are not directly observable from gaze behavior or the written product alone (Lehr & Hvelplund, 2020; Weng et al., 2022).

Against this background, a broader research project was designed to examine flow experience in Chinese–English translation across three studies, focusing respectively on the experiential profile of flow, gaze-based patterns of effort allocation and ST–TT coordination, and physiological responses. The present paper synthesizes these studies to ask how self-reported flow is related to observable patterns in translators' online processing and, on that basis, what form of task engagement flow represents in translation.

3 Project design and methods

The following section outlines the design of the three studies included in this synthesis. Study 1 examined post-task flow experience in a task-based setting, while Studies 2 and 3 used the same laboratory tasks and participant sample, but focused on different data streams. Table 1 provides an overview.

Evidence source	Participants	Task and measures	Analytic focus
Study 1: task-based self-report study	108 second-year undergraduate English majors	Chinese–English (C-E) translation task; Short Flow State Scale (Martin & Jackson, 2008); Perceived task difficulty	Flow profile; Relation between perceived difficulty and flow
Studies 2–3: laboratory process studies	32 fourth-year undergraduate English majors	Two C-E translation tasks; Positive/neutral affective priming; EyeLink Portable Duo eye-tracking; Empatica E4 physiological recording	Attention allocation; Cognitive effort; Physiological responses

Table 1: Overview of the evidence sources.

The findings below are organized around three strands of evidence: self-report, gaze behavior, and physiological responses.

4 Results

Study 1 depicted the subjective profile of flow in C-E translation (Weng & Li, 2025). K-means clustering distinguished a flow group from a non-flow group, with the flow group accounting for 37.04% of the sample. This proportion suggests that flow can occur during C-E translation, but it is not the default state for student translators. Figure 1 shows the score distributions of the nine flow dimensions for the flow and non-flow groups. The clearest group difference concerned challenge-skill balance: participants in the flow group reported a stronger match between their perceived abilities and the demands of the task. Perceived task difficulty was negatively associated with both challenge-skill balance and flow, and challenge-skill balance mediated the relation between perceived difficulty and flow. The flow group also reported a stronger sense of control and greater loss of self-consciousness, suggesting that they felt able to manage the task and became more absorbed in the translation activity. By contrast, concentration and action-awareness merging were less

discriminative, possibly because translation itself requires sustained attention and student translators still rely on conscious checking. This pattern suggests that flow experience in translation is grounded in perceived manageability and control, rather than in heightened concentration alone.

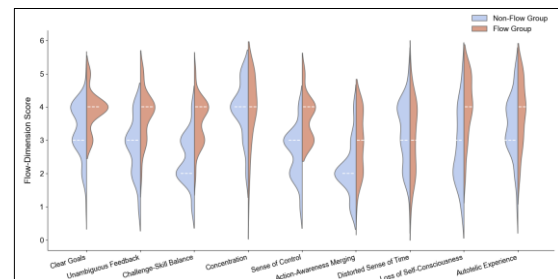


Figure 1: Distributions of flow scores across the nine dimensions in the flow and non-flow groups.

The gaze-based findings refined this subjective profile by showing that flow experience was associated with specific processing patterns, mainly in the positive affective condition (Li et al., forthcoming). For ST–TT transitions, the interaction between affective context and flow was significant ($p = .026$): in the positive condition, higher flow was associated with fewer transitions between ST and TT, a pattern consistent with less fragmented ST–TT coordination. A similar, although marginal, pattern was observed for saccade amplitude ($p = .076$), with higher flow tending to correspond to larger saccades in the positive condition, which may indicate broader visual uptake.

Fixation-based measures provided further evidence for the relation between flow and processing effort. The interaction between affective context and flow was significant for overall mean fixation duration ($p = .035$): in the positive condition, higher flow was associated with shorter fixation durations, indicating lower processing effort. The ST–TT fixation duration ratio also showed a marginal interaction ($p = .074$), with higher flow tending toward a more balanced distribution of attention between ST and TT. By contrast, regression count, the monitoring-related indicator, did not show a stable relation with flow. This pattern is theoretically informative because it suggests that stronger flow was not accompanied by a clear reduction in rechecking or monitoring-related behaviors, cautioning against interpreting flow as a simple decrease of control during translation. Figure 2 summarizes these affective

context $\times$ flow interactions across the eye-tracking measures.

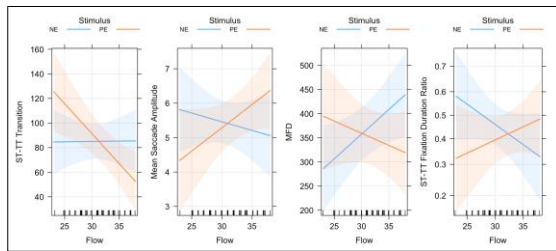


Figure 2: Affective context $\times$ flow interactions across eye-tracking measures.

The physiological findings further extended the process-level account of flow (Li & Weng, *forthcoming*). Higher flow was associated with lower SCR ($p = .003$), suggesting reduced sympathetic arousal. It was also related to higher baseline-corrected RMSSD ($p = .002$), which points to stronger autonomic regulation. Pupil size showed a marginal negative association with flow ($p = .073$), indicating a possible reduction of cognitive load. Taken together, these physiological patterns are consistent with a more regulated working state: translators reporting stronger flow showed lower arousal and stronger autonomic regulation, while the marginal pupil size was in the same direction as the fixation-based evidence for reduced processing demand. Figure 3 exhibits the associations between flow and the three physiological indicators.

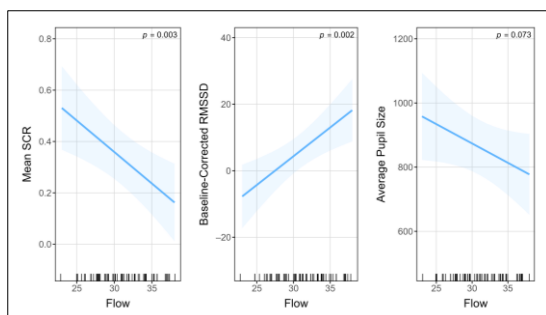


Figure 3: Flow-related physiological patterns.

5 Discussion and Conclusion

Across the three strands of evidence, flow in C-E translation emerged as a regulated mode of task engagement. The self-report findings showed that flow was anchored in challenge-skill balance, a sense of control, and reduced self-consciousness, indicating that translators experienced the task as

manageable rather than merely attention-demanding. The eye-tracking findings located this experience in online processing: under positive affective priming, higher flow was associated with fewer ST–TT transitions, shorter mean fixation durations, and tendencies toward broader visual uptake as well as more balanced ST–TT attention allocation. The physiological findings further showed lower sympathetic arousal and stronger autonomic regulation among translators reporting stronger flow. These results suggest that flow is reflected not only in post-task evaluations of the translation experience but also in the organization of effort, attention, and physiological regulation during the translation process.

Meanwhile, understanding flow as regulated engagement does not mean that monitoring demands disappear from translation. Translation remains a monitoring-intensive activity in which translators need to check ST comprehension, solve emerging problems, and evaluate TT formulations. In this respect, the lack of a significant association between flow and regression count is informative. It prevents simply equating flow with reduced control. Flow may therefore involve a selective reorganization of effort: reducing fragmented or excessive processing demand while preserving the checking and control operations required for translation quality.

This account also has implications for current technology-mediated translation practices. In machine translation and LLM-supported workflows, translators' effort is often redistributed. Less effort may be spent on generating initial solutions, while more attention is required for evaluating suggestions, detecting context-sensitive problems, integrating information across interfaces, and deciding when to accept, revise, or reject machine output. The present findings suggest that flow may offer a useful lens for examining how such workflows shape translators' cognitive engagement. A flow-sensitive perspective can therefore complement product- and system-oriented evaluations of translation workflows by considering whether these workflows help translators maintain controlled, sustainable engagement during human-machine collaboration.

References

Erik Angelone. 2010. Uncertainty, uncertainty management and metacognitive problem solving in the translation task. In Gregory M. Shreve and Erik

- Angelone, editors, *Translation and Cognition*, pages 17–40. John Benjamins.
- Michael Carl, Yuxiang Wei, Sheng Lu, Longhui Zou, Takanori Mizowaki, and Masaru Yamada. 2024. Hesitation, orientation, and flow: A taxonomy for deep temporal translation architectures. *Ampersand*, 12:100164. <https://doi.org/10.1016/j.amper.2024.100164>.
- Mihaly Csikszentmihalyi. 1975. *Beyond Boredom and Anxiety*. Jossey-Bass Publishers.
- Mihaly Csikszentmihalyi. 1990. *Flow: The Psychology of Optimal Experience*. HarperCollins.
- Jean-Marc Dewaele and Peter D. MacIntyre. 2022. “You can’t start a fire without a spark”: Enjoyment, anxiety, and the emergence of flow in foreign language classrooms. *Applied Linguistics Review*, 15(2):403–426. <https://doi.org/10.1515/applirev-2021-0123>.
- Barbara Dragsted. 2010. Coordination of reading and writing processes in translation: An eye on uncharted territory. In Gregory M. Shreve and Erik Angelone, editors, *Translation and Cognition*, pages 41–62. John Benjamins.
- Kristian Tangsgaard Hvelplund. 2014. Eye tracking and translation process: Reflections on the analysis and interpretation of eye tracking data. *MonTI*, Special Issue 1:201–223. <http://dx.doi.org/10.6035/MonTI.2014.ne1.6>.
- Kristian Tangsgaard Hvelplund. 2020. Translation, attention and cognition. In Fabio Alves and Arnt Lykke Jakobsen, editors, *The Routledge Handbook of Translation and Cognition*, pages 279–293. Routledge.
- Marcel Adam Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354. <https://doi.org/10.1037/0033-295X.87.4.329>.
- Caroline Lehr and Kristian Tangsgaard Hvelplund. 2020. Emotional experts: Influences of emotion on the allocation of cognitive resources during professional translation. In Ricardo Munoz Martin and Sandra L. Halverson, editors, *Multilingual Mediated Communication and Cognition*, pages 44–68. Routledge.
- Yanze Li, Yu Weng, and Xianlong Zhu. Forthcoming. Translating in the zone: The joint impact of positive emotions and flow on behavioral and cognitive dynamics in translation. *Target*.
- Yanze Li and Yu Weng. Forthcoming. Flow and resource-efficient bilingual processing: Behavioral and psychophysiological evidence from translation.
- Andrew J. Martin and Susan A. Jackson. 2008. Brief approaches to assessing task absorption and enhanced subjective experience: Examining short and core flow in diverse performance domains. *Motivation and Emotion*, 32:141–157. <https://doi.org/10.1007/s11031-008-9094-0>.
- Mehdi Mirlohi, Joy Egbert, and Behzad Ghonsooly. 2011. Flow in translation: Exploring optimal experience for translation trainees. *Target*, 23(2):251–271. <https://doi.org/10.1075/target.23.2.06mir>.
- Jeanne Nakamura and Mihaly Csikszentmihalyi. 2021. The experience of flow: Theory and research. In C. R. Snyder, S. J. Lopez, L. M. Edwards, and S. Marques, editors, *The Oxford Handbook of Positive Psychology*, pages 279–296. Oxford University Press.
- Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422. <https://doi.org/10.1037/0033-2909.124.3.372>.
- Mika P. Tarvainen, Juha-Pekka Niskanen, Jukka A. Lipponen, Perttu O. Ranta-aho, and Pasi A. Karjalainen. 2014. Kubios HRV: Heart rate variability analysis software. *Computer Methods and Programs in Biomedicine*, 113(1):210–220. <https://doi.org/10.1016/j.cmpb.2013.07.024>.
- Tra&Co Group, editor. 2021. *Translation, Interpreting, Cognition: The Way out of the Box*. Language Science Press, Berlin. <https://doi.org/10.5281/zenodo.4544686>.
- Yu Weng and Yanze Li. 2025. Hanying fanyi zhong de yizhe xinliu tiyan: Jiji xinli jizhi yu renzhi tezheng tanxi [Flow experience in Chinese–English translation: Positive psychological mechanisms and cognitive characteristics]. *Foreign Language Teaching and Research*, 57(2):290–303. <https://doi.org/10.19923/j.cnki.fltr.2025.02.013>.
- Yu Weng and Bingham Zheng. 2025. Automaticity in translation: Effects of time pressure and work experience. *Meta*, 69(2):355–379. <https://doi.org/10.7202/1118382ar>.
- Yu Weng, Bingham Zheng, and Yanping Dong. 2022. Time pressure in translation: Psychological and physiological measures. *Target*, 34(4):601–626. <https://doi.org/10.1075/target.20148.wen>.

Development of an Uzbek–English Parallel Corpus Search System: A Case Study of Literary Works

Shahlo Khamroyeva Mirdjanovna

Tashkent State University of Uzbek
language and literature
Tashkent, Uzbekistan,

shaxlo.xamrayeva@navoiy-uni.uz

Nuriddinova Gulirano Baxtiyor kizi

Tashkent State University of Uzbek
language and literature
Tashkent, Uzbekistan

guliranonuriddinova@gmail.com

Kilicheva Mehriniso Rajabovna

Department of English Literature
Translation Studies
Bukhara State University
Uzbekistan,

m.r.kilicheva@buxdu.uz

Qurbonova Nodira Roziqovna

Department of English Literature and
and Translation Studies
Bukhara State University Bukhara,
Bukhara, Uzbekistan

n.r.qurbonova@buxdu.uz

Boltaeva Gulrukh Erkin kizi

Department of English Literature
and Translation Studies
Bukhara State University,
Bukhara, Uzbekistan,

g.e.boltayeva@buxdu.uz

Kamalova Aziza Islomovna

Department of English philology
Samarkand state institute of
foreign languages
Samarkand, Uzbekistan

aziza19866@icloud.com

Abstract

This article addresses the development of a search system based on an Uzbek–English parallel corpus. In the study, original and translated versions of literary works were collected, and their segmentation and alignment were carried out. Sentence-level alignment methods were applied in the construction of the parallel corpus, and sentence-level alignment, indexing techniques, and exact-match retrieval methods were employed to improve the efficiency of the search system. As a result, a system model was developed that enables users to quickly retrieve equivalent translations between the two languages. The findings demonstrate that the use of parallel corpora plays a significant role in automating the translation process and advancing linguistic analysis.

Keywords: parallel corpus, Uzbek–English languages, search system, NLP, Python, literary texts, alignment, translation, corpus linguistics

1 Introduction

In recent years, the rapid advancement of artificial intelligence and modern technologies has brought several fields within applied philology to a new stage of development. Working with multilingual texts—particularly improving machine translation and information retrieval systems—has become one of the key directions in this domain. In this context, parallel corpora serve as an important linguistic and technological resource. They enable the identification of equivalence between texts in different languages, the automation of translation processes, and the in-depth analysis of linguistic units. For low-resource languages such as Uzbek, however, high-quality parallel corpora remain insufficiently developed. Moreover, most existing resources are not fully systematized or lack efficient search capabilities. In particular, the scarcity of parallel corpora based on literary texts, as well as the absence of effective search systems for working with them, represents a significant scientific and practical challenge.

The object of this study is a parallel corpus composed of literary texts in Uzbek and English. The main objective of the research is to develop an efficient and user-friendly search system based on this corpus. To achieve this goal, several tasks were defined, including

the collection and preparation of parallel texts, their sentence-level alignment, the development of search algorithms, and the evaluation of system performance.

The scientific novelty of this research lies in the proposal of a search system model for an Uzbek–English parallel corpus based on literary texts, implemented using modern Natural Language Processing (NLP) approaches. This approach contributes not only to the optimization of translation processes but also to the advancement of new research outcomes within the field of corpus linguistics.

The results of the study are expected to play an important role in deepening the understanding of linguistic relationships between Uzbek and English, improving translation systems, and advancing multilingual information retrieval technologies.

2 Related Work

Parallel corpora and search systems built upon them constitute one of the key directions in the field of Natural Language Processing (NLP). In early studies, parallel corpora were primarily used for developing machine translation systems, enabling the identification of semantic and syntactic correspondences between different languages. In particular, statistical machine translation models laid the foundational framework for the effective utilization of parallel corpora (Brown et al., 1993). At a later stage, the problem of aligning parallel texts emerged as a distinct area of research. The algorithm proposed by William A. Gale and Kenneth W. Church became widely adopted as an effective method for aligning parallel texts based on sentence length (Gale and Church, 1993). Furthermore, the work of Franz Josef Och and Hermann Ney significantly contributed to improving alignment quality through statistical approaches (Och and Ney, 2003).

Sinclair emphasized corpus-based approaches in linguistics (Sinclair, 1991). In addition, Tony McEnery and Andrew Wilson systematically outlined the methodology of corpus linguistics and emphasized the practical applications of parallel corpora (McEnery and Wilson, 2001). In modern NLP research, parallel corpora are widely used not only in translation but also in information retrieval systems. In particular, neural network-based approaches, such as the Transformer model, have significantly improved the efficiency of working with parallel texts (Vaswani et al., 2017). Multilingual systems developed on the basis of this model, such as BERT and mBERT, have initiated a new stage in the use of parallel corpora (Devlin et al., 2019). Despite these advancements, research on the creation and utilization of parallel corpora for the Uzbek language has mainly focused on building small-scale corpora or addressing specific linguistic tasks. The development of fully functional search systems, however, remains insufficiently explored. In particular,

studies on parallel corpora based on literary texts and the search systems designed to operate on them are limited. From this perspective, the creation of an Uzbek–English parallel corpus based on literary texts and the development of a corresponding search system represent a relevant and significant research challenge.

Recent studies have increasingly focused on multilingual neural architectures, large language models (LLMs), and generative AI systems for low-resource language processing. Transformer-based multilingual models such as XLM-R and mT5 have significantly improved cross-lingual retrieval, semantic alignment, and machine translation quality for underrepresented languages. These developments demonstrate the growing importance of parallel corpora in multilingual NLP and information retrieval research (Conneau et al., 2020; Xue et al., 2021).

3 Methodology

In this study, a multi-stage methodological approach was employed to develop a search system based on an Uzbek–English parallel corpus. The process encompasses several stages, including data collection, preprocessing, alignment, and the construction of the search mechanism.

At the initial stage of the research, literary texts in Uzbek and English were selected and compiled into a parallel corpus. To ensure semantic equivalence, the texts were collected based on their original and translated versions. In the subsequent stage, the texts were cleaned and subjected to preprocessing procedures. This process involved the removal of irrelevant characters, tokenization, normalization, and sentence segmentation.

Sentence alignment methods were applied to align the parallel texts. During the alignment process, algorithms based on statistical approaches were utilized to identify corresponding sentence pairs. This stage represents one of the key factors determining the overall quality of the parallel corpus.

In developing the search system, indexing techniques and information retrieval methods were employed. Indexes were created for each sentence or segment, and a mechanism was developed to efficiently retrieve relevant translation pairs based on user queries. The search process incorporated both keyword-based matching and approaches that account for semantic similarity.

The current implementation primarily relies on sentence-level alignment, indexing techniques, and exact-match retrieval methods. Statistical alignment approaches were used to identify equivalent sentence pairs between Uzbek and English texts. The integration of multilingual semantic models is considered a potential direction for future enhancement of the system. Due to the prototype-oriented nature of the

current study, the evaluation primarily focused on functional testing of bilingual retrieval, exact-match search performance, and usability of the system interface. Future work will include a quantitative evaluation using precision, recall, and F1-score metrics on benchmark datasets.

4 Dataset and Corpus Construction Process

The development of the search system represents the final stage of the study, and its performance largely depends on the quality of the preceding processes. The construction of a parallel corpus requires a precise and systematic approach. In this research, the corpus development was based on the sequential implementation of the following stages:

ID	Original Work	Translated Work	Aligned Sentences
1	Ikki eshik orasi	Between the doors	3289
2	Garri Potter and cursed child	Garri Potter va la'natlanggan bola	3963
3	Al Kimyogar	The Alchemist	1773
4	Atomic habits	Atom odatlar	2001
5	Marten Eden	Martin Iden	8115
6	White Fang	Oqso'yloq	3221
7	Garri Potter and the Goblet of Fire	Garri Potter va olov Kubogi	2032

Table 1: List of Literary Works Selected for the Corpus

1. Data selection and collection
2. Text cleaning and preprocessing
3. Alignment process
4. Annotation process
5. Corpus evaluation and quality control
6. Development of the search system based on the corpus

The selected literary works represent different genres, writing styles, and translation characteristics. The corpus was intentionally limited to seven books in the current stage in order to ensure manual quality control, accurate sentence-level alignment, and consistency between original and translated versions. Future work will focus on expanding the corpus with additional literary and non-literary materials.

All selected literary works were manually revised prior to the alignment stage and subsequently aligned at the sentence level. The resulting corpus contains a total of 24,394 aligned sentence pairs.

4.1 Data Preprocessing

To ensure consistency and uniformity within the corpus, the following preprocessing steps were carried out:

- All texts were converted into string format.
- Cyrillic script was normalized into Latin script.
- Additional fields were created to facilitate efficient search operations: `uzbek_latin`, `english_norm`.

4.2 Search Algorithm

The search system employs an exact match approach. For this purpose, Python's `re` library was used with the following pattern:

```
pattern = r'\b' + re.escape(query) + r'\b'
```

This approach ensures that the query "kitob" matches only the exact word "kitob", while variants such as "kitobim" or "kitobi" are excluded.

4.3 Highlight Mechanism

To enhance user experience, matched words are visually emphasized using the following method:

```
re.sub(pattern, '<mark>...</mark>', text)
```

This allows users to easily identify search results within the text.

4.4 Web Application Development

The backend of the system was developed using the Flask framework. The frontend consists of HTML (with Jinja2 templating) and CSS (responsive design). The core functionalities of the system include: bilingual search (Uzbek and English), Cyrillic ↔ Latin script toggle, filtering by literary work, and highlighting of matched terms.

The system was deployed on the cloud platform Render. During deployment, the project was uploaded to GitHub, a `requirements.txt` file was created, and the application was configured via Procfile.

5 Results

As a result of the study, a parallel corpus search system was developed, consisting of a medium-scale parallel corpus for a low-resource language based on seven literary works and a database containing over 20,000 sentence pairs. The system provides functionalities such as real-time search, highlighting, and exact match retrieval. It enables users to view Uzbek and English sentences side by side, filter results by specific literary works, and efficiently locate queried terms.

The proposed system contributes to the development of parallel corpora for the Uzbek language. It is built on a simple and extensible architecture, allowing for further enhancements in several directions, including the integration of KWIC (Key Word in Context) functionality, lemmatization, statistical analysis, corpus expansion through an administrative panel, and integration with machine translation models.

Additionally, the system can be developed into a national linguistic resource by deploying it under a .uz domain.

You can explore the developed parallel corpus search system via the following link: <https://korpus-parallel.onrender.com>.

Despite the promising results, the current version of the corpus has several limitations. The corpus is limited to literary texts and contains a relatively moderate number of aligned sentence pairs. In addition, certain translation ambiguities and stylistic differences between source and target texts may affect alignment quality. Future research will focus on expanding the corpus, improving annotation procedures, and integrating semantic retrieval techniques.

6 Conclusion

This study resulted in the creation of an Uzbek–English parallel corpus based on literary texts and the development of an efficient search system built upon it. The research systematically covered key stages, including data collection, preprocessing, sentence-level alignment, and the design of search algorithms. As a result, a system was developed that enables users to quickly and conveniently retrieve translation equivalents between the two languages.

The proposed system expands the possibilities for working with parallel corpora in Uzbek and contributes to the partial automation of translation processes. Furthermore, it serves as a practical platform for future research in corpus linguistics and NLP. With further improvements and extensions, the system has the potential to evolve into a comprehensive national linguistic resource.

7 Bibliographical References

- Brown, P. F., Della Pietra, V. J., Della Pietra, S. A., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.
- Conneau, A., et al. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.
- Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.
- McEnery, T. and Wilson, A. (2001). *Corpus Linguistics: An Introduction* (2nd ed.). Edinburgh University Press, Edinburgh.

- Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

- Vaswani, A., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

- Xue, L., et al. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL*.

Benchmarking Multilingual Terminology Translation for UNDRR-ISC Hazard Information Profiles

Maria Carmen Staiano

Eliza Research Center
University of Macerata

Johanna Monti

Eliza Research Center
University of Naples
“L’Orientale”

Francesca Chiusaroli

Eliza Research Center
University of Macerata

Slava Tykhonov

Committee on Data
of the ISC (CODATA)

Ken Hawkins

United Nations Office for
Disaster Risk Reduction (UNDRR)

Hélène Jacot Des Combes

International Science
Council (ISC)

Saini Yang

Integrated Research on
Disaster Risk (IRDR)

Urbano Fra Paleo

University of Extremadura

Virginia Murray

University College London

{m.staiano, f.chiusaroli}@unimc.it, jmonti@unior.it
slava@codata.org, kenneth.hawkins@un.org, hjdc.resilience.consulting@gmail.com,
saini.yang@irdrinternational.org, upaleo@unex.es, v.murray@ucl.ac.uk

Abstract

This paper presents the results of the multilingual translation of the 281 UNDRR-ISC Hazard Information Profiles (HIPs) terms from English into French, Spanish, and Chinese, in response to a request from UN partners, member states, and the scientific community. We compare three translation setups: (1) gpt-oss-20b as an open-weight single-LLM baseline; (2) ChatGPT-5.5 as a state-of-the-art proprietary baseline; and (3) a Voter-Arbitrator multi-agent architecture, which combines three models through consensus-based arbitration, served locally via Ollama. Outputs are aligned with a SKOS-based hazard knowledge organisation system, enabling machine-actionable multilingual terminology interoperable with semantic disaster risk infrastructures. Evaluation against a human-validated gold standard shows that ChatGPT-5.5 achieves the highest Exact Match accuracy across all languages, while the Voter-Arbitrator system consistently outperforms the open-weight baseline on Exact Match (58.0% vs 45.6% for French) and achieves higher cosine similarity, although human evaluation identifies a higher rate of conceptually incorrect terms for Chinese. The findings suggest that consensus-based multi-agent arbitration offers a reproducible, sovereign, and infrastructure-independent alternative to proprietary systems, with performance gains over single open-weight model inference.

1 Introduction

The UNDRR-ISC Hazard Information Profiles (2025), developed within the Sendai Framework for Disaster Risk Reduction 2015-2030 (2015) constitute a standardised classification, definition and description of hazards designed to support global disaster risk monitoring, management and governance.

The Sendai Framework identifies four priority areas for action during the period concerned, namely: (1) Understanding disaster risk; (2) Strengthening disaster risk governance to manage disaster risk; (3) Investing in disaster risk reduction for resilience; and (4) Enhancing disaster preparedness for effective response and to “Build Back Better” in recovery, rehabilitation and reconstruction (UNISDR, 2015).

These priorities informed subsequent efforts to standardise disaster risk terminology, leading to the publication in 2016 of the Intergovernmental Open Indicators and Terminology (United Nations General Assembly, 2016), later adopted by the United Nations General Assembly in 2017¹.

Building on the UNDRR-ISC Hazard Definition and Classification Review (2020) and the UNDRR-ISC Hazard Information Profiles (Murray et al., 2021), the updated HIPs provide

¹<https://www.undrr.org/quick/11605>

structured descriptions of 281 hazards, compiled through an extensive consultative process involving community, national and international scientific and policy experts (2025). Considerable efforts were made to establish a coherent and comprehensive taxonomy of HIPs that included extensive discussions related to the terms used for the hazard types, clusters, 'chapeau' groups, and individual hazard groups. This consensus process uncovered areas where the terminology needed to be clarified in order to ensure more consistent categorisation of hazards.

Disaster risk researchers and policymakers frequently rely on disaster risk reduction (DRR) term definitions to ensure the consistent application of key concepts across institutional, scientific, and operational contexts.

In this landscape, the HIPs serve as a trusted source of scientifically grounded, standardised hazard information used by governments, international organisations, researchers, and educators worldwide. They support data collection, disaster preparedness, loss and damage tracking, and multi-sectoral planning. Above all, they represent a practical resource, valued for being usable, useful, and widely used².

As part of broader efforts to strengthen global disaster risk knowledge and support frameworks such as the Sendai Framework for Disaster Risk Reduction 2015–2030, the UNDRR has identified the multilingual dissemination of hazard knowledge as a strategic priority. Ensuring the availability of HIPs beyond English is essential not only to standardise hazard terminology at regional and national levels, but also to increase usability, foster inclusive risk communication, facilitate international coordination, and support risk-informed decision-making across diverse linguistic and institutional contexts.

However, translating highly specialised hazard terms presents significant challenges, including terminological precision, conceptual alignment, and cultural adaptation. These constraints make the development of scalable, reliable multilingual workflows a critical yet underexplored task. To address this gap, this paper introduces The Minority Report³, an open-source pipeline adopting a multi-model Voter–Arbitrator framework in which candidate translations are proposed, evalu-

ated, and resolved through consensus-based arbitration, integrating terminology-augmented generation with ontology alignment to support scalable, machine-actionable multilingual hazard communication. The contributions of this work are:

- **Multilingual Hazard Terminology Resource:** We publish the full set of 281 UNDRR-ISC hazard terms translated into French, Spanish, and Chinese, with synonyms, validated by professional linguists and by domain experts, available in Table C.
- **Automated Ontology Alignment:** We develop a vector-based retrieval framework that links translated terms to persistent semantic identifiers (e.g., Wikidata Q-IDs), ensuring conceptual grounding and cross-lingual terminological consistency.
- **End-to-End Reproducibility:** We provide a fully reproducible pipeline spanning data acquisition, processing, and publication, including ML-ready Croissant metadata generation (Akhtar et al., 2024) and dataset release via Dataverse.

2 Related Work

Breakthroughs in generative artificial intelligence have demonstrated that LLMs can be useful tools to generate fluent, and contextually relevant text across diverse subject areas, especially in scientific and technical domains (Moslem et al., 2023b; Stokel-Walker and Van Noorden, 2023). Nevertheless, their outputs frequently show terminological inconsistencies, stylistic errors, and limited transparency regarding the linguistic and conceptual sources of the generated content, especially in the risk communication domain (Castaldo et al., 2026; Staiano et al., 2025).

To mitigate these challenges, domain adaptation techniques such as terminology constraints (Zhu et al., 2025) and glossary injection (Moslem et al., 2023a; Briva-Iglesias and Ferre-Fernández, 2026) have been proposed to guide generation and improve terminological consistency. More recently, research has explored the use of multi-agent LLM architectures in MT, showing that these systems might significantly enhance contextual awareness and factual reliability (Briva-Iglesias, 2025). Approaches based on consensus-driven generation, arbitration protocols, and role-

²<https://www.preventionweb.net/drr-glossary/hips>

³<https://github.com/4tikhonov/the-minority-report>



Figure 1: Overview of the UNDRR-ISC Hazard Information Profiles 2025, showing 281 hazards organised into 8 macro-categories and multiple clusters. Source: (United Nations Office for Disaster Risk Reduction (UNDRR) and International Science Council (ISC), (2025)).

specialised agents show promise in improving output precision by reframing text generation as a collaborative process that mimics human reasoning and negotiation process rather than a single-pass inference task (Yao et al., 2022; Ng, 2025; Sin et al., 2025; Wu et al., 2025). The system proposed in this paper can be framed within the emerging paradigm of Terminology-Augmented Generation (TAG) (Di Nunzio, 2025), which emphasises the integration of structured terminological knowledge to guide generation. In contrast to Retrieval-Augmented Generation (RAG), TAG prioritises controlled vocabularies, multilingual consistency, and expert-defined term usage. While TAG has been introduced as a conceptual framework, its practical implementation and evaluation remain largely unexplored. The present work operationalises this paradigm by combining terminology-augmented generation with a consensus-based arbitration mechanism and ontology alignment, providing a concrete application in a high-stakes domain. Finally, the adoption of Semantic Web standards such as SKOS has en-

abled the representation of multilingual terminologies in machine-actionable, interoperable formats aligned with FAIR data principles. Positioned at the intersection of terminology-augmented translation, agentic AI, and semantic knowledge engineering, this work proposes an arbitration-based pipeline aligned with SKOS hazard ontologies.

3 Methodology

3.1 Arbitration-Based Translation Method

Conventional single-pass LLM prompting, when applied without access to domain-specific terminology resources, may produce outputs that are linguistically fluent yet misaligned with controlled vocabularies or expert-defined term usage (Di Nunzio, 2025). In terminology-intensive contexts, this can introduce lexical variability and reduce conceptual accuracy. To investigate whether consensus-driven arbitration offers advantages over single-model translation in this setting, we implement a multi-model Voter-Arbitrator framework combining definition-grounded generation

with consensus-based adjudication.

Each source terminology unit, provided in English, is enriched with its corresponding *Scope Note* (i.e., official definition) prior to translation. This contextual grounding step ensures that candidate translations are generated with explicit access to definitional constraints. By incorporating structured semantic context, the system aims to reduce ambiguity and promote conceptually accurate outputs.

The grounded terminology entry is submitted to 3 independent LLMs operating in parallel: gemma3:27b, deepseek-r1:14b, and gpt-oss:20b, all served locally via Ollama⁴. Each model generates a candidate translation of the term. This multi-model configuration introduces diversity in lexical selection while mitigating individual model bias. Generated candidates undergo a normalisation procedure involving case-insensitive comparison and minor orthographic standardisation. A simple majority threshold is applied: if at least two out of three voter models converge on the same normalised term, the translation is accepted without arbitration. In such cases, the confidence score assigned to the winning translation is the original score returned by the model that generated it.

In cases where candidate translations diverge, the arbitration stage is dispatched to the same voter models, which are re-prompted with the full list of candidates and their confidence scores. Rather than relying on a separate reasoning model, the arbitration process leverages the same LLMs used in the voting stage, prompted to select the most accurate candidate in relation to the provided Scope Note. It should be noted that this evaluation is prompt-driven: while the models may implicitly apply domain knowledge during generation, no structured validation mechanism formally enforces specific quality criteria at the arbitration stage. The selected translation is accompanied by a concise explanation of the decision, increasing interpretability of the process. All prompts used in the voting and arbitration stages are reported in Appendix A and Appendix B.

3.2 Ontology Alignment via Vector Search

To ensure interoperability with existing Semantic Web infrastructures, the pipeline integrates an ontology alignment component based on vector

⁴<https://ollama.com>

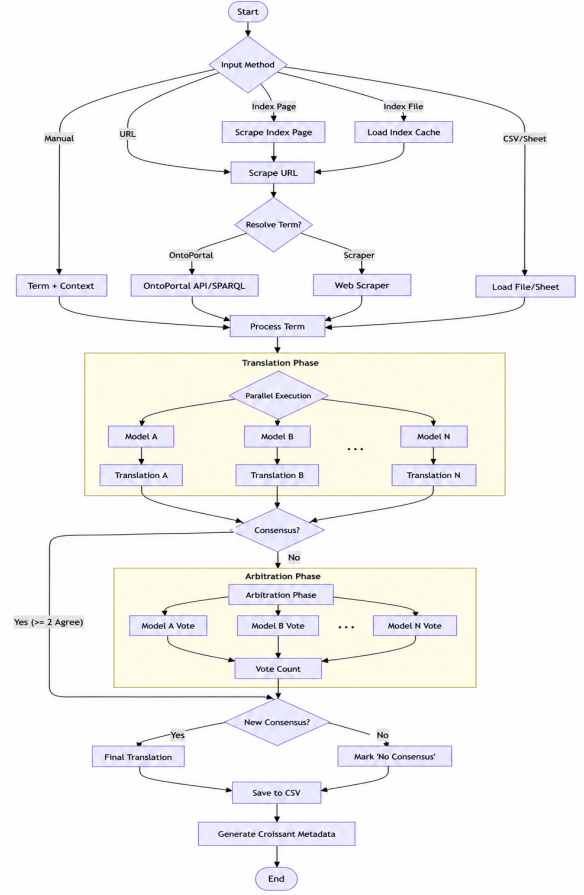


Figure 2: End-to-end architecture of *The Minority Report*.

similarity search. Source terms and definitions are encoded into a space and used to query external knowledge bases including Wikidata (via SparqlMuse) and Wikipedia. Candidate entities are ranked by cosine similarity, and matches below a predefined threshold ($\cos(\theta) < 0.75$) are discarded to mitigate false-positive mappings. Aligned entities are incorporated into the structured terminology record, enabling linkage to external knowledge graphs. To ensure AI-readiness and reproducibility, the pipeline generates structured dataset metadata following the MLCommons Croissant standard (v1.0), including provenance tracking and Wikidata identifiers for each aligned concept.

4 Experimental Investigation

To assess the quality of the multilingual translations generated by the proposed Voter-Arbitrator framework, we conduct a structured evaluation against a human-validated gold standard. We compare three translation setups: (1) gpt-oss-20b,

an open-weight single-LLM baseline operating without definitional context; (2) ChatGPT-5.5, a state-of-the-art proprietary system used as a commercial reference baseline; and (3) our Voter–Arbitrator multi-agent architecture, which combines gemma3:27b, deepseek-r1:14b, and gpt-oss-20b through consensus-based arbitration served locally via Ollama. This setup allows us to test both the contribution of arbitration over single-model inference and the performance gap between open-weight and proprietary systems.

4.1 Corpus Description

The corpus consists of 281 entries from the UNDRR-ISC Hazard Information Profiles (2025), each corresponding to a distinct hazard type. Each entry includes eleven fields: term, synonyms, definition, annotations, metrics and numeric limits, key relevant UN convention or multilateral treaty, drivers, impacts, multi-hazard context, risk management, and monitoring. Metadata fields without translatable content (identifiers, URLs, references, classifications) were excluded. Table 1 summarises the key quantitative measures of the corpus. The quantitative evaluation reported in this paper focuses on the translation of the 281 hazard terms, for which a human-validated gold standard is available. Evaluation of full definition translations is planned as part of the ongoing human evaluation campaign described in Section 7.

Measure	Value
Number of terminological entries	281
Total tokens	550,562
Total types	16,614
Type-Token Ratio (TTR)	0.0302
STTR ($w = 1,000$)	0.3657
Avg. tokens per entry	1,959.30

Table 1: Quantitative profile of the PreventionWeb HIPs corpus.

4.2 Evaluation Metrics

To evaluate translation quality, we adopt a multi-metric evaluation framework combining lexical and semantic similarity measures. Specifically, we employ Exact Match, Levenshtein similarity, and cosine similarity, following recent approaches to LLM-assisted terminology translation in the context of SKOS-based resources (Kraus et al., 2025). All metrics are computed against a gold standard

validated by professional linguists who are native speakers of the target languages.

Exact Match (EM). We compute string-level equality between system outputs and gold standard translations for hazard terms. Prior to comparison, both strings are normalised for case to avoid penalising capitalisation differences. Exact Match provides a strict indicator of terminological alignment but does not account for valid synonyms or acceptable paraphrases.

Lexical Similarity. To capture near-matches and minor surface variations, we calculate normalised Levenshtein distance (Lev). A translation is considered a near-match if the similarity score exceeds 0.8, allowing for minor orthographic variation while maintaining lexical proximity.

Semantic Similarity. To evaluate conceptual equivalence beyond surface-level string matching, we compute sentence embeddings using the paraphrase-multilingual-MiniLM-L12-v2 model (Reimers and Gurevych, 2019), which maps terms from all target languages into a shared multilingual vector space. Cosine similarity (Cos) is then calculated between the system-generated and gold standard embeddings. Following (Zhang and Liu, 2024), a translation is considered semantically acceptable if:

$$\cos(\theta) \geq 0.90$$

This threshold aims for legitimate lexical variation while ensuring high semantic fidelity. We also report mean cosine similarity as a continuous measure of overall semantic proximity.

4.3 Evaluation Results

Table 2 reports the evaluation results obtained by comparing the two baselines and the proposed Voter–Arbitrator framework against the human-validated gold standard across the three target languages.

ChatGPT-5.5 achieves the highest Exact Match accuracy across all languages, confirming the strong terminological alignment of state-of-the-art proprietary systems. The Voter–Arbitrator framework consistently outperforms the gpt-oss-20b baseline on Exact Match across all three languages, demonstrating that consensus-based arbitration improves terminological precision over single-model inference when using open-weight models.

Levenshtein similarity follows a similar pattern: the Voter–Arbitrator outperforms gpt-oss-20b in French and Spanish, while gpt-oss-20b achieves a marginally higher score for Chinese. For cosine similarity, the Voter–Arbitrator achieves its strongest result in Chinese, outperforming gpt-oss-20b, while performing below it for French and Spanish.

5 Human Evaluation

To complement the automatic evaluation, we conducted a qualitative assessment of the mismatches produced by the three systems. Following the recommendations of Briva-Iglesias and Ferre-Fernández (2026) and drawing on a restricted error taxonomy of the Multidimensional Quality Metrics (MQM) framework (Lommel and Melby, 2018), the same domain experts in disaster risk reduction who validated the gold standard terminology, as described in Section 1, manually performed a binary classification of the non-matching outputs. Each annotator was assigned exclusively to their native language: French, Spanish, and Chinese respectively. The two categories are: (1) *acceptable variant or synonym*, including diatopic variants recognised in the target language community; and (2) *wrong term*, where the translation uses a form that a domain expert would not recognise as appropriate, or introduces a conceptual mismatch with respect to the source term. This analysis covers the full set of mismatches across the three languages and the systems used in the current study.

5.1 Mismatch Analysis

Table 3 reports the binary mismatch classification across the three systems and languages. As discussed in Section 4.2, Exact Match is a strict surface-level metric that does not account for valid lexical alternatives such as synonyms or diatopic variants. This limitation becomes evident when comparing EM scores with embedding-based measures. For the Voter–Arbitrator, French and Spanish achieve Exact Match rates of 58.0% and 55.5% respectively, yet their Levenshtein near-match rates exceed 65% and mean cosine similarity scores reach 0.884 and 0.896, suggesting that a substantial proportion of non-matching cases correspond to semantically equivalent translations rather than errors. The human evaluation confirms this: for French and Spanish, the ma-

jority of mismatches were classified as acceptable variants rather than wrong terms (FR: 91 vs. 27; ES: 88 vs. 37), consistent with the high embedding-based scores.

Chinese presents a distinct evaluation profile. Exact Match (47.0%) and Levenshtein similarity (55.5%) are the lowest across all three languages, yet cosine similarity reveals a different trend: 77.6% of translations exceed the $\cos(\theta) \geq 0.90$ threshold. This divergence is partly structural: in character-based writing systems, orthographic distance between semantically equivalent translations is inherently amplified. The two renderings of *sinkhole* (塌陷坑 vs. 天坑) illustrate this well. Both terms are correct, yet their Levenshtein similarity is low due to character-level differences, while their cosine score remains high.

However, the Chinese results also expose a limitation of embedding-based evaluation. Unlike French and Spanish, where acceptable variants consistently outnumber wrong terms, the human evaluation for Chinese reveals the opposite pattern. For the Voter–Arbitrator system, 125 mismatches were classified as incorrect terms, compared to only 24 acceptable variants (16%), indicating a clear contrast with French (77%) and Spanish (70%). This suggests that for Chinese, a non-trivial share of mismatches represent systematic translation errors that high cosine scores fail to penalise. The translation of *mercury* is a clear example: the Voter–Arbitrator system rendered 水星 (the astronomical body) rather than the chemically correct 汞, yet obtained a cosine score of 0.828, below the acceptance threshold, but still misleadingly high for a conceptually incorrect term.

In conclusion, these findings confirm that automatic metrics alone are insufficient for the rigorous evaluation of specialised terminology translation, and underscore the indispensable role of human evaluation, especially across morphologically and orthographically distant languages.

6 Conclusions

This paper presented a comparative evaluation of three translation systems for specialised disaster risk terminology: gpt-oss-20b as an open-weight single-LLM baseline, ChatGPT-5.5 as a commercial reference baseline, and our consensus-based Voter–Arbitrator multi-agent architecture, applied to the 281 UNDRR-ISC Hazard Information Profile terms across French, Spanish, and Chinese.

System	Lang	Correct terms	Accuracy (%)	Lev ≥ 0.8 (%)	Cos ≥ 0.9 (%)
gpt-oss-20b	FR	128/281	45.6	65.1	64.4
	ES	123/281	43.8	65.5	69.0
	ZH	127/281	45.2	56.9	76.2
Voter–Arbitrator	FR	163/281	58.0	71.9	61.2
	ES	156/281	55.5	67.6	64.8
	ZH	132/281	47.0	56.6	77.6
ChatGPT-5.5	FR	193/281	68.7	82.2	84.0
	ES	172/281	61.2	77.2	80.8
	ZH	160/281	56.9	68.7	84.3

Table 2: Evaluation results against human-validated gold standard for 281 hazard terms. Bold values indicate the best-performing open-weight system. Shaded rows indicate the proprietary commercial reference baseline (ChatGPT-5.5).

System	Lang	Acceptable	Wrong	Total
gpt-oss-20b	FR	104	49	153
	ES	91	67	158
	ZH	38	116	154
Voter–Arbitrator	FR	91	27	118
	ES	88	37	125
	ZH	24	125	149
ChatGPT-5.5	FR	84	4	88
	ES	87	22	109
	ZH	57	64	121

Table 3: Binary mismatch classification. Acceptable = acceptable variant or synonym; Wrong = wrong term; Total = total error spans. Shaded = ChatGPT-5.5 proprietary baseline.

Empirical assessment against a human-validated gold standard shows that ChatGPT-5.5 achieves the highest Exact Match accuracy across all languages (68.7% FR, 61.2% ES, 56.9% ZH). The Voter–Arbitrator consistently outperforms the gpt-oss-20b baseline on Exact Match across all three languages, demonstrating that consensus-based arbitration improves terminological precision over single-model inference when using open-weight models. However, performance alone does not capture the full value of the proposed approach. Beyond accuracy, the Voter–Arbitrator offers a sovereign and infrastructure-independent alternative to proprietary systems: by running exclusively on locally-served open-weight models via Ollama, it avoids reliance on cloud providers for which comprehensive environmental reporting remains limited (Riemland, 2026).

The integration of ontology alignment and Croissant metadata further ensures interoperability within Semantic Web infrastructures and AI-ready data ecosystems, supporting scalable multilingual hazard communication beyond the three languages evaluated in this study.

7 Future Work

The present study is scoped to term-level translation evaluation; while this provides a rigorous and reproducible benchmark, it does not capture the full complexity of translating complete hazard definitions, which represent the next critical stage of the HIPs multilingual workflow.

Future work will extend the system along two main directions. First, we plan to apply the same Voter–Arbitrator pipeline to the translation of full HIPs definitions, introducing a human-in-the-loop interface to resolve arbitration edge cases and provide structured feedback in cases of persistent model disagreement. This will allow domain experts to validate not only terminological choices but also definitional accuracy and conceptual consistency across languages. Second, a structured human evaluation campaign will assess the quality of these full definition translations across all 281 HIPs, following MQM guidelines (Lommel and Melby, 2018). Human evaluators will work against a gold standard that includes the validated term translations produced in this study. The evaluation framework draws on established shared tasks in terminology and LLM evaluation, including CLEF 2024 SimpleText (Di Nunzio et al., 2024) and CLEF 2025 BioASQ GutBrain Infor-

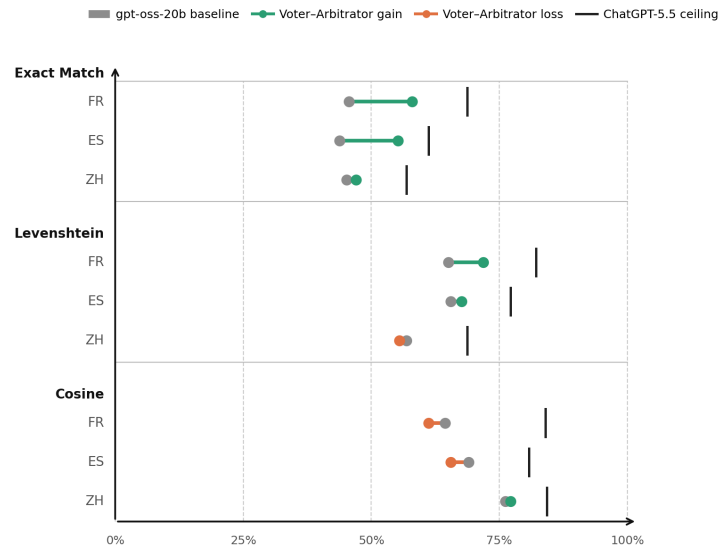


Figure 3: Translation accuracy gains and losses from consensus-based arbitration across nine metric–language combinations. Each row compares the open-weight single-model baseline (grey dot) against the Voter–Arbitrator system (coloured dot); the connecting line shows the direction and magnitude of change — green for improvement, coral for regression. The vertical tick marks the ChatGPT-5.5 proprietary reference ceiling.

mation Extraction (Martinelli et al., 2025). Results from this campaign will inform further refinement of the arbitration strategies and provide a more comprehensive assessment of system performance beyond term-level accuracy.

8 CO2 Emission Related to Experiments

Experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.432 kgCO₂eq/kWh. A cumulative of 23 hours of computation was performed on hardware of type A100 PCIe 40/80GB. Total emissions are estimated to be 2.48 kgCO₂eq, of which 0% were directly offset. Estimations were conducted using the Machine Learning Impact calculator presented in Lacoste et al. (2019).

Acknowledgments

This work was partially supported by the PhD Programme in Humanities and Technologies funded by the University of Macerata under D.R. No 253/2023. This work was also partially enabled as an activity linked to the Climate-Adapt4EOSC project (Horizon Europe Framework Programme grant agreement No 101188248).

The authors would like to thank the UNDRR-ISC HIPs 2025 update Machine Actionability Group for the insightful discussions and constructive feedback throughout this work, including Marcus Elten, Matti Heikkurinen, Simon Hod-

son. In addition, the authors would like to thank all those who have helped with this work, including Xuan Che, Iria Touzon Calle, Animesh Kumar, Ms. Lian Fang from IRDR-IPO, Prof. Du Juan from Beijing Normal University, Rubén del Campo Hernández from the Spanish Meteorological Agency (AEMET), Miguel Llorente Isidro from the Geological and Mining Institute of Spain (IGME), Antonio Castaldo, René Alberto García Taboada, Sebastián Ortega Jiménez, and Bonnie Xinyue Ren.

References

- Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Luca Foschini, Pieter Gijsbers, Joan Giner-Miguel, Sujata Goswami, Nitisha Jain, Michalis Karamousadakis, Satyapriya Krishna, et al. 2024. [Croissant: A metadata format for ml-ready datasets](#). *Advances in Neural Information Processing Systems*, 37:82133–82148.
- Vicent Briva-Iglesias. 2025. [Are AI agents the new machine translation frontier? Challenges and opportunities of single- and multi-agent systems for multilingual digital communication](#). In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 365–377, Geneva, Switzerland. European Association for Machine Translation.
- Vicent Briva-Iglesias and María Ferre-Fernández. 2026. [AI-assisted cultural heritage dissemination: Comparing NMT and glossary-augmented LLM translation in rock art documents](#). In *Proceedings of*

- the 26th Annual Conference of the European Association for Machine Translation, Tilburg, Netherlands. European Association for Machine Translation.
- Antonio Castaldo, Maria Carmen Staiano, Johanna Monti, Sheila Castilho, and Francesca Chiusaroli. 2026. [Translating under pressure: Domain-aware LLMs for crisis communication](#). In *Proceedings of the 26th Annual Conference of the European Association for Machine Translation*, Tilburg, Netherlands. European Association for Machine Translation.
- Giorgio Maria Di Nunzio. 2025. [Terminology-augmented generation \(TAG\): Foundations, use cases, and evaluation paths](#). *Journal of Digital Terminology and Lexicography*, 1(1):97–104.
- Giorgio Maria Di Nunzio, Federica Vezzani, Hosein Azarbonyad, Jaap Kamps, Liana Ermakova, et al. 2024. [Overview of the CLEF 2024 SimpleText task 2: identify and explain difficult concepts](#). In *CEUR Workshop Proceedings*, volume 3740, pages 3129–3146. CEUR-WS.
- Felix Kraus, Nicolas Blumenröhr, Danah Tonne, and Achim Streit. 2025. [Mind the language gap in digital humanities: LLM-aided translation of SKOS thesauri](#). *Anthology of Computers and the Humanities*, 3:883–904.
- Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. [Quantifying the carbon emissions of machine learning](#).
- Arle Lommel and Alan Melby. 2018. [Tutorial: MQM-DQF: A Good Marriage \(Translation Quality for the 21st Century\)](#). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, Boston, MA. Association for Machine Translation in the Americas.
- Marco Martinelli, Gianmaria Silvello, Vanessa Bonato, Giorgio Maria Di Nunzio, Nicola Ferro, Ornella Irera, Stefano Marchesin, Laura Menotti, Federica Vezzani, et al. 2025. [Overview of GutBrainIE@CLEF 2025: gut-brain interplay information extraction](#). In *CEUR Workshop Proceedings*, volume 4038. CEUR-WS.
- Yasmin Moslem, Rejwanul Haque, John Kelleher, and Andy Way. 2023a. [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. European Association for Machine Translation.
- Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John Kelleher, Rejwanul Haque, and Andy Way. 2023b. [Domain terminology integration into machine translation: Leveraging large language models](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 902–911, Singapore. Association for Computational Linguistics.
- V. Murray, J. Abrahams, C. Abdallah, K. Ahmed, L. Angles, D. Benouar, A. Brenes Torres, C. Chang Hun, S. Cox, J. Douris, L. Fagan, U. Fra Paleo, Q. Han, J. Handmer, S. Hodson, W. Khim, L. Mayner, N. Moody, L. Moraes, M. Nagy, J. Norris, P. Peduzzi, A. Perwaiz, K. Peters, J. Radisch, M. Reichstein, J. Schneider, A. Smith, C. Souch, A.-S. Stevance, A. Triyanti, M. Weir, and N. Wright. 2021. [Hazard Information Profiles Supplement to UNDRR-ISC Hazard Definition and Classification Review](#). United Nations Office for Disaster Risk Reduction and International Science Council, Geneva and Paris.
- Andrew Y. Ng. 2025. Translation-agent: A multi-agent translation framework. <https://github.com/andrewyng/translation-agent>. GitHub repository.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Matt Riemland. 2026. [Eco-translation practice in resisting ai’s ecological harms: Towards a preliminary action framework](#). *Encounters in Translation*, 5.
- King-kui Sin, Xi Xuan, Chunyu Kit, Clara Ho-yan Chan, and Honic Ho-kin Ip. 2025. [Solving the unsolvable: Translating case law in Hong Kong](#). *arXiv preprint arXiv:2501.09444*.
- Maria Carmen Staiano, Lifeng Han, Johanna Monti, and Francesca Chiusaroli. 2025. [ITALERT: Assessing the quality of LLMs and NMT in translating Italian emergency response text](#). In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 566–577, Geneva, Switzerland. European Association for Machine Translation.
- Chris Stokel-Walker and Richard Van Noorden. 2023. [What ChatGPT and generative AI mean for science](#). *Nature*, 614(7947):214–216.
- UNISDR. 2015. [Sendai Framework for Disaster Risk Reduction 2015–2030](#). United Nations Office for Disaster Risk Reduction, Geneva.
- United Nations General Assembly. 2016. [Report of the Open-ended Intergovernmental Expert Working Group on Indicators and Terminology Relating to Disaster Risk Reduction](#). United Nations, New York.
- United Nations Office for Disaster Risk Reduction and International Science Council. 2020. [UNDRR-ISC Hazard Definition and Classification Review: Technical Report](#). United Nations Office for Disaster Risk Reduction and International Science Council, Geneva and Paris.

- United Nations Office for Disaster Risk Reduction (UNDRR) and International Science Council (ISC). 2025. *UNDRR-ISC Hazard Information Profiles – 2025 Update*. Geneva: United Nations Office for Disaster Risk Reduction; Paris: International Science Council.
- Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wan, Weihua Luo, and Kaifu Zhang. 2025. (Perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Transactions of the Association for Computational Linguistics*, 13:901–922.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2022. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.
- Hanlu Zhang and Wenyan Liu. 2024. A comparative study on large language models’ accuracy in cross-lingual professional terminology processing: An evaluation across multiple domains. *Journal of Advanced Computing Systems*, 4(10):55–68.
- Lichao Zhu, Maria Zimina-Poirot, Stephane Patin, and Cristian Valdez. 2025. Fine-tuning NMT models and LLMs for specialised EN-ES translation using aligned corpora, glossaries, and synthetic data: MultiTan at WMT25 terminology shared task. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1284–1291, Suzhou, China. Association for Computational Linguistics.

A Voter Agent Prompt

We report the prompt used in the voting stage of the Voter-Arbitrator pipeline described in Section 3.1, to ensure the replicability of the study. The prompt is dispatched in parallel to the three voter models (gpt-oss:latest, gemma3:27b, deepseek-r1:14b).

Prompt for Voting

You are a highly specialized technical translator agent. Your task is to translate the given technical term from English into the target language.

Context: The term is part of a Controlled Vocabulary for Machine Learning and AI. A “Scope Note” will be provided to give you the exact context of the term.

Instructions:

1. Analyze the provided “Scope Note” carefully.
2. Translate the term into the following languages: `{{target_languages}}`.
3. **CRITICAL:** Provide ONLY the single most standard, officially recognized technical term for this concept.
 - Do NOT provide synonyms, lists of options, or descriptive explanations.
 - Do NOT provide “literal” translations if a specific technical term exists.
 - If there are multiple variants, choose the one most used in scientific or official government contexts (e.g., WMO, UN-DRR, ISO).
4. Output ONLY valid JSON where keys are the language codes (e.g. “fr”, “es”, “de”).
5. Do not include markdown formatting.

Input:

- Term: `{{term}}`
- Scope Note: `{{scope_note}}`

Output JSON:

```
{
  "fr": {"translation": "...",
        "confidence_score": 0.95},
  "es": {"translation": "...",
        "confidence_score": 0.98}
}
```

Prompt for Arbitration

You are the Arbitrator Agent. You have received translations from multiple sources (or just one verified source) for a technical term. Your job is to select the BEST translation or synthesize a better one if all are flawed.

Source Term: `{{term}}` **Scope Note:** `{{scope_note}}`

Candidates: `{{candidates}}`

Instructions:

1. Review the candidates and their confidence scores.
2. Select the winning translation.
3. If the consensus is weak (different translations), use your reasoning capabilities to decide the most accurate technical term.
4. Assign a final confidence score.

Output Format (JSON):

```
{
  "selected_translation": "...",
  "reasoning": "...",
  "final_confidence_score": 0.98,
  "winning_model": "...",
  "rai_flags": []
}
```

B Arbitrator Agent Prompt

We report the prompt used in the arbitration stage of the Voter-Arbitrator pipeline described in Section 3.

C Multilingual Terminology for the HIPs

Table C lists the 281 terms in English, French, Spanish, and Chinese, organised according to the eight hazard categories defined in the 2025 HIPs (2025). Synonyms are shown in smaller font below the main term.

English	French	Spanish	Chinese
Meteorological and Hydrological			
avalanche	avalanche	avalancha	雪崩
blizzard	blizzard	nevasca nevazón	暴风雪
coastal flooding	submersions marines submersions côtières, inondations côtières	inundación costera	海岸洪水
cold wave	vague de froid	ola de frío	寒潮
depression or cyclone (low pressure area)	dépession ou cyclone (zone de basse pression)	depresión o ciclón (zona de baja presión) borrasca	低气压或气旋（低压区）
derecho	derecho	derecho	雷暴大风
downburst	rafale descendante	reventón	下击暴流
drought	sécheresse	sequía	干旱
dust storm or sandstorm	tempête de poussière ou tempête de sable	tormenta de polvo o tormenta de arena	沙尘暴
dzud	dzud	dzud	冬季严寒/“白灾”/极寒 雪灾
estuarine (coastal) flooding	inondations estuariennes (côtières)	inundación de estuario	河口（海岸）洪水
extra-tropical cyclone	cyclone extratropical	ciclón extratropical	温带气旋
flash flooding	inondations éclair crues soudaines	inundación relámpago inundación repentina	山洪 / 暴洪
flooding	inondations crues	inundación crecida, avenida, riada	洪水
fluvial (riverine) flooding	inondations fluviales (riveraines) crues fluviales	inundación fluvial crecida	河道（河流）洪水
fog	brouillard	niebla	雾
freeze	gel	helada	冻结
freezing rain (ice storm)	pluie verglaçante (tempête de glace)	lluvia engelante lluvia gélida	冻雨（冰风暴）
frost (hoar frost)	givre (givre blanc)	escarcha	霜（白霜）
gale (strong gale)	vent fort (coup de vent)	racha (rágafa muy fuerte) galerna	大风（烈风）

English	French	Spanish	Chinese
glacial lake outburst flooding	crue due à la vidange d'un lac glaciaire vidange d'un lac glaciaire	inundación por desbordamiento de lago glaciario	冰湖溃决洪水
glaze	verglas	hielo transparente	雨凇
ground frost	gel au sol gel du sol	suelo helado	地面霜冻
groundwater flooding	inondations par remontée de nappe phréatique	inundación por aguas subterráneas	地下水洪水
hail	grêle	granizada	冰雹
haze	brume sèche	calima	霾
heatwave	canicule vague de chaleur	ola de calor	热浪
ice flow	écoulement glaciaire	hielo flotante	冰流
ice-jam flooding including debris	inondations dues aux embâcles, y compris les débris	inundación por represamiento de hielo y detritos	凌汛（凌汛洪水）
icing (including ice)	givrage	engelamiento	结冰（含冰）
lightning (electrical storm)	foudre (orage électrique)	rayo (tormenta eléctrica)	闪电(雷暴)
marine heatwave	vague de chaleur marine canicule marine	ola de calor marina	海洋热浪
ponding (drainage) flooding	inondation par saturation du sol	inundación por encharcamiento	内涝
rain	pluie	lluvia precipitación	雨
rogue wave	vague scélérate	ola gigante	异常巨浪 奇异波
sand haze	brume de poussière	calima calina	沙霾
sea ice (icebergs)	glace de mer (icebergs)	hielo marino iceberg	海冰（冰山）
seiche	seiche	seiche onda de resonancia, onda larga estacionaria	假潮
smoke	fumée	humo	烟
snow	neige	nieve	雪
snow storm	tempête de neige	tormenta de nieve	暴风雪
snowmelt flooding	inondation due à la fonte des neiges	inundación por deshielo	融雪洪水
squall	bourrasque	turbonada	飑
storm surge	onde de tempête	marea ciclónica marea meteorológica, surgencia ciclónica, marea de tempestad, mar de leva	风暴潮
storm tides	marées de tempête	marea de tormenta	风暴潮位

English	French	Spanish	Chinese
sub-tropical cyclone	cyclone subtropical	ciclón subtropical	亚热带气旋
surface water flooding	inondations par ruissellement inondations torrentielles	inundación por agua superficial	地表水洪水
thaw	dégel	deshielo	解冻
thunderstorm	orage	tormenta	雷雨
tornado	tornado	tornado	龙卷风
tropical cyclone	cyclone tropical	ciclón tropical	热带气旋
tsunami	tsunami	tsunami	海啸
wind	vent	viento	风

Extraterrestrial

airburst	explosion en vol	explosión aérea	空爆
geomagnetic disturbance	perturbations géomagnétiques orages magnétiques	perturbación geomagnética	地磁扰动
ionospheric disturbances	perturbations ionosphériques	perturbación ionosférica	电离层扰动
meteorite impact	impact de météorites	impacto de meteorito	陨击作用
near-Earth object	objet géocroiseur astéroïde	objeto cercano a la Tierra	近地天体
solar energetic particles	particules énergétiques solaires	partículas energéticas solares	太阳高能粒子
solar flares	éruptions solaires	llamarada solar	太阳耀斑
space debris	débris spatiaux	basura espacial	空间碎片
space hazard / accident	risque spatial / accident	peligro / accidente espacial amenaza / accidente espacial	空间危险 / 事故
ultraviolet radiation	rayonnement ultraviolet	radiación ultravioleta	紫外线辐射

Geological

ash/tephra fall (including volcanic ballistic projectiles)	chute de cendres et de téphra (y compris les projectiles balistiques volcaniques) pluie de cendres, retombées de cendres, projections de cendres et de téphra	caída de piroclastos	火山灰 / 火山碎屑沉降 (含火山弹道抛射物)
coastal erosion and accretion	érosion et accrétion côtières	erosión y acreción costera	海岸侵蚀与淤积
compressive soils	sols compressibles	suelos compresibles	压缩土壤
debris and earth (mud) flows and rock avalanches	coulées de débris et de boue et avalanches de roches	colada de derrubios, de barro y avalancha de rocas flujo de derrubios, flujo de detritos	泥石流和岩崩

English	French	Spanish	Chinese
earthquake	séisme tremblement de terre	terremoto seísmo, sismo, temblor	地震
gravitational mass movement ('landslide')	mouvement de masse gravitaire (glissement de terrain)	movimiento gravitacional en masa (deslizamiento/deslave)	重力性斜坡物质运动（滑坡） 重力滑坡
ground gases (CH ₄ , Rn, etc.)	gaz du sol (CH ₄ , Rn, etc.)	gases del subsuelo (CH ₄ , Rn, etc.) emanación de gases	地面气体（CH ₄ 、Rn等）
lahars	lahars	lahar	火山泥流
lava flows and domes	coulées et dômes de lave	colada y domos de lava flujo y domo de lava	熔岩流与熔岩穹丘
liquefaction	liquéfaction	licuefacción fluidificación del suelo	液化
pyroclastic density current	courant de densité pyroclastique coulées pyroclastiques, écoulements pyroclastiques, nuées ardentes	corriente de densidad piroclástica nube de piroclastos, nube ardiente	火山碎屑密度流
river erosion and accretion	érosion et accrétion fluviale	erosión y acreción fluvial	河流侵蚀与淤积
rock, debris and earth (mud) slide	glissement de roches, de débris et de terre (boue)	deslizamiento de rocas, detritos y tierra (alud de barro) huaico, huayco, lloclla	岩石、碎石和泥土（泥浆）滑坡
rock, debris and earth falls	chutes de pierres, de débris et de terre	desprendimiento de rocas, suelo y derrubios caída de rocas, descalce de ladera	岩石、碎石和泥土崩塌 崩塌
rock, debris and earth spreads (including landscape creep)	éboulements de roches, de débris et de terre (y compris glissements de terrain)	corrimiento de rocas, detritos y expansión lateral del terreno (incluyendo la reptación)	岩石、碎石和土方扩散（包括景观蠕变）
rock, debris and earth topples	basculement de roches, de débris et de terre	derrumbe de rocas, tierra, detritos vuelco de rocas	岩石、碎石和泥土倾倒
sand encroachment	ensablement	intrusión de arena	沙地侵蚀 沙侵
shrink-swell subsidence	retrait-gonflement des argiles	subsistencia por contracción o expansión del suelo	收缩膨胀下沉
sinkhole	doline	dolina	天坑
soil degradation	dégradation des sols	degradación del suelo	土壤退化
soil erosion	érosion des sols	erosión del suelo	土壤侵蚀 水土流失
submarine gravitational mass movement ('landslide')	mouvement de masse gravitaire sous-marin (glissement de terrain)	movimiento gravitacional en masa submarino deslizamiento submarino	海底重力性斜坡物质运动（滑坡） 海底滑坡
subsidence and uplift	affaissement et soulèvement	subsistencia y levantamiento	沉降和隆起
surface rupture and fissuring	rupture et fissuration de la surface	fisuración y ruptura de la superficie	地表破裂与地裂缝

English	French	Spanish	Chinese
volcanic gases and aerosols	gaz et aérosols volcaniques fumerolles	gases y aerosoles volcánicos	火山气体与气溶胶
Environmental			
acid rain	pluie acide	lluvia ácida	酸雨
air pollution (point source)	pollution atmosphérique (sources ponctuelles)	contaminación atmosférica (fuente puntual) polución atmosférica (fuente puntual)	空气污染（点光源/点能 源/点辐射源）
ambient (outdoor) air pollution	pollution de l'air ambiant (extérieur)	contaminación del aire exterior	环境（室外）空气污染
biodiversity loss	perte de biodiversité érosion de la biodiversité, déclin de la biodiversité	pérdida de biodiversidad	生物多样性丧失
black carbon	carbone noir carbone suie	carbón negro	黑碳/黑碳芯
coral bleaching	blanchissement des coraux	blanqueamiento del coral decoloración del coral	珊瑚白化
deforestation	déforestation	deforestación	毁林 森林滥伐
desertification	désertification	desertificación	荒漠化
eutrophication	eutrophisation	eutrofización	富营养化
forest declines and diebacks	déclin et dépérissement des forêts	decaimiento y mortalidad regresiva forestal	森林衰退与枯死
forest disturbances	perturbations forestières naturelles	perturbación forestal	森林扰动
forest invasive species	espèces forestières envahissantes espèces exotiques envahissantes en forêt	especie invasora forestal	森林入侵物种
household air pollution	pollution de l'air intérieur	contaminación del aire doméstico polución del aire doméstico	室内空气污染
land degradation	dégradation des sols	degradación del suelo	土地退化
loss of mangroves	perte des mangroves disparition des forêts de mangrove	pérdida de manglares	红树林丧失
ocean acidification	acidification des océans	acidificación de los océanos	海洋酸化
permafrost loss	dégel du pergélisol fonte du pergélisol	pérdida de permafrost	多年冻土退化
runoff / nonpoint source pollution	ruissellement / pollution diffuse	contaminación difusa por escurrentía	径流/非点源污染
salinity & sodicity	salinité & sodicité	salinidad y sodicidad	盐碱化
saltwater intrusion	intrusion saline biseau salé, intrusion d'eau salée	intrusión salina	盐水侵入

English	French	Spanish	Chinese
sand mining	extraction de sable surexploitation du sable	extracción de arena	采砂
sea level rise	élévation du niveau de la mer	subida del nivel del mar aumento del nivel del mar	海平面上升
wetland loss/degradation	perte/dégradation des zones humides	degradación y pérdida de humedales	湿地丧失/退化
wildfires	incendies de forêt	incendio forestal	野火

Chemical

aflatoxins	aflatoxines	aflatoxinas	黄曲霉毒素 黄曲霉素
ammonia	ammoniac	amoníaco	氨
ammonium nitrate	nitrate d'ammonium	nitrato de amonio	硝酸铵
arsenic	arsenic	arsénico	砷
asbestos	amiante	amianto	石棉
asphyxiant gases	gaz asphyxiants	gases asfixiantes	窒息性气体
benzene and hydrocarbons	benzène et hydrocarbures	benceno y hidrocarburos	苯和碳氢化合物
cadmium	cadmium	cadmio	镉
carbon monoxide	monoxyde de carbone	monóxido de carbono	一氧化碳
chemical warfare agents	agents de guerre chimique	agentes de guerra química	化学战剂
chlorine	chlore	cloro	氯
corrosive substances	substances corrosives	sustancias corrosivas	腐蚀性物质
dioxins and dioxin-like substances	dioxines et substances apparentées aux dioxines	dioxinas y sustancias similares a las dioxinas	二噁英及类二噁英物质
fluoride and iodine/iodide excess or inadequate intake	fluorure et excès ou apport inadéquat en iode/iodure	ingesta excesiva o inadecuada de yodo/yoduro o flúor	氟和碘/碘摄入过量或不足
heavy metals and other trace elements	métaux lourds et autres oligo-éléments	metales pesados y otros elementos traza	重金属和其他微量元素
lead	plomb	plomo	铅
levels of contaminants in food and feed	concentrations de contaminants dans les denrées alimentaires et les aliments pour animaux	presencia de contaminantes en alimento humano y animal	食品和饲料中的污染物含量
marine toxins	biotoxines marines	toxinas marinas	海洋毒素
mercury	mercure	mercurio	汞
methanol	méthanol	metanol	甲醇
microplastics	microplastiques	microplásticos	微塑料
opioids and other psychoactive substances	opioïdes et autres substances psychoactives	opioides y otras sustancias psicoactivas	阿片类及其他精神活性物质

English	French	Spanish	Chinese
perfluoroalkyl and polyfluoroalkyl substances	substances perfluoroalkyles et polyfluoroalkyles	sustancias perfluoroalquiladas y polifluoroalquiladas	全氟和多氟烷基物质
persistent organic pollutants	polluants organiques persistants	contaminantes orgánicos persistentes	持久性有机污染物
pesticides	pesticides	pesticidas	农药
substandard and falsified medical products	produits médicaux de qualité inférieure et contrefaits produits médicaux de qualité inférieure et falsifiés	productos médicos de baja calidad y falsificados	伪劣医疗产品
toxic gases	gaz toxiques	gases tóxicos	有毒气体

Biological

African swine fever	peste porcine africaine	peste porcina africana	非洲猪瘟/非洲猪瘟病毒
airborne diseases	maladies à transmission aérienne maladies transmises par l'air	enfermedades de transmisión aérea	空气传播疾病
anthrax	anthrax	ántrax	炭疽
antimicrobial resistance	résistance aux antimicrobiens	resistencia antimicrobiana	抗微生物药物耐药性
avian influenza	grippe aviaire	gripe aviar gripe aviaria	禽流感
bacterial plant disease	maladie bactérienne des plantes	enfermedades bacterianas de las plantas	植物细菌性病害
biological agents	agents biologiques	agentes biológicos	生物战剂
bloodborne viruses	virus à transmission hématogène virus à diffusion hématogène, risque viral hématogène	virus de transmisión sanguínea	血源性病毒
brucellosis	brucellose	brucelosis	布鲁氏菌病
chikungunya	chikungunya	chicunguña chikungunya	基孔肯雅热
cholera	choléra	cólera	霍乱
classical swine fever	peste porcine classique	peste porcina clásica	猪瘟 / 经典猪瘟
contagious bovine pleuropneumonia (CBPP)	pleuropneumonie contagieuse bovine (PPCB)	pleuroneumonía contagiosa bovina (PCB) perineumonía contagiosa bovina (PCB)	牛传染性胸膜肺炎 (CBPP)
contagious caprine pleuropneumonia (CCPP)	pleuropneumonie contagieuse caprine (PPCC)	pleuroneumonía contagiosa caprina (PCC) perineumonía contagiosa caprina (PCC)	山羊传染性胸膜肺炎 (CCPP)
COVID-19 (SARS-CoV-2)	COVID-19 (SARS-CoV-2)	COVID-19 (SARS-CoV-2)	新型冠状病毒感染 (SARS-CoV-2)
Crimean-Congo haemorrhagic fever	fièvre hémorragique de Crimée-Congo	fiebre hemorrágica de Crimea-Congo	克里米亚-刚果出血热

English	French	Spanish	Chinese
cryptosporidium	cryptosporidium cryptosporidies	cryptosporidium	隐孢子虫
cysticercosis	cysticercose	cisticercosis	囊虫病
dengue	dengue	dengue	登革热
diarrhoeal diseases	maladies diarrhéiques diarrhées	enfermedades diarreicas	腹泻病
diphtheria	diphtérie	difteria	白喉
ebola	ebola	ébola	埃博拉
Escherichia coli (STEC)	Escherichia coli (STEC)	Escherichia coli (STEC)	产志贺毒素大肠埃希菌 (STEC)
foodborne diseases	maladies d'origine alimentaire	enfermedades transmitidas por alimentos	食源性疾病
foot-and-mouth disease	fièvre aphteuse	fiebre aftosa	口蹄疫
fungal plant disease	maladie fongique des plantes maladies cryptogamiques	enfermedad fúngica de las plantas	真菌病植物病害
harmful algal blooms	efflorescences algales nuisibles	proliferación de microalgas nocivas marea roja	有害藻华
hepatitis A	hépatite A	hepatitis A	甲型肝炎
hepatitis B	hépatite B	hepatitis B	乙型肝炎
hepatitis C	hépatite C	hepatitis C	丙型肝炎
HIV and AIDS	VIH et SIDA	VIH y SIDA	艾滋病毒与艾滋病
human-wildlife conflict	conflits homme-faune sauvage	conflictos entre humanos y fauna silvestre	人类与野生动物的冲突
infectious animal diseases (not zoonoses)	maladies animales infectieuses (hors zoonoses)	enfermedades infecciosas animales (no zoonosis)	动物传染病（非人畜共患病）
insect pest infestations	infestations d'insectes nuisibles	infestación de plagas de insectos	虫害
invasive species, including weeds	espèces envahissantes, y compris adventices	especies invasoras, incluyendo malas hierbas	入侵物种（含杂草）
Lassa fever	fièvre de Lassa	fiebre de Lassa	拉沙热
leprosy	lèpre	lepra enfermedad de Hansen	麻风病
leptospirosis	leptospirose	leptospirosis enfermedad de Weil	钩端螺旋体病
listeriosis	listériose	listeriosis	李斯特菌病
locust upsurge	criquets ravageurs invasions de locustes, invasions de criquets ravageurs, invasion de criquets pèlerins	infestación de langostas	蝗虫肆虐
lumpy skin disease	dermatose nodulaire contagieuse	dermatosis nodular contagiosa	牛结节性皮肤病

English	French	Spanish	Chinese
malaria	paludisme	malaria paludismo	疟疾
Marburg virus disease	maladie du virus de Marburg	fiebre hemorrágica de Marburg	马尔堡病毒病
measles	rougeole	sarampion	麻疹
meningococcal meningitis	méningite à méningocoques	meningitis meningocócica	脑膜炎球菌性脑膜炎
Middle East respiratory syndrome (MERS)	syndrome respiratoire du Moyen-Orient (MERS)	síndrome respiratorio de Oriente Medio (MERS)	中东呼吸综合征 (MERS)
mpox	mpox	mpox	猴痘
neglected tropical diseases	maladies tropicales négligées	enfermedades tropicales desatendidas	被忽视的热带病
New World screwworm (NWS)	ver du Nouveau Monde (NWS)	gusano barrenador del ganado	新世界螺旋蝇 (NWS)
Newcastle disease	maladie de Newcastle	enfermedad de Newcastle	新城疫
oyster disease aquaculture	maladies des huîtres en aquaculture	patologías en ostras cultivadas	水产养殖牡蛎疾病
pandemic influenza (human)	grippe pandémique (humaine)	gripe pandémica (humana)	大流行性流感 (人类)
paratyphoid fever	fièvre paratyphoïde	fiebre paratifoidea	副伤寒
pertussis (human)	coqueluche (humaine)	tos ferina (humana) tos convulsa (humana), coqueluche	百日咳 (人类)
peste des petits ruminants	peste des petits ruminants	peste de los pequeños rumiantes	小反刍兽疫
plague	peste	peste	鼠疫
polio	poliomyélite	poliomielitis polio	脊髓灰质炎
prion diseases	maladies à prions	enfermedades priónicas	朊病毒病
Q fever	fièvre Q	fiebre Q	Q热
rabies	rage	rabia	狂犬病
Rift Valley fever	fièvre de la vallée du Rift	fiebre del Valle del Rift	裂谷热
rinderpest	peste bovine	peste bovina	牛瘟
rotavirus	rotavirus	rotavirus	小儿肠胃炎病毒
seasonal influenza	grippe saisonnière	gripe estacional	季节性流感
severe acute respiratory syndrome (SARS)	syndrome respiratoire aigu sévère (SRAS)	síndrome respiratorio agudo severo (SRAS) síndrome respiratorio agudo grave (SRAG)	严重急性呼吸综合征 (SARS)
sexually transmitted infections	infections sexuellement transmissibles	infección de transmisión sexual (ITS) enfermedad de transmisión sexual (ETS)	性传播感染
shigellosis	shigellose	shigelosis	志贺氏菌病

English	French	Spanish	Chinese
shrimp disease (bacterial) – acute hepatic pancreatic necrosis	maladie bactérienne de la crevette – nécrose hépatopancréatique aiguë	enfermedad bacteriana del camarón cultivado – necrosis hepatopancreática aguda	虾病（细菌性）– 急性肝胰腺坏死
smallpox	variole	viruela	天花
snakebite envenoming	envenimation par morsure de serpent	envenenamiento por mordedura de serpiente	蛇咬中毒
trypanosomiasis	trypanosomiase	tripanosomiasis	锥虫病
tuberculosis	tuberculose	tuberculosis	结核病
typhoid fever	fièvre typhoïde	fiebre tifoidea fiebre entérica	伤寒
vaccine-preventable diseases	maladies évitables par la vaccination	enfermedades prevenibles mediante vacunación	疫苗可预防疾病
varicella and herpes zoster	varicelle et herpès zoster	varicela y herpes zóster lechina y culebrilla	水痘和带状疱疹
vector-borne diseases (VBD)	maladies à transmission vectorielle (MVC)	enfermedades transmitidas por vectores (ETV)	病媒传染疾病（VBD）
viral haemorrhagic fevers	fièvres hémorragiques virales	fiebres hemorrágicas virales	病毒性出血热
viral, phytoplasma and viroid plant disease outbreaks	foyers de maladies végétales virales, à phytoplasme et à viroïde	brote de enfermedades de las plantas causadas por virus, fitoplasmas y viroides	病毒、植原体和类病毒引起的植物病害暴发
waterborne diseases	maladies transmises par l'eau	enfermedades transmitidas por el agua	水传播疾病 / 水源性疾病
West Nile fever	fièvre du Nil occidental infection due au virus du Nil occidental	fiebre del Nilo Occidental	西尼罗热
yellow fever	fièvre jaune	fiebre amarilla	黄热病
Zika virus	virus Zika	virus del zika	寨卡病毒
zoonotic diseases	zoonoses maladies zoonotiques	enfermedades zoonóticas	人畜共患病

Technological

advanced persistent threat	menace persistante avancée	amenaza persistente avanzada	高级持续性威胁
air transportation accident	accident de transport aérien	accidente de transporte aéreo	航空运输事故 航空事故
bridge failure	effondrement de pont	fallo estructural de un puente	桥梁失效 桥梁损坏
building collapse	effondrement d'un bâtiment	colapso de un edificio	建筑物坍塌
building, highrise, cladding	bâtiment, gratte-ciel, revêtement	incendio del revestimiento de un rascacielos	高层建筑外墙覆面
critical infrastructure failure	défaillance des infrastructures critiques	fallo de una infraestructura crítica	关键基础设施失效 关键基础设施损坏

English	French	Spanish	Chinese
cyberbullying	cyberharcèlement	ciberacoso	网络霸凌 网暴
dam failure	rupture de barrage	fallo estructural de una presa fallo estructural de una represa	溃坝
data breach & PII breach	fuite de données et violation des données personnelles	filtración de datos y de datos personales	数据泄露与个人身份信息泄露
denial of service	déni de service	denegación de servicio	拒绝服务
disaster and conflict waste	déchets de catastrophes et de conflits déchets issus de catastrophes et de conflits, déchets post-catastrophes et post-conflits	residuos resultantes de desastres y conflictos	灾害和冲突造成的废弃物
drain and sewer flooding	inondations des réseaux d'assainissement	inundación por desbordamiento de desagües y alcantarillas	排水管网与下水道内涝
electronic waste (e-waste)	déchets électroniques (e-waste)	residuos electrónicos	电子废物 (E-Waste)
emergency telecommunications failure	défaillance des télécommunications d'urgence	fallo de las telecomunicaciones de emergencia	应急通信中断
explosion	explosion	explosión	爆炸
explosive agents	agents explosifs	agentes explosivos	爆炸物
fire	incendie	incendio	火
hazardous waste	déchets dangereux	residuos peligrosos	危险废物/有害废弃物
health-care waste	déchets d'activité de soins déchets médicaux, déchets de soins de santé	residuos sanitarios	医疗废物/医疗废弃物
inland water ways transportation accident	accident de transport sur voies navigables intérieures	accidente de transporte fluvial	内河航运事故
landfilling	mise en décharge enfouissage	vertedero de residuos tiradero, relleno sanitario, basurero	垃圾填埋
leaks and spills	fuites et déversements	fugas y vertidos fugas y derrames	泄漏与溢流
malware	logiciels malveillants	programa malicioso	恶意软件
marine debris	débris marins	residuos marinos desechos marinos, basura marina	海洋废弃物 海洋垃圾
maritime accident	accident maritime	accidente marítimo	海事事故
mining hazards	risques miniers	peligros mineros amenazas mineras	矿山灾害
natech	natech	natech	自然灾害诱发灾害事件
nuclear agents	agents nucléaires	agentes nucleares	核生化剂
nuclear plant failure	défaillance d'un réacteur nucléaire défaillance d'une centrale nucléaire	fallo de una central nuclear	核电站事故

English	French	Spanish	Chinese
plastic waste	déchets plastiques	residuos plásticos	塑料废弃物
pollution	pollution	contaminación polución	污染
power outage / or blackout	panne d'électricité / ou panne de courant	apagón eléctrico corte de luz	停电 / 大面积停电
radio and other telecommunication failures	pannes de radio et autres télécommunications	fallo de radio y otras telecomunicaciones	无线电及其他通信中断
radioactive agents & material	substances et matières radioactives	agentes y material radiactivos	放射性物质和材料
radioactive waste	déchets radioactifs	residuos radiactivos	放射性废物
rail accident	accident ferroviaire	accidente ferroviario	铁路事故
reservoir flooding	inondations de réservoirs	inundación por desembalse	水库泄洪
road traffic accident	accident de la route	accidente de tráfico	道路交通事故
safety hazards associated with oil and gas	risques liés à l'extraction de pétrole et de gaz	peligros para la seguridad asociados al petróleo y el gas amenazas para la seguridad asociados al petróleo y el gas	油气安全隐患
social engineering – phishing	ingénierie sociale – hameçonnage	ingeniería social – phishing	社会工程 – 网络钓鱼
soil pollution	pollution des sols	contaminación del suelo polución del suelo	土壤污染
solid waste	déchets solides	residuos sólidos	固体废弃物
construction/structural failure	défaillance structurelle ou de construction	fallo de construcción o estructural	施工/结构失效
supply chain attack	attaque de la chaîne d'approvisionnement	ataque a la cadena de suministro	供应链攻击
supply chain failure	défaillance de la chaîne d'approvisionnement	fallo en la cadena de suministro	供应链中断 / 供应链断裂
tailings	résidus miniers déchets miniers, stériles miniers, résidus de traitement minier	relave minero cola minera	尾矿
tunnel failure	effondrement de tunnel	fallo estructural de túneles	隧道失效 隧道坍塌
waste treatment lagoons	lagunes de traitement des déchets lagunage	balsa de tratamiento de residuos	废水处理塘
wastewater	eaux usées	aguas residuales	废水
water supply failure	rupture de l'approvisionnement en eau	fallo en el suministro de agua	供水中断

Societal

civil unrest	troubles à l'ordre public atteintes à la paix publique	disturbio social	社会动荡
--------------	---	------------------	------

English	French	Spanish	Chinese
environmental degradation from conflict	dégradation de l'environnement due aux conflits	degradación ambiental resultante de conflictos	冲突造成的环境退化
explosive ordnance	munitions explosives non explosées restes explosifs de guerre	artefacto explosivo	爆炸性弹药
financial shock	choc financier crise financière	shock financiero choque financiero	金融冲击
international armed conflict (IAC)	conflit armé international (CAI)	conflicto armado internacional (CAI)	国际武装冲突 (IAC)
non-international armed conflict (NIAC)	conflit armé non international (CANI)	conflicto armado no internacional (CANI)	非国际性武装冲突 (NIAC)
stampede or crushing (human)	bousculade ou écrasement (humain) mouvements de foule	estampida o aplastamiento humano	踩踏或挤压 (人群)
suicide cluster	grappe de suicides suicides en grappes, suicides collectifs	cúmulo de suicidios	丛集性自杀 / 自杀群发
violence	violence	violencia	暴力

Translation Quality Assessment of legal translations by MateCat and its use in future translation didactics

Edyta Żralka

The University of Silesia, Katowice, Poland

edyta.zralka@us.edu.pl

Abstract

The paper introduces the principles of research in progress undertaken to enrich MateCat's database with the legal terminology of English, Polish and Romance languages (French, Italian, Spanish and Portuguese) and, based on it, to perform specialised legal translations with post-editing (PE). The introductory stage of the research (a pilot study of ENG-PL translations described here) led to the development of criteria for a manual metric necessary to carry out the Translation Quality Assessment (TQA) and PE processes rather than performing traditional translation without the involvement of electronic tools. The research data will be used to develop a teaching programme for educating students in TQA and PE.

1 Introduction

Translation has undergone transformations, fulfilling the needs of the times and target audiences (TAs). The extended demand for translations in all possible languages in the progressive globalisation and fast pace of life enforcing instant responses to translation commissions caused specialists in linguistics, translation, IT and web developers to create modern translation tools. Nowadays, they are even available online for free, like Google Translate (GT) or DeepL. When machine translation (MT) tools are considered, they are based on automatic renderings of source texts (STs) into target languages (TLs) using data stored in databases selected by special algorithms. The outcomes, especially in the case of the latest neural machine translations, are increasingly acceptable and gain popularity among different kinds of users. However, the combination of this method with computer-assisted translation (CAT), using translation memories (earlier human translations stored in the system), is even more promising, as

observed in MateCat's offer. The tool, as defined by Federico et al. (2014: 129), is:

[...] a new web-based CAT tool providing translators with a professional work environment, integrating translation memories, terminology bases, concordancers, and machine translation. The tool is completely developed as open source software and has been already successfully deployed for business, research and education.

To be precise, MateCat offers its users a built-in MT system called ModernMT (ModernMT Lite version), providing contextually based suggestions (the engine's Context Awareness). It adapts to users' corrections in real-time, systematically improving its output. It is also integrated with MyMemory, the largest existing public translation memory (TM). For any language pairs not supported by ModernMT, the system automatically switches to GT. .

Given MateCat's state-of-the-art, the fundamental need is to develop specialised databases, e.g., for legal texts, based on automatic and manual metrics of translation quality assessment (TQA) and post-editing (PE) the errors discovered. Even if automatic metrics are used on a global scale by companies especially nowadays (e.g. BLEU), when PE is planned translators use manual metrics (Multidimensional Quality Metrics – MQM, and TAUS Dynamic Quality Framework – DQF, a subset of MQM analysing error typology), which will also be the case here. Database development will require applying manual metrics worked out for the research based on the analysed texts' real needs in to control lexis, grammar, orthography and syntax when introducing data. The data gathered and recorded can be prospectively used to introduce modern programmes of students' training emphasising post-editing abilities within legal texts rather than traditional translation. The research involves human participants (focus groups, student

observations), that is why the principles of Ethics Approval will be used,

The shift in translation process is noted by Pym (2021: 39), discussing a gradual transformation of translators' duties from translation into PE and even giving prospects of translators' remuneration dispersion, mainly caused by a wider use of MT tools in the future.

1.1 General research goals

This research aims to gather possible data within a legal language, sources of errors in MateCat translations, and post-edit them while creating an enlarged database. The aim of the research team, consisting of experts in Polish, English, French, Italian, Spanish and Portuguese (native speaker linguists dealing with translation), would be to raise the quality of MateCat legal translations in a selected language group to the highest standards possible through post-editing MateCat renderings of a text collection proposed and creating a richer professional database referring to the errors encountered. And finally, the research aims to propose a possible modification in the system of translators' training that gradually evolves into the abilities of PE rather than mere translation.

The effect reached by the research will be a well-designed database of legal terminology, phraseology, syntactic issues and orthography in the selected languages, with the perspective of raised quality of translations performed automatically, but controlled by well-trained translators.

1.2 Research problems

There are still more research problems connected with the goals stated. First, the idea of TQA is one of the most prominent issues in Translation Studies now. Since Julian House introduced the background of her theory in the 1990s (1997) promoting functional equivalence and then developed it in her book *Translation Quality Assessment: Past and Present* published in 2015, her views were analysed, discussed, adopted and developed in reference to translation in general (e.g. Basil Hatim), or in the legal language (e.g. Prieto Ramos).

Second, the problem of quality control gained particular importance when MT engines occurred, and institutions dealing with the translation business started to transform the translation market into MT + TQA + PE mode (incorporating MT tools, then automatic or manual TQA tools, and optionally, PE) instead of traditional translations by human translators. It was due to the swelling demand for both translations and the rapidity to perform them, unfortunately accompanied by decreasing requirements for correctness. Even if standards

are specified for translations (e.g. DIN 2345 and ISO 9000, Brunette 2000: 169, or ISO 17100 and ISO 18587, Biel 2021:13), first, they concern rather a process than a product, and second, some companies believe that what they need is just to feel safe with the satisfying percentage of translation correctness (measured by e.g. BLEU, or METEOR), without wasting money on PE. Such a model cannot be accepted for legal texts. It is due to the burden of responsibility for wrongly performed translations and the subsequent consequences to companies and individuals. What scholars need to add is their research in the ways to raise the correctness of automatic translations in a systematic way, proposing the contents of metrics of quality control and criteria for PE.

1.3 Research questions and hypotheses

Research questions necessary to be posed are the following:

- Is MateCat quality performance a trustworthy basis for carrying out academic research?
- Will PE lead to the systematic enhancement of the MateCat performance?
- How much can the quality of MateCat translations be improved by the enrichment of the tool's database?
- Will preparing parallel entries for the MateCat database in all the languages proposed result in better outcomes when Polish is juxtaposed with Romance languages through direct links, not the agency of English?
- Will it be possible to use the data gathered and experience gained to start a new concept of students' education, namely, teaching PE instead of the traditional translation?

The hypothesis to be tested is that creating the enriched database should result in raised correctness along with higher efficiency. In the era of intensified globalisation, when communication on both professional and amateur levels is sometimes needed in an instant, the existence of such tools serves as a time- and effort-saver and solves many time-critical problems. The target, when reached, would be a milestone in public access translation tools' reliability and a record of necessary data to be used for future software systems' development (e.g. training programmes for university students). It can also serve as a professional reference for linguists for further research in the field.

2 Categories and criteria in TQA

As for the linguistic theories of TQA, scholars have referred to both categories taken into consideration in evaluation (House, Hatim) and criteria based on which evaluation can be performed (Ramos, Brunette, Mossop, Colina, and Angelelli), the second of which being of more practical use for the research. These are two different perspectives, the understanding of which needs elaboration. Definitions of the two terms are taken from the most reliable English dictionaries: *Oxford English Dictionary* (OED), *Merriam-Webster Dictionary* (MWD), and *The Britannica Dictionary* (BD). A noun *category* receives the following definitions:

- OED: “A class, or division, in any general scheme of classification. spec. in Linguistics.”
 - MWD: “1: any of several fundamental and distinct classes to which entities or concepts belong; 2: a division within a system of classification.”
 - BD: “A group of people or things that are similar in some way.”
- On the other hand, the word *criterion* is defined as:
- OED: “A test, principle, rule, canon, or standard, by which anything is judged or estimated.”
 - MWD: “1: a standard on which a judgment or decision may be based; 2: a characterizing mark or trait.”
 - BD: “Something that is used as a reason for making a judgment or decision.”

“Class”, “division” and “classification” are the most commonly used references to the word *category* in the definitions quoted. It is then some set of elements that represent “similarities” in any aspect and allow for differentiating them from other collections of similar components.

Categories taken into account when searching for evaluation criteria of any linguistic content, including translation, could reasonably be the following:

- terminology,
- grammar features,
- orthography
- consistency.

This is because translations normally comprise such or similar classes of linguistic components or features according to which they are performed.

Based on the definitions quoted, we can see that typical attributes of a noun *criterion* are adjusting to some standards and being subject to judgements. It is then all selected elements that undergo evaluation.

According to the definitions, categories seem to represent a broader construct from which one can choose a criterion. When manual metrics for MT evaluation are considered, “Such metrics could be based on the criteria including:

- term selection and consistency,
- grammar subtleties: declensions, concord of pronouns, structure of phrases, tenses concerning different meanings (prescriptive and descriptive) of a modal *shall*,
- redundancies,
- spelling in different types of names” (Żrałka 2023).

The fundamental discoveries in the previous research reveal that the most typical for MT (GT and DeepL) are:

1. Errors within lexis:

- incorrect or partly correct terms,
- retaining ST terms without any translation,
- homonyms given as equivalents,
- redundancies,
- word-for-word renderings,
- terminological inconsistencies;

2. Errors within grammar:

- using a wrong case (applied to Polish as an inflected language),
 - using the wrong gender,
 - using wrong pronouns,
 - using a wrong part of speech,
 - using a wrong tense in the interpretation of the meanings of modal *shall* (descriptive or prescriptive),
 - lack of prepositions,
 - parting phrases,
 - omissions of some syntactic contents, especially in syntactically complicated complex sentences;
- ### 3. Errors within orthography:
- spelling inconsistencies,
 - wrongly used capital or small letters for names referring to entities or legal instruments (Żrałka 2023).

The percentage of particular error types in a sample of legal text translations from English into Polish compared with the sample wordcount and the ratio of errors percentage-wise presented in the charts below shows how important the particular categories are and how much attention the PE of them will require from post-editors (Żrałka 2023).

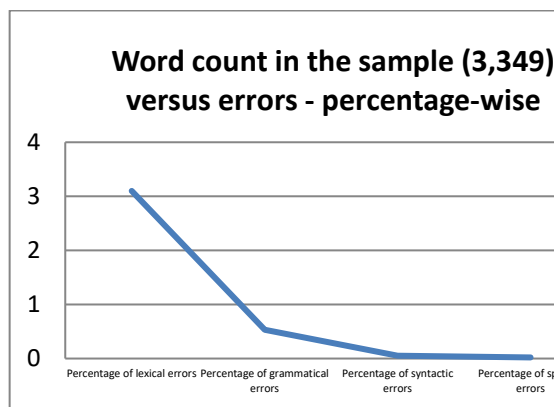


Figure 1. Percentage of different error categories in the sample

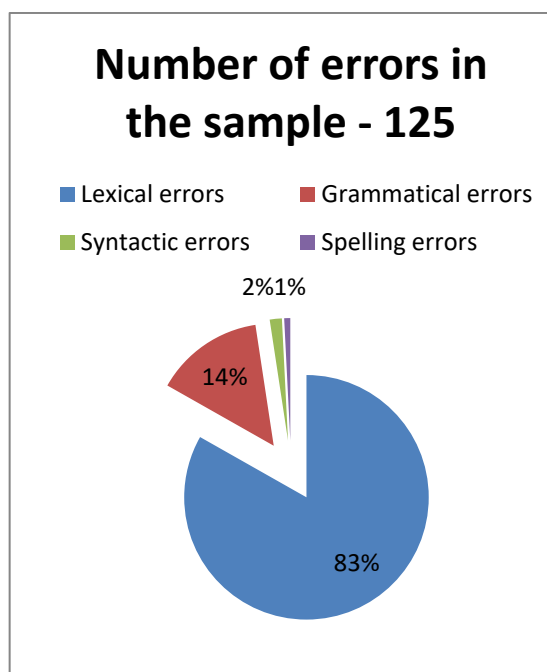


Figure 2. Ratio of the error categories in the sample

Apart from just analysing the errors, a post-editor of automatic translation renderings has to be aware of more than just linguistic content. To set the stage for PE, he or she has to consider what House (2015) refers to in her theory of TQA based on Halliday's Systemic-Functional Linguistics (1978). According to it, texts are always set in some situation, and this situation conditions the texts' functions. The notion of situation is characterised by specific situational dimensions:

- Field: refers to the subject matter of the text (here, it is law);
- Tenor: incorporates the people involved in the communication and the relationships between them (level of formality – here, it is formal);
- Mode: refers to the form of language, spoken or written, used in the interaction (here, it is a written language).

An important element of the Hallidayan theory is that it goes beyond the mere level of language and incorporates the idea of functional attitude to the translation phenomenon. It is necessary to start translation evaluation by realising such basic features as register, style, and the general function of a translated text according to the defined needs.

Taking that idea into account, House gathers and groups views on the quality of translation into several categories based on equivalence and the functional treatment of STs and TTs in translation. The author, instead of traditionally understood equivalence, allows the functionalistic approach to the TQA derived not only from Halliday but also from Skopos theory of Katherine Reiss and Hans Vermeer of the 1970s, later developed by Christiane Nord. She even claims that "The skopos or purpose is the most important factor in translation, the original text being downgraded to a mere 'offer of information' and the translator often seen as a type of 'co-author'." (House 2015: 11)).

Hatim (1998) observes that, in House's theory, a function of translation is considered a mixture of language functions and text functions (based on the views of Reiss 1971, Bühler 1990 and Halliday 1978). Hatim (1998: 95) states that a full reproduction of specific ST functions is not possible in translation because, according to House (1997: 67), "the ST is tied to a specific non-repeatable historic event in the source culture [...] or because of the unique status that the source text has in the source culture". Hatim (1998: 95) concludes that if "situationality" existing between the two texts cannot be fully reproduced, "a second-level function" combining ways of perception by ST and TT readers is required to be reached by a translation. Such a function "must hold not only for the contemporary target language readers but also for their counterparts in the source culture" (House 1997: 191).

Hatim's and House's observations are crucial to the attitude towards legal texts' TQA. A translator and post-editor have to be aware of the functions and the specificity of legal texts in both languages comprising their linguistic features (terminology, phraseological structures, grammar, syntax, punctuation, etc.), and cultural dependence. These are all elements that imply a source-text-oriented translation, with the TT receiver kept on mind to incorporate the views of House and Hatim.

Ramos (2015) and the authors mentioned by him refer to the criteria of translation evaluation incorporated into legal translations. The most systematic of them, at the same time sharing the features of other classifications, seems to be the one by Brunette. The author

differentiates the following criteria of TQA (2000: 175–177):

- Logic (depending on coherence and cohesion), directed to the target audience, not the ST,
- Purpose (effect and intention),
- Context, defined as “Non-linguistic circumstances surrounding the production of the discourse to be assessed.” (2000: 178),
- Language norm (rules and conventions of the language).

Through addressing the purpose and context, she refers to the criteria introduced by House. Language rules and logic remain standard linguistic components of assessing any text.

The classifications by Mossop (2007: 125–139), Colina (2008: 103–106), and Angelelli (2009: 40–41) mostly refer to typical comparative elements of source and target texts. Mossop additionally includes layout and organisation among the evaluative criteria. Angelelli sees “translation skill” as a criterion of evaluation in translation, which seems an element encapsulating all other criteria.

For this research, the criteria introduced by Żrałka (2023), mentioned earlier, will be incorporated into the analysis on the condition that additional needs will not be revealed. The principle of assessing the quality of translations and PE will be orientation to the ST, taking into account the needs of the target recipient.

3 Research concept and plan

The research aims to perform translations of a selected corpus of legal texts (those which are of value for natural persons) their evaluation based on two kinds of metrics – automatic and manual, and post-editing the translations with classifying errors to prepare entries to the tool’s database. The database entries will be a source to use in developing possible software dedicated to the automatic translation of legal texts with a possibility to improve subsequent outcomes by CAT tools options. It will give bases for a course in post-editing automatic translations of legal texts, proposed to modify translation teaching at the university level.

The translations into Polish will be consulted with the professionals employed as academic teachers at the Silesian University – Romance Languages Philologies, English Philology (also native speakers). All observations, suggestions, solutions for the PE effectiveness will be included in planning a university course in PE dedicated to the field of law. Even if the courses of PE have gained an interest among Polish universities recently (e.g. Jagiellonian

University in Kraków, Nicolaus Copernicus University in Toruń), the PE should be centered around particular technolects’s needs to be effective. The legal technolect and its translation requires particular knowledge, so post-editors should acquire also skills in its specialised content, not just general abilities of PE.

Finally, the data gathered can be used to develop legal translation software (MT, CAT, or PE educational programmes).

4 Methodology of research

Our main concern in the research will be the analysis of a translation product - corpus texts’ first renderings performed by MateCat from English in all languages selected, starting with the ENG-PL pair analysed here. Contracts of four different types have been selected (Open-ended Employment Agreement, Mandate Agreement, Real Estate Sales Contract, and Donation Agreement), each comprising between 1,000 and 2,500 words (1,452, 1,123, 2,347, 1,365, respectively). The choice was motivated by their functional importance in legal translation for a typical client (the agreements are documents that freelance translators often deal with). The main method will be then a corpus research and comparative study, but also discourse analysis, critical discourse analysis will be incorporated within the assessment of Fluency (as opposed to Accuracy in a metric proposed for TQA).

After the translation, the BLEU metric test evaluating quality will be performed before the PE. The method used will be a comparative study of STs and TTs. In the analysis of translation quality performed according to the criteria established both quantitative and qualitative methods will be included. First, the errors will be evaluated and classified under the criteria enumerated, then categories will be ascribed to them, finally, errors counted by the BLEU evaluation test will be grouped according to their types, and quantitative measures will be introduced, and finally conclusions concerning general translation quality and suggestions for the PE process will be given. The BLEU metric will be used after the PE to control the quality enhancement.

When a satisfying level of correctness is reached and shown by BLEU (if not, PE will be repeated), all errors corrected will be referred to in the process of database creation. The entries will include the lexical and grammatical information derived from the three dictionaries referred to before (OED, MWD, BD), the IATE platform (Inter-Active Terminology for Europe), comprising the EU legal terminology, and the WordNet platform, which is a large lexical database of English. Its particular advantage is the

fact that parts of speech are grouped into the so-called synsets - sets of cognitive synonyms. Discovering commonly functioning synonyms will help to give variants of terms in case the particular legal system prefers using a different equivalent than some other system. A typical example could be the use of words “petitioner”, “plaintiff” “claimant”, “pursuer” for *powód/powódka* vs. “respondent”, “defendant” for *pozwany/pozwana*. New equivalents will be introduced into the MateCat TM, collocations will be given proper equivalents within their lexical and grammatical content, typical grammatical phrases will be introduced into the database, and orthography will be corrected in equivalents. The database will be developed in all languages included in the research in their translation from English, but also directly between the languages in pairs Polish-English, Polish-French, Polish-Spanish, Polish-Italian, Polish-Portuguese..

The database will be published and the stages of error analysis and corrections in PE will be the basis for creating future syllabuses in PE courses. Recipient-oriented research will be performed by means of questionnaires, interviews, focus group research, observations in a case study including qualitative and quantitative research encompassing the participants acting as post-editors. Participants will be questioned according to the needs and difficulties in PE in particular languages to refer to the course needs based on a questionnaire distributed via the mediation of translation agencies. Here, also qualitative and quantitative research will be performed.

At last, the database will be published as the final record and a basis for any further academic or technical activity, e.g., creating a possible software dedicated to legal translations, including the languages analysed in the research.

References

- Claudia V Angelelli. 2009. “Using a rubric to assess translation ability.” *Testing and assessment in translation and interpreting studies*, 13-49.
- Łucja Biel. 2021. “Postedycja tłumaczeń maszynowych.” *Lingua Legis*, 1(29).
- Louise Brunette. 2000. “Towards a terminology for translation quality assessment: A comparison of TQA practices.” *The Translator*, 6(2): 169–182.
- Karl Bühler. 1990. *Theory of language. The representational function of language*. John Benjamins Publishing Company, Amsterdam.
- Sonia Colina. 2008. “Translation quality evaluation: Empirical evidence for a functionalist approach.” *The translator*, 14(1), 97-134.
- Marcello Federico, Nicola Bertoldi, Marco Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, and Ulrich Germann. 2014, August. The MateCat tool. In *COLING (Demos)* (pp. 129-132).
- Mike L. Forcada. 2010. “Machine translation today.” In *Handbook of translation studies*, 1, 215-223. John Benjamins Publishing Company.
- Han, L., G., JF Jones and Alan F. Smeaton. 2021. “Translation quality assessment: A brief survey on manual and automatic methods.” *arXiv preprint arXiv:2105.03311*. https://www.researchgate.net/publication/351448577_Translation_Quality_Assessment_A_Brief_Survey_on_Manual_and_Automatic_Methods
- M. A. K Halliday. 1978. *Language as Social Semiotic*. Edward Arnold, London, UK.
- M.A.K. Halliday. 1985. *An Introduction to Functional Grammar*. Edward Arnold, London, UK.
- Basil Hatim. 1998. “Translation quality assessment: Setting and maintaining a trend.” *The Translator* 4 (1): 91-100.
- Juliane House. 1977. *A Model for Translation Quality Assessment*. Narr, Tübingen.
- Juliane House. 1997. *Translation Quality Assessment: A Model Re-visited*. Narr, Tübingen.
- Juliane House. 2015. *Translation Quality Assessment: Past and Present*. Routledge, New York, NY.
- Christiane Nord. 1997. *Translating as a Purposeful Activity. Functionalist Approaches Explained*. St. Jerome, Manchester, UK.
- Brian Mossop. 2007. *Revising and editing for translators*, 2nd ed. St. Jerome Publishing. Manchester, UK.
- Mirjam Sapesy Maučec. and Gregor Donaj. 2020. “Machine translation and the evaluation of its quality.” In *Recent Trends in Computational Intelligence*, ed. by Ali Sadollah, and Tilendra Shishir Sinha, .143-162. IntechOpen.
- Anthony Pym and Ester Torres-Simón. 2021. “Is automation changing the translation profession?” *International Journal of the Sociology of Language* 2021(270): 39-57.
- Prieto Fernando Ramos. 2015. “Quality assurance in legal translation: Evaluating process, competence and product in the pursuit of adequacy.” *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, 28, 11-30.
- Katherine Reiss. 1971. *Möglichkeiten und Grenzen der Übersetzungskritik*. Hueber, Munchen.

- Katherine Reiss, Hans J. Vermeer, Christiane Nord, and Matthew Dudenhöfer, 2015. *Towards a General Theory of Translational Action: Skopos Theory Explained* (1st ed.). Routledge Taylor & Francis Group, London, UK.
- Lawrence Venuti. 1995. *The translator's invisibility: A history of translation*. Routledge, London, UK.
- Hans. J. Vermeer. 1996. *A skopos theory of translation: (some arguments for and against)*. Wissenschaft, Heidelberg.
- Hans J. Vermeer. 1978. "Ein Rahmen für eine allgemeine Translationstheorie". *Lebende Sprachen*, 23(3), 99-102.
- Edyta Żrałka. 2023. "Quality assessment and post-editing of Google Translate output in specialised legal Translation." In *Approaches on Machine Translation: Quality Assessment, Acceptance and Language Trends, Translating and Translanguaging in*

Multilingual Contexts. John Benjamins. (unpublished – publication planned for December 2023)

Dictionaries:

Merriam-Webster Dictionary
Oxford English Dictionary
The Britannica Dictionary

Websites:

<https://iate.europa.eu>
<https://wordnet.princeton.edu/>
<https://www.britannica.com/dictionary>
<https://www.matecat.com/>
<https://www.merriam-webster.com/>
<https://www.oed.com/>

COPECO-Speech: Integrating Speech and LLMs for Post-Editing in Translation Workflows

Jeevanthi Liyana Pathirana¹, Pierrette Bouillon², Jonathan Mutal², Sabrina Girletti², Lise Volkart²

University of Geneva

Jeevanthi.Liyana@etu.unige.ch

{Pierrette.Bouillon, Jonathan.Mutal, Sabrina.Girletti, Lise.Volkart}@unige.ch

Abstract

In this paper, we explore how large language models (LLMs) can be integrated into translation workflows through multimodal interaction. Recent developments in generative Artificial Intelligence (AI) are influencing translation workflows, yet current uses of LLMs in translation and post-editing (PE) remain mostly output-oriented, providing limited opportunities for users to guide the editing process.

We present COPECO-Speech¹, a multimodal teaching platform for translation and PE. It allows users to interact with the translation process through typing, speech-based navigation commands and spoken natural-language instructions interpreted by an LLM. The system is web-based and combines automatic speech recognition (ASR), machine translation (MT), and LLM-based editing within a single workflow, allowing more interactive translation and PE.

COPECO-Speech was designed for both research and teaching purposes, and supports the collection of interaction data such as speech logs, keystrokes, timing information, and PE annotations for subsequent analysis.

1 Introduction

Recent advances in LLMs are influencing translation technologies, shifting workflows from traditional MT towards more interactive forms of human-AI collaboration (Sánchez-Gijón and Palenzuela-Badiola, 2023; Qian and Kong, 2024). Studies have explored the use of LLMs as post-editors of machine-translated text, demonstrating improvements in translation quality and stylistic refinement through prompt-based editing (Raunak et al., 2023; Chen et al., 2024; Farrell, 2023; Ki and Carpuat, 2024; Kunilovskaya et al., 2024; Macken, 2024; Li et al., 2025).

¹copeco.unige.ch

At the same time, speech technologies have been introduced in Computer-Assisted Translation (CAT), primarily for dictation and simple commands (Ciobanu, 2016). However, these approaches do not fully support natural-language interaction or more complex editing intentions.

This paper introduces COPECO-Speech, a multimodal approach to translation and PE that combines speech recognition, machine translation, and LLM-based editing within a single environment. Users can interact through typing, speech-based navigation commands, and spoken natural-language instructions.

The main contribution of the present work is the integration of LLM-based editing through spoken editing instructions. Unlike previous COPECO versions which focused on speech-based text entry along with navigation, COPECO-Speech enables users to express revision intentions interpreted by an LLM and applies them directly to the translation output. Unlike earlier multimodal PE systems such as MMPE (Herbig et al., 2020), COPECO-Speech combines multimodal interaction with LLM-mediated PE within a single workflow.

The remainder of this paper is organised as follows. Section 2 reviews related work, Section 3 introduces multimodal PE with LLM-based interaction, Sections 4 and 5 present the COPECO-Speech platform and its use in practice, and Section 6 discusses limitations and future work.

2 Related Work

Speech has been explored in a range of translation workflows, including respeaking and other approaches (e.g. interlingual subtitle voicing, where spoken reformulations are transcribed by ASR systems) (Radić, 2024). Speech has also been used in translation and PE processes, both as an input

modality (Ciobanu, 2016) and as an output modality through text-to-speech systems that influence revision behaviour (Ciobanu et al., 2024). More generally, PE has been studied as a form of human-machine interaction in which users refine machine-generated output rather than producing translations from scratch (O’Brien, 2010; Koponen, 2016).

Several existing CAT tools (MateCat, memoQ, and Trados Studio ²) use speech, but are mainly used for dictation or simple predefined commands. While this improves speed, it does not allow users to express complex editing intentions. More recent multimodal approaches combine speech, touch, pen and typing to support PE tasks (Herbig et al., 2020). There, MMPE enables translators to interact with machine-translated text using multiple input modalities and was evaluated with professional translators. However, these approaches primarily use speech as an alternative input modality for editing actions and commands rather than as a means of expressing natural-language editing intentions. Furthermore, they do not integrate recent developments in generative AI.

Another area of research has explored using LLMs as post-editors of machine-translated text, often framed as automatic PE (APE). Studies have shown that LLMs can improve machine-translated output through prompt-based revision and iterative refinement strategies (Raunak et al., 2023; Chen et al., 2024; Farrell, 2023; Ki and Carpuat, 2024). Beyond translation quality, researchers have examined the linguistic and stylistic effects of LLM-based post-editing, including increased lexical richness (Macken, 2024), mitigation of translationese through targeted prompting (Kunilovskaya et al., 2024), and document-level PE enhancements using monolingual data (Li et al., 2025).

LLMs have also enabled more interactive and human-centred translation workflows. Prompt-based approaches demonstrate how translation quality can be influenced through natural-language instructions (Zhang et al., 2023), while recent work has combined LLM prompting and translation memory to support greater user control over MT systems (Qian and Kong, 2024). Studies have further explored how systems such as ChatGPT may reshape translator-computer interaction (Sánchez-Gijón and Palenzuela-Badiola, 2023). At the same time, multimodal PE research has investi-

gated speech, touch, pen and typing as alternative input modalities (Herbig et al., 2020).

However, these two strands of research have largely developed separately, with limited work combining multimodal interaction and LLM-mediated PE within a single workflow.

3 Multimodal PE with LLM-Based Interaction

COPECO-Speech builds on the idea of allowing users to provide spoken editing instructions that guide how translations are modified. Unlike traditional dictation, where speech is used primarily to provide text input, speech in COPECO-Speech is mostly used to express PE intentions. Spoken instructions are first transcribed by an ASR system and then passed to an LLM, which performs the requested modification. This enables users to request actions such as rephrasing, restructuring, simplification, or regeneration through natural-language instructions.

In contrast to multimodal PE systems such as MMPE (Herbig et al., 2020), where speech is primarily used to execute commands, COPECO-Speech uses speech to communicate editing goals to an LLM. Similarly, while recent LLM-based PE approaches focus on automatically improving machine-translated output (Raunak et al., 2023; Ki and Carpuat, 2024; Macken, 2024), COPECO-Speech enables users to guide and control the editing process through multimodal interaction. This design reflects a shift from direct text manipulation towards intent-based interaction.

4 System Overview

COPECO-Speech is a web-based platform built on the COPECO framework (Mutal et al., 2020). Building on previous COPECO research on speech-based PE and speech-command navigation (Liyana-pathirana et al., 2024), the present version introduces LLM-based editing through spoken natural-language instructions.

The platform supports translation and post-editing through typing, speech, or a combination of both, while recording interaction data such as keystrokes, speech logs, task duration, and translation or PE mode. Figure 1 provides an overview of the system workflow. Further implementation details, including the backend prompt and its capabilities, are presented in the Appendix.

²<https://site.matecat.com>, <https://www.memoq.com>,
<https://www.trados.com/>

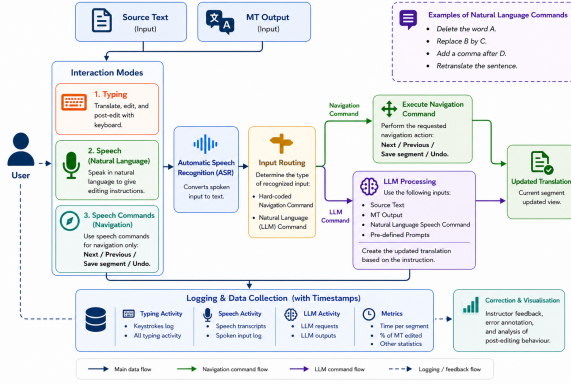


Figure 1: Multimodal PE workflow in COPECO-Speech. The system supports typing, (speech based) navigation commands, and natural-language speech input. LLM-based processing is applied to natural-language instructions, enabling flexible PE.

Currently COPECO-Speech supports two types of speech-based interaction: predefined navigation commands and natural-language editing instructions. Navigation commands are filtered through a predefined rule-based command set with fuzzy matching to accommodate ASR errors and variations in phrasing. These commands are executed directly within the interface, whereas unmatched inputs are treated as editing instructions and processed by GPT-4.1 mini.³ GPT-4.1 mini and Azure Speech Services were selected as practical choices for the current prototype, balancing language-processing capabilities with cost and deployment considerations.

For LLM-based processing, the system combines the source text, MT output, and spoken instructions within a structured prompt, which is then processed by the LLM to generate an updated translation. Depending on the instruction, the output may involve local edits or full segment retranslation.

The current version supports English and French and can be extended to additional languages.

5 COPECO-Speech: Interaction in practice

Figure 2 illustrates speech-driven post-editing in COPECO-Speech.

Example

Source (EN): We often talk about the future.

Incorrect MT (FR): Nous souvent parlons de l’avenir.

³<https://openai.com/>

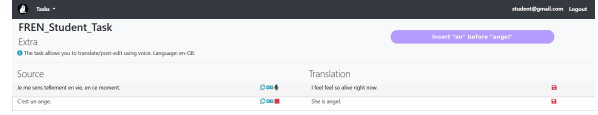


Figure 2: Example of speech-driven PE in COPECO-Speech. A (natural-language) French spoken instruction (“Inverser souvent et parlons”) is being applied through the LLM to correct the target French translation.

Spoken Instruction (FR): Inverser souvent et parlons

Output (FR): Nous parlons souvent de l’avenir.

5.1 Backend Prompt Design for LLM-Based Editing

The backend prompt acts as a controlled transformation mechanism rather than a free-form generation prompt. It combines the source text, MT output, and user instruction while enforcing language, output, and fallback constraints. The complete prompt and an overview of its key features are provided in the Appendix (Sections B, C).

5.2 Interaction Logging and Analysis

All interactions are recorded, including keystrokes, speech logs, LLM logs, and processing time. The system also allows tasks to be sent to instructors for review, enabling correction and error annotation for further analysis.

Figure 3 illustrates examples of speech logs, task statistics, and annotation data collected by COPECO-Speech. These data support process-oriented analyses of user behaviour and multimodal post-editing (Carl et al., 2016).

6 Discussion and Future Work

The proposed approach shows how LLMs can support flexible and interactive PE workflows. Combining speech and natural-language instructions, COPECO-Speech allows users to express editing intentions more directly compared to traditional interaction methods. Being a work in progress, future work will involve analysing how different interaction modalities influence post-editing behaviour, as well as the impact of prompt design on translation quality and user interaction.

As a preliminary test, the current setup was tested with four students performing English-to-French PE tasks (consisting of ten sentences) using speech-based interaction. The study confirmed the technical feasibility of the approach and provided

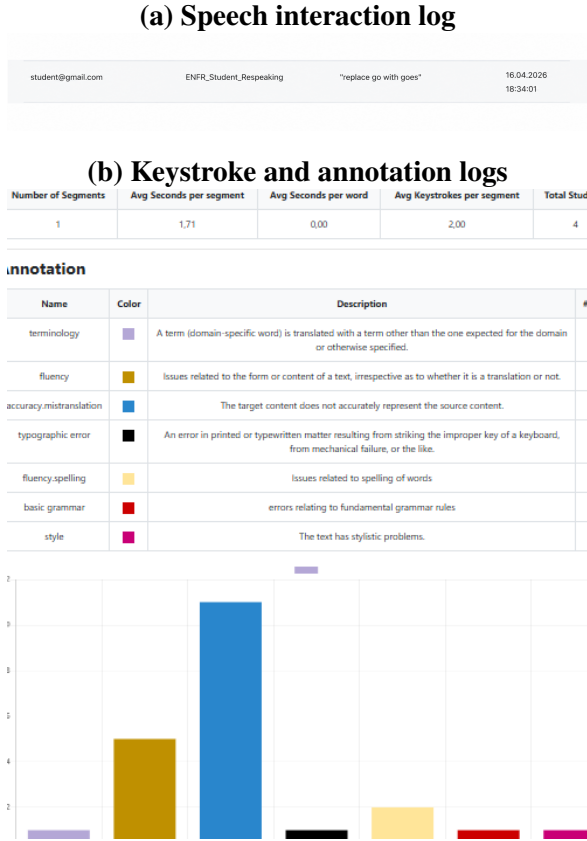


Figure 3: Examples of interaction logs collected by COPECO-Speech. (a) Speech interaction logs showing spoken instructions and timestamps. (b) Keystroke, task-level, and annotation statistics recorded after task completion. Together, these logs support process-oriented analyses of multimodal post-editing behaviour.

examples of spoken editing instructions, some of which are included in the Appendix (section D). Observations also suggested that the LLM was sometimes able to correct ambiguous ASR transcriptions by making use of the translation context and the user’s editing intent. While these observations are preliminary, they highlight an interesting area for future work.

Future work will involve larger-scale user studies and process-oriented analyses based on the interaction data collected by COPECO-Speech, including speech logs, timestamps, keystroke activity, LLM interactions, and PE correction annotations. Such analyses will help answer questions concerning the temporal effort, interaction patterns, and user experience associated with speech-based and typing-based PE.

Future research will also examine the extent to which LLMs can recover from ASR errors, the influence of prompt design on translation quality and user behaviour, and using the setup for addi-

tional language pairs. Finally, we plan to evaluate the feasibility of using alternative ASR and LLM solutions, including open-source and privacy-preserving deployments.

Currently COPECO-Speech relies on external cloud-based services for speech recognition and LLM processing. Due to this, privacy and data governance are important considerations. In the current deployment, users are informed about the use of external services and participation is voluntary. Interaction data used for research purposes are anonymised prior to analysis and handled according to institutional research ethics and data-protection requirements.

7 Conclusion

This paper presents COPECO-Speech, a multimodal PE platform that combines speech interaction and LLM-based editing. The system allows users to modify translation output using natural-language instructions, going beyond traditional typing and command-based (spoken) interaction. The system is currently used in teaching settings, where students can perform PE using traditional typing and speech with LLM support, enabling instructors to analyse interaction patterns and support reflection on translation and PE strategies and the role of AI in translation.

References

- Michael Carl, Moritz Schaeffer, and Srinivas Bangalore. 2016. The critt translation process research database. In *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*, pages 13–54. Springer.
- Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 181–190.
- Dragoş Ciobanu. 2016. Automatic speech recognition in the professional translation process. *Translation Spaces*, 5(1):124–144.
- Dragoş Ciobanu, Miguel Rios, Alina Secară, Julia Brockmann, R. M. Chereji, and Christian Wiesinger. 2024. The impact of using text-to-speech (tts) in post-editing machine translation (pemt) workflows on translators’ cognitive effort, productivity, quality, and perceptions. *Tradumàtica: Tecnologies de la Traducció*, 22:323–354.
- Michael Farrell. 2023. [Preliminary evaluation of chatgpt as a machine translation engine and as an au-](#)

- tomatic post-editor of raw machine translation output from other machine translation engines. In *Proceedings of the International Conference on Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2023)*, pages 108–113.
- Nico Herbig, Tim Düwel, Santanu Pal, Kalliopi Meladaki, Mahsa Monshizadeh, Antonio Krüger, and Josef van Genabith. 2020. Mmpe: A multi-modal interface for post-editing machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1691–1702.
- Dayeon Ki and Marine Carpuat. 2024. Guiding large language models to post-edit machine translation with error annotations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4253–4273.
- Maarit Koponen. 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.
- Maria Kunilovskaya, Koel Dutta Chowdhury, Heike Przybyl, Cristina España-Bonet, and Josef Genabith. 2024. Mitigating translationese with gpt-4: Strategies and performance. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation*, pages 411–430.
- Zongyao Li, Zhiqiang Rao, Hengchao Shang, Jiaxin Guo, Shaojun Li, Daimeng Wei, and Hao Yang. 2025. Enhancing large language models for document-level translation post-editing using monolingual data. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8830–8840.
- Jeevanthi U. Liyanapathirana, Perrine Schumacher, Pierrette Bouillon, and Jonathan David Mutal. 2024. Enhanced multilingual speech-based post-editing based on user feedback with copeco-speech. In *Proceedings of Translating and the Computer 46*, page 161.
- Lieve Macken. 2024. Machine translation meets large language models: Evaluating chatgpt’s ability to automatically post-edit literary texts. In *Proceedings of the First Workshop on Creative-text Translation and Technology*, pages 71–87.
- Jonathan Mutal, Pierrette Bouillon, Perrine Schumacher, and Johanna Gerlach. 2020. Copeco: A collaborative post-editing corpus in pedagogical context. In *Proceedings of the Workshop on Post-Editing of Machine Translation and Translation Technology (PEMDT@AMTA)*, pages 61–78.
- Sharon O’Brien. 2010. Introduction to post-editing: Who, what, how and where to next? In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas (AMTA): Tutorials*.
- Ming Qian and Chuiqing Kong. 2024. Enabling human-centered machine translation using concept-based large language model prompting and translation memory. In *Artificial Intelligence in HCI: 5th International Conference, AI-HCI 2024, Held as Part of the 26th HCI International Conference (HCII 2024)*, pages 118–134. Springer Nature Switzerland.
- Zeljko Radić. 2024. *ISV—Interlingual Subtitle Voicing: A New Technique for the Translation of Interlingual Subtitles—A Case Study in Croatian*. Ph.D. thesis, University of Surrey.
- Vikas Raunak, 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.
- Pilar Sánchez-Gijón and Leire Palenzuela-Badiola. 2023. Analysis and evaluation of chatgpt-induced hci shifts in the digitalised translation process. In *Proceedings of the International Conference on Human-Informed Translation and Interpreting Technology (HiT-IT 2023)*, pages 227–267.
- Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language models for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 41092–41110. PMLR.

A Prepared Workflow

The workflow consists of the following steps:

1. Loading: Source text and MT output are loaded into COPECO-Speech.

2. PE via typing: Users perform direct keyboard-based corrections to the MT output.

3. Speech interaction:

- **Navigation (predefined commands):** “Next/Previous/Save segment”, “Undo”
- **Editing via LLM:** Users provide natural-language instructions: e.g. “insert A before B”, “delete the word C”, “add a comma after D”, “retranslate the sentence”, etc

4. Output generation: The system applies the requested modifications and produces an updated translation.

5. Logging and analysis: COPECO-Speech records user activity, including keystrokes, speech input, LLM requests and outputs, and processing time. It also sends tasks (upon completion) to instructors for review, enabling correction and error annotation for further analysis.

B Prompt Capabilities

Capability	Description
Instruction interpretation	Distinguishes between navigation commands and editing instructions
Editing modes	Supports local editing operations and full retranslation
Context integration	Combines source text, MT output, and user input to be context aware.
Language control	Ensures the correct expected output language
Output constraints	Produces only the updated sentence
Fallback behaviour	Preserves the original output if instructions cannot be applied

Table 1: Key features of the prompt used for LLM-based editing

C Backend Prompt

The following prompt is currently used by the LLM-based editing component of COPECO-Speech.

You are a text transformation engine.

EXPECTED LANGUAGE (output language):
{lang}

INPUTS:

SOURCE SENTENCE (may be empty):
{sourceSentence}

TARGET SENTENCE (may be empty):
{targetSentence}

INSTRUCTION:
{instruction}

MODES

RETRANSLATE_MODE:

- If INSTRUCTION contains the token **RETRANSLATE_MODE**:
 - Translate **SOURCE SENTENCE** into the **EXPECTED LANGUAGE**.
 - Use **TARGET SENTENCE** only as context to preserve preferred terminology, tone, punctuation, and style where helpful.
 - Output **ONLY** the new **TARGET** sentence in the **EXPECTED LANGUAGE**.
 - Do **NOT** output **SKIP_LANGUAGE_MISMATCH** in this mode.

DEFAULT MODE (normal edit)

DETECT WHETHER THE INSTRUCTION IS A COMMAND OR DICTATION:

- A **COMMAND** requests a modification of the text (e.g., replacing, deleting, adding, correcting, rephrasing, shortening, expanding, or formalizing content, including equivalent expressions in French).
- **DICTATION** is plain content to insert as text, not asking for a transformation.

LANGUAGE CHECKS (default mode):

a) If TARGET SENTENCE is non-empty:

1. Detect the primary language of **TARGET SENTENCE**.
2. Detect the primary language of the **INSTRUCTION**.
3. If **TARGET** language is **NOT** primarily the **EXPECTED LANGUAGE**, output **EXACTLY**:

SKIP_LANGUAGE_MISMATCH

4. If INSTRUCTION is a COMMAND and its language is NOT primarily the same as TARGET language, output EXACTLY:

SKIP_LANGUAGE_MISMATCH

b) If TARGET SENTENCE is empty:

1. If INSTRUCTION is DICTATION:

- Detect the primary language of the INSTRUCTION.
- If it is NOT primarily the EXPECTED LANGUAGE, output EXACTLY:

SKIP_LANGUAGE_MISMATCH

- Otherwise output the INSTRUCTION as the new TARGET sentence.

2. If INSTRUCTION is a COMMAND:

- If SOURCE SENTENCE is empty: output an empty string.
- Otherwise create a good TARGET draft in the EXPECTED LANGUAGE from SOURCE SENTENCE, then apply the instruction to that draft.

DEFAULT MODE RULES:

- Return ONLY the resulting sentence text.
- Do NOT explain your reasoning.
- Do NOT include quotes, labels, or extra text.
- NEVER translate between languages, except in RETRANSLATE_MODE (only into the EXPECTED LANGUAGE).
- If the instruction cannot be applied, return the TARGET SENTENCE unchanged (or the drafted/inserted target if TARGET was empty).
- Preserve meaning unless the instruction explicitly changes it.
- If the instruction asks to replace some words (replace X with Y / remplace X par Y), perform an in-sentence replacement, NOT a full sentence replacement, unless REPLACE_MODE is present.

D Speech-driven Post-Editing: Additional Examples

The following examples were collected during a preliminary English–French post-editing experiment involving four students (see Section 6). They illustrate both predefined navigation commands and natural-language editing instructions recorded during the tasks.

D.1 Navigation Commands

The following spoken commands were used for navigation and interface control. These commands are matched against a predefined rule-based command set and executed directly within the interface.

- **Sauver segment** — save the current segment
- **Segment suivant** — move to the next segment
- **Segment précédent** — move to the previous segment
- **Arrêter** — stop the current task

D.2 LLM-Based Editing Instructions

Source (EN): One vomits what they do not digest.

MT Output (FR): On vomit ce qu'on digère.

Spoken Instruction (FR): Ajouter négation

Corrected Output (FR): On vomit ce qu'on **ne** digère.

Source (FR): Elle a quitté la pièce sans dire au revoir.

MT Output (FR): Elle a quitté pièce sans dire au revoir.

Spoken Instruction (FR): Ajouter l'article devant pièce

Corrected Output (FR): Elle a quitté **la** pièce sans dire au revoir.

Translation Memory Systems: Evolution, Retrieval Methods, and Human-Centred Perspectives

Alicia Picazo-Izquierdo

University of Alicante
alicia.picazo@ua.es

Salima Lamsiyah

Luxembourg University
salima.lamsiyah@uni.lu

Ruslan Mitkov

Lancaster University
University of Alicante
r.mitkov@lancaster.ac.uk

Abstract

This paper traces the evolution of translation memory systems (TMS) from early lexical and edit-distance matching to contemporary semantic, neural, and AI-augmented retrieval approaches. It reviews TMS as both technical infrastructures and human-computer interaction (HCI) systems within computer-aided translation (CAT) workflows. Situating itself within current debates on machine translation, large language models, and generative AI, the paper argues that TMS should no longer be viewed only as segment-reuse databases, but as hybrid retrieval infrastructures connecting human translators, institutional language resources, and AI-assisted translation workflows. The paper examines TM structure, match types, translator interaction, public resources, shared tasks, ethical issues of memory reuse, and user adoption. Finally, it discusses current limitations and future directions involving TM cleaning, explainability, provenance-aware metadata, hybrid lexical-semantic retrieval, and human-centred AI-assisted translation workflows.

1 Introduction

Translation memories (TM) are databases of aligned source–target segments that allow translators to retrieve and reuse previous translations within computer-aided translation (CAT) workflows (Bowker and Fisher, 2010; Samson, 2005; Reinke, 2018; Mitkov, 2022). Embedded in translator workbenches, they support consistency, productivity, and cost estimation by returning exact, fuzzy, contextual, or subsegment matches (Reinke, 2018; Christensen et al., 2024). Their usefulness, however, depends on memory quality, domain fit, metadata, and interface presentation, not only on match scores (da Silva Cardoso Teixeira, 2014; Krüger, 2016; Vargas-Sierra, 2019).

TMS have evolved from early work on bilingual concordancing, text reuse, and memory-based translation (Arthern, 1979; Kay, 1980; Melby,

1981; Sato and Nagao, 1990; Hutchins, 1998) to mature CAT technologies studied in terms of adoption, usability, translator cognition, and human-computer interaction (HCI) (Lagoudaki, 2006; Schjoldager and Christensen, 2010; O’Brien, 2012, 2020). In parallel, following the first work by Pekar and Mitkov (2007), NLP research has expanded TM retrieval beyond edit distance through paraphrasing, linguistic transformations, sentence encoders, semantic retrieval, and automatic cleaning (Bloodgood and Strauss, 2014; Gupta et al., 2016; Timonera and Mitkov, 2015; Ranasinghe et al., 2020; Barbu et al., 2016; Bhardwa et al., 2022; Estevanell Valladares et al., 2026).

Recent work further integrates TMs into machine translation (MT), neural machine translation (NMT), and LLM-assisted translation, where retrieved segments provide contextual, domain-specific, and human-validated evidence (Cao and Xiong, 2018; He et al., 2021; Cheng et al., 2022; Hao et al., 2023; Mu et al., 2023; Qian and Kong, 2024). Although previous surveys examined TM systems and translator perceptions (Lagoudaki, 2006; Schjoldager and Christensen, 2010; Reinke, 2018; Mitkov, 2022), an updated synthesis is needed to connect classical TM research with semantic retrieval, TM cleaning, public resources, ethical reuse, and AI-assisted translation workflows. This paper reviews translation memory systems (TMS) as technical, social, and human-centred infrastructures, covering their history, structure, HCI dimensions, social ecosystem, retrieval technologies, and future directions for explainable and provenance-aware AI-augmented TMS.

2 CAT Tools, Machine Translation, and Translation Memories

CAT tools are integrated environments that support file preparation, segmentation, terminology management, quality assurance, TM retrieval, MT, review, and delivery (Bowker and Fisher, 2010;

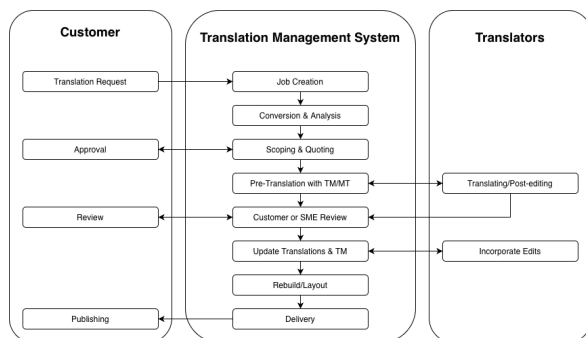


Figure 1: Translation management workflow integrating customer requests, TM/MT pre-translation, translator post-editing, review, TM updating, and delivery (Esselink, 2020).

Samson, 2005; Esselink, 2020). Within these environments, MT and TM often appear together, but they rely on different principles: MT generates new translations automatically, while TM retrieves previously translated segments which can be further edited by the user (translator) (Reinke, 2018; Mitkov, 2022).

Figure 1 shows how MT and TM are embedded in a broader workflow, from customer request and project analysis to pre-translation, post-editing, review, TM updating, and delivery.

Machine translation is a natural language processing task in which a system automatically produces a target text from a source text. In professional workflows, MT output is usually treated as a draft requiring post-editing, terminology control, and quality verification (O’Brien, 2012; Screen, 2017). Recent neural and LLM-based systems can also use external resources, including terminology databases and TM examples, to improve domain consistency and contextual adequacy (Cao and Xiong, 2018; He et al., 2021; Hao et al., 2023; Mu et al., 2023; Qian and Kong, 2024).

Translation memories are CAT databases of past human translations stored as source–target segment pairs. Unlike MT, TM does not generate a translation from scratch; it retrieves human-approved segments that may be reused or adapted in the current task (Reinke, 2018; Mitkov, 2022). TM suggestions therefore require match evaluation: translators must judge whether a retrieved segment is contextually appropriate, up to date, terminologically consistent, and reusable (da Silva Cardoso Teixeira, 2014; Krüger, 2016; Vargas-Sierra, 2019). This distinction matters because TM and MT create different forms of cognitive effort, trust,

and interaction within CAT environments (O’Brien, 2012, 2020).

3 Translation Memories as Human-Centred Technologies

Beyond retrieval performance, TMS must be understood as human-centred technologies because their value depends on how translators interact with suggestions, manage cognitive effort, and build trust in the system.

3.1 Translator Interaction and Cognitive Effort

When translators work with translation memories, they engage in an interaction loop in which the TM engine retrieves and ranks stored segments, the interface presents suggestions with metadata such as match percentage, origin, and date, and the translator accepts, edits, or rejects them before the memory is updated (Lagoudaki, 2006; da Silva Cardoso Teixeira, 2014). This makes TMS mixed-initiative environments in which the system proposes reusable translations, while the translator evaluates their adequacy and contextual fit (Green et al., 2014; O’Brien, 2012).

The interaction between the translator and TMs is a distributed cognitive task, because the translator’s decisions are partly mediated by TM suggestions rather than holding sole authority over target-language choices (Dragsted, 2006).

TM use does not automatically reduce cognitive effort. Even high-percentage fuzzy matches may require careful reading, comparison, and editing when terminology, style, or context differs (O’Brien, 2007, 2008; Mellinger and Shreve, 2016). Eye-tracking, keystroke, and pause-based studies show that match quality and interface cues influence translators’ attention, revision behaviour, and productivity (O’Brien et al., 2010; O’Brien, 2006, 2013; Screen, 2017). Thus, the value of a TM suggestion depends not only on similarity, but also on the effort needed to verify and adapt it.

3.2 Usability, Trust, and Metadata

Within this interaction loop, the design of the TM interface has been examined through classic HCI lenses such as usability, efficiency, and cognitive ergonomics. Studies of specific interface components, such as the concordance search window or the arrangement of match lists, show how layout, typography, colour coding and shortcut design shape

translators’ visual attention and navigation strategies (O’Brien et al., 2010; Krüger, 2016; Vargas-Sierra, 2019). A usable TM interface should therefore help translators rapidly understand what has been retrieved, why it is relevant, and whether it can be trusted in the current context (da Silva Cardoso Teixeira, 2014; Taravella and Villeneuve, 2013).

Metadata plays a central role in this process. Information such as client, domain, creation date, previous translator, approval status, and match origin can help translators judge whether a suggested translation is reliable or outdated (da Silva Cardoso Teixeira, 2014; Reinke, 2018). Cognitive-ergonomic approaches argue for TM designs that support selective attention, make uncertainty and origin information legible, and allow translators to tailor the interface to their workflows rather than forcing a one-size-fits-all configuration (Krüger, 2016; Taravella and Villeneuve, 2013). This is particularly important in AI-assisted workflows, where translators may need to distinguish between human-translated, post-edited MT, and LLM-assisted segments. Provenance-aware metadata can therefore support trust calibration, responsible reuse, and better human control over TM suggestions (Park, 2024; Qian and Kong, 2024).

4 Social, Professional, and Ethical Dimensions of TMS

Translation memory systems are not only CAT retrieval tools; they are also professional and institutional infrastructures shaped by research communities, public resources, industry practices, user adoption, and ethical questions of reuse. This section examines these social, professional, and ethical dimensions of TMS.

4.1 Research Communities, Workshops, and Shared Tasks

TM research has evolved from database-oriented work toward NLP-enhanced retrieval, cleaning, and TM–MT integration. Dedicated venues such as Beyond Translation Memories at MT Summit 2009 (BMT, 2009) and the NLP4TM workshops at RANLP 2015 and LREC 2016 (Orasan and Gupta, 2015; Orasan et al., 2016) helped establish TM as a focused research area. The 1st Shared Task on Automatic Translation Memory Cleaning further introduced a benchmark for detecting and removing low-quality TM units (Barbu et al., 2016). Together,

these venues show a shift from describing TM tools to evaluating retrieval quality, cleaning methods, and NLP-enhanced reuse. See also Mitkov (2022), section ‘Recent developments: New-generation TM tools – can Language Technology make them better?’.

Broader initiatives have also shaped TM research. The M3TRA workshop addressed TM–MT integration and fuzzy-match repair for professional translation workflows (Bulté et al., 2018). Recent work on knowledge-enhanced MT and TM-augmented generation further connects TMS with neural and AI-assisted translation (Tezcan et al., 2024; Hao et al., 2023; Mu et al., 2023; Qian and Kong, 2024).

4.2 Public Translation Memory Resources

Translation memories are often proprietary assets curated for specific clients, domains, or tasks, but public TM infrastructures have also become important for research. A key example is the European Commission’s DGT-TM, released by the Directorate-General for Translation, which contains millions of translation units across EU languages (Steinberger et al., 2012). Related resources, such as JRC-Acquis for EU law and ECDC-TM for public health, provide large-scale multilingual data for TM retrieval, MT training, domain adaptation, and evaluation.

Beyond institutional repositories, the OPUS repository acts as a critical infrastructure for the research community by aggregating and converting parallel corpora into standardized formats. Similarly, MyMemory represents one of the largest collaborative TM infrastructures, pooling professional contributions and web-crawled data. However, public and collaborative resources also raise questions about quality control, provenance, licensing, and the reliability of reused segments, especially when they are integrated into commercial or AI-assisted workflows (Barbu et al., 2016; Negri et al., 2017; Park, 2024).

4.3 Professional Uptake and User Perception

An important dimension of TMs is how translators perceive and use them. User surveys, especially Lagoudaki’s, show that TM impact cannot be assessed only through retrieval performance, but must also consider adoption, perceived benefits, and limitations (Lagoudaki, 2006; Vargas-Sierra, 2019).

Translators value TMs for consistency, productivity, and terminology control, especially in repet-

itive domains, but report problems with noisy matches, weak contextualisation, interface overload, and the effort needed to judge reusability (Schjoldager and Christensen, 2010; Taravella and Villeneuve, 2013). This shows that professional uptake depends not only on retrieval accuracy, but also on usability, trust, metadata transparency, and integration with translators’ actual workflows (da Silva Cardoso Teixeira, 2014; Krüger, 2016; Mellinger and Shreve, 2016).

4.4 Ethical and Legal Issues of TM Reuse

The reuse of translation memories also raises ethical and legal issues. TMs often contain client-specific, domain-specific, or institutionally owned translations, which makes ownership, consent, confidentiality, and copyright central concerns (Park, 2024). In professional settings, a segment stored in a TM may be technically reusable but ethically problematic if it comes from a restricted client project, contains sensitive information, or lacks clear licensing and provenance.

These concerns become more pressing in AI-assisted workflows. If MT or LLM-generated outputs are stored in a TM without explicit labels, future translators may unknowingly reuse segments that were only weakly verified, stylistically homogenized, or affected by machine-generated errors. For this reason, future TMS should include provenance-aware metadata distinguishing human translation, post-edited MT, and LLM-assisted output, as well as quality and approval indicators to support responsible reuse (Park, 2024; Qian and Kong, 2024; Mu et al., 2023).

Semantic retrieval also raises professional implications for pricing. Conventional fuzzy-match bands rely on surface similarity and are often used for billing, but embedding-based retrieval may identify useful matches that are semantically close while lexically distant. This may reshape pricing bands in translation projects including embedding-based TMs.

5 Retrieval Technologies and Quality Control

Translation memory retrieval ranks stored bilingual segments for a new source segment using methods ranging from surface similarity to semantic and quality-aware approaches. The main challenge is to retrieve not only similar segments, but reusable and trustworthy ones for real translation workflows.

5.1 Edit-Distance and Linguistically Aware Matching

The classical retrieval logic in TMs is based on surface similarity between the new source segment and stored translation units. Most systems have historically relied on edit-distance variants, often computed at word level and normalized by segment length (Bloodgood and Strauss, 2014). These methods are transparent, easy to threshold, and useful for pricing and effort estimation, but they fail when segments are semantically close yet lexically different. See Mitkov (2022) for a more detailed discussion.

In practical CAT workflows, even subsegment or fragment matches provide reusable phrases or shorter bilingual units when no full-segment match is available, and some try to prepair fuzzy matches automatically, such as Trados Studio’s upLIFT Fuzzy Match Repair or memoQ’s MatchPatch.

This limitation motivated linguistically-aware methods that use lemmas, part-of-speech information, paraphrases, clause splitting, or other transformations to improve matching beyond surface overlap (Planas and Furuse, 2000; Gupta and Orasan, 2014; Chatzitheodorou, 2015; Timonera and Mitkov, 2015; Gupta et al., 2016; Vanallemeersch and Vandeghinste, 2015). Such methods make retrieval more flexible while preserving part of the interpretability of fuzzy matching.

5.2 Semantic and Embedding-Based Retrieval

Embedding-based retrieval represents segments as dense vectors and retrieves nearest neighbours according to semantic similarity. Sentence encoders can therefore retrieve paraphrases or meaning-preserving variants that edit distance may miss (Reimers and Gurevych, 2019; Ranasinghe et al., 2020). This is useful when translators need conceptually similar examples rather than near-identical repetitions.

SmartMatch illustrates this direction by combining sentence encoders, lexical baselines, and a vector database in a real-time TM retrieval toolkit, allowing users to compare retrieval backends, inspect top-*k* matches, and evaluate quality–latency trade-offs for multilingual TM data (Estevanell Valldares et al., 2026). However, semantic similarity does not guarantee reusability. Embedding-based retrieval should therefore be combined with lexical scores, metadata, domain filters, and interface cues that help translators judge whether a match is

trustworthy.

5.3 TM Cleaning and Quality-Aware Retrieval

Retrieval quality also depends on the quality of the memory itself. Noisy TMs may contain misaligned pairs, partial translations, wrong-language segments, inconsistent terminology, encoding errors, or outdated translations. TM cleaning addresses this by removing or down-ranking unreliable translation units using heuristics, feature-based models, unsupervised methods, or quality-estimation-inspired approaches (Barbu et al., 2016; Jalili Sabet et al., 2016b,a; Negri et al., 2017; Wolff, 2016). Recent work shows that even refining a mostly clean TM can improve downstream MT quality (Bhardwa et al., 2022).

Recent developments in AI-based quality estimation (QE) and linguistic quality assessment (LQA) also offer promising directions for TM cleaning and maintenance. Such models could flag source–target pairs that appear mistranslated, incomplete, misaligned, terminologically inconsistent, stylistically divergent, or otherwise unreliable. In professional TMS, these scores could be used to exclude low-confidence units from retrieval, down-rank them, route them for human review, or display quality warnings to translators.

Quality-aware retrieval shifts the goal from finding the most similar segment to finding the most reliable and useful one. This requires attention to alignment quality, domain relevance, recency, approval status, provenance, and translator feedback, especially in AI-assisted workflows where TMs may contain human translations, post-edited MT, and LLM-assisted segments.

6 Translation Memories in the Neural and Generative AI Era

With the rise of NMT, LLMs, and generative AI, translation memories are increasingly integrated with MT systems in both directions: TMs can provide domain-specific evidence for MT, while MT can assist TM workflows by offering suggestions when no useful fuzzy match is available.

6.1 TM–MT Integration in Professional Workflows

The integration of translation memories and machine translation has developed in two complementary directions. The first uses TM to improve

MT or NMT, where retrieved translation units provide similar source segments, target translations, or fuzzy-match signals during decoding (Wang et al., 2013; Li et al., 2014; Liu et al., 2019). In neural architectures, TM matches can be encoded as additional inputs, integrated through gating mechanisms, or used to improve domain consistency and lexical choice (Cao and Xiong, 2018; Eriguchi et al., 2019; He et al., 2021). Recent work further extends this idea through graph-based, monolingual, contrastive, and non-autoregressive TM integration (Xia et al., 2019; Cai et al., 2021; Cheng et al., 2022; Hao et al., 2023; Xu et al., 2023).

The second direction, which is more visible in professional CAT workflows, uses MT or NMT to assist TM systems when no exact or sufficiently useful fuzzy match is available. In this setting, MT output is presented alongside TM suggestions, or used to repair low-quality fuzzy matches, so that translators can compare, edit, or post-edit the most useful option (Kanavos and Kartsaklis, 2010; Koehn and Senellart, 2010; Caro Quintana, 2021). This external integration became increasingly common after the improvement of NMT, as CAT tools began to combine TM matches, MT suggestions, terminology, and post-editing within the same interface (Zaretskaya et al., 2015; Sánchez-Gijón et al., 2019; Screen, 2017). For translators, TM–MT integration therefore means not only using TM to improve MT, but also using MT as a fallback or repair mechanism when TM retrieval alone cannot provide a useful match. This practical perspective is central to understanding TMS in the neural and generative AI era.

6.2 LLM-Assisted Translation and Retrieval-Augmented Workflows

LLM-assisted translation further expands the role of TMs by using retrieved segments as examples, constraints, or contextual evidence for generation and revision (Mu et al., 2023; Qian and Kong, 2024). This resembles retrieval-augmented generation: relevant bilingual evidence is first retrieved from the TM and then used to guide translation, terminology choice, or post-editing. Semantic retrieval is central here because LLMs may benefit from conceptually similar examples even when surface similarity is low. SmartMatch illustrates this direction by combining lexical baselines, sentence encoders, vector databases, and real-time inspection for CAT-oriented TM retrieval (Estevanell Val-

ladares et al., 2026). Still, retrieved examples require human evaluation, since semantic similarity does not guarantee contextual, terminological, or legal reusability.

6.3 Provenance, Explainability, and Human Control

AI-assisted workflows also introduce risks. If MT or LLM-generated outputs are stored without clear labels, future translators may unknowingly reuse machine-generated, weakly verified, or stylistically homogenized segments. Future TMS should therefore include provenance-aware metadata distinguishing human translation, post-edited MT, and LLM-assisted output (Park, 2024; Qian and Kong, 2024).

Explainability is also needed so that translators can understand why a segment was retrieved, where it came from, and how reliable it is. Interfaces should expose match origin, quality indicators, domain information, approval status, and retrieval rationale. In this sense, the future of TMS is not full automation, but human-centred retrieval that supports translator judgment while preserving control, accountability, and trust.

7 Conclusion

This paper has examined translation memory systems as evolving technical, social, and human-centred infrastructures. The discussion shows that TMS remain central to CAT workflows because they support consistency, productivity, and reuse, but their effectiveness depends on more than retrieval accuracy. Memory quality, metadata, domain fit, usability, and translator trust are equally important. Recent developments in semantic retrieval, TM cleaning, neural MT, and LLM-assisted translation further reposition TMS as retrieval resources for AI-supported translation rather than simple segment-reuse databases. Future work should empirically evaluate these hybrid retrieval designs with professional translators, since the practical reusability of semantically similar matches depends on how translators interpret, verify, and adapt them in real CAT workflows.

Several challenges remain. Classical fuzzy matching is transparent but limited by surface similarity, while semantic retrieval improves coverage but may return segments that are not directly reusable. Noisy, outdated, or poorly documented memories can also reduce reliability, especially

when human translations, post-edited MT, and LLM-assisted outputs are stored together without clear provenance. Future TMS should therefore move toward hybrid lexical-semantic retrieval, quality-aware cleaning, provenance-aware metadata, and explainable interfaces that help translators understand why a match is suggested and whether it can be trusted. In this sense, the future of TMS is not full automation, but human-centred AI support that strengthens translator judgment, accountability, and control.

Acknowledgements

This research has been funded by the Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport) through the project "The limits and future of data-driven approaches: A comparative study of deep learning, knowledge-based and rule-based models and methods in Natural Language Processing" (CIDEXG/2023/13).

References

- Peter Arthern. 1979. Machine translation and computerized terminology systems: A translator's viewpoint. In *Translating and the Computer, Proceedings of a Seminar*. North-Holland Publishing Company.
- Eduard Barbu, Carla Parra Escartín, Luisa Bentivogli, Matteo Negri, Marco Turchi, Constantin Orasan, and Marcello Federico. 2016. The first automatic translation memory cleaning shared task. *Machine Translation*, 30(3):145–166.
- Shivendra Bhardwa, David Alfonso-Hermelo, Philippe Langlais, Gabriel Bernier-Colborne, Cyril Goutte, and Michel Simard. 2022. [Refining an almost clean translation memory helps machine translation](#). In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 215–226, Orlando, USA. Association for Machine Translation in the Americas.
- Michael Bloodgood and Benjamin Strauss. 2014. [Translation memory retrieval methods](#). In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 202–210, Gothenburg, Sweden. Association for Computational Linguistics.
- BMT. 2009. [Mt summit](#). In *Beyond Translation Memories: New Tools for Translators Workshop*.
- Lynne Bowker and Des Fisher. 2010. Computer-aided translation. In Y. Gambier and L. V. Doorslaer, editors, *Handbook of translation studies*, volume 1, pages 60–65. John Benjamins Publishing Company.

- Bram Bulté, Tom Vanallemeersch, and Vincent Vandeghinste. 2018. [M3TRA: integrating TM and MT for professional translators](#). In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 89–98, Alicante, Spain.
- Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 2021. [Neural machine translation with monolingual translation memory](#). In *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 7307–7318, Online. Association for Computational Linguistics.
- Qian Cao and Deyi Xiong. 2018. [Encoding gated translation memory into neural machine translation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3042–3047, Brussels, Belgium. Association for Computational Linguistics.
- Rocío Caro Quintana. 2021. [Integration of machine translation and translation memory: Post-editing efforts](#). In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 161–166.
- Konstantinos Chatzitheodorou. 2015. Improving translation memory fuzzy matching by paraphrasing. In *Proceedings of the Workshop Natural Language Processing for Translation Memories*, pages 24–30.
- Xin Cheng, Shen Gao, Lemao Liu, Dongyan Zhao, and Rui Yan. 2022. [Neural machine translation with contrastive translation memories](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3591–3601, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tina Paulsen Christensen, Kristine Bundgaard, Anne Schjoldager, and Helle Dam Jensen. 2024. Translation memory systems. *Encyclopedia of translation and interpreting*.
- Carlos da Silva Cardoso Teixeira. 2014. *The impact of metadata on translator performance: how translators work with translation memories and Machine translation*. Ph.D. thesis, Universitat Rovira i Virgili.
- Barbara Dragsted. 2006. Computer-aided translation as a distributed cognitive task. *Pragmatics and Cognition*, 2:443–464.
- Akiko Eriguchi, Spencer Rarrick, and Hitokazu Matsushita. 2019. Combining translation memory with neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 123–130.
- Bert Esselink. 2020. Multinational language service provider as user. In Minako OHagan, editor, *The Routledge Handbook of Translation and Technology*, pages 109–126. Routledge.
- Ernesto Luis Estevanell Valladares, Salima Lamsiyah, Alicia Picazo-Izquierdo, Tharindu Ranasinghe, Ruslan Mitkov, and Rafael Munoz. 2026. [SmartMatch: Real-time semantic retrieval for translation memory systems](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 527–536, Rabat, Morocco. Association for Computational Linguistics.
- Spence Green, Jason Chuang, Jeffrey Heer, and Christopher D Manning. 2014. Predictive translation memory: A mixed-initiative system for human language translation. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 177–187.
- Rohit Gupta and Constantin Orasan. 2014. Incorporating paraphrasing in translation memory matching and retrieval. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 3–10.
- Rohit Gupta, Constantin Orăsan, Marcos Zampieri, Mihaela Vela, Josef van Genabith, and Ruslan Mitkov. 2016. Improving translation memory matching and retrieval using paraphrases. *Machine Translation*, 30(1):19–40.
- Hongkun Hao, Guoping Huang, Lemao Liu, Zhirui Zhang, Shuming Shi, and Rui Wang. 2023. [Rethinking translation memory augmented neural machine translation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2589–2605, Toronto, Canada. Association for Computational Linguistics.
- Qiuxiang He, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. 2021. [Fast and accurate neural machine translation with translation memory](#). In *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 3170–3180, Online. Association for Computational Linguistics.
- John Hutchins. 1998. The origins of the translator’s workstation. *Machine Translation*, 13:287–307.
- Masoud Jalili Sabet, Matteo Negri, Marco Turchi, and Eduard Barbu. 2016a. [An unsupervised method for automatic translation memory cleaning](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 287–292, Berlin, Germany. Association for Computational Linguistics.
- Masoud Jalili Sabet, Matteo Negri, Marco Turchi, José G. C. de Souza, and Marcello Federico. 2016b. [TMop: a tool for unsupervised translation memory cleaning](#). In *Proceedings of ACL-2016 System Demonstrations*, pages 49–54, Berlin, Germany. Association for Computational Linguistics.

- Panagiotis Kanavos and Dimitrios Kartsaklis. 2010. [Integrating machine translation with translation memory: A practical approach](#). In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 11–20, Denver, Colorado, USA. Association for Machine Translation in the Americas.
- Martin Kay. 1980. The proper place of men and machines in language translation. *Research Report CSL-80-11*.
- Philipp Koehn and Jean Senellart. 2010. [Convergence of translation memory and statistical machine translation](#). In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 21–32, Denver, Colorado, USA. Association for Machine Translation in the Americas.
- Ralph Krüger. 2016. Contextualising computer-assisted translation tools and modelling their usability. *transkom*, pages 114–148.
- Elina Lagoudaki. 2006. Translation memories survey 2006: Users’ perceptions around tm use. *Translating and the Computer*, 28:1–29.
- Liangyou Li, Andy Way, and Qun Liu. 2014. A discriminative framework of integrating translation memory features into smt. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, pages 249–260.
- Yang Liu, Kun Wang, Chengqing Zong, and Keh-Yih Su. 2019. A unified framework and models for integrating translation memory into phrase-based statistical machine translation. *Computer Speech & Language*, 54:176–206.
- Alan Melby. 1981. A bilingual concordance system and its use in linguistic studies. In *The Eight LACUS Forum*, pages 541–549. Hornbeam Press.
- Christopher D Mellinger and Gregory M Shreve. 2016. Match evaluation and over-editing in a translation memory environment. In *Reembedding translation process research*, pages 131–148. John Benjamins Publishing Company.
- Ruslan Mitkov. 2022. Translation memory systems. In *The Routledge Handbook of Translation and Memory*, pages 364–380. Routledge.
- Yongyu Mu, Abudurexiti Reheman, Zhiqian Cao, Yuchun Fan, Bei Li, Yinqiao Li, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023. [Augmenting large language model translators via translation memories](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10287–10299, Toronto, Canada. Association for Computational Linguistics.
- Matteo Negri, Duygu Ataman, Masoud Jalili Sabet, Marco Turchi, and Marcello Federico. 2017. Automatic translation memory cleaning. *Machine Translation*, 31(3):93–115.
- Sharon O’Brien. 2006. [Pauses as indicators of cognitive effort in post-editing machine translation output](#). *Across Languages and Cultures*, 7(1):1 – 21.
- Sharon O’Brien. 2007. [Eye-tracking and translation memory matches](#). *Perspectives*, 14(3):185–205.
- Sharon O’Brien. 2008. Processing fuzzy matches in translation memory tools : an eye-tracking analysis. In *Looking at Eyes: Eye-Tracking studies of Reading and Translation Processing*, pages 79–102. Samfundslitteratur.
- Sharon O’Brien. 2012. Translation as human–computer interaction. *Translation Spaces*, page 101–122.
- Sharon O’Brien. 2013. [The borrowers: Researching the cognitive aspects of translation](#). *Target. International Journal of Translation Studies*, 25(1):5–17.
- Sharon O’Brien. 2020. Translation, human–computer interaction and cognition. In *The Routledge Handbook of Translation and Cognition*. Routledge.
- Sharon O’Brien, Minako O’Hagan, and Marian Flanagan. 2010. Keeping an eye on the ui design of translation memory: how do translators use the "concordance" feature? *ECCE ’10: Proceedings of the 28th Annual European Conference on Cognitive Ergonomics*, pages 187 – 190.
- Constantin Orasan and Rohit Gupta. 2015. 1st workshop on natural language processing for translation memories (nlp4tm). In *Recent Advances in Natural Language Processing Conference*.
- Constantin Orasan, Carla Parra, Eduard Barbu, and Marcello Federico. 2016. 2nd workshop on natural language processing for translation memories (nlp4tm). In *Language Resources and Evaluation*.
- Jiyoung Park. 2024. Ethical approach to translation memory reuse: discussions from copyright and business ethics perspectives. *Translation Studies*, 17(1):37–52.
- Emmanuel Planas and Osamu Furuse. 2000. [Multi-level similar segment matching algorithm for translation memories and example-based machine translation](#). In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*.
- Ming Qian and Chuiqing Kong. 2024. Enabling human-centered machine translation using concept-based large language model prompting and translation memory. In *International Conference on Human-Computer Interaction*, pages 118–134. Springer.
- Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. [Intelligent translation memory matching and retrieval with sentence encoders](#). In *Proceedings of the 22nd Annual Conference of the European*

- Association for Machine Translation, pages 175–184, Lisboa, Portugal. European Association for Machine Translation.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Uwe Reinke. 2018. State of the art in translation memory technology. *Language Technologies for a Multilingual Europe*; Rehm, G., Stein, D., Sasaki, F., Witt, A., Eds, pages 55–84.
- Richard Samson. 2005. Computer-assisted translation. In M. Tennent, editor, *Training for the new millennium : Pedagogies for translation and interpreting*, pages 101–126. John Benjamins Publishing Company.
- Pilar Sánchez-Gijón, Joss Moorkens, and Andy Way. 2019. Post-editing neural machine translation versus translation memory segments. *Machine Translation*, 33(1):31–59.
- Satoshi Sato and Makoto Nagao. 1990. Toward memory-based translation. In *COLING 1990 Volume 3: Papers Presented to the 13th International Conference on Computational Linguistics*.
- Anne Gram Schjoldager and Tina Paulsen Christensen. 2010. Translation-memory (tm) research: what do we know and how do we know it? *Hermes*, 44:89–101.
- Benjamin Screen. 2017. Productivity and quality when editing machine translation and translation memory outputs: an empirical analysis of english to welsh translation. *Studia Celtica Posnaniensia*, 2(1):113–136.
- Ralf Steinberger, Andreas Eisele, Szymon Kloczek, Spyridon Pilos, and Patrick Schlüter. 2012. [DGT-TM: A freely available translation memory in 22 languages](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 454–459, Istanbul, Turkey. European Language Resources Association (ELRA).
- AnneMarie Taravella and Alain Villeneuve. 2013. Acknowledging the needs of computer-assisted translation tools users: The human perspective in human-machine translation. *The Journal of Specialised Translation*, 19:62–74.
- Arda Tezcan, Víctor M. Sánchez-Cartagena, and Miquel Esplà-Gomis, editors. 2024. [Proceedings of the First International Workshop on Knowledge-Enhanced Machine Translation](#). European Association for Machine Translation (EAMT), Sheffield, United Kingdom.
- Katerina Raisa Timonera and Ruslan Mitkov. 2015. Improving translation memory matching through clause splitting. In *Proceedings of the Workshop Natural Language Processing for Translation Memories*, pages 17–23.
- Tom Vanallemeersch and Vincent Vandeghinste. 2015. [Assessing linguistically aware fuzzy matching in translation memories](#). In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, Antalya, Turkey. European Association for Machine Translation.
- Chelo Vargas-Sierra. 2019. Usability evaluation of a translation memory system. *Quaderns de Filologia: Estudis Lingüístics*, 24:171–195.
- Kun Wang, Chengqing Zong, and Keh-Yih Su. 2013. [Integrating translation memory into phrase-based machine translation during decoding](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–21, Sofia, Bulgaria. Association for Computational Linguistics.
- Friedel Wolff. 2016. Combining off-the-shelf components to clean a translation memory. *Machine Translation*, 30(3):167–181.
- Mengzhou Xia, Guoping Huang, Lema Liu, and Shuming Shi. 2019. Graph based translation memory for neural machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7297–7304.
- Jitao Xu, Josep Crego, and François Yvon. 2023. [Integrating translation memories into non-autoregressive machine translation](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1326–1338, Dubrovnik, Croatia. Association for Computational Linguistics.
- Anna Zaretskaya, Gloria Corpas Pastor, and Miriam Seghiri. 2015. Integration of machine translation in CAT tools: State of the art, evaluation and user attitudes. *SKASE Journal of Translation and Interpretation*, 8(1):76–89.

Project HRVOJE-M: integration of Croatian LLM with a TBMS

Marko Tadić

University of Zagreb, Faculty of
Humanities and Social Sciences /
Ivana Lučića 3, Zagreb, Croatia
marko.tadic@ffzg.hr

Mladen Stojak

Ciklopea Ltd. / Međimurska 21,
Zagreb, Croatia

mladen.stojak@ciklopea.com

Abstract

The paper describes the project HRVOJE-M that was selected for funding under EU-programme Competitiveness and Cohesion within the IRI S3 call. The first goal of the project is to produce a LLM trained exclusively on Croatian data of at least 20 billion tokens in size. The second goal is the integration of that LLM in the proprietary TBMS Orchestra and its offer to different users through API.

1 Introduction

In this paper we are presenting the work-in-progress, or more precisely, work-at-its-beginning, on a project HRVOJE-M (Hrvatski Resursi za Veliki Obučeni Jezični Model / Croatian Resources for a Large Trained Language Model) that was selected for funding from the Competitiveness and Cohesion EU-programme, within IRI S3 call PK.1.1.12 in Croatia. Although the project has been approved for funding, we are still waiting for the allocation of funds and the GA to be signed.

The project coordinator is translation and localisation company Ciklopea Ltd.¹ from Zagreb, Croatia as an industrial partner, while the academic partner is University of Zagreb, Faculty of Humanities and Social Sciences.

The first goal of the project is to produce a monolingual LLM for Croatian (HR-VojeM) following the experience from the previous projects in building and using LLMs. The second goal is to

integrate this LLM in the proprietary Translation Business Management System (TBMS) Orchestra, developed by Ciklopea in order to offer the usage of Croatian LLM through this TBMS and to boost up its efficiency.

TBMSes have become common and expected tools for management of a workflow in translation and localisation companies, i.e. LSPs. Ciklopea developed the TBMS Orchestra in 2016 and it has been used extensively by several enterprise users and dozens of individual users on daily basis.

In the very same period, LLMs demonstrated significant advancements in processing natural language tasks and here we want to present a case of integration of LLM in the translation process through its integration in the TBMS Orchestra.

The paper is organized as follows: after the introduction the section 2 presents the Croatian monolingual LLM through the previous work and through the planned upgrading of the first HR-GPT Beta LLM. In the section 3 the TBMS Orchestra is being presented, its features and capabilities listed and explained. In the section 4 the planned integration is described, while the paper ends with conclusions in the section 5.

2 Croatian monolingual LLM

2.1 Previous work

Croatian language appeared in many multilingual LLMs so far, but the real foundation for the project HRVOJE-M is the previous project Croatian

¹ <https://ciklopea.com/>

Extended Reality Extensions (HR-XR-XTEND²) that in 2024 produced the first Croatian monolingual LLM (see for more details in Štefanec et al. 2024, Thakkar et al. 2024).

Most of the LLMs created so far are multilingual, i.e., they were trained using texts in different languages, and the number of languages varies from 2 to 200+. One inherent problem with multilingual LLMs is the unbalanced dataset used for training (Lai et al., 2023). The languages with large amounts of training data are usually well represented in the model, whereas languages with limited resources are constrained by the paucity of training data (Kahana et al., 2024). Although there is a dominant opinion that multilingual LLMs perform better in all NLP tasks, some studies (Martin et al., 2020; Papadimitriou et al., 2023; Virtanen et al., 2019) have shown that models trained in only few languages (bilingual/trilingual LLMs) or in a single language (monolingual LLMs) perform better in a number of NLP tasks than multilingual LLMs. The most recent contribution to this claim comes additionally from (Chang et al., 2026) where in the collection of monolingual LLMs for 350 languages, for many under-resourced languages monolingual LLMs surpassed several popular multilingual LLMs in benchmarks FLORES-200 log perplexity and MultiBLiMP. There are also recent attempts to distil the monolingual model from a larger multilingual LLM in order to come up with a smaller and more efficient monolingual LLM (Singh et al., 2023). That was the main reason for instigating the project HR-XR-XTEND where we proposed the creation of a monolingual LLM tailored to address the needs of lower-resourced languages, focusing on those with less than 6 million speakers (Language Report Croatian, 2022).

The main objective of the project HR-XR-XTEND, a FSTP subproject of a Horizon Europe-funded project UTTER³, was to build a large monolingual generative pre-trained transformer model for Croatian (HR-GPT). The relevant project goals were to collect at least 6 billion tokens of Croatian texts and to train the LLM for Croatian using monolingual data only.

A large number of tokens were collected from various sources: corpora and text collections from

the existing repositories such as CC100-Hr Dataset, Corpus of Croatian News Feeds, Parallel data for En-Hr in OPUS resources, ParaCrawl, etc. (although often under-performing for Croatian, e.g. results from the Oscar project⁴), but also newly collected data (online and offline) that were not available in previous data collection campaigns (e.g. Croatian Scientific Corpus). In total, more than 14.5 billion tokens were collected, but due to the short duration of the project, only 8.9 billion were selected for cleaning up. After the cleaning up and deduplication, where repeating paragraphs were removed from the training data, the total number of tokens used for training was downsized to 7.72 billion. For that reason we still name the resulting LLM for Croatian HR-GPT Beta and it will reach its full version only when we will manage to clean all collected data and train the model again.

The architecture used for producing the HR-GPT Beta with training from the scratch was open source GPT-NeoX. However, with the same data collection we additionally trained two openly available models (Gemma and Pythia) and produced HR-Gemma Beta and HR-Pythia Beta. All three models are accessible under permissive licenses from the repository⁵ at the HR-CLARIN, a Croatian node in the European research infrastructure CLARIN ERIC⁶.

2.2 Upgrading the HR-GPT Beta

Despite the significant results of previous research, numerous limitations have been observed – both in terms of quality of results, robustness of models in real conditions, and in possibilities of integration into business and public digital systems. These challenges have been recognised as an opportunity to explore new LLM architectures and develop a new generation innovative solution. Therefore, the applicant and the partner within the project HRVOJE-M decided to develop a large generative language model which would be trained on a much larger amount of texts (20-25 billion tokens), it would include more complex training mechanisms and could be available through API interfaces integrated into existing digital platforms. The aim of the project is to develop a LLM trained exclusively in Croatian,

² <https://hr-xr-xtend.ffzg.unizg.hr/>

³ <https://he-utter.eu/>

⁴ <https://oscar-project.org/>

⁵ <https://repository.clarin.hr/>

⁶ <https://clarin.eu>

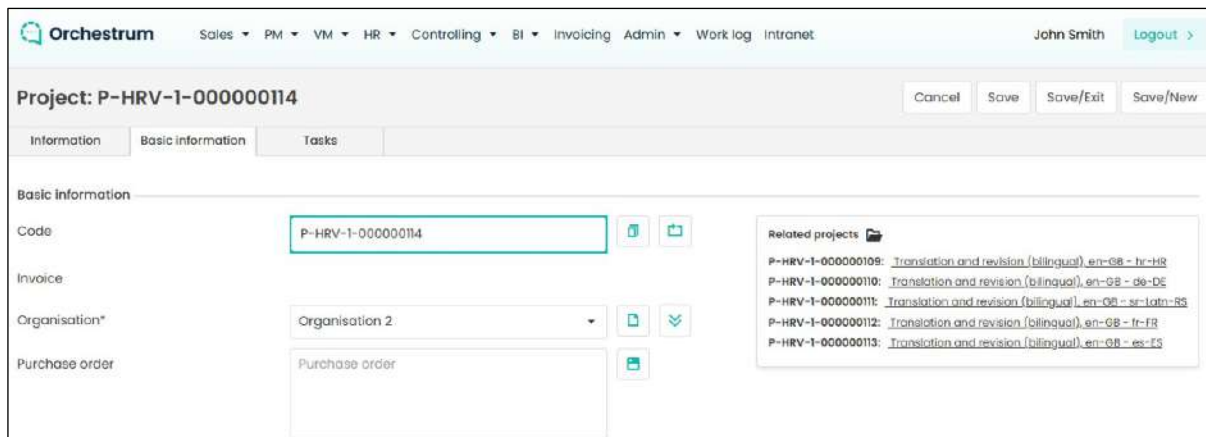


Figure 1: A screenshot of the Orchestra with the basic information about the translation project

capable of widespread application in the language services industry, public administration, health care, education and various other sectors. Such a model would be the first of its kind and size in Croatia.

The project will be realized in four phases:

1. Collection, processing and structuring of data – from various sources including scientific, business, legislative and social domains.
2. Development and training of models – using state-of-the-art machine learning methods and deep neural networks as well as advanced EuroHPC and Sree HPC computing facilities. Apart of existing GPT-NeoX open-source architecture for training from the scratch, two already mentioned openly accessible multilingual LLMs (Gemma and Pythia) will be trained with collected data. Additionally, the experiment with fine-tuning other openly accessible architectures, such as Llama 4 and Mistral Large 3 is planned. Since during the HR-GPT Beta training and evaluation the first version of our specific Croatian tokeniser yielded poorer results, we used the GPT-NeoX's native tokeniser. The experiments with the second version of the Croatian tokeniser are also planned thus contributing with a new solution.
3. The first level of evaluations of the LLMs will be conducted using FLORES-200 log perplexity (Costa-jussà et al., 2022). Perplexity is applicable to non-annotated text data, is predictive of performance of a number of

downstream tasks (Xia et al., 2023), and was used to measure LLMs quality in previous work (Lin et al., 2024). Grammatical linguistic knowledge will be evaluated using MultiBLiMP (Jumelet et al., 2025). The second level of evaluation will include benchmarks for LLMs' proficiency in knowledge completion in Croatian by utilising established datasets such as MMLU (Hendrycks et al., 2021), ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), Belebele (Bandarkar et al., 2023), and a dataset for seven NLP-tasks (Dac Lai et al., 2023). Since these benchmarks were used to evaluate previous HR-GPT Beta, their usage will assure the straight-forward comparison.

4. Integration and commercialisation – through the Orchestra digital platform that will enable the application of the model in a real business environment.

The expected result is a high-quality innovative product: a generative LLM HR-VojeM available through API, and adaptable for different domains and business needs.

3 Orchestra

Orchestra is a state-of-the-art TBMS developed by Ciklopea, designed specifically to meet the needs of modern language service providers (LSPs). The product was developed within the EU-funded project NPOO.C1.1.1.R6-I1.04-V1.0004 for enhancing the competitiveness of European SMEs. Orchestra follows the principle of Software as a Service (SaaS) and is built on the foundation of over 20 years of industry expertise

in translation and localisation. As a cloud-based SaaS platform, Orchestra removes the burden of costly infrastructure and on-going maintenance. Translation business can focus entirely on delivering quality translation services without worrying about server management or technical upkeep. Orchestra offers a robust suite of advanced features tailored to the specific needs of LSPs. From automated workflows that eliminate manual tasks, to business intelligence tools that provide critical insights into translators' performance, Orchestra empowers the team to work more efficiently. The platform also includes powerful document management capabilities, ensuring the content that is always organized, accessible, and secure.

Orchestra includes a built-in client portal that enhances transparency and communication with the clients. Real-time updates and easy access to the status of projects keep everyone on the same page. Orchestra also provides a comprehensive vendor portal designed to streamline communication and collaboration with translation vendors. With easy access to project details, task assignments, and real-time updates, the vendor portal helps improve the efficiency of vendor management. Whether the internal teams or external vendors are managed, Orchestra ensures that all stakeholders are aligned and that workflows run smoothly.

Orchestra's powerful Business Intelligence (BI) component is designed to generate detailed reports on key metrics like project performance, resource utilization, and feedback of client satisfaction.

Orchestra is fully compatible with industry standards such as ISO 9001, ISO 17100, ISO 18587, and ISO 27001, ensuring the highest quality, security, and consistency levels in your translation processes.

4 Integration

The HR-VojeM will be available through API and will be offered by Ciklopea as a commercial service. It will be integrated in the Orchestra TBMS as an additional service wherever a Croatian LLM could be applicable. One example of usage is the application of HR-VojeM in the process of translation (either by NMT and/or CAT) where LLM can serve for checking the quality of the output, i.e. accuracy and fluency.

HR-VojeM will also be fine-tuned for selected domains (e.g. health, law and administration, fi-

nances, etc.) and will be used to suggest corrections and avoid terminological inconsistencies in translations to Croatian for these domains. Parts of training data will be set aside for the fine-tuning and evaluation of LLMs in specific domains.

API integration will enable the extension of Orchestra services to new user segments (e.g. legal, financial or public institutions).

In addition, the result will include openly available versions of HR-VojeMs in the HR-CLARIN repository, which will make a significant contribution to the scientific community and open science.

5 Conclusions

We presented a project HRVOJE-M where a producing of a monolingual Croatian LLM (HR-VojeM), several additionally trained openly accessible LLMs and their integration in the TBMS Orchestra is planned. This project was already selected for funding within the EU-programme Competitiveness and Cohesion and we are waiting to sign the GA with the Government of the Republic of Croatia.

Acknowledgments

This research was funded by European Union: a) Next Generation EU programme, b) Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436) and c) Competitiveness and Cohesion programme, IRI S3 (Grant to be announced).

References

- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. *The belebele benchmark: a parallel reading comprehension dataset in 122 language variants*. [arXiv:2308.16884v2](https://arxiv.org/abs/2308.16884v2).
- Taylor A. Chang, Catherine Arnett, Zhouwen Tu, and Benjamin K. Bergen. 2026. Goldfish: Monolingual Language Models for 350 Languages. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference: LREC 2026*, European Language Resources Association (ELRA), pages 3750–3781. <http://www.lrec-conf.org/proceedings/lrec2026/pdf/2026.lrec2026-1.300.pdf>.

- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord. 2018. *Think you have solved question answering? try arc, the ai2 reasoning challenge*. [arXiv:1803.05457v1](https://arxiv.org/abs/1803.05457v1).
- Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. [arXiv:2207.04672](https://arxiv.org/abs/2207.04672). Version 3
- Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Association for Computational Linguistics, pages 13171–13189. <https://aclanthology.org/2023.findings-emnlp.878.pdf>
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. *Measuring massive multitask language understanding*. [arXiv:2009.03300v3](https://arxiv.org/abs/2009.03300v3) Version 3
- Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. *MultiBLiMP 1.0: A Massively Multilingual Benchmark of Linguistic Minimal Pairs*. [arXiv:2504.02768v4](https://arxiv.org/abs/2504.02768v4) Version 4
- Adar Kahana, Jaya Susan Mathew, Said Bleik, Jeremy Reynolds, and Oren Elisha. 2024. *Evaluation Methodology for Large Language Models for Multilingual Document Question and Answer: Computation and Language Repository*, [arXiv:2402.01065v1](https://arxiv.org/abs/2402.01065v1). Version 1
- Language Report Croatian (2022). European language equality. https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_7_Language_Report_Croatian_.pdf
- Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André Martins, and Hinrich Schütze. 2024. *MaLA-500: Massive language adaptation of large language models*, [arxiv.org:2401.13303v2](https://arxiv.org/abs/2401.13303v2) Version 2
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. *CamemBERT: a Tasty French Language Model*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pages 7203–7219. <https://aclanthology.org/2020.acl-main.645.pdf>
- Isabel Papadimitriou, Kezia Lopez, and Dan Jurafsky. 2023. *Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models*. In *Findings of the Association for Computational Linguistics: EACL 2023*, Association for Computational Linguistics, pages 1194–1200. <https://aclanthology.org/2023.findings-eacl.89.pdf>
- Pranaydeep Singh, Aaron Maladry, and Els Lefever. 2023. *Too Many Cooks Spoil the Model: Are Bilingual Models for Slovene Better than a Large Multilingual Model?*. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, Association for Computational Linguistics, pages 32–39. <https://aclanthology.org/2023.bsnlp-1.5.pdf>
- Vanja Štefanec, Daša Farkaš, Gaurish Thakkar, and Marko Tadić. 2024. Building a Large Language Model for Croatian. In *Proceedings of the Conference New Trends in Translation and Technology 2024*. Incoma Ltd., pages 204–209. https://doi.org/10.26615/issn.2815-4711.2024_017.
- Gaurish Thakkar, Vanja Štefanec, Daša Farkaš, and Marko Tadić. 2024. Building a Large Language Model for Moderately Resourced Language: A Case of Croatian. In *Proceedings of the 35th Central European Conference on Information and Intelligent Systems - CECIS 2024*. University of Zagreb, Faculty of Organization and Informatics, pages 225–229. https://ceciis.foi.hr/sites/default/files/Proceedings_CECIIS_2024.pdf.
- Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. *Multilingual is not enough: BERT for Finnish. Computation and Language Repository*, [arXiv:1912.07076v1](https://arxiv.org/abs/1912.07076v1). Version 1
- Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Veselin Stoyanov. 2023. Training trajectories of language models across scales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, pages 13711–13738. <https://aclanthology.org/2023.acl-long.767.pdf>

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pages 4791–4800. <https://aclanthology.org/P19-1472.pdf>

Author Index

- Al-Huthali, Najat B., 146
Al Sharou, Khetam, 157
Alsulaiman, Abied, 130
Alvarez Vidal, Sergi, 193
Alves, Diego, 54
- Bellemont, Florence, 46
Boltaeva, Gulrukh Erkin kizi, 203
Bouillon, Pierrette, 238
Brasoveanu, Mihai Cristian, 46
- Calvo Del Barrio, Sofía, 46
Calvo del Barrio, Sofía, 167
Cannavina, Valeria, 119
Chernigovskaya, Maria, 84
Cheung, Andrew K. F., 68
Chiusaroli, Francesca, 207
Corpas Pastor, Gloria, 138
- Dumančić Acar, Ana, 41
- Farrell, Michael, 1
Fra Paleo, Urbano, 207
- Gene, Viveta, 109, 119
Girletti, Sabrina, 238
- Hao, Xindi, 17
Hawkins, Ken, 207
HE, Sihui, 34, 93
- Jacot des Combes, Hélène, 207
Julaiti, Kaifusai, 68
- Kadirovna, Mehriniso Kilicheva, 203
Kamalova, Aziza Islomovna, 203
Kokkinidou, Anna, 46
- Lamsiyah, Salima, 245
Lee, Jinam, 63
Li, Ailin, 79
Li, Jingyi, 189
Li, Yanze, 198
Li, Zhenhai, 17
Liu, Kanglong, 98
Liyanapathirana, Jeevanthi Uthpala, 238
- Macken, Lieve, 98
Mirdjonovna, Shahlo Khamroyeva, 203
Mitkov, Ruslan, 245
Molés Cases, Teresa, 9
Monti, Johanna, 207
Murray, Virginia, 207
Mutal, Jonathan David, 238
- Nahhas, Abdulrahman, 84
Neumann, Ksenia, 84
Noriega, Laura, 138
Nuriddinova, Gulirano Baxtiyor qizi, 203
- Oliver Gonzalez, Antoni, 193
Orasan, Constantin, 146
Öztaş, Seyda, 193
- Picazo-Izquierdo, Alicia, 245
- Qurbonova, Nodira Roziqovna, 203
- Rebechi, Rozane, 54
Ryu, Jaehee, 63
- Sánchez Gijón, Pilar, 167
Sánchez-Gijón, Pilar, 46
Shahnazari, Masoud, 130
SHI, Jiawei, 34, 93
Sosoni, Vilelmini, 109
Staiano, Maria Carmen, 207
Stojak, Mladen, 254
- Tadić, Marko, 254
Theuerkauf, Lukas, 84
Tykhonov, Slava, 207
- Valdez, Susana, 46
Volkart, Lise, 238
- Walia, Damanpreet Singh, 84
Wang, Lulu, 98
Weng, Yu, 189, 198
- Xu, Han, 29
- Yang, Saini, 207
Yu, Janet Sohlhee, 63

Yujie, Huang, 29

Zhang, Hengtian, 98

Zhang, Shuyin, 17

Zou, Yuan, 146

Żrałka, Edyta, 231