Translating and the Computer 46

18-20 November 2024

European Convention Center, Luxembourg

Proceedings

ISBN 978-2-9701733-2-8

9 782970 173328

March 2025. Editions Tradulex, Geneva

This document is downloadable from www.tradulex.com and www.asling.org

Acknowledgements

AsLing wishes to thank and acknowledge the support of the sponsors of TC46:

Gold Sponsors

Silver Sponsors

Preface

The evolution of digital computing since the advent of the ENIAC in 1946 has been remarkable, paving the way for advancements in artificial intelligence (AI) that were scarcely conceivable a few decades ago. For over 46 years the Translating and the Computer (TC) conference series has consistently documented and analysed the impact of these technological developments on the field of translation and interpreting. After 41 sessions in London, UK, and 2 conferences on the web, the series has now moved to Luxembourg since 2022, where it benefits from the sponsorship of the Publications Office of the European Union

This year's TC46 conference proceedings showcase a diverse range of articles that address the ongoing concerns and interests of both academics and language professionals. The central themes explored include:

- Training and empowering future language professionals
- Driving language innovation
- Advances in AI for translation
- Practical applications of AI
- Driving innovation with quality
- AI-driven innovation in terminology management
- Machine translation (MT) and quality assurance
- Innovations in interpretation and localization
- Language data and machine translation at work
- Enhancing translation through technology

The articles in this volume offer both theoretical and practical perspectives on these themes, contributed by researchers and practitioners from various institutions and organizations, as well as from the industry.

On the academic side, contributors represent a wide range of institutions, like the School of Translation and Interpretation (University of Ottawa, Canada), International University of Languages and Media (IULM, Milan, Italy), UCLouvain (Belgium), Leiden University (UK), University of Turku (Finland), University of Alicante (Spain), University of Manchester (UK), Department of Translation and Interpreting of the University of Granada (Spain), Universidad Politécnica de Madrid (UPM, Spain), Ionian University (Greece), University of Málaga (Spain), University of Algiers 2 (Algeria), Faculty of Translation and Interpreting (FTI, University of Geneva, Switzerland), Centre for Translation Studies at University of Surrey (UK), University of Melbourne (Australia), School of Translation and Interpretation of University of Ottawa (Canada), University of La Rioja (Spain), and University of Osijek (Croatia). While not all presenters provided full papers to these proceedings, all contributed to the success of the conference.

TC46 participants also benefited from contributions by high ranking practitioners representing the Digital Technologies Research Centre (DT, National Research Council of Canada (NRC-CNRC) and European Commission's Directorate-General for Translation (DGT), Languages Service of the Division of Conference Management at the United Nations Office at Geneva (UNOG), Interpreting Solutions at Acolad, European Parliament’s translation service (DG TRAD), XTM International, International Telecommunication Union (ITU, Geneva, Switzerland), and Translation Centre for the Bodies of the European Union (Luxembourg), as well as various independent researchers, consultants or freelance translators or interpreters.

The diverse contributions to both the Conference and this volume reflect the evolving landscape of translation and interpreting in the digital age. As AI continues to develop, it is crucial to consider its implications for the future of language professionals and the articles in these proceedings provide valuable insights into the challenges and opportunities that lie ahead.

In the spirit of the inaugural TC conference in 1978, we are reminded of the ongoing dialogue between MT and AI. As we navigate the complexities of this dynamic field, it is essential to maintain a critical perspective on the capabilities and limitations of technology and the contributions of TC46 manage to do exactly that as they serve as a testament to the enduring importance of human expertise and ingenuity in the face of technological advances.

We wish to thank all those who devoted their time to prepare and present contributions at TC46, in particular our keynote and invited speakers who shed ample light on crucial areas of translation and technology: Maria Aretoulaki elucidated the role of human translators in the age of Generative AI, Christos Ellinides, the Director-General for the Directorate-General for Translation (DGT) at the European Commission, addressed the use of AI in public administration language services and Ildikó Horváth, the Director of the Translation Centre for the Bodies of the European Union, who explored the synergy between human expertise and AI.

Furthermore, we wish to acknowledge the valuable contribution of all those who submitted proposals and developed their proposals and presentations into full papers for the proceedings. We extend special thanks to the editors, whose dedication was instrumental in the publication's completion. We are also grateful to the practitioners, researchers and academics who accepted to join the Programme Committee for their meticulous review of submissions and for their indispensable contribution to the conference's success. Last but not least, we would like to express our sincere gratitude to the sponsors, whose support was crucial for the realization of TC46.

The TC46 Organizing Comittee

The Executive Committee of AsLing establishes several bodies each year, to organise and carry out the annual conference. Membership in these bodies overlap. The tables below show membership in these bodies for TC46.

Conference Organising Committee:
Denis Dechandon, European Union
Amal Haddad Haddad, University of Granada
Valentini Kalfadopoulou, Ionian University
Ruslan Mitkov, Lancaster University & University of Alicante
Maria Recort Ruiz, International Labour Office
Vilelmini Sosoni, Ionian University
Olaf-Michael Stefanov, United Nations (ret.)
Nelson Verástegui, International Telecommunications Union (ret.)
Coordinator: Denis Dechandon

Editors of the Proceedings:
João Esteves-Ferreira, Tradulex
Amal Haddad Haddad, University of Granada
Vilelmini Sosoni, Ionian University
Olaf-Michael Stefanov, United Nations (ret.)

Programme Committee:
Juan José Arevalillo, Hermes Traducciones
Lynne Bowker, University of Ottawa
Félix do Carmo, University of Surrey
Gloria Corpas Pastor, University of Malaga
Ayten Dersan, World Trade Organization
Jorge Diaz-Cintas, Centre for Translation Studies (CenTraS) & University College London,
Gökhan Doğru, Universitat Autònoma de Barcelona & Dublin City University
Joanna Drugan, Heriot-Watt University
María Fernandez Parra, Swansea University
Amal Haddad Haddad, University of Granada
Ramon Inglada Heriot-Watt University
Valentini Kalfadopoulou, Ionian University
Alina Karakanta, Leiden Centre for Linguistics
Laila Karlsson, International Telecommunication Union
Raquel Lazaro-Gutierrez, University of Alcala
Elizabeth Marshman, University of Ottawa
Johanna Monti, L'Orientale University of Naples
Joss Moorkens, Dublin City University/ADAPT Centre
Elena Murgolo, Localisation Technology Consultant Custom MT
Constantin Orăsan, University of Surrey
Michail Panagopoulos, Ionian University
Rozane Rebechi, University of Rio Grande do Sul
Vilelmini Sosoni, Ionian University
Tharindu Ranasinghe, University of Lancaster
Mark Shuttleworth, Hang Seng University
Leena Salmi, University of Turku
Paola Valli, Project Manager, Tamedia
Nelson Verástegui, International Telecommunications Union (ret.)
Michal Ziemski, World Intellectual Property Organization

Contents

AI-assisted translation-friendly writing:
A pilot study of potential and effects

Elizabeth Marshman

School of Translation and Interpretation, University of Ottawa/OLST

elizabeth.marshman@uottawa.ca

Hana Nessakh

School of Translation and Interpretation, University of Ottawa

hness095@uottawa.ca

Ting Liu 劉婷

School of Translation and Interpretation, University of Ottawa

tliu109@uottawa.ca

Ahmed Elhuseiny Bedeir

School of Translation and Interpretation, University of Ottawa

aelhu037@uottawa.ca

Haifa Ben Naji

School of Translation and Interpretation, University of Ottawa

hbenn070@uottawa.ca

Anwar Alfetlawi

School of Translation and Interpretation, University of Ottawa

aalfe022@uottawa.ca

Abstract

The negative effects of English as a lingua franca in scholarly communication are increasingly evident. One mitigating strategy is translating scholarly literature (e.g., journal abstracts) using machine translation (MT) or generative artificial intelligence (genAI) tools. However, common challenges in English writing may lead to translation problems and interfere with effective information transmission. In this project, we test the potential of the genAI tool ChatGPT by editing a sample of 10 scholarly abstracts to make them more "translation-friendly" and then translating them to Arabic and French. We analyse the number, types and impact of the English edits, as well as the number and types of translation errors observed. We conclude that ChatGPT may help users produce more translation-friendly abstracts, particularly given specific guidelines, but that the nature and impact of changes are highly variable. The quality of the translations was similarly unpredictable in the two languages. While ChatGPT editing had a positive effect overall for translations to French, the Arabic translations of these versions were often not as good as those of the originals. Thus, we advise using ChatGPT for translation-friendly writing with caution and only when the user can verify the end product, and recommend further investigation.

1 Introduction

The importance of sharing information world-wide through scholarly communication is clear. However, the dominance of English as a lingua franca in scholarship means that many researchers must read about and publish new discoveries in a second or foreign language (Bowker and Buitrago Ciro, 2019; Goulet et al., 2017; Lin, 2024; O'Brien et al., 2018). One alternative to English-only publication is translating at least key elements of research (e.g., abstracts) to make them more broadly accessible (Bowker and Buitrago Ciro, 2019; Fecher et al., 2023; Steigerwald et al., 2022). Given the volume of research and limited time and budgets, the use of machine translation (MT) and generative artificial intelligence (genAI) tools based on large language models (LLMs) is almost inevitable (Donlon and Tiernan, 2023; Macken, De Wilde, and Tezcan, 2024; O'Brien et al., 2018; Steigerwald et al., 2022). While these tools may be promising for boosting productivity and dissemination of research (e.g., Garrido-Merchan, 2023; Lin, 2024), the potential for translation errors in MT and genAI-translated text is widely recognised (Fecher et al., 2023; Lund et al., 2023; Macken, De Wilde, and Tezcan, 2024; O'Brien et al., 2018). One option for minimizing risk of errors is *translation-friendly writing*, using guidelines that aim to avoid potentially problematic lexical and syntactic choices (Bowker and Buitrago Ciro, 2019; Simonova and Patiniotaki, 2022; Steigerwald et al., 2022). Guidelines can be used by authors themselves, but with increasing access to tools such a ChatGPT, we may wonder whether they can be implemented (semi-)automatically.

In this project, we evaluate ChatGPT's implementation of translation-friendly writing guidelines with a minimalist (basic) and a more detailed (explicit) editing prompt. We then evaluate whether the result has affected readability, compliance with the guidelines, and meaning and form, as well as the number and type of translation errors in two target languages, Arabic and French.

We hypothesize that ChatGPT's editing will improve readability of the abstracts and reduce the frequency of translation errors, but that a human editor will still produce better results. Moreover, we hypothesize that including explicit translation-friendly writing guidelines in prompts will more effectively avoid translation errors.

2 Background and context

2.1 Scholarly communication

Scholarly communication is the process by which academics, researchers, and students share and publish their findings, ensuring the creation, evaluation, dissemination, and preservation of knowledge within the research community (Bowker and Buitrago Ciro, 2019: 7). While a central scientific language in some ways facilitates global collaboration, the use of English as a lingua franca is increasingly criticised because of the obstacles and inequalities it creates for non-English-speaking scholars and communities (O'Brien et al., 2018; Steigerwald et al., 2022). Steigerwald et al. (2022) suggest that translation can make science more inclusive and representative and that, though imperfect, it offers a solution for fostering a multilingual academic network.

2.2 Generative AI

Generative AI has been identified as a promising tool in scholarly communication (e.g., Fecher et al., 2023; Garrido-Merchan, 2023; Lin, 2024; Lund et al., 2023), particularly in the production and adaptation of written texts. However, much research is still needed to fully understand the potential of these tools, particularly in the humanities (Lozić and Štular, 2023; O'Brien et al., 2018).

These tools are certainly not without weaknesses and pitfalls. The literature (e.g., Arnold, 2003; Koehn, 2020; Koehn and Knowles, 2017) abounds with descriptions of challenges for MT, and many of these are also problematic for genAI tools (Lund et al., 2023): lexical, part-of-speech and structural ambiguities; long and complex sentences; idiomatic expressions and metaphors; under-specification of information in the source language; and rare or highly specialized words.

Despite their subject field expertise, neither the original author nor the end user of a translation may be able to identify where there are problems (especially with accuracy or completeness), even if they are aware that a text has been machine translated. Easily anticipated harms may occur, including the transmission of incorrect information and damage to a researcher's reputation and the dissemination of their work.[1]

2.3 Translation-friendly writing

Translation-friendly writing, while beneficial for human comprehension and translation (Bowker and Buitrago Ciro, 2019: 55–56; Steigerwald et al., 2022: 993), is often intended specifically to avoid linguistic phenomena known to challenge MT (and genAI tools), and thus to reduce errors and harms such as those described above.

Intervening in the source text (*pre-editing*) can improve output from MT or similar tools in a range of target languages, an efficient (and often more feasible) alternative to correcting the errors (*post-editing*) in each target language. In some use cases, controlled languages that impose strict limitations on vocabulary, syntax and sentence length may be used (Bowker and Buitrago Ciro, 2019; O'Brien et al., 2018). However, these are often challenging, time-consuming and expensive to implement. Guidelines that inform choices while offering more flexibility are more realistic for researchers.

Such guidelines, e.g., from Bowker and Buitrago Ciro (2019: 63–70; Appendix A) and Simonova and Patiniotaki (2022: 261–62), help authors to avoid known challenges: lexical difficulties can be decreased by avoiding polysemous words and idiomatic expressions; structural ambiguities can be reduced by avoiding long stacks of nouns with no indication of how they are related; problems in agreements and collocations arising from long sentences can be reduced by splitting sentences; and problems with under-specified information can be avoided by using nouns instead of pronouns and avoiding passive constructions. Non-language specialists with a good grasp of basic grammar and writing skills should often be able to implement these without much difficulty.

3 Methodology

The data collection and analysis for this project were carried out by the authors, specialists in translation technologies and related tools with at least some PhD-level education in Translation Studies and professional translation experience. Two members of the team had Arabic and

[1] Many other issues surrounding genAI tools and their potential impact in scholarly communication, research and education are important and relevant (Fecher et al., 2023; Lund et al., 2023; Schmidt and Meir, 2023; Steigerwald et al., 2022; Zielinski et al., 2023), but beyond the scope of this paper.

English as working languages; two had Arabic, English and French; one had Hakka, Mandarin and English; and one had English and French.

3.1 Languages of the project

We chose to work on English, a working language of all of the researchers, to explore tools for researchers who write in it as a second or foreign language. While much translation in scholarly communication is currently towards English, the potential to use MT to disseminate research in a wider variety of languages is increasingly gaining attention (e.g., Steigerwald et al., 2022). For the target languages of translation, we selected Arabic and French, as there were at least three members of the team with each of these as a working language who could contribute to data annotation.

3.2 Corpus design and building

For our corpus, we selected *The Journal of Specialised Translation* (https://www.jostrans.org/) from the field of Translation Studies (TS), a domain with which the authors were familiar. This journal includes articles in various sub-fields of TS, and is available in open access.

To design a representative sample, we began with readability scores as an indicator of complexity. We first surveyed the abstracts of all articles from issues 38 (July 2022) to 41 (January 2024), analysing them for length and for indicators of readability (e.g., Bailin and Grafstein, 2016) available in Microsoft Word: number of words per sentence, percentage of passive sentences, Flesch Reading Ease, and Flesch-Kinkade Grade Level. While limited in their predictive power and less reliable for short texts such as abstracts (Collins-Thompson, 2014), these measures are widely used, and were considered adequate for comparison here. We calculated the range for each indicator within this reference sample, and for our corpus, retained only abstracts with indicators in the middle quartiles of the reference scores.

To select our corpus, we began with abstracts from the second and fourth articles of each of the journal issues, working backwards from July 2022 (thus, before the widespread availability of ChatGPT). If an abstract was rejected because it fell outside of the allowed readability values, we moved on to the next article in that issue, and so on, until we obtained a sample of 10 abstracts that fell within the acceptable range (see the References).

3.3 Choice and use of the genAI tool

We chose the widely available genAI tool ChatGPT 4o for the project. On July 5, 2024, we generated separate queries for each task in the project (basic and explicit edits, translations of each version into each language). We deactivated the memory function of the system, to ensure that it would not carry over instructions or data from one query to another. The edited/translated abstracts were then saved to individual Word files for annotation.

3.4 Abstract editing and annotation

For editing, we used the translation-friendly writing guidelines provided by Bowker and Buitrago Ciro (2019: 63–70). In a first stage, a member of the research team with English as a third language (L1 Hakka), was provided with a list of the writing guidelines and access to the full text discussing these guidelines. She then edited the abstracts to make them more translation friendly (in Microsoft Word with track changes activated). In a second pass, she annotated her edits with comments indicating the guideline or other reason(s) that motivated each one.

We then submitted the original abstracts to ChatGPT. The first prompt, to produce the edited version referred to as *ChatGPT basic*, was "I am going to give you an abstract from a scholarly journal in English. Please revise this abstract to make it easier for you to translate it into another language." The second version, *ChatGPT explicit*, added "…, using the guidelines below:" followed by the translation-friendly writing guidelines as they appear in Appendix A.

Two members of the team annotated the abstracts separately, using two main sets of tags. The first, identifying the nature of the changes, included tags for each of the translation-friendly writing guidelines, plus others created *ad hoc* for additional changes (e.g., additions or omissions, word changes [e.g., substitution of synonyms], reordering, changes in punctuation, tense or number). The second tag set added a value judgment about the impact of the changes made on the abstract's content and form: improvement (i.e., changes reflecting the guidelines, or, in the case of other types of changes, considered to maintain meaning and improve language quality), degradation of the text (i.e., changes violating the guidelines, or considered to add, distort or omit meaning or degrade language quality), neutral changes, and changes that were ambivalent (i.e., that improved the text in one or more respects, but degraded it in one or more others). For example, we tagged as improvements divisions of long sentences into two shorter ones, and word substitutions that increased cohesion, but tagged as degrading cases of replacing an unambiguous word with an ambiguous one, and omitting elements of relevant meaning. Edits were segmented by the nature of changes, to provide the finest granularity possible for description. The two annotators then discussed and came to a consensus, and the finalized annotations were added to an Atlas.ti[2] project to facilitate analysis.

3.5 Abstract translation and annotation

A similar process was followed for the translations: translations of each of the abstract versions (original, human edit, ChatGPT basic and ChatGPT explicit) were generated using ChatGPT, using the prompt "Please translate the abstract I will give you from English to French/Arabic." Two evaluators who had the target language as their L1 independently identified and annotated errors in the translations using the MQM error typology (MQM Council, 2024). (At this stage, only error types were annotated.) Annotators were asked to annotate only errors that would need to be corrected for the abstracts to be published in a scholarly context similar to that of the journal, and that they would be able to defend to a client (e.g., the author). The annotations were then compared, and any discrepancies resolved by consensus. (In a few cases for French, a third evaluator with French as a working language independently evaluated the problematic sentences to resolve differences in annotation decisions.) The resulting annotations were then entered into an Atlas.ti project.

Once the annotations were completed, basic quantitative and qualitative analysis of the results was conducted to investigate the project hypotheses.

[2] https://atlasti.com

4 Results

Below, we summarize the results in terms of readability, changes made in the editing process, and translation errors observed. The Arabic and French translation results are also compared, and some possible links between readability, edits and translation quality explored. While space constraints prohibit the inclusion of detailed data in this paper, they are available here.

4.1 Readability evaluation

We began by comparing the readability of the abstracts, anticipating that the edits would result in increased readability. The expected trend was generally observed (Table 1), although variations were relatively small, and there were exceptions in results for individual texts.

Readability indicator[3]	**Original**	**Human**	**Basic**	**Explicit**
Words per sentence	24.9	16.4	21.1	13.9
Flesch Reading Ease	17.3	21.6	18.1	21.4
Flesch-Kinkaid Grade Level	16.9	14.3	15.9	13.8

Table 1. Average of readability indicators for English abstract versions

Our readability hypotheses appear to be largely borne out, suggesting that ChatGPT editing can achieve results similar to human editing in terms of these metrics when explicit instructions are included in the prompt. Nevertheless, a more focused look at the results is merited.

4.2 Results of the editing comparison

In the annotation, 43 edits on average were made in the ChatGPT basic version and 34 in ChatGPT explicit.[4] The human edits correspond more often to the translation-friendly writing guidelines (85% overall) than ChatGPT explicit (55% overall) and ChatGPT basic (44% overall).[5] A similar trend also held for 8 of the 10 documents individually. This suggests that ChatGPT's intervention is moderated and more focused when explicit instructions about the nature of desired changes is provided.

Given the considerable number of edits, it is important to also evaluate their effects on the quality of the product (Table 2).

Edit result	**Basic**	**Explicit**
Improved	35%	46%

[3] For Flesch Reading ease, higher scores indicate better (easier) readability, while for the Flesch-Kinkaid Grade Level, lower scores indicate better readability.

[4] Because of the difference in annotation approach, the numbers for the human abstract are not strictly comparable.

[5] Percentages may not sum to 100 because of rounding.

Ambivalent	7%	5%
Neutral	36%	26%
Degraded	23%	24%

Table 2. Qualitative results of ChatGPT edits

Approximately one-third to one-half of the edits were judged to improve the abstract (e.g., shorter sentences, unambiguous words, and less wordy constructions) and an additional one-quarter to one-third were neutral (e.g., substitution of words with synonyms, changes in order or punctuation), while another quarter in fact were found to degrade quality (e.g., omissions of relevant content). In a few cases, edits improved an abstract in one respect, but degraded it in another, and were considered ambivalent. For example, wordy constructions in the original abstracts were often made more compact in edited versions by converting them into noun stacks, which are avoided in translation-friendly writing.

For the edits specifically relating to the translation-friendly guidelines, the picture was similar. Overall, in the ChatGPT basic edits, about 69% of the changes were considered to align with the guidelines, and 31% to violate them. For the explicit edit, this shifted to a somewhat better 75%/25% distribution. However, for individual guidelines, there was more variation. In both basic and explicit versions, ChatGPT was relatively good at using the active voice (80% and 87% alignment with the guidelines, respectively) and short sentences (67% and 71%), and at reducing wordiness (75% and 93%); however, it appeared to struggle with avoiding noun stacks (47% and 9%), and with using nouns instead of pronouns (33% and 55%). In some cases, there appeared to be problems balancing multiple rules. For example, the phenomenon described above doubtless contributed to the considerable improvement in edits affecting wordiness (from 75% to 93%) coupled with a negative shift in the presence of noun stacks (from 47% to only 9%). The latter guideline is nevertheless one of only two where the explicit versions were worse than the basic, the other concerning the use of unambiguous words (50% and 69%, respectively).

While there was little difference between basic and explicit ChatGPT versions in the proportion of edits that degraded the abstracts, there was a noticeable shift in the proportion of neutral edits in the basic version towards changes that were considered to improve the explicit version. This suggests that giving ChatGPT explicit instructions for translation-friendly writing may be preferable to more generic prompting not only for readability, but also in terms of the number and nature of changes. However, the prevalence of problematic changes in both ChatGPT edits remains cause for concern, encouraging us to investigate the impact of the changes on the translations.

4.3 Analysis of the translations

Below, we discuss the findings of the translation analysis and compare them to our hypotheses. To facilitate comparison, we discuss the results using the number of errors per 1,000 words in the English abstracts.

Arabic analysis: For the Arabic translations, the results were almost the opposite of our hypotheses. Translations of the human-edited abstracts contained more errors than those of the original versions in 6 of the 10 cases and for the set overall. The translations of the explicit ChatGPT edits contained more errors than those of the basic ChatGPT versions in 5 of the 10 cases and overall, but contained fewer errors than the translations of the human edits in 9 of the 10 cases.

As shown below in Table 3, many of the errors observed (over 44%) were in Terminology. Interestingly, translations of ChatGPT versions showed fewer such errors compared to those of the original and the human-edited versions. For Language conventions, about one quarter of the annotated errors, the results diverged substantially from expectations: translations of the ChatGPT basic version had fewest errors, and those of the explicit version just slightly more than the others. In the category of Accuracy, also just over one quarter of the errors annotated, the translations of the human edits had most errors, followed by those of ChatGPT basic, the original, and finally ChatGPT explicit. (Other categories showed so few occurrences that useful comparisons were not possible.)

These results suggest that ChatGPT editing with explicit instructions might help to reduce problems with Accuracy (likely to be problematic for dissemination of knowledge) in Arabic, but that the benefits may not be very noticeable compared to the original. Left to its own devices, ChatGPT did appear to have a positive effect at the level of linguistic conventions for translation to Arabic in this case, but unfortunately this effect did not appear to persist with more explicit prompts. This may suggest that ChatGPT, left undirected, is helpful for linguistic polishing, but less so for more specific and content-related tasks.

In light of the unexpected results, we investigated some of the possible causes. In a qualitative analysis, ChatGPT seemed to have problems with some terms in the original abstracts (e.g., *subtitling, interpreters*). When abstracts were human-edited, sometimes such problematic terms were repeated for clarity, resulting in a higher number of terminology errors, and also appearing to increase inconsistent terminology. The differences in grammatical structures between English and Arabic were also found to influence the findings. For example, in Arabic, the equivalents of “or” and “and” in Arabic are repeated before each element in a list, unlike in English. In reducing repetition, the human-edited version sometimes favoured the production of translations that violated this textual convention. Splitting and shortening complex sentences also provided the tool with less intra-sentential context, and appeared to contribute to errors such as subject-verb disagreement in gender, especially with acronyms (e.g., *LSP*).

French analysis: In contrast to the findings for Arabic, the results for French support our hypothesis that human-edited texts would be the most translation-friendly. The translations of the human-edited abstracts had the fewest errors overall, followed by the ChatGPT explicit versions, then the basic, and finally the originals. Notably, a difference of over 25 fewer errors per 1,000 words in 4 of the 10 translations of the human-edited abstracts (compared to those of the original versions) appeared to account for the decrease in the average number of errors, while in 4 other texts the number of errors in fact increased. This demonstrates notable variability in the potential benefits of human editing.

For French, Accuracy errors, accounting for just over a third of the errors annotated, also followed the expected pattern: they were numerous in the translations of the original versions and lowest in the translations of human edits, with the translations of edits by ChatGPT falling between the two (although with the basic version slightly better than the explicit) (Table 3). For this language, Style was the next most prevalent error type in the original version's translations, with values staying relatively consistent across versions; findings were similar for Language conventions except for a small increase in the ChatGPT basic versions' translations. Finally, Terminology errors accounted for a much smaller percentage in the French translations than in the Arabic (just over 10%), but interestingly were relatively stable except for another small increase, this time in the translations of the human edits. As in Arabic, some problems with terminological consistency were noted.

Thus, for French, our hypotheses were supported: ChatGPT did appear to be useful for reducing some translation errors, although not as much as a human edit, and explicit instructions did appear to improve its performance. Effects were nevertheless modest, reducing the number of errors overall by just over 8 per 1,000 words.

Comparative analysis: In the translations of the human and ChatGPT explicit versions, the numbers of errors in the two languages were not dramatically dissimilar; however, the original abstract and ChatGPT Basic translations contained notably more annotated errors in French than in Arabic (Table 3).

In both languages, most errors fell into the categories of Terminology, Accuracy, and Language conventions. However, for the originals, Terminology appeared to be consistently more problematic in Arabic than in French, whereas in French, most errors were concentrated in the area of Accuracy.

Terminological problems remained relatively stable across versions, but with a considerable difference between languages. In this respect, the human edit was unexpectedly the most problematic for translation to French and also contained more errors than the ChatGPT versions in Arabic. ChatGPT's performance for terminology can thus be considered to be comparatively good, although it was somewhat prone to inconsistency.

Abstract	**Original**		**Human**		**Basic**		**Explicit**	
	Arabic	*French*	*Arabic*	*French*	*Arabic*	*French*	*Arabic*	*French*
Terminology	20.4	5.5	19.4	8.2	17.5	5.0	18.7	4.8
Accuracy	9.9	23.7	14.1	11.2	11.9	18.1	9.0	18.7
Language conventions	12.1	12.7	12.9	12.9	5.6	15.0	13.1	11.8
Style	0.6	13.8	1.8	11.7	0.6	11.9	3.5	12.4
Overall[6]	43.1	55.8	48.1	45.2	35.6	50.6	44.3	47.7

[6] Categories with only a few errors identified have been excluded to simplify the presentation.

Table 3. Total errors per 1,000 words in Arabic and French translations of abstract versions, per category

Accuracy, likely the most concerning error type for scholarly communication, again showed considerable interlinguistic differences, but in favour of Arabic. Among these translations, the ChatGPT explicit edit was the best, and the human edit most problematic. French showed almost the opposite tendency, with the human edit the best and the ChatGPT explicit edit second only to the original in the prevalence of problems. The explicit ChatGPT prompting option did consistently improve on the accuracy of translations compared to the original, but the variability in performance for both human and basic ChatGPT editing was striking.

In sum, ChatGPT editing for translation-friendliness does not appear to pose problems for Terminology, and may in fact help to avoid some errors, although the best approach for prompting in this respect remains unclear. The choices made in pre-editing (or not pre-editing) the abstracts did not appear to have a predictable effect on errors related to Language conventions. ChatGPT editing may be one option for reducing the prevalence of Accuracy errors in some cases. Given the variability observed, however, more investigation is necessary, particularly in regard to performance across target languages.

5 Discussion and implications

For increasing readability, the use of ChatGPT does appear to be moderately promising, particularly when explicit instructions are provided. However, there is certainly no guarantee that ChatGPT will follow instructions consistently, or that these will be the only modifications made. Each modification introduces risk (e.g., of unintended changes in meaning), and a considerable proportion of changes annotated were considered to degrade abstract quality. From the perspective of the English text, therefore, ChatGPT appears to be a tool that may be used with caution, by authors who are able to at least review the changes to ensure that the meaning has been appropriately preserved (cf. Fecher et al., 2023; Lin, 2024). (This may be even more important given the discussion surrounding responsibility for content generated by genAI (Zielinski et al., 2023).) The comparison between the ChatGPT basic and explicit edits from a qualitative perspective, while it suggests that the explicit versions may reflect translation-friendly writing guidelines slightly better overall, does not indicate a strong, predictable performance difference. ChatGPT may be more successful at implementing certain kinds of changes than others. More purely formal changes (e.g., shorter sentences) may be more successfully implemented than others (e.g., avoiding ambiguous words and noun stacks). Of course, it must be kept in mind that human edits may not always be entirely effective and predictable in their effects, especially for non-language specialist researchers working in a second or foreign language.

In scholarly writing, terminology is central, but may also be one of the areas where errors are less problematic, as subject-field experts with knowledge of the target language may be able to identify and correct terms in their fields more easily than other types of errors. ChatGPT-edited versions of the abstracts did not appear to be particularly prone to terminological translation errors, as compared to the originals. In the qualitative annotation of Arabic, multiple occurrences of incorrect terms may have played a role in the high number of errors noted; future research could explore the differences in error prevalence when unique errors (as opposed to error occurrences) are counted.

The potential effect of editing on accuracy is concerning, particularly given the high risks this kind of error entails. More investigation, particularly of the severity and potential impact of Accuracy errors, is necessary. This will also advance the analysis of Style errors that were observed mainly in French. In contrast to Accuracy errors, while numerous, these may not be considered to be particularly important for the dissemination of scholarship, although they may have indirect effects on perceptions of the research and researcher, as noted, e.g., by Bowker and Buitrago Ciro (2019) and Steigerwald et al. (2022).

More investigation of interlinguistic differences in ChatGPT's performance in the translation of scholarly work (original or edited) is clearly merited. Overall differences in translation quality may be linked to several well-known factors, including the availability of high-quality and appropriate training data (e.g., Koehn, 2020; Koehn and Knowles, 2017) and differences in language structures and conventions (Arnold 2003). Interlinguistic differences in the translation of edited abstracts could also potentially be linked to the fact that the translation-friendly writing rules used were proposed by researchers who typically work with Romance languages. However, they were based on observations of various sources (Bowker and Buitrago Ciro, 2019: 62), and similar guidelines were found to be very useful for Russian translation (Simonova and Patiniotaki, 2022). This thus bears more in-depth investigation.

Limitations: Of course, we must acknowledge that this project reflects a restricted sample of abstracts in a single subject field, edited and evaluated by a small group of individuals. Much more study is required to determine whether the trends observed are generalized. In addition, human evaluation of phenomena such as translation errors is notoriously difficult and prone to subjectivity and disagreement between annotators (e.g., Al Sharou and Specia, 2022). In this project, the classification of intralingual edits posed even more challenges for inter-annotator agreement. Despite our efforts towards rigorous annotation methods, it is possible that inter-annotator differences may have affected the findings. In future work, one option would be to have the members of the team who annotated the French independently annotate the Arabic as well, for comparison. Nevertheless, we believe that while imperfect, this research can contribute to the overall picture of ChatGPT's potential usefulness for translation-friendly writing, and can inspire further investigation.

6 Concluding remarks

While our initial hypotheses that ChatGPT could be helpful for implementing translation-friendly writing strategies to facilitate multilingual scholarly communication were partly supported—at least for translations to French—the considerable variability in the findings demonstrates that this tool is far from a panacea for achieving reliable translations. Without appropriate verification of the output, the results may not always be helpful, and could even be damaging. Much more investigation is needed to better analyse the strengths and weaknesses of this approach and whether (and how) it can be made more useful and reliable in future.

Future work: Planned work includes more in-depth quantitative and qualitative analysis of the editing changes and translation errors, as well as the links between them. Additional human edits, as well as additional target languages, may be added to expand the comparison. The analysis of translation error severity is also planned.

Acknowledgements

The authors wish to thank the Journal of Specialised Translation for making the corpus texts available for analysis, and the anonymous reviewers for their helpful feedback.

References

Al Sharou, Khetam, 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*, pages 171–80.

Arnold, Doug. 2003. Why Translation Is Difficult for Computers. In Harold Somers, editor, *Computers and Translation: A Translator's Guide*. John Benjamins, Amsterdam/Philadelphia, pages 119–142.

Bailin, Alan, and Ann Grafstein. 2016. Readability Formulas. In Alan Bailin and Ann Grafstein, editors, *Readability: Text and Context*. Palgrave Macmillan UK, London, pages 10–64.

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

Collins-Thompson, Kevyn. 2014. Computational Assessment of Text Readability: A Survey of Current and Future Research. *ITL. Instituut Voor Toegepaste Linguistik* 165(2): 97–135.

Donlon, Enda, and Peter Tiernan. 2023. Chatbots and Citations: An Experiment in Academic Writing with Generative AI. *Irish Journal of Technology Enhanced Learning* 17(2): 75–87.

Fecher, Benedikt, Marcel Hebing, Melissa Laufer, Jörg Pohle, and Fabian Sofsky. 2023. Friend or Foe? Exploring the Implications of Large Language Models on the Science System. *AI & Society*.

Garrido-Merchan, Eduardo C. 2023. Best Uses of ChatGPT and Generative AI for Computer Science Research. arXiv.

Goulet, Marie-Josée, Michel Simard, Carla Parra Escartín, and Sharon O'Brien. 2017. La traduction automatique comme outil d'aide à la rédaction scientifique en anglais langue seconde : résultats d'une étude exploratoire sur la qualité linguistique. *ASp. la revue du GERAS*, 72: 5–28.

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

Koehn, Philipp, and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Kruger, Jan-Louis, Marc Orlando, Pam Peters, Chloe Liao, and Helen Sturgess. 2022. Assessing the Impact of Readability on Translation Quality and Productivity. North Ryde, NSW: Macquarie University.

Lin, Zhicheng. 2024. Techniques for Supercharging Academic Writing with Generative AI. *Nature Biomedical Engineering*.

Lozić, Edisa, and Benjamin Štular. 2023. Fluent but Not Factual: A Comparative Analysis of ChatGPT and Other AI Chatbots' Proficiency and Originality in Scientific Writing for Humanities. *Future Internet* 15(10): 336.

Lund, Brady D., Ting Wang, Nishith Reddy Mannuru, Bing Nie, Somipam Shimray, and Ziang Wang. 2023. ChatGPT and a New Academic Reality: Artificial Intelligence-written Research Papers and the Ethics of the Large Language Models in Scholarly Publishing. *Journal of the American Society for Information Science and Technology*.

Macken, Lieve, Vanessa De Wilde, and Arda Tezcan. 2024. Machine Translation for Open Scholarly Communication: Examining the Relationship between Translation Quality and Reading Effort. *Information* 15(8): 427.

MQM Council. 2024. The MQM Error Typology. MQM (Multidimensional Quality Metrics). https://themqm.org/error-types-2/typology/. [last accessed October 5, 2024].

O'Brien, Sharon, Michel Simard, and Marie-Josée Goulet. 2018. Machine Translation and Self-Post-Editing for Academic Writing Support: Quality Explorations. In Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, editors, *Translation Quality Assessment: From Principles to Practice*. Springer International Publishing, Cham, pages 237–262.

Schmidt, Paul G., and Amnon J. Meir. 2023. Using Generative AI for Literature Searches and Scholarly Writing: Is the Integrity of the Scientific Discourse in Jeopardy? arXiv.

Simonova, Valentina, and Emmanouela Patiniotaki. 2022. Pre-Editing for the Translation of Life-Science Texts from Russian into English via Google Translate. In *Proceedings of the New Trends in Translation and Technology 2022*, pages 259–265.

Steigerwald, Emma, Valeria Ramírez-Castañeda, Débora Y. C. Brandt, András Báldi, Julie Teresa Shapiro, Lynne Bowker, and Rebecca D. Tarvin. 2022. Overcoming Language Barriers in Academia: Machine Translation Tools and a Vision for a Multilingual Future. *Bioscience* 72(10): 988–98.

Zielinski, Chris, Margaret A. Winker, Rakesh Aggarwal, Lorraine E. Ferris, Markus Heinemann, Jr Lapeña, Sanjay A. Pai, et al. 2023. Chatbots, Generative AI, and Scholarly Manuscripts: WAME Recommendations on Chatbots and Generative Artificial Intelligence in Relation to Scholarly Publications. *Colombia Medica (Cali, Colombia)* 54(3): e1015868.

Corpus texts

Chen, Ya-mei. 2022. An activity theory perspective on the TEP model replicated in translation crowdsourcing: A case study of Global Voices Lingua. *Journal of Specialised Translation*, 38: 320-347.

Cui, Ying, and Jie Li. 2021. Imagery packed with poetic qualities: A case study on English-Chinese translation of Apple’s advertisements. *Journal of Specialised Translation*, 36a: 79-98.

de los Reyes Lozano, Julio. 2020. Straight from the horse’s mouth: children’s reception of dubbed animated films in Spain. *Journal of Specialised Translation*, 33: 233-258.

Jeanmaire, Guillaume, and Jeong-yeon Kim. 2022. Traduire des jeux vidéo thérapeutiques pour enfants : entre localisation et vulgarisation médicale; Translating therapeutic video games for children: between localisation and medical vulgarisation. *Journal of Specialised Translation*, 38: 232-253.

Ragni, Valentina. 2020. More than meets the eye: an eye-tracking study of the effects of translation on the processing and memorisation of reversed subtitles. *Journal of Specialised Translation*, 33: 99-128.

Robert, Isabelle S., Amaury De Meulder, and Iris Schrijver. 2021. Live subtitling for access to education: A pilot study of university students’ reception of intralingual live subtitles. *Journal of Specialised Translation*, 36: 53-78.

Ruiz Rosendo, Lucía. 2022. Interpreting for the military: Creating communities of practice. *Journal of Specialised Translation*, 37: 16-34.

Tatar Anđelić, Jasmina. 2022. Traduire les médias dans une communauté de pratique virtuelle : Expérience du portail francophone le Courrier des Balkans / Translating media in a virtual community of practice: The experience of the French-language news portal Le Courrier des Balkans. *Journal of Specialised Translation*, 37: 55-74.

Tenbrink, Thora, and Kate Lawrence. 2021. ’Omnibus’: A cross-modal experience between translation and adaptation. Journal of Specialised Translation, 35: 186-208.

Wang, Xiangyu, and Xiangdong Li. 2020. The market’s expectations of interpreters in China: A content analysis of job ads for in-house interpreters. *Journal of Specialised Translation*, 34: 118-149.

Appendix A: Translation-friendly writing guidelines (Bowker and Buitrago Ciro, 2019: 63–70)

1. Use short sentences.
2. Use the active voice rather than the passive voice.
3. Avoid long noun strings or modifier stacks.
4. Use relative pronouns such as “that” and “which”.
5. Avoid wordiness.
6. Use nouns instead of personal pronouns.
7. Use terminology consistently.
8. Choose unambiguous words.
9. Avoid abbreviated forms.
10. Avoid idiomatic expressions, humor, and cultural references.

Survey on the use of generative artificial intelligence by professional translators

Michael Farrell

IULM University, Milan, Italy

michael.farrell@iulm.it

Abstract

This paper presents the findings of an anonymous online survey conducted in early 2024 on the use of generative artificial intelligence (GenAI) among professional translators. The survey revealed that 29.4% of professional translators incorporate GenAI into their workflow, in line with the results of another recent study. There is a significant association between the use of machine translation (MT) and GenAI, with MT users more likely to also use GenAI. Translators primarily use GenAI for writing-related tasks, such as finding contextual meanings, rephrasing sentences, shortening, summarizing and simplifying, and finding metaphors, synonyms and definitions. This suggests that GenAI enhances translation quality rather than productivity. Only 28.8% of GenAI users use it more than 50% of the time, implying that it is just one of several tools. ChatGPT is the most popular GenAI system, used by 80.8% of GenAI users, followed by Microsoft Copilot at 29.6%. However, only 20% of GenAI users pay for premium services. Many professional translators do not use GenAI (70.6%), often due to strong negative attitudes. GenAI's role as an alternative to traditional MT followed by post-editing is less common than might be expected.

1 Introduction

The first Generative Pre-trained Transformer (GPT) model was launched in 2018 (Radford et al., 2018). However, it was not until November 2022, with the release of GPT 3.5, that generative artificial intelligence (GenAI) garnered widescale public attention, reaching a staggering more than 1.7 billion users a year later (DeVon, 2023). Moreover, a report by Eloundou et al. (2023) noted that interpreting and translating are among the professions most exposed to AI in the US job market. Given this, it was virtually inevitable that stakeholders in the translation profession would start looking into ways of using this new natural language processing tool to their benefit.

This paper presents the results of an anonymous online survey designed to gain insight into the proportion of professional translators who currently use GenAI during their work and the various ways they do so. At least two other surveys have already sought to measure the use of GenAI among translators: the 2024 annual European Language Industry Survey, published by ELIS Research (2024), and the survey on generative AI conducted by the Society of Authors (2024). Additionally, Tavares et. al. (2023) conducted a survey that assessed the awareness and knowledge of both machine translation (MT) and GenAI among Language Service Providers in Portugal, in which they asked language professionals about the usefulness of GenAI in their work in general terms. However, to this author's knowledge, there have been no surveys designed to obtain details of precisely how professional translators choose to integrate GenAI into their workflow from among the whole host of options available to them. This paper intends to fill that gap.

2 Methods

The online survey, hosted by EUSurvey,[1] was written in English due to its international scope and anonymized to minimize social desirability bias (Larson, 2019), which occurs when respondents provide answers they believe to be more socially acceptable or desirable than their actual beliefs or behaviour. The questions were based on various uses mentioned in blogs, ezines and websites (Goldsmith, 2023; Nader, 2023, to name just the most systematically organized sources) and built into some computer-aided translation (CAT) tools[2].

The various uses that emerged from a review of the literature allowed closed-ended survey questions, with the advantage of simplifying result analysis and making the survey quicker to complete. However, since it is very hard to predict every possible way such a flexible tool as GenAI might be used, an additional *Other (please specify)* option was also included.

The survey link was emailed to 96 professional translators' associations on 15 February 2024, almost all of which were members of the International Federation of Translators. It was also shared via social media (Facebook, LinkedIn, X and ProZ.com). In a previous survey conducted in 2022 on the use of MT by professional translators (Farrell, 2022), responses from translators contacted through professional associations were initially kept separate from responses received from those who found the link on social media, based on the assumption that social media users might be more tech-savvy and therefore more likely to use technology like MT. However, no such difference was observed. For this reason, no attempt was made in this survey to distinguish between the two kinds of respondents.

At the beginning of the survey, all respondents were given the following definition and asked to confirm that they clearly understood the distinction between GenAI and MT:

> *Generative AI (GenAI) systems like ChatGPT and Gemini (formerly Bard) differ from conventional machine translation (MT) engines such as Google Translate and DeepL in various ways. One important distinction is that, although GenAI systems can be used to translate text between languages like conventional MT engines, they are primarily designed to perform tasks such as answering questions, writing texts or simulating conversations. Unless otherwise specified, all the questions in this survey refer to GenAI systems and not to conventional MT engines. For this reason, it is essential to understand what is meant by GenAI system in this survey and why this does not include conventional MT engines.*

All but one of the variables measured in the survey are non-numeric, non-parametric, categorical variables which can only take on a limited number of values. The only continuous numerical variable — proportion of working time during which GenAI is used — was analysed in bands of values and therefore transformed into a categorical variable.

The widely used chi-square ($\chi 2$) test was chosen for the statistical analysis to determine whether respondents are more likely to use MT in their workflows (MT users) if they work with higher-resource languages, where MT output quality is generally considered better. It was also used to assess whether MT users are more likely to incorporate GenAI at some stage in their translation workflow (GenAI users). The significance level was set to .05, as is standard, to ensure a 95% confidence level. The chi-square test was performed using an online calculator provided by Stangroom (2018). The results are reported in the format required by the American

[1] https://ec.europa.eu/eusurvey/

[2] Wordscope Translator's Assistant (https://pro.wordscope.com) and RWS AI Professional plugin for Trados Studio (https://appstore.rws.com/Plugin/200)

Psychological Association (APA): χ2 (degrees of freedom, N = sample size) = chi-square statistic value, p = p value.

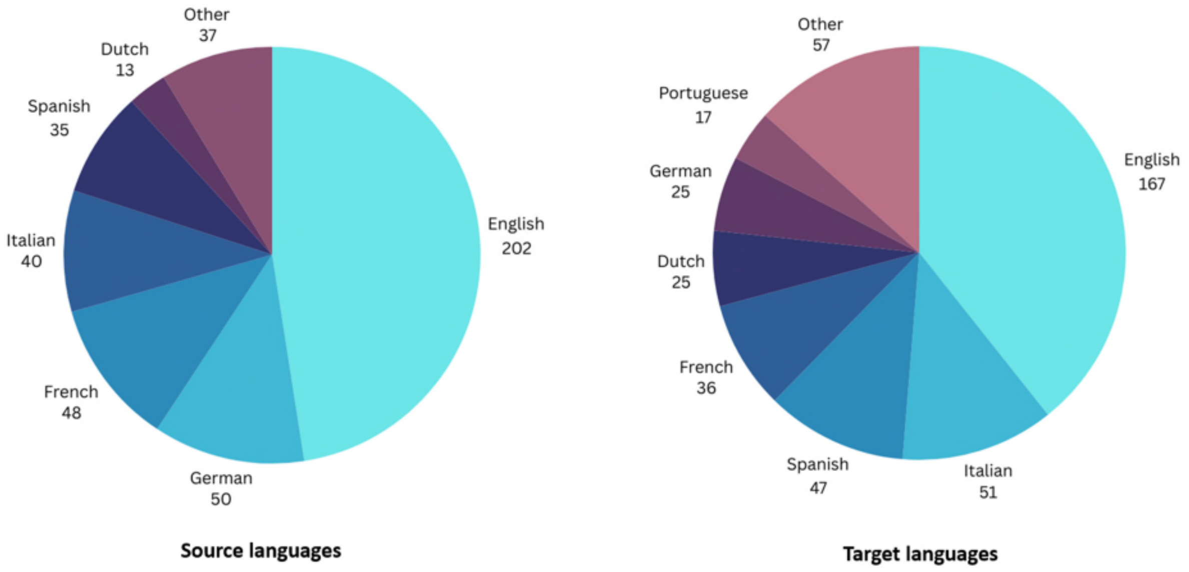
The Digital Language Equality Metric (technological factors) was used as a measure of language resource richness (Gaspari, 2022). The responses were iteratively divided into two categories according to the respondent's main source language. Initially, the first category consisted of respondents working with the most resource-rich language, while the second category included all the others. Then, the first category was expanded to include respondents working with the top two most resource-rich languages, with the second category comprising the remaining respondents. This process continued, with the first category progressively including respondents working with the top N most resource-rich languages, and so on. At each stage, a two-by-two contingency table was drawn up and the chi-square (χ2) test was used to determine if there was a statistical difference between the way the two sets responded. The purpose was to find a threshold value after which the two categories consistently responded differently, disregarding any isolated instances where they temporarily differed and then reverted to their previous behaviour in the next iteration. This procedure was repeated for the main target language.

3 Results

3.1 Survey population

The survey was originally scheduled to close on 20 March, but the deadline was extended to 31 March 2024 in order to exceed 385 responses, the calculated number for achieving a 95% confidence level with a large population, assuming the sample is truly random (Stangroom, 2018). Out of the 96 professional associations contacted, seven confirmed that they had shared the survey link with their members, although others may have done so without replying to the invitation. Survey responses were received from 437 people. A total of 12 were disqualified: 4 because they answered that they were *not* professional translators and 8 because they stated that they did not clearly understand the difference between conventional MT and GenAI. The remaining 425 responses were analysed.

3.2 Translation languages



Source languages

Target languages

Chart 1. Main source and target languages

3.2.1 Main source language

The main source languages are shown in Chart 1. The other main source languages mentioned were Finnish (7), Japanese (5), Portuguese (4), Chinese (2), Czech (2), Danish (2), Greek (2), Norwegian (2), Romanian (2), Swedish (2), Catalan (1), Estonian (1), Polish (1), Russian (1), Serbian (1), Thai (1) and Vietnamese (1).

3.2.2 Main target language

The main target languages are shown in Chart 1. The other main target languages reported were Greek (12), Norwegian (11), Finnish (10), Polish (4), Romanian (3), Russian (3), Swedish (3), Arabic (1), Bulgarian (1), Catalan (1), Chinese (1), Croatian (1), Czech (1), Galician (1), Hindi (1), Serbian (1), Turkish (1) and Ukrainian (1).

3.2.3 Conventional MT

The number of respondents who use conventional MT in their translation workflow is shown in Table 1.

	n.	%
Yes	312	73.4
No	113	26.6

Table 1. Conventional MT

3.2.4 Difference in MT use between high and low resource languages

Not all the languages reported by survey respondents are rated on the European Language Grid Dashboard. However, the languages that are included account for 98% of the overall source language data and 99% of the overall target language data gathered in this survey.

Using the previously described method, it was determined that there was no statistically significant association between the respondents' propensity to use MT in their workflows and resource richness of their working languages.

3.3 Use of GenAI

A total of 29.4% (n=125, "Yes, always" + "Yes, but not always") of respondents reported using GenAI at some point in their translation workflow (GenAI users), as shown in Table 2.

	n.	%
Yes, always	12	2.8
Yes, but not always	113	26.6
No	300	70.6

Table 2. Use of GenAI

As shown in Table 3, there was a statistically significant association between the MT users and those who use GAI (χ2 (2, N = 425) = 31.35, p = < .00001).

The likelihood that professional translators use GenAI was also found *not* to be associated with the resource-richness of their working languages. The respondents (n=300) who indicated that they never use GenAI at any point in their translation workflow provided the reasons in Table 4. Multiple answers were allowed.

	GenAI yes	**GenAI no**
MT yes	115	197
MT no	10	103

Table 3. Contingency table between MT and GenAI users

	n.	%
Because it may lead to a loss of human creativity and critical thinking skills	164	54.7
I have never tried to integrate it into my workflow	151	50.3
Because of GDPR/privacy issues	137	45.7
Because it harms the quality of the final translation	112	37.3
Because the kinds of texts I translate do not lend themselves to GenAI	106	35.3
Because it may lead to job displacement and economic hardship for some people	95	31.7
Because it is unprofessional	76	25.3
I have experimented with it but do not find it useful	70	23.3
On account of environmental concerns	68	22.7
Because my employer/client(s) specifically ask(s) me not to use it	54	18.0
Other reason (please specify)	36	12.0

Table 4. Reasons for not using GAI

Nineteen quite diverse additional reasons were given that did not fit well into the fixed categories. The most frequently mentioned were an absence of perceived benefit over existing MT technology (5), translation being faster without it (4), various ethical and moral concerns (3), inability to see any potential use for translators (3), the use of client-provided tools that do not incorporate GenAI (2) and risks associated with translating safety-critical texts (2).

Two unexpected responses, each mentioned once, were the difficulty of using GenAI when translating hard copy documents and never having heard of GenAI before. One highly detailed 520-word response listed a wide variety of disruptive or potentially disruptive factors, including AI, ageism and Brexit.

The majority of translators who use GenAI (n=125) do not use it all the time (see Table 5).

A total of 71.2% of GenAI users use it less than 50% of the time, with an overall average usage of 32.6% of the time. The translators who do not use it all the time (n=113) were also asked to specify the situations in which they chose not to use it. Multiple answers were allowed.

	n.	%
Less than 10% of the time	27	21.6
10 to 19% of the time	25	20.0
20 to 29% of the time	19	15.2
30 to 39% of the time	11	8.8
40 to 49% of the time	7	5.6
50 to 59% of the time	13	10.4
60 to 69% of the time	4	3.2
70 to 79% of the time	3	2.4
80 to 89% of the time	3	2.4
90 to 99% of the time	1	0.8
100% of the time	12	9.6

Table 5. Frequency of GenAI use

I do not use GenAI when:	**n.**	**%**
I do not think it would be useful	94	83.2
There are GDPR/privacy issues	73	64.6
My employer/client(s) specifically ask(s) me not to use it	60	53.1
Other reason (please specify)	8	7.1

Table 6. Reasons for not always using GenAI

Other reasons given included difficulties using GenAI with text that is broken up by complex formatting and its unavailability in some places.

3.4 GenAI systems

The respondents (n=125) listed the GenAI systems they use (Table 7). Multiple answers were allowed.

The other systems specified were clients' or agencies' own systems (3), Mistral AI, an open-source large language model (1), CoTranslator AI, which links with ChatGPT (1), MemoQ AGT (marketed as a proprietary AI system) (3), AI-driven search assistants like Perplexity AI (2) and You.com (1), Grammarly, which is an AI-driven writing tool (1), and DeepL Translator (1) and Globalese AI (1), which are not GenAI systems.

	n.	%
ChatGPT	101	80.8
Microsoft Copilot (Bing Chat)	37	29.6
ChatGPT Plus	19	15.2
Gemini (formerly Bard)	19	15.2
Other GenAI systems (please specify)	10	8.0
Microsoft Copilot Pro	6	4.8

Table 7. GenAI systems used

3.5 How professional translators use GenAI

Table 8 shows the uses reported by the 125 GenAI users, in order of popularity, starting with the most frequent.

	Total users	**Often**	**Occasionally**	**Rarely**	**No longer**	**Never tried**
Finding the meaning of words or terms in specific contexts	106 (84.8%)	37 (29.6%)	49 (39.2%)	20 (16.0%)	2 (a) (1.6%)	17 (13.6%)
Rephrasing sentences	94 (75.2%)	36 (28.8%)	37 (29.6%)	21 (16.8%)	7 (b) (5.6%)	24 (19.2%)
Finding context-specific translations of words or expressions	93 (74.4%)	31 (24.8%)	44 (35.2%)	18 (14.4%)	14 (a) (11.2%)	18 (14.4%)
Searching for synonyms	91 (72.8%)	32 (25.6%)	41 (32.8%)	18 (14.4%)	7 (a) (5.6%)	27 (21.6%)
Looking up the definitions of words or terms	91 (72.8%)	23 (18.4%)	44 (35.2%)	24 (19.2%)	4 (a) (3.2%)	30 (24.0%)
Finding words or terms from their definitions	85 (68.0%)	16 (12.8%)	38 (30.4%)	31 (24.8%)	2 (a) (1.6%)	38 (30.4%)
Shortening or summarizing texts	76 (60.8%)	18 (14.4%)	36 (28.8%)	22 (17.6%)	2 (b) (1.6%)	47 (37.6%)
Simplifying texts	71 (56.8%)	13 (10.4%)	32 (25.6%)	26 (20.8%)	3 (b) (2.4%)	51 (40.8%)
Finding collocations and common word groupings	66 (52.8%)	19 (15.2%)	31 (24.8%)	16 (12.8%)	13 (a) (10.4%)	46 (36.8%)
Translation of sentences, paragraphs or entire texts with a specific style or tone	61 (48.8%)	7 (5.6%)	26 (20.8%)	28 (22.4%)	10 (b) (8.0%)	54 (43.2%)
Searching for metaphors	55 (44.0%)	7 (5.6%)	25 (20.0%)	23 (18.4%)	2 (a) (1.6%)	68 (54.4%)
Proofreading, correcting typos and grammar	32 (25.6%)	11 (8.8%)	9 (7.2%)	12 (9.6%)	10 (b) (8.0%)	83 (66.4%)

Translation of entire text for subsequent post-editing (PEMT)	26 (20.8%)	3 (2.4%)	13 (10.4%)	10 (8.0%)	13 (b) (10.4%)	86 (68.8%)
Revision of human translation	26 (20.8%)	1 (0.8%)	17 (13.6%)	8 (6.4%)	7 (b) (5.6%)	92 (73.6%)
Automated PEMT	24 (19.2%)	2 (1.6%)	12 (9.6%)	10 (8.0%)	10 (b) (8.0%)	91 (72.8%)
Identifying typos in the source text	20 (16.0%)	3 (2.4%)	8 (6.4%)	9 (7.2%)	6 (a) (4.8%)	99 (79.2%)
Avoiding gender bias	16 (12.8%)	2 (1.6%)	10 (8.0%)	4 (3.2%)	7 (b) (5.6%)	102 (81.6%)
Raw MT output quality estimation	15 (12.0%)	3 (2.4%)	5 (4.0%)	7 (5.6%)	3 (a) (2.4%)	107 (85.6%)

Table 8. How professional translators use GenAI

The first column of Table 8 shows the total number of GenAI users that use it for the listed task and is equal to the sum of the second (*often*), third (*occasionally*) and fourth (*rarely*) columns. The precise wording of the option in the fifth column (*no longer*) was different according to the task. For those marked (a), the wording was “I have tried, but more conventional tools are better. So, I never use GenAI this way now.” For those marked (b), it read “I have tried, but I was not satisfied with the results. So, I never use GenAI this way now.”

3.5.1 Other uses of GenAI in the translation workflow

The most commonly mentioned other uses were brainstorming for alternatives or inspiration (8), terminology mining (5), understanding poorly written, highly technical or complex source texts (3), and researching concepts or background information to better understand the context (3). All other uses were mentioned only once each, and included tasks like search engine optimization, harmonizing the style of source texts written by multiple authors, using it as a search tool, displaying images of machinery components or architectural styles, improving writing in a second language (e.g., emails to clients), checking consistency of figures between source and target, fact-checking the source text, writing regex for CAT tools, searching for idioms based on definitions and writing macros to automate parts of the translation process.

4 How GenAI is accessed

Professional translators access GenAI as shown in Table 9 (n=125). Multiple answers were allowed.

The three additional ways of accessing GenAI mentioned were opening a browser window (2), and as an external tool or activating with an MT tool in order to get suggestions (1).

	n.	**%**
I write my own prompts (instructions/questions)	117	93.6
GenAI functions built into CAT tools[3]	37	29.6
As described below	3	2.4

Table 9. GenAI access

	n.	**%**
Trados Studio via plugin	4	36.4
memoQ AGT	3	27.3

[3] This does not include using GenAI in a browser window as an external tool.

Only 11 of the 37 respondents who said they used built-in GenAI functions in their CAT tool specified which tool they used. Unfortunately, due to a flaw in the survey design, this question was not mandatory. Multiple answers were allowed.

SmartCAT	2	18.2
Wordscope	2	18.2
CafeTran Espresso	1	9.1
Other (CotranslatorAI)	1	9.1
Other (Wordfast)	1	9.1

Table 10. GenAI via CAT tools

No version of Wordfast has built-in GenAI functions. Therefore, the respondent either misunderstood the question or has found an undocumented way of setting up a GenAI system as one of Wordfast's built-in MT engines via an API.

5 Transparency

	n.	**%**
Never	100	80.0
Sometimes	13	10.4
Always	12	9.6

Table 11. Clients or employer informed

Eighty percent of GenAI users do not inform their clients or employer that they use GenAI.

Eleven of the 125 respondents specified that they tell their clients or employer that they use GenAI when the client or employer asks if they use it (5), when the client or employer asks them to use GenAI (1), for technical translations (1), when the client already knows because the translator works in-house (1), when the text will be published (1), at the beginning of the working relationship (1), and when the client knows already (1). One respondent said they have only told one client so far, and another has told only one particularly concerned client that they only use GenAI for research purposes.

6 Training

Thirty-seven of the 125 GenAI users (29.6%) have received training on GenAI, as shown in Table 12. The 37 respondents were allowed to give multiple answers.

The three respondents who selected *Other* specified that they had been shown how to use GenAI in a different context with a focus on prompt usage; received training on confidentiality and ethical issues, building translation glossaries, prompt engineering, and using GenAI for audio transcription and subtitles; and learned how marketers and SEO experts use GenAI.

7 Discussion

The proportion of respondents who use conventional MT systems at some point in their translation workflow (73.4%) is higher than the 69.54% reported in the 2022 survey (Farrell, 2022), in line with the expectation that MT is becoming more widely used. It is also similar to the 76.87% recorded in the 2024 survey conducted by ELIS Research, which also reported an increase in MT usage among independent professionals over recent years.

	n.	**%**
On GenAI in general	27	73.0
Specifically on how translators may use GenAI	22	59.5
Specifically on how language professionals may use GenAI	11	29.7
Specifically on how interpreters may use GenAI	3	8.1
Other (please specify)	3	8.1

Table 12. GenAI training

Professional translators might be expected to be more likely to use MT in their workflows if they work with higher-resource languages, for which the quality of MT output is normally considered better. However, while the 2022 survey identified a resource-richness threshold below which professional translators were less likely to accept PEMT assignments, no such threshold was found for the use of MT simply at some point in the workflow. This absence of threshold was confirmed again in this latest survey.

The proportion of survey respondents who use GenAI during their translation work (29.4%) is virtually identical to the 29% reported in the 2024 European Language Industry Survey (ELIS Research, 2024). In January 2024, the Society of Authors (SoA), the UK's largest writers' union, reported an even higher figure of 37% for translators. However, this figure is based on responses from only 78 people (just under 10% of the total survey population) who self-identified as translators[4].

As might be expected, there was a strong association between professional translators who reported using MT and those who use GenAI (χ^2 (2, N = 425) = 31.35, p = < .00001). Moreover, the likelihood of using GenAI was also found to be independent of the resource-richness of the translator's main working languages.

Regarding the 70.6% of respondents who do not use GenAI at all, it was evident — especially from the comments left under *other* — that some professional translators have strong negative feelings towards it. Similarly, in a survey conducted in Portugal by Tavares et al. (2023), negative perceptions of GenAI also emerged as prominent.

On average, GenAI users use it 32.6% of the time, with only 28.8% using it more than 50% of the time. This suggests that most of these professional translators view GenAI as just one of several tools available to them.

The most widely used system is ChatGPT (80.8% of GenAI users), followed by Microsoft Copilot, which trails significantly behind at 29.6%. Only 20% of GenAI users reported paying for the systems they used (such as ChatGPT Plus or Microsoft Copilot Pro). The fact that two translators mentioned DeepL Translator and Globalese AI as *other* GenAI systems shows that a few respondents did not have the distinction between *traditional* MT and GenAI entirely clear. DeepL Translator is a widely used neural MT (NMT) engine and Globalese AI is also an NMT engine, which can be used to build custom MT systems.

Many of the uses of GenAI mentioned in this paper may at first seem more useful to writers than to translators, such as rephrasing sentences, searching for synonyms, shortening, summarizing, simplifying and finding metaphors. However, if we consider translation as rewriting a text in another language, this is not at all surprising. It is also immediately apparent that, in several cases, professional translators simply use GenAI instead of other existing tools, like thesauruses, spellcheckers, monolingual, bilingual and reverse dictionaries, and concordancers.

However, there is a risk in using GenAI indiscriminately to look up information like word definitions since it is not an information retrieval system but rather a system that generates new content based on patterns and the data it has been trained on. This can sometimes result in it providing incorrect information, a well-known phenomenon called hallucination (Xu, 2024). Other problems arise in the case of rare or unusual terms, which may not be present in the training data. For instance, if you ask ChatGPT 4 for the definition of the antiquated term *discrutator*, it will repeatedly claim in separate chats that the user has probably mistyped the word[5].

Using GenAI to generate MT output for post-editing ranks much lower on the list (a little over 20% of GenAI users) than might be expected. It should be kept in mind, however, that

[4] Unpublished data courtesy of the SoA.

[5] Discrutator, n. A person who disputes or doubts to an extreme or excessive degree; a caviller. Oxford English Dictionary Word of the Day on 28 December 2017. Tested four consecutive times in separate ChatGPT 4 chats on 19 April 2024.

GenAI was not originally designed for text translation but rather for autonomously generating new content. Nevertheless, both GPT and other GenAI systems can perform tasks they were not specifically trained for, known as emergent abilities, with MT being one of them. Some evidence suggests that the quality of their translation output is inferior to that of some existing NMT engines, at least for certain languages and kinds of text (Ding, 2024; Farrell, 2023; Jiao et al., 2023; Xiang, 2024). However, one important difference between GenAI systems and traditional MT engines is the ability to use prompts to assign a persona or provide a brief with the aim of improving translation quality and applying the appropriate style or tone — this is the highest-ranking explicitly translation-related use among the GenAI users in this survey. In this regard, He (2024) found that only the translator persona offered any advantage over a basic prompt, while Gao et al. (2023) observed that providing domain-specific information enabled GenAI to outperform traditional MT engines. Gao et al. also noted similar improvements with few-shot prompting (Brown et al., 2020; Chen et al., 2021), where ChatGPT was provided with up to five translation examples.

There is some overlap between "translation of sentences, paragraphs or entire texts with a specific style or tone" and "translation of entire text for subsequent PEMT". However, GenAI users who use it to translate the entire text with a specific style or tone for subsequent PEMT will presumably have indicated both uses.

The use of GenAI to perform other professional tasks, like proofreading, revision and post-editing also ranks near the bottom of the list (less than 26% of GenAI users in each case). At the very bottom are specialist tasks like avoiding gender bias and MT quality evaluation.

Eight respondents used the *other* option to mention using GenAI for brainstorming for alternatives or *inspiration* and five, for terminology mining. These uses should be included as selectable options in any future editions of this survey to obtain more precise figures for these activities, also bearing in mind that their absence from the list has inevitably caused a bias against them.

The most frequent uses of GenAI among professional translators seem more focussed on improving the quality of their work than increasing productivity. This stands in stark contrast to the purpose of PEMT, whose main aim is to produce translations more quickly, thereby reducing costs. Nevertheless, professional translators also use MT in other ways besides for PEMT (Farrell, 2022), which are not dissimilar to the ways GenAI is reported to be used in this survey.

The vast majority of GenAI users write their own prompts (93.6%) and 29.6% of them have received some form of training to do so, despite GenAI being a relatively new field. Additionally, 29.6% of them use or also use GenAI functions built into CAT tools.

As was observed for MT in the 2022 survey, a large majority of GenAI users (80%) consider it to be just another tool that their clients do *not need* to know about.

8 Conclusion

While the adoption of GenAI by professional translators might seem rapid at 29.4%, this is still less than half the proportion of conventional MT users, which stands at 73.4%.

Eight of the top nine uses are more closely related to the task of writing than to pure translation, with generating MT output for post-editing ranking only thirteenth on the list. This suggests that the professional translators who use GenAI primarily use it to enhance the quality of their work rather than to boost productivity.

Funding

This research has received no funding.

References

All hyperlinks last accessed 11 September 2024.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems. https://doi.org/10.48550/arXiv.2005.14165

Chen, Yiming, Yan Zhang, Chen Zhang, Grandee Lee, Ran Cheng and Haizhou Li. 2021. Revisiting self-training for few-shot learning of language model. https://doi.org/10.48550/arXiv.2110.01256

DeVon, Cheyenne. 2023. On ChatGPT's one-year anniversary, it has more than 1.7 billion users—here's what it may do next. CNBC. https://www.cnbc.com/2023/11/30/chatgpts-one-year-anniversary-how-the-viral-ai-chatbot-has-changed.html

Ding, Lijie. 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, Vol. 14, No. 9, pp. 2823-2833, September 2024. https://doi.org/10.17507/tpls.1409.18

ELIS Research. 2024. European Language Industry Survey 2024. https://elis-survey.org/wp-content/uploads/2024/03/ELIS-2024-Report.pdf

Eloundou, Tyna, Sam Manning, Pamela Mishkin and Daniel Rock. 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. https://doi.org/10.48550/arXiv.2303.10130

Farrell, Michael. 2022. Do translators use machine translation and if so, how? Results of a survey held among professional translators. In Proceedings of the 44th Conference Translating and the Computer, pages 49–60. https://asling.org/tc44/wp-content/uploads/TC44-luxembourg2022.pdf#page=49

Farrell, Michael. 2023. Preliminary evaluation of ChatGPT as a machine translation engine and as an automatic post-editor of raw machine translation output from other machine translation engines. International Conference HiT-IT 2023, 108-113. https://doi.org/10.26615/issn.2683-0078.2023_009

Gao, Yuan, Ruili Wang and Feng Hou. 2023. How to Design Translation Prompts for ChatGPT: An Empirical Study, School of Mathematical and Computational Science, Massey University, New Zealand. https://doi.org/10.48550/arXiv.2304.02182

Gaspari, Federico, Owen Gallagher, Georg Rehm, Maria Giagkou, Stelios Piperidis, Jane Dunne and Andy Way. 2022. Introducing the Digital Language Equality Metric: Technological Factors. In LREC 2022 Workshop Language Resources and Evaluation Conference, page 1. http://www.lrec-conf.org/proceedings/lrec2022/workshops/TDLE/pdf/2022.tdle-1.1.pdf

Goldsmith, Josh. 2023. Eight Ways to Power Up Your Vocabulary with ChatGPT. The Slator Tool Box, October 2023. https://mailchi.mp/slator/tool-box-october-2023#mctoc6

He, Sui. 2024. Prompting ChatGPT for Translation: A Comparative Analysis of Translation Brief and Persona Prompts. Proceedings of the 25th Annual Conference of the European Association for Machine Translation (EAMT) (Volume 1). Sheffield, UK 2024. Publisher: European Association for Machine Translation, pp. 316-326. https://aclanthology.org/2024.eamt-1.27/

Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi and Zhaopeng Tu. 2023. Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. Tencent AI Lab. http://dx.doi.org/10.48550/arXiv.2301.08745

Larson, Ronald B. 2019. Controlling Social Desirability Bias. International Journal of Market Research. 2019, Volume 61, Issue 5 https://doi.org/10.1177/147078531880530

Nader, Antonios. 2023. Revolutionize Your Translations: 4 ChatGPT Prompts for Perfect Results. https://www.linkedin.com/pulse/revolutionize-your-translations-4-chatgpt-prompts-perfect-nader

Radford, Alec, Narasimhan, Karthik, Salimans, Tim, and Sutskever, Ilya. 2018. Improving language understanding by generative pre-training. Technical report. OpenAI. https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

Society of Authors. 2024. SoA survey reveals a third of translators and quarter of illustrators losing work to AI. https://www2.societyofauthors.org/2024/04/11/soa-survey-reveals-a-third-of-translators-and-quarter-of-illustrators-losing-work-to-ai/

Stangroom, Jeremy. 2018. Chi-Square Test Calculator. Social Science Statistics. https://www.socscistatistics.com/tests/chisquare2/default2.aspx

Tavares, Célia, Luciana Oliveira, Pedro Duarte and Manuel Moreira da Silva. 2023. Artificial Intelligence: A Blessing or a Threat for Language Service Providers in Portugal. Informatics 2023, 10, 81. https://doi.org/10.3390/informatics10040081

Xiang, Cailing. 2024. Study on the Effectiveness of ChatGPTin Translating Forestry Sci-tech Texts. International Journal of Linguistics, Literature and Translation ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print) http://dx.doi.org/10.32996/ijllt.2024.7.9.11

Xu, Ziwei, Sanjay Jain, Mohan Kankanhalli. 2024. Hallucination is Inevitable: An Innate Limitation of Large Language Models. https://doi.org/10.48550/arXiv.2401.11817

Editing synthetic text from generative artificial intelligence: two exploratory case studies

Michael Farrell

IULM University, Milan, Italy

michael.farrell@iulm.it

Abstract

As the use of generative artificial intelligence (GenAI) becomes more mainstream, an increasing number of authors may turn to this technology to write directly in a second language, bypassing traditional translation methods. Consequently, professional editors may have to develop new skills: shifting from correcting translation and non-native errors to editing AI-assisted texts. This study includes several stages: participant selection, text planning, prompt engineering, text generation and text editing. The recruited authors provided prompts for GPT-4 to generate texts, edited the output as they desired and then passed them on to professional editors for a final edit. All participants reported their experiences and described the nature of their interactions. The findings reveal that, while GenAI significantly improved the grammatical accuracy of the non-native English texts, it also introduced anomalies. In conclusion, although AI was useful in these two cases, it did not fully replace the human editors, and professional translators — with their language skills — may like to consider offering this additional service. The study also suggests that both authors and editors should be trained in synthetic-text editing to fully harness the benefits of AI-assisted writing, and that further research should be conducted with diverse texts and authors to generalize the findings.

1 Introduction

With the advent of generative artificial intelligence (GenAI), an increasing number of authors may be tempted to bypass traditional translation and craft their texts directly in a second language with the aid of GenAI prompts. This practice might be termed AI-assisted second-language authoring. Some authors have always preferred to write in their second language and many professional editors already make a career out of correcting these texts.

The consequent shift presents new challenges and opportunities for professional editors, who will need to transition from correcting translation and non-native errors to editing synthetic texts (STs) generated by GenAI, hereon in referred to as synthetic-text editing (STE). While machine translation (MT) output can also be considered a form of ST since it is created artificially, it is useful to limit the term ST to output generated by systems based on large language models (LLMs) (Farrell, 2024). In contrast, traditional MT output is created by AI systems trained using parallel corpora, such as Google Translate or DeepL Translate.

If the envisaged scenario becomes reality, there may be a slight decrease in traditional translation work and an increase in demand for synthetic-text editors. Professional translators, bilingual post-editors and author's editors, with their language skills, could be ideally positioned to offer this new service.

2 Aim and limitations

The experiment aims to explore the feasibility of using GenAI as a tool to allow authors to write directly in a second language, bypassing traditional translation methods. If the results of this limited experiment are promising, it should ideally be repeated with different kinds of text on a wider variety of subjects in different languages by authors from diverse backgrounds.

3 Method

Figure 1 shows the method schematically. Refer to the sections for details.

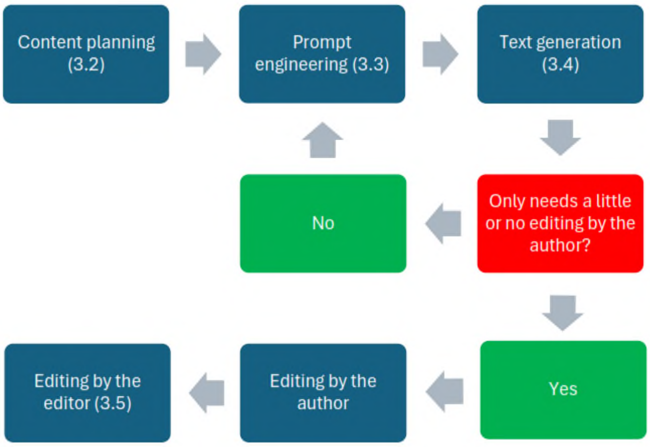


Figure 1. Experimental working method

3.1 Recruitment

A call for participants was published on the internet and distributed through social media channels (LinkedIn and Facebook). The Mediterranean Editors & Translators[1] association was also asked to share the call with its members since they belong to two of the three potentially affected professions.

The applicant authors were asked about their experience, their native language and other languages they knew, the subject areas they would like to write about for the experiment, whether the text they would write would be real or a simulation, and to provide any other information they considered important.

The candidate editors were asked about their experience as English-language editors, particularly with non-native authors, the languages they were proficient in besides English, their preferred subject areas, and any experience they had of post-editing (MTPE), STE, translation and human translation revision, as well as any other information they considered relevant.

The authors were also asked to provide a sample of at least 100 words they had previously written in English without the aid of AI, MT, computer tools (except for dictionaries) or other people, on the same or a similar subject as the text they intended to write. This text served two purposes: firstly, so that the editor could gauge the author's knowledge of English and, secondly, as a sample that could be used during GenAI prompt engineering.

3.2 Content planning

The authors were asked to provide the precise subject or a provisional title for the text they planned to write in English using GPT-4's web interface (500 to 2000 words). They were also asked about their usual approach to planning a text of this kind, which computer tools they usually used to write in English, whether their text would have a particular structure (such as an introduction, discussion, etc.), and if their text needed to comply with a style guide or specific writing conventions.

[1] www.metmeetings.org

3.3 Prompt engineering

A prompt engineering technique based on the automatic prompt engineer (APE) (Zhou et al., 2022) was used. Essentially, GPT-4 was asked to *reverse engineer* its own prompt from the sample text the author had provided.

Firstly, GPT-4 was instructed to correct the English of the sample text. Then, without starting a new chat, it was asked to summarize the sample text as a list of short notes. Lastly, again in the same chat, GPT-4 was asked to write four different prompts, in order from best to worst, which — together with the notes — would cause it to generate the corrected sample text.

The short notes and four prompts were then sent to the author as a model on which to base a prompt which could be used to generate the text they wanted to write. The authors were told that, if they found it easier, they could write their prompt and/or notes in their native language (or any other language), and even mix languages.

If the author wanted to organize their text into sections, they were instructed to divide the notes into the same sections with headings, and if they had to follow a style guide or specific conventions, this too had to be added to the prompt.

3.4 Text generation

The researcher checked the prompt provided by the author for completeness. He then fed it to GPT-4, took the generated text and sent it back to the author, together with the actual prompt used to generate it. The author was also sent a feedback questionnaire asking their opinion of the output and how they wished to proceed. They could choose to edit the prompt to see if better results could be obtained, including by breaking the task down into steps and using prompt chaining techniques (Wu et al., 2021), or they could use the GenAI output as a base for the text they had in mind. The authors were allowed to make as many edits to the AI-generated text as they felt were required to achieve the desired result, including rewriting, deleting or adding entire paragraphs. All edits were marked using *Track Changes* in Microsoft Word.

Once the researcher received the author's final draft, he asked them to give examples of edits they had made and explain why they were necessary. The researcher then sent the file to the editor with all edits hidden (*Accept All Changes*) so that they could not tell which parts the author had edited.

If the authors required, the researcher and editors were willing to sign nondisclosure agreements, but they were warned that it is not advisable to share sensitive or unpublished data on online platforms.

3.5 Text editing

After giving their initial impressions, the editors were asked to do the work they would normally do, while noting the changes they made on an *errors and textual anomalies* form. This questionnaire suggested several error categories, including second-language authoring errors (Corder, 1975; James, 1998), MT errors (Popović, 2018) and some commonly reported hallmarks of GenAI (Dondoni Braz, 2024; Dou, 2022; Gillham, 2024; Gluska, 2023; OpenAI, 2022). The editor and author were asked to interact as they normally would during the editing process, without any interference from the researcher. Once editing was complete and the author approved the final text, the researcher sent the authors and editors final feedback forms.

4 Results

Three authors and three editors answered the call. Each author was paired with an editor based on the subject matter and type of article the author intended to write: Author 1 (A1) with Editor 1 (E1), Author 2 (A2) with Editor 2 (E2), etc. Each pair was considered a separate case study. The third case study did not reach conclusion due to participant dropout.

4.1 Participants

A1 and A2 know each other: A1 suggested that A2 answer the call for participants. However, they did not consult each other during the experiment. None of the other participants had met before.

4.1.1 Author 1

A1 is in his twenties and writes in Italian (his native language) for a local newspaper while studying History of Art at university. He has no professional experience of writing in English, which he knows to B2 level (Council of Europe, 2001). However, he does write in English for his studies and can also speak Spanish. He chose History of Art as his preferred subject area and decided to write an article solely for the purpose of this experiment. He normally writes newspaper articles and academic papers/research reports in his native language for professional or study reasons.

4.1.2 Editor 1

E1 is in her thirties and has over fifteen years of experience as an English-language editor. In addition to her native English, she speaks Dutch. For this experiment, she was willing to edit texts on any subject. She has extensive experience of correcting non-native English and some experience with MTPE. She mentioned having done STE and revised human translations and has considerable experience as a translator. Her typical editing work includes blog posts, business plans or reports, non-fiction books, marketing materials and web copy.

4.1.3 Author 2

A2 writes in her native Italian for a different local newspaper than A1 while studying Art and Literature at the same university. She is in her twenties and has no experience of composing articles, academic papers or other short texts in English, except for blog posts, which she has been writing both professionally and for fun for a year. English is her only second language, and she knows it to B2 level (Council of Europe, 2001). Her preferred subjects are art, literature, cinema and poetry. She initially considered using the article after the experiment but ultimately decided against it. She normally writes newspaper articles, academic papers/research reports, novels, short stories, poetry, and scripts for films, television and theatre in Italian for professional or academic purposes.

4.1.4 Editor 2

E2 is in her thirties with three years' experience as an English-language editor. She normally edits newspaper articles, academic papers, research reports, blog posts, technical manuals, scripts for films, television or theatre, business plans or reports, and essays. Besides her native English, she is fluent in Italian. For this experiment, she was willing to edit texts on professional development, health, career transitions, fashion and AI. She has a lot of experience in correcting non-native English and no experience in MTPE. She has also done a fair amount of STE, translation and human translation revision. She is currently a full-time content writer in Italy and has worked as a teacher of English as a second language for many years.

4.2 Content planning

4.2.1 Author 1

Since his editor was not an expert in the History of Art, A1 planned a blog post suitable for laypeople. He chose the title *How to Guide People to Look at a Work of Art*. He typically plans such texts through brainstorming and by researching relevant sources. While writing in English, he usually uses DeepL Translate[2], WordReference.com[3], PONS[4] and Google Translate[5].

4.2.2 Author 2

A2 chose the title *The Universal Language of Art* for her piece, which she classified as a newspaper article/blog post/short essay. Regarding her approach to planning such texts, she said that she just writes them and fixes them at the end, stating, "I plan only interviews." She normally consults WordReference.com[6], the Oxford English Dictionary[7] and Merriam-Webster[8] while writing in English.

4.3 Prompt engineering

Neither author chose to organize their text into sections or specified a style guide or writing conventions. None of the participants were asked to sign nondisclosure agreements.

4.3.1 Author 1

A1 said that the structure of the prompt he was asked to write was as he expected and found the process laborious but not overly difficult. He added that it was a helpful way to clarify his thoughts before writing. The notes he provided to GPT-4 were mostly in English and partly in Italian.

4.3.2 Author 2

A2 said that the structure of the prompt was as she expected, and quick and easy to write. The notes she provided to GPT-4 were entirely in English.

4.4 Text generation

Neither author chose to edit the prompt and try again. A2 opted to keep the raw GenAI output exactly as it came, while A1 decided to make some changes.

4.4.1 Author 1

A1 noted that the generated text was better than he expected, awarding it a score of eight out of ten. He observed that there was no content that was not implicit in the prompt and identified no serious errors. He commented that the raw output resembled something a human might write and was surprised by its accuracy, describing it as "a good base, especially for the lexicon," since only a few things needed editing. One change he made was to replace an example provided by GPT-4 (a painting by Leonardo da Vinci) with one he considered more appropriate (a fresco by Michelangelo Buonarroti). Although A1 stated in the feedback form that there was no missing content, he added a whole sentence which he defined as "the main message of the article" and another to help the reader "understand the fact that art is something close to each of us."

[2] www.deepl.com
[3] www.wordreference.com
[4] www.pons.com
[5] https://translate.google.com
[6] www.wordreference.com
[7] www.oed.com
[8] www.merriam-webster.com

4.4.2 Author 2

A2 noted that the generated text was better than she expected and also gave it a score of eight out of ten. She observed that there was no content that was not implicit in the prompt and there was nothing missing. Moreover, she commented that the GenAI output looked like something a human could write and found no serious errors.

4.5 Text editing

4.5.1 First impressions

4.5.1.1 Editor 1

E1 said that the text she received was very much better than the sample of her author's English written without the aid of AI or other tools. She classified it as an academic paper/research report or short essay, gave it a score of eight out of ten, and said it showed no serious issues. However, she added, "on first reading it seems quite high level/vague." Although the subject matter was different from what she usually edits, she felt comfortable working with it.

4.5.1.2 Editor 2

E2 noted that the sample of her author's English written without the aid of tools displayed a high level of creativity. However, it contained spelling and grammar mistakes typical of native Italian speakers, which were absent in her GPT-4-generated text. She gave the GenAI-assisted text a score of six out of ten and said that it was repetitive and redundant, over-reliant on common phrases and lacked novelty and creativity. She added that the text was typical of GenAI, stating, "a couple of sentences in you begin to think 'Wow, this has been written well.' When you reach the second paragraph, it becomes dull — it lacks the human touch. Phrases like the *language of art* are repeated, all sentences are long, in fact of a similar length. What's more, they are all highly descriptive and fanciful. It doesn't speak to the reader." Despite this criticism, she concluded that, on the whole, the information was very interesting and that it just needed tweaking. The subject matter was in line with the kind of thing she normally edits, and she classified the text as an academic paper/research report or short essay.

4.5.2 Errors and textual anomalies detected

4.5.2.1 Editor 1

4.5.2.1.1 Introduced by the author's edits

E1 found a calque of an Italian expression (*the major part* instead of *the majority*), an incorrect or inconsistent verb tense (*is* instead of *was*) and an improper use of articles (*a* on-site installation). She also flagged a part that might benefit from being made more gender-neutral (*criticism is man's response to man, and we are all human*). There were also some misused prepositions (*the same of* and *see throughout the former*).

4.5.2.1.2 In the GenAI raw output

E1 noted a little redundancy in one expression (*composed of [...] composition*) and the non-existent word *grasitating* (see discussion below). The expression *the journey through art*, which E1 replaced with *looking at a work of art*, also came from GPT-4. She said, "I felt it could be clearer and tie in more with the actual topic of the piece."

4.5.2.2 Editor 2

4.5.2.2.1 Introduced by the author’s edits

The author chose not to edit the raw GenAI output.

4.5.2.2.2 In the raw GenAI output

E2 said that there was an over-reliance on common phrases noting that most sentences started with *this* or *the*, and that the word *through* and the expressions *the language of art* and *the universal language* were used too much. She added, “AI is a giveaway with this text due to the overly long and descriptive sentences that all tend to follow the same structure.”

4.5.2.3 Summary of textual anomalies

Anomaly	Description or effect	Also seen in MT
Excessive repetition of words or phrases	Poor lexical variety	Yes[9]
Redundancy	Repetition of information without adding new meaning or value	No
Non-existent words	*See discussion*	Yes[10]
Blandness	Absence of emotion, creativity or engagement	No
Verbosity	Overly long, highly descriptive, fanciful sentences	No
Low burstiness	Most sentences start in the same way and have uniform structure and length	No
Lack of complex analysis	Superficial, vague and lacking specificity	No
Perfect grammar and spelling	Grammatical mistakes and typos are more typical of human-written copy	No

Table 1. ST textual anomalies reported in this study

4.5.3 Author-editor interaction

4.5.3.1 Author 1

A1 was unable to judge if the kind and frequency of interaction with the editor were different due to the use of GenAI, as he does not have sufficient experience working with this type of editor.

4.5.3.2 Editor 1

E1 stated that the interaction with the author during the editing was more or less the same as normal. Regarding differences, she said, “I've never worked in a situation where I know that the text was written with the help of AI, so that was the only difference — that the author *blamed* a few things on ChatGPT.” She added, “when editing under these kinds of circumstances, there's a kind of *third party* involved, which is a bit odd. When I ask an author for clarification, I want to know what it is that *they* (not the AI) meant or wanted to say.”

[9] Vanmassenhove et al., 2021

[10] Macken, 2019

4.5.3.3 Author 2

A2 said that her level of interaction with the editor during the editing was more or less the same as normal. However, she did not feel that she had sufficient experience of working with this kind of editor to say whether the interaction was in any way different from normal.

4.5.3.4 Editor 2

E2 remarked that there was no real interaction with the author during the editing, which is not the norm.

4.5.4 Authors' and editors' final opinions

4.5.4.1 Author 1

A1 rated the likelihood of using GenAI again as a tool for writing in English at eight out of ten, although he would never use it for academic papers, poems or essays. He found GenAI effective but noted that excessive use might affect a writer's ability to "feel the text." He rated the likelihood of employing an editor again to correct his English at ten out of ten, stating that he would have expected to pay 25 euros for their service in this experiment.

Originally, his preferred method for producing texts in English was to write in his native language and then use a professional translator. However, after this experiment, he said he would now write newspaper articles, blog posts and "light" texts directly in English with the aid of GenAI. He found this method useful for addressing his main difficulty with English (vocabulary) and believed GenAI could be a good tool for editing, although "it cannot replace a human editor."

4.5.4.2 Editor 1

E1 rated her likelihood of accepting future STE assignments as ten out of ten. She said she would have charged 35 euros for her work, had it been a real job. E1 was surprised at how much of the raw GenAI output her author had "taken wholesale." She expected her author to use GenAI more as a starting point and then tweak the output to make it their own. She reiterated her initial impression that the text seemed bland and not concrete enough, especially considering it was about art. She remarked that the text could have been much more engaging if it had focused on a specific work of art as a visual aid example, stating, "but this is — in my opinion — one of the limitations of AI at the moment: it's never specific enough to what you're trying to achieve." Assuming her author were an art expert, she would have recommended he write his article in his native Italian, where he could fully express himself, and then have it translated for a better overall text. "Editing it in this way myself would have been beyond the scope of a copyeditor, but something a development editor — and an art expert — might consider for a longer piece, like a book."

E1 does not believe there would have been much advantage in understanding her author's native language or having experience as a translator because "generally speaking, the quality of the language was very high." She noted one calque where she could immediately tell A1 had translated directly from Italian (see section 4.5.2.1.1), but added, "any monolingual English speaker could have spotted that and worked it out." However, she added that this text was a straightforward example. "Across a longer text where more edits have been made by the author (rather than coming directly from ChatGPT), it might get more annoying for the editor as you'll spend more time trying to figure out what the author meant/asking for clarification."

4.5.4.3 Author 2

A2 rated the likelihood of using GenAI in her English writing process again at only 50%, despite acknowledging its effectiveness in helping her. She rated the probability of employing an editor to correct her English again at ten out of ten and said that she would have expected to pay her editor 20 euros for her service. A2's preferred method for producing texts in English was already to write directly in English before the experiment. She added that using GenAI as an English writing tool was an interesting approach and "is useful when you have to write an article or an essay with a huge number of words."

4.5.4.4 Editor 2

E2 rated the likelihood of accepting STE assignments again in the future at ten out of ten. She said she would have charged 50 euros for the editing she did, had it been a real job. In her opinion, the challenge was less about correcting grammar errors and more about making the text more engaging and giving it a more *human* voice.

She added that she believed a monolingual English speaker could edit ST just as well as a translator who knew the author's language although it might make the process quicker depending on the extent of the author's errors, "but having knowledge of their language is not a guarantee that the editor will produce great work."

5 Discussion

Since LLMs and modern MT engines are both artificial neural networks, one might naively imagine that the kinds of errors that occur in GenAI output might be similar to those commonly found in MT output, particularly when mixed-language prompts are used. However, the only two anomalies reported they have in common were the coining of non-existent words (Macken, 2019), and an over-reliance on common phrases, which manifests itself in MT output as poorer lexical variety (Vanmassenhove et al., 2021) and normalization (Toral, 2019). Interestingly, two of the excessively repeated expressions were found in the prompt, specifically in the title the author provided. This phenomenon is reminiscent of what is known in web copywriting as *keyword stuffing*[11].

The GenAI-assisted text also exhibited anomalies not normally associated with raw MT output (see section 4.5.2.3), such as redundancy, lack of engagement and complex analysis, and low burstiness, a feature also measured by the automatic AI content detector GPTZero (Chaka, 2023). Consequently, STE differs from MTPE more than one might initially suspect. No cases of hallucination (Xu et al., 2024) were identified, probably because — in this experiment — GenAI was used to write up notes provided by the author rather than create new content.

The authors were impressed with the GenAI output, noting that it closely resembled human-written text, whereas the editors immediately recognized it as different. The authors' opinion is consistent with the observations of Clark et al. (2021), who noted that "untrained evaluators are not well equipped to detect machine-generated text". Even with training, Clark et al. found that the detection success rate only marginally improved, reaching about 55%. Dou et al. (2022) proposed a framework that could potentially improve this rate, which was validated in a subsequent study (Dugan et al., 2023). However, the ten error categories identified in their framework do not align with the anomalies reported in this study, except for redundancy. In fact, Clark et al. found that style-related aspects were not reliable detection criteria. Nevertheless, these are issues that an editor has to address. In future studies, it would be

[11] https://developers.google.com/search/docs/essentials/spam-policies?hl=en&visit_id=638602790106513155-3530511413&rd=1#keyword-stuffing

interesting to investigate whether GPT-4 can be explicitly prompted to generate text devoid of the reported anomalies.

Both editors acknowledged that GenAI improved their author's English grammar and spelling, which are known to be infrequent error types in ST[12] (Dou, 2022; Gillham, 2024).

Regarding the non-existent word *grasitating*, this error has already been noted several times[13]. The intended word is *grasping* in all previous reports and in this experiment. The precise cause of this error is unclear, but it may be related to the tokenization of non-existent words found in the training data, probably resulting from optical character recognition (OCR) errors. For example, *gravitation* is sometimes read as *grasitation* by OCR software[14]. Interestingly, *grasitating* appears to be gaining traction as a neologism, as a simple Google search shows[15], and may one day appear in the dictionary.

The use of GenAI seems to alter the author-editor dynamic. In this experiment, when the author accepted the unaltered GenAI output, interaction with the editor was minimal. In the other case the editor perceived a sort of *third party* whose work the author could not clarify.

This experiment showed that authors are likely to use GenAI when writing in a second language, especially if they have relied on translation in the past. However, one editor in this experiment suggested that her author may have obtained better results if he had written in his native language and had his text translated, instead of using GenAI.

Human editing, by both authors and editors, remains crucial in refining and enhancing GenAI output. However, the authors in this experiment would have been willing to pay only 40 to 71% of what the editors would have typically charged for the service they provided. Despite the editors' work, the final texts still have a high likelihood of being recognized as GenAI output (79% and 95%, respectively, according to the Plagramme AI detector[16]). Therefore, if the hypothetical publishers had a strict no-AI rule, more extensive editing might be necessary, potentially making this working method uneconomical.

Besides using GPT-4 as a drafting tool, as in this experiment, one author suggested using it as an editing tool too. When prompted to correct the grammatical errors he had unintentionally introduced into his draft, GPT-4 successfully removed all of them. However, it erroneously corrected the previously mentioned calque to *the fact that many*, instead of *the majority*. In this case, the author would probably have reinstated his original error to correct GPT-4's misinterpretation. Interestingly, it also corrected *grasitating*. Upon examining this new output, his editor remarked that it had helped with "the part of editing that takes the smallest chunk of my time." Nevertheless, the partial results of the unconcluded third case study further support this use, suggesting that for complex texts, it might be more fruitful for the author to draft the paper in their second language unaided and then use GPT-4 as an editing tool to refine the rough draft.

The editors did not see any significant advantage in understanding the author's native language or having translation experience. However, since both editors were bilingual, they may not fully appreciate the potential difficulties monolingual editors might face in correcting English as a second language.

The working method presented in this paper can replace translation only when there is no need for an original text in the author's native language. Moreover, if the same text is required in multiple languages, it is clearly more cost-effective to write in one of those languages and

[12] They are not so infrequent in other languages.

[13] https://www.reddit.com/r/ChatGPT/comments/1ai9cfi/chatgpt_made_up_a_word_typo/

[14] https://www.govinfo.gov/content/pkg/GPO-CRECB-1916-pt4-v53/pdf/GPO-CRECB-1916-pt4-v53-14.pdf — Search for the word *grasitation* in the text with CTRL+F (mentioned in discussion at footnote 13).

[15] www.google.com/search?q=grasitating

[16] https://www.plagramme.com

translate it into the others. Consequently, any decrease in translation work due to GenAI-assisted second-language authoring is not likely to be substantial.

6 Conclusion

We should be very cautious about generalizing the conclusions of these two exploratory case studies since they both concern similar types of text on very similar topics in the same language by authors with similar profiles. It would be advisable to repeat the experiment with different kinds of text on a wide variety of subjects in different languages by authors from diverse backgrounds.

The errors and textual anomalies found were either human errors introduced by the authors writing in a second language or typical GenAI anomalies, such as verbosity and excessive repetition of words or phrases (OpenAI, 2022), probably excluding hallucination (see discussion above). A summary of the detected anomalies is shown in Table 1 in section 4.5.2.3.

Although the authors found prompt engineering intuitive, providing some basic training in this area might be beneficial. This may lead them to adjust the initial prompt to try to produce better base GenAI output for editing. Moreover, both second-language authors and editors should be trained to discern and enhance AI-generated content through STE. This study reveals the importance of human editors in adding creativity and engagement to AI-generated texts.

References

All hyperlinks last accessed 18 September 2024.

Chaka, Chaka. 2023. Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. Journal of Applied Learning & Teaching, July 2023. https://doi.org/10.37074/jalt.2023.6.2.12

Clark, Elizabeth, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. 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). https://aclanthology.org/2021.acl-long.565/

Corder, Stephen Pit. 1975. Error Analysis, Interlanguage and Second Language Acquisition. Cambridge University Press. http://dx.doi.org/10.1017/S0261444800002822

Council of Europe. 2001. Council for Cultural Co-operation. Education Committee. Modern Languages Division. Common European framework of reference for languages: Learning, teaching, assessment. Cambridge University Press.

Dondoni Braz, Ana Clara. 2024. Can You Spot AI-Generated Text? Learn How To Recognise It. Growth Tribe. https://growthtribe.io/blog/spotting-ai-generated-text

Dou, Yao, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, Yejin Choi. 2022. Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text. https://doi.org/10.48550/arXiv.2107.01294

Dugan, Liam, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, Chris Callison-Burch. 2023. Real or Fake Text?: Investigating Human Ability to Detect Boundaries between Human-Written and Machine-Generated Text. https://doi.org/10.48550/arXiv.2212.12672

Farrell, Michael. 2024. On the various kinds of post-editing - Machine translation post-editing, translation memory match editing, hybrid post-editing, monolingual post-editing, stealth post-editing and synthetic-text editing. Presented at Scenari Multimediali e Didattica della Traduzione - Teaching Translation for Multimedia Scenarios, Milan - 14-15 December 2023, currently undergoing peer review.

Gillham, Jonathan. 2024. How To Identify AI-Generated Text?. Blog of Originality.ai AI & Plagiarism Detector. https://originality.ai/blog/identify-ai-generated-text

Gluska, Justin. 2023. How to Check If Something Was Written with AI. Gold Penguin, https://goldpenguin.org/blog/check-for-ai-content/

James, Carl. 1998. Errors in Language Learning and Use: Exploring Error Analysis. Routledge. ISBN 9780582257634

Macken, Lieve, Laura Van Brussel, Joke Daems. 2019. NMT's Wonderland Where People Turn into Rabbits. A Study on the Comprehensibility of Newly Invented Words in NMT Output. Computational Linguistics in the Netherlands Journal 9:67–80. https://www.clinjournal.org/clinj/article/view/93/84

OpenAI. 2022. Introducing ChatGPT. https://openai.com/index/chatgpt/#fn-1

Popović, Maja. 2018. Error Classification and Analysis for Machine Translation Quality Assessment. In Translation Quality Assessment. Machine Translation: Technologies and Applications, vol 1. Springer, Cham. https://doi.org/10.1007/978-3-319-91241-7_7

Toral, Antonio. 2019. Post-editese: an exacerbated translationese. In Proceedings of Machine Translation Summit XVII Volume 1: Research Track, p. 273–281, Dublin, Ireland. European Association for Machine Translation. https://doi.org/10.48550/arXiv.1907.00900

Vanmassenhove, Eva, Dimitar Shterionov, 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, April 19-23, 2021. https://doi.org/10.48550/arXiv.2102.00287

Wu, Tongshuang, Michael Terry, Carrie J. Cai. 2021. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. https://doi.org/10.48550/arXiv.2110.01691

Xu, Ziwei, Sanjay Jain, Mohan Kankanhalli. 2024. Hallucination is Inevitable: An Innate Limitation of Large Language Models. https://doi.org/10.48550/arXiv.2401.11817

Zhou, Yongchao, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, Jimmy Ba. 2022. Large language models are human-level prompt engineers. Published as a conference paper at ICLR 2023. https://doi.org/10.48550/arXiv.2211.01910

Appendix: Funding

The research project reported in this paper has received funding from the International Center for Research on Collaborative Translation at the IULM University and a Small Grant for Research from Mediterranean Editors and Translators.

The authors and editors were offered a token participation fee, but only three of them chose to accept it. They received 125 euros each.

Empowering future language professionals: Findings from a classroom experiment on MT Quality Evaluation in collaboration with DGT

Romane Bodart
UCLouvain
romane.bodart@uclouvain.be

Christine Pasquier
UCLouvain
christine.pasquier@uclouvain.be

Marie-Aude Lefer
UCLouvain
marie-aude.lefer@uclouvain.be

Abstract

Recent advances in translator education encompass technology training, including machine translation (MT) (e.g. the updated 2022 version of the European Master's in Translation competence framework). In this paper, we present the findings of a hands-on teaching unit focused on human evaluation of MT quality through error annotation. The teaching unit was part of a larger international project led by the European Commission's Directorate-General for Translation (DGT). DGT's objectives for this project were to disseminate DGT's methodology for translation quality evaluation, gather empirical data on the quality of eTranslation outputs, and collect feedback for further enhancing eTranslation (DGT 2024). Our objectives as lecturers were to train translation students in machine translation evaluation using a real-life error taxonomy and, more broadly, to equip them as language professionals who will interact with MT output in their future careers. Specifically, we aimed to make them aware of the current limitations of neural MT systems through extensive hands-on experience. Thirty-two students took part in the project and worked in pairs to assess the quality of the machine-translated texts they were assigned. In this article, we analyze the results obtained and we take stock of the teaching unit.

1 Introduction: machine translation quality evaluation in translator education

Recent developments in translator education have increasingly incorporated technology training, including machine translation (MT). As MT continues to evolve, driven by advances in artificial intelligence and large language models, and becomes increasingly prominent in the language services industry, it is crucial for translation students to acquire the skills needed to effectively post-edit machine-generated translations. Building on O'Brien's work (2002), several authors, such as Doherty and Kenny (2014), Koponen (2015), Mellinger (2017), and Guerberof Arenas and Moorkens (2019), have designed course frameworks that incorporate MT and post-editing (PE) into translation curricula.

O'Brien (2002) proposed a structured approach that integrates both theoretical and practical MT components into translator training. In her view, students must not only understand how MT works but also be able to identify and address its shortcomings. This foundational work set the stage for subsequent researchers, such as Doherty and Kenny (2014), Koponen (2015), Mellinger (2017), and Guerberof Arenas and Moorkens (2019), who developed similar course frameworks. These frameworks feature a clear division between theoretical instruction and practical exercises. The theoretical portion often covers a wide array of topics, including the history and development of MT and PE, controlled languages, pre-editing for MT, common MT errors, MT quality evaluation, PE types, post-editing quality, PE effort, and PE skills.

On the practical side, researchers consistently emphasize the importance of hands-on experience with MT and PE, both during classroom hours and through independent work. Practical activities range from comparing different MT outputs to performing pre-editing tasks and evaluating MT quality. O'Brien (2002) and Doherty and Kenny (2014) place particular emphasis on students developing programming skills to understand the mechanics of MT. For example, Doherty and Kenny (2014) require students to compute automatic evaluation metrics, an exercise that helps them gain a deeper understanding of MT performance. Meanwhile, Guerberof Arenas and Moorkens (2019) and Koponen (2015) focus more on the linguistic aspects of post-editing, such as identifying errors in the MT output and correcting them accurately.

Moorkens (2018) conducted a classroom experiment comparing statistical and neural MT systems, using three quality assurance metrics: adequacy, post-editing productivity (measured by temporal effort), and error annotation. This experiment underscored the complexity of evaluating MT quality. It is especially relevant as new MT types, including systems based on large language models (LLMs), emerge. As the technology progresses, it has become clear that regardless of the type of MT system – whether statistical, neural, or LLM-based – students need to develop a strong capacity for MT quality evaluation.

The importance of teaching MT quality evaluation is further highlighted in a 2021 study by Ginovart Cid and Colominas Ventura. This study surveyed 53 translation educators from European Master's in Translation (EMT) institutions to explore current practices related to PE training. Of the respondents, 45 indicated that they teach human MT evaluation as part of their courses, while 27 reported that they also cover automatic evaluation metrics. This consensus reflects a broad recognition among trainers that students must be able to assess MT quality and detect MT errors, irrespective of the specific MT technology being used.

Automatic quality metrics, such as BLEU (Papineni et al., 2002), TER (Snover et al., 2006), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), chrF (Popovic, 2015), and COMET (Rei et al., 2020), are frequently introduced to students as tools for evaluating MT output. These metrics are often praised for their objectivity and efficiency, as they provide a quick, automated assessment of translation quality. However, they are also criticized for being less detailed and less accurate than manual evaluation. While automatic metrics can provide a useful starting point, they do not replace the need for human evaluation, which requires students to engage more deeply with the translation output and to identify specific error types, which in turn improves their post-editing performance.

Recognizing the importance of error detection, researchers have developed a variety of error classification schemes tailored specifically to MT. These include the schemes proposed by Llitjós et al. (2005), Farrús et al. (2010), Bojar (2011), Kirchoff et al. (2012), Comelles et al. (2012), Federico et al. (2014), and Castilho et al. (2017). Although these taxonomies were primarily designed for MT, they can also be applied to human translation. They typically classify errors into categories such as grammar, terminology, and content. One widely used taxonomy is the Multidimensional Quality Metrics (MQM) framework (2014), which is popular among language service providers, even though it was not specifically created for MT error annotation. These error classification schemes are valuable tools for translation trainers aiming to address human MT evaluation, as they offer structured frameworks for assessing quality.

Several applied studies have examined the ability of translation students to detect and correct errors in MT output, revealing varying levels of accuracy and highlighting key challenges in post-editing across different language pairs and text types. For instance, Koponen and Salmi (2017) conducted a study to assess the accuracy and necessity of post-editing corrections made by five Finnish-speaking translation students. The participants, four master's students and one bachelor's student, were tasked with post-editing English-to-Finnish machine translations. The researchers found that students failed to detect and correct 3% of the MT errors, a relatively low rate compared with other studies. However, among the errors that were identified, 9% were incorrectly edited. In these cases, students either failed to correct the errors

(34% of the time) or introduced new errors (66% of the time). This highlights the dual challenge students face: detecting errors in MT output and making the appropriate corrections.

In contrast, Pavlović and Antunović (2021) conducted a study with 44 students working with the English-Croatian language pair and reported a much higher error detection failure rate. Students missed 30% of the errors in the MT output, and 12% of the corrections they made failed to improve the translation. This rate of missed errors is comparable to that found by Kübler et al. (2022), who analyzed how students handled complex noun phrases in specialized texts during post-editing. Their study showed that students failed to detect errors 31% of the time, often due to overconfidence in the MT output. Additionally, students failed to correct the MT appropriately in 49% of cases.

Bodart, Piette, and Lefer (2024) reported an even higher error detection failure rate in their study, which involved analyzing 30 post-edited texts produced by master's students working with the French-English language pair in the financial and legal domains. The researchers found that students missed 47.5% of MT errors and failed to correct errors appropriately in 4.7% of cases. These findings show that error detection remains a persistent challenge for students, particularly when working with complex or specialized texts.

Overall, this overview highlights two key trends in empirical research. First, students consistently struggle with detecting errors in MT output. This difficulty is compounded by the fact that errors in MT can be subtle and may not be immediately obvious. Second, once students successfully identify an error, they are generally able to correct it effectively. The exception to this trend is seen in Kübler et al.'s study, where students detected more errors but were unable to correct them appropriately.

Given these empirical findings, it is clear that MT error detection and classification is a critical component of technology training in translator education. Translation lecturers must ensure that students are equipped with the tools and skills necessary to identify MT errors and to correct them accurately. By focusing on error detection in both theoretical and practical training, translation programs can help students overcome one of the most significant challenges they face in post-editing machine translations.

2 Background: a collaborative project with the European Commission's DGT

In this paper, we present the results of a practical teaching unit designed to engage students in the human evaluation of MT quality through error annotation. This teaching unit, conducted during the spring semester of 2023, formed part of a broader international initiative coordinated by the Directorate-General for Translation (DGT) of the European Commission. The project involved collaboration between DGT and six European Master's in Translation (EMT) universities, including ours, with a focus on assessing the quality of MT outputs produced by eTranslation's General Text engine, specifically in connection with news articles published on the website of the Joint Research Centre (JRC).

The objectives set by DGT for this project included dissemination of its established methodology for translation quality evaluation, collection of empirical data on the quality of eTranslation outputs, and gathering feedback to inform future improvements to the system (DGT 2024). As lecturers, our primary goal was to train translation students in the evaluation of machine translation by applying a real-world error taxonomy. Additionally, we aimed to equip them with the skills needed to critically assess and work with MT output, an increasingly vital aspect of their future roles as language professionals. Through this hands-on experience, we sought to deepen students' understanding of the current limitations of neural machine translation systems, thus preparing them for the challenges they may face as translators in a technology-driven environment.

3 Data and methodology

3.1 MT error taxonomy used

At the start of the project, DGT provided materials related to its internal error taxonomy, including relevant documentation and access to two e-learning modules. The DGT's taxonomy, largely based on the MQM framework, categorizes errors into six types: accuracy, terminology, linguistic norms, job-specific style, general style, and design (DGT 2020) (see Figure 1).

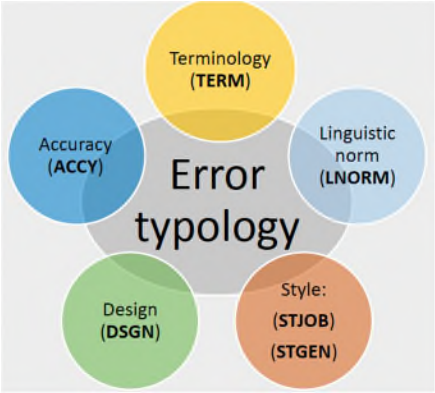


Figure 1: DGT's error typology (DGT 2020)

Accuracy errors (ACCY) relate to the content or meaning between the source and target text, such as mistranslations (e.g. distortion, mismatched names of places or numbers, ambiguity), additions, omissions, or untranslated content. Terminological errors (TERM) arise when the translation does not adhere to accepted terminology within a specific domain or does not comply with a term base or reference document provided by the contractor. DGT defines a term as "a lexical unit comprising one or more words that corresponds to a concept in a particular subject field or application area. Terms are used for expert communication and in that sense are different from purely linguistic and/or stylistic expressions" (DGT 2020:4). Linguistic norm errors (LNORM) concern the linguistic "well-formedness" of the text, which can be evaluated independently of whether the text is a translation. These errors involve formal language aspects, such as grammar, punctuation, and spelling, which are governed by linguistic norms. Style errors refer to grammatically and linguistically correct formulations that are nevertheless inappropriate because they deviate from organizational style guides or job-specific instructions (STJOB), or they exhibit an inappropriate general style, such as tone, register, and stylistic appropriateness (STGEN). Design errors (DSGN) pertain to the presentation of the translated product, including issues such as text or paragraph formatting, layout, the proper integration of graphical elements, and mark-up. This category excludes typographical and stylistic errors. Design errors can be identified either within a document (e.g. a second-level heading formatted as a first-level heading) or in relation to the source text (e.g. inconsistently formatted headings between the source and target). DGT developed a decision tree to assist users in selecting the appropriate error category. It is graphically represented in Figure 2. Examples of each category are provided in Table 1.

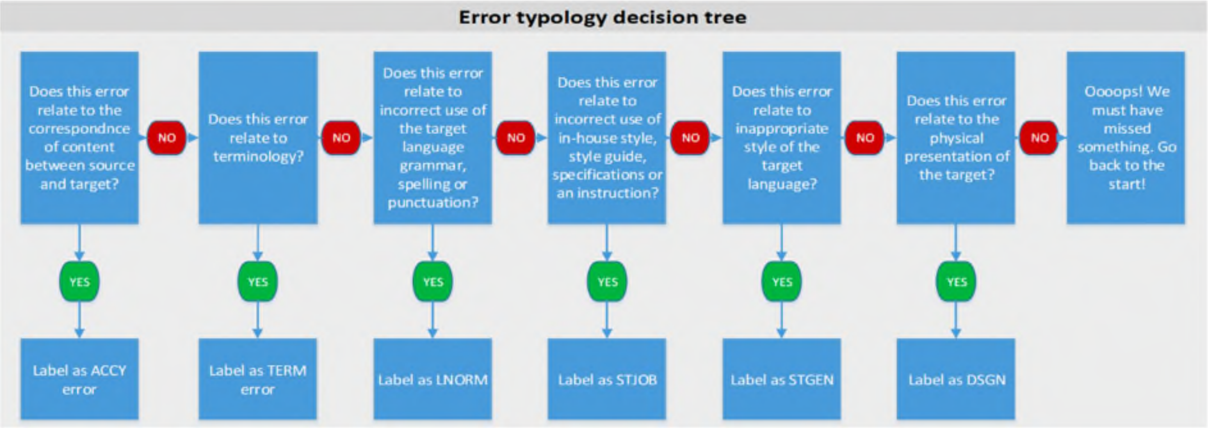

Error typology decision tree
Does this error relate to the correspondnce of content between source and target?
NO
Does this error relate to terminology?
NO
Does this error relate to incorrect use of the target language grammar, spelling or punctuation?
NO
Does this error relate to incorrect use of in-house style, style guide, specifications or an instruction?
NO
Does this error relate to inappropriate style of the target language?
NO
Does this error relate to the physical presentation of the target?
NO
Oooops! We must have missed something. Go back to the start!
YES
YES
YES
YES
YES
YES
Label as ACCY error
Label as TERM error
Label as LNORM
Label as STJOB
Label as STGEN
Label as DSGN

Figure 2: DGT's error typology decision tree (DGT 2020)

Error category	Example	Correct version
ACCY	*ST: The first Zero Pollution Outlook by the Joint Research Centre analyses whether the EU is on track to reach its zero pollution targets **with current and newly proposed EU policies**.* *MT: Les premières perspectives de pollution zéro par le Centre commun de recherche analysent si l'UE est sur la bonne voie pour atteindre ses objectifs de zéro pollution **avec les politiques actuelles et proposées par l'UE**.*	*[…] l'UE est sur la bonne voie pour atteindre ses objectifs de zéro pollution **avec les politiques actuelles et les nouvelles politiques proposées par l'UE**.*
TERM	*ST: In the new CAP (starting in 2023), the share of **UAA** that will receive CAP support for organic farming is higher.* *MT: Dans la nouvelle PAC (à partir de 2023), la part des **UAA** qui bénéficieront d'un soutien de la PAC pour l'agriculture biologique est plus élevée.*	*UAA = utilised agricultural area* ***SAU** = superficie agricole utile*
LNORM	*ST: […] 27 national digital contact tracing apps (21 Member States, 2 EAA countries, Switzerland and United Kingdom (3 apps**)**) were examined in this study […]* *MT: […] 27 applications nationales de traçage numérique des contacts (21 États membres, 2 pays de l'EEE, la Suisse et le Royaume-Uni (3 applications**)** ont été examinées dans cette étude […]*	*[…] 27 applications nationales de traçage numérique des contacts (21 États membres, 2 pays de l'EEE, la Suisse et le Royaume-Uni (3 applications)**)** ont été examinées dans cette étude […]*
STJOB	*ST: More than **100** organisations provided feedback on the interim report and the feedback received was generally positive.* *MT: Plus de **100** organisations ont fourni des commentaires sur le rapport provisoire et les commentaires reçus ont été généralement positifs.*	*Plus de **cent** organisations …* Cf. Code de rédaction interinstitutionnel (2022), section 10.4

STGEN	*ST: Organic farms have lower yields* ***on average*** *[…]* *MT: Les exploitations biologiques ont des rendements* ***en moyenne*** *inférieurs […]*	***En moyenne,*** *les exploitations biologiques ont des rendements inférieurs […]*
DSGN	*ST: All EU Member States stand united in the determination that any form of* ***racism, antisemitism*** *and* ***hatred*** *have no place in Europe.* *MT: Tous les États membres de l'UE sont unis dans leur détermination à ce que toute forme de racisme, d'antisémitisme et de haine n'ait pas sa place en Europe.*	Bold face is missing in the MT

Table 1: Examples of DGT's error categories

3.2 Data used: workbooks prepared by DGT

At the start of the project, it was decided to provide students with full texts for analysis, allowing MT quality to be evaluated at text level rather than segment level, as is often the case in MT evaluation campaigns. The JRC news texts selected for assessment by DGT were systematically organized into eight separate Excel workbooks, each dedicated to a specific topic: biodiversity, creating digital society, organic farming, security, trade agreements, public health, digital education, and sustainable finance. Within each workbook, news items were provided on separate spreadsheets, totaling approximately 200 segments. In total, across all eight workbooks, there were 56 full texts comprising 1,514 segments.

For ease of comparison and error identification, the English source texts were aligned side by side with their corresponding French machine translations. Alongside these, additional columns were provided for annotating any errors using the DGT's error typology, as well as for providing comments or clarifications as needed. This layout facilitated detailed tracking and assessment of MT quality.

Each workbook also included a set of comprehensive instructions, a translation brief, and links to the original news items available online, enabling students to consult the broader context of each article as necessary. Additionally, to encourage accountability, the first spreadsheet in each workbook was devoted to a logbook where students were required to indicate the amount of time they had spent on the project. This time log also captured whether the students had worked individually or collaborated. This structure aimed to ensure a well-documented approach to the MT quality evaluation task.

3.3 Outline of the teaching unit

The teaching unit was embedded within a first-year master's course focused on revision and post-editing. Thirty-two students participated in Spring 2023, working in pairs. At the outset of the project, all students signed an informed consent form. Each Excel workbook was assigned to two pairs of students who worked independently on the same material. The structure of the teaching unit followed a multi-step process, represented in Figure 3: (1) a short e-learning module on DGT's error taxonomy, (2) an in-class introductory session led by the three lecturers, which included the project presentation, a detailed overview of DGT's taxonomy, and an introductory MT quality evaluation exercise, (3) a more comprehensive e-learning module, (4) pair work on MT outputs, (5) an in-class Q&A session to address challenges in identifying and annotating MT errors, and (6) pair work to finalize the assignment.

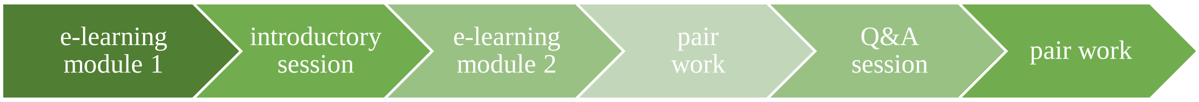


Figure 3: Workflow of the teaching unit

In the first phase (1), students were introduced to DGT's error taxonomy through a brief e-learning module, which provided an overview of its core principles and concepts. This module, designed to be completed in 15 minutes, included short explanatory videos and quizzes to test comprehension. Students were required to complete this module before attending the in-class session (2), during which the three lecturers presented the project and provided further insights into DGT's taxonomy, complementing the e-learning material. They also organized an introductory evaluation exercise that mirrored the students' final task: identifying MT errors, categorizing them using DGT's taxonomy, and performing post-editing (this last PE step was not part of the collaborative project with DGT, but was added by the three lecturers). Following this, students were assigned a second e-learning module (3), estimated to take around one hour, which delved deeper into DGT's taxonomy. Each error category was covered in a dedicated video, with examples provided for clarity, followed by exercises to check understanding. Once students felt confident in their grasp of the material, they were expected to collaborate in pairs on their assigned workbooks (4). A Q&A session (5) was held in class a few weeks later, providing an opportunity to address any difficulties encountered during the initial stages of pair work. Students shared the challenges they faced in identifying and annotating errors, allowing for a collective discussion and clarification of concepts. Finally, students were given several weeks to complete their assignments outside of class (6), ensuring they had sufficient time to apply the knowledge gained and finalize their work before submission.

4 Results and discussion

The full dataset comprised 56 texts, consisting of 1,514 segments and a total of 30,800 words. Based on the students' evaluations, the proportion of machine-translated segments without any errors, as assessed by the two pairs of students annotating the same workbook, was 20.5% (n=311/1,514; see Table 2). For the remaining 1,203 segments, students identified 2,653 errors in total, i.e. an average of 2.2 errors per segment.

Total number of MT segments	1,514 (100%)
Error-free MT segments	311 (20.5%)
Erroneous MT segments	1,203 (79.5%)

Table 2: Error-free vs erroneous MT segments in the workbooks (full dataset)

As shown in Figure 4, when considering all annotated MT errors, regardless of whether they were identified by one or both pairs of students, stylistic errors emerged as the most frequent, accounting for roughly one-third of all MT errors. These were followed by errors related to linguistic norms and accuracy, with terminology errors ranking just behind them. Design-related errors remained marginal, representing only a small fraction of the total errors.

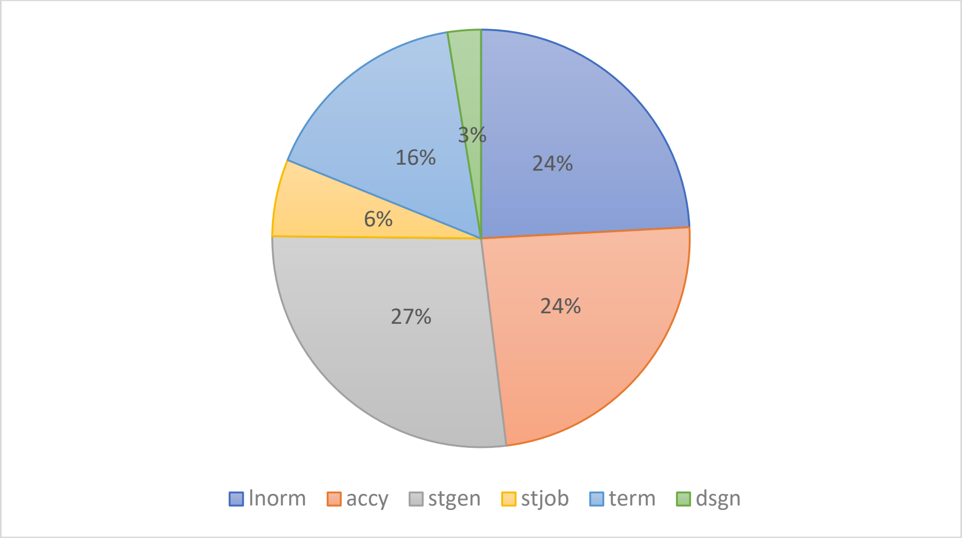


Figure 4: Categories of MT errors (full dataset; n=2,653 MT errors)

However, it is crucial to highlight that of these 2,653 errors, only about one-fourth (i.e. 648 errors) were consistently identified by both pairs of students analyzing the same workbook. These 648 errors were spread across 501 MT segments, representing one-third of the MT data analyzed (see Table 3). Unlike the higher figure presented in Table 2 (79.5% of erroneous MT segments), here, we can assert with a high level of confidence that these 501 segments are indeed erroneous, given students' agreement.

Total number of MT segments	1,514 (100%)
Error-free MT segments	1,013 (66.9%)
Erroneous MT segments	501 (33.1%)

Table 3: Error-free vs erroneous MT segments in the workbooks (cases of agreement across student pairs)

As can be seen in Figure 5, among the 648 errors consistently identified by both pairs of students evaluating the same MTs, approximately one-third were related to accuracy (30%), meaning that source-text content was not transferred appropriately. This was followed by errors relating to linguistic norms (23%) and terminological errors (23%). Errors related to general style accounted for 14%, while task-specific stylistic errors made up 8%, and design-related errors were the least frequent at 2%. These findings, combined with the trends represented in Figure 4 above, suggest that while there was some consistency in identifying critical errors such as accuracy, linguistic norms, and terminology, style-related categories were less consistently annotated.

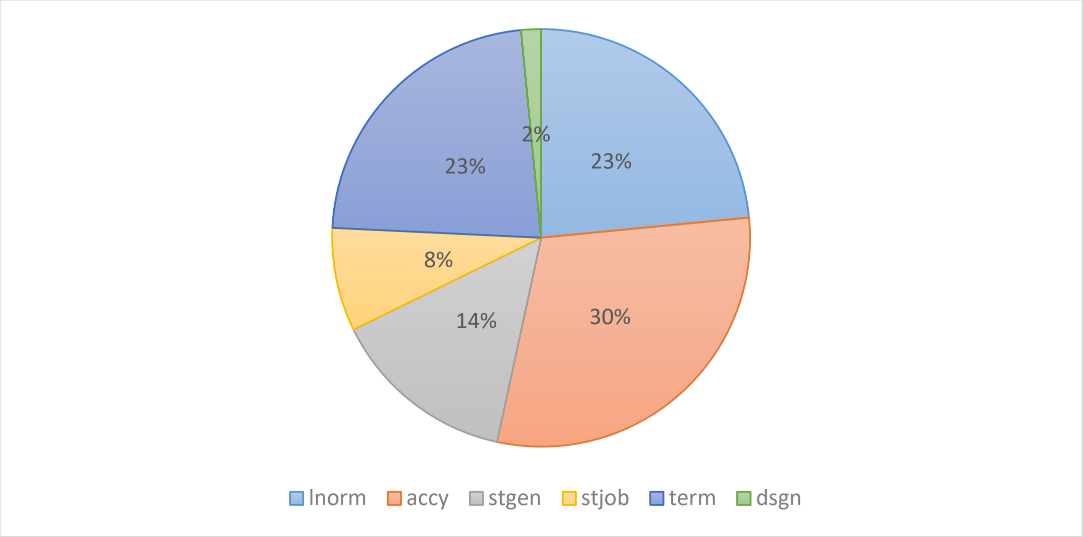


Figure 5: Categories of the errors identified by both pairs of students evaluating the same workbook (n=648 MT errors)

Interestingly, as shown in Table 4, there was substantial variation in the results obtained for the different workbooks provided to students. Depending on the workbook (and, hence, the topic at hand), different error types – linguistic norms, accuracy, terminology, or general style – occupied the top position. This variability suggests that eTranslation generated different error types depending on the sets of source texts used as input. For instance, terminological errors rank first in the economic texts related to trade agreements and sustainable finance, while accuracy is the most problematic area for the texts that deal with organic farming and security.

	Biodiversity	Creating digital society	Organic farming	Security	Trade agreements	Public health	Digital education	Sustainable finance
LNORM	10	27	7	9	16	11	66	6
ACCY	8	25	32	17	15	34	39	24
STGEN	11	9	21	6	5	11	20	10
STJOB	0	0	6	0	0	3	13	30
TERM	4	15	15	6	21	36	14	36
DSGN	0	3	0	4	0	0	3	0

Table 4: Breakdown per workbook of the errors identified by both student pairs (n=648 MT errors)

Our dataset includes 2,005 MT errors identified by only one of the two student pairs. These cases fall into two main categories: (1) MT errors that one pair overlooked, and (2) words, expressions, or constructions incorrectly identified as erroneous by one pair, despite being appropriate in French. Among these 2,005 errors, those related to general style were the most frequent, comprising nearly one-third of the discrepancies (31%). Errors associated with linguistic norms ranked second, accounting for 25%, while accuracy-related errors followed closely at 22%. Together, these three categories represent the most common types of error observed by only one student pair (see Figure 6). In addition to these, terminological errors made up 14% of the remaining discrepancies, while job-specific stylistic errors accounted for 5%, and design-related errors were the least frequent, constituting 3%. These findings reveal the varying degrees of student attention to different aspects of translation quality, particularly those involving general style, linguistic norms, and accuracy. The variability in detecting stylistic errors probably reflects the inherently subjective nature of style, making it more prone to variation in assessment. By contrast, the high number of segments identified as erroneous by only one of the two student pairs in the other error areas confirms the critical need for thorough training in MT quality evaluation.

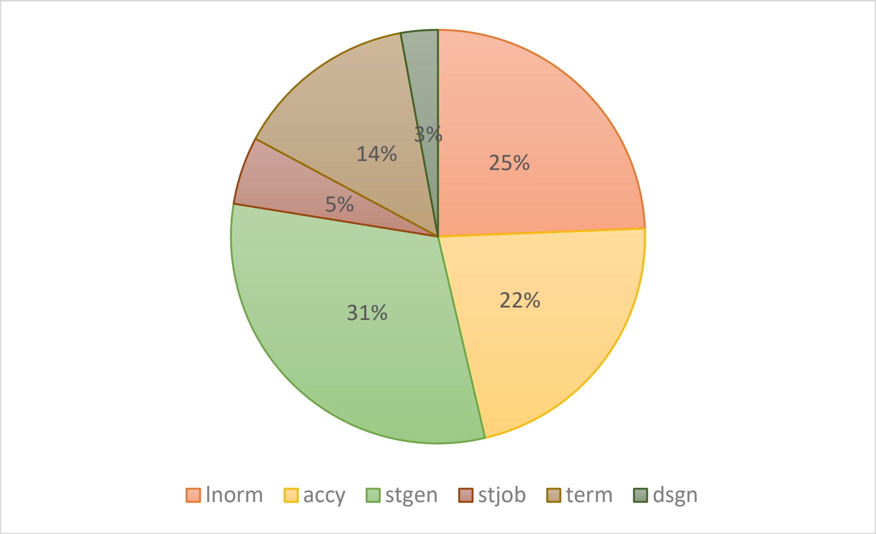


Figure 6: Categories of the errors identified by only one of the two student pairs evaluating the same workbook (n=2,005 MT errors)

Overall, our results show that human MT quality evaluation remains a challenging task for first-year master's students. They further indicate that it is necessary to integrate large-scale hands-on training into translation curricula in order to help students develop the critical evaluation skills required to interact effectively with MT outputs.

5 Student experience: retrospective questionnaire

Following the project, students were invited to complete a questionnaire, which we developed collaboratively with DGT, to provide feedback on their experience. Sixteen of the 32 UCLouvain students who participated in the project responded to the questionnaire. All respondents found the project to be engaging and valuable. They reported that it helped them learn to distinguish between different error types. Additionally, they noted the significance of evaluating whether certain edits are truly necessary during the post-editing process, among other insights.

However, one commonly reported drawback was the project's time demands. Many students felt that it was too time-consuming, with most reporting that they spent over 20 hours on completing the project (average time spent: 24 and a half hours; see Table 5). One group, in particular, reported spending nearly 46 hours on the evaluation task.

Biodiversity	Pair 1	19h25min
	Pair 2	16h58min
Creating digital society	Pair 1	21h55min
	Pair 2	14h45min
Organic farming	Pair 1	26h15min
	Pair 2	25h45min
Security	Pair 1	15h05min
	Pair 2	13h40min
Trade agreements	Pair 1	34h17min
	Pair 2	19h23min
Public health	Pair 1	22h25min
	Pair 2	38h54min
Digital education	Pair 1	19h40min
	Pair 2	26h54min
Sustainable finance	Pair 1	31h25min
	Pair 2	45h54min
Average		**24h32min**

Table 5: Time spent on the evaluation task (self-reported, on the basis of a project logbook)

In the questionnaire, students were also asked whether before participating in the project they had feared that MT engines might replace human translators, and if their views had changed after completing the project. Almost half of respondents (7 out of 16) indicated that they had not been concerned about MT replacing human translators, and their opinion remained unchanged after the project. However, nine participants replied that they were initially concerned that MT engines could eventually replace human translators. Three of these changed their minds after completing the project, while six maintained their initial concern (see Table 6). The students who see a promising future for human translators provided several reasons for their optimism. They highlighted the poor quality of the MT outputs, noting that MT engines are not yet sufficiently advanced to compete with human translators, particularly because they cannot fully grasp the nuances of language in the way a human can. Additionally, students pointed out that MT engines still produce a considerable number of errors, including gender bias, incoherence, and repetitions, further emphasizing the ongoing need for human oversight and expertise in the translation process.

Number of students	Opinion before and after the project
7	No → No
3	Yes → No
6	Yes → Yes

Table 6: Students' opinions as to whether MT engines will replace human translators

The primary reason cited by the six students who believe that MT engines will eventually replace human translators is the expectation that MT technology will continue to improve over time. However, upon closer examination of their comments, it appears that two of these students held a more nuanced view. One student clarified that the translator’s role is unlikely to disappear entirely but will evolve to meet market needs and technological advances, resulting in more post-editing assignments rather than traditional translation work. Similarly, another student noted that post-editing is particularly appealing to clients seeking to reduce costs, suggesting that while the profession may shift toward more post-editing, it is unlikely to vanish altogether. This indicates that both students foresee an adaptation of the profession rather than its complete replacement by MT.

6 Concluding remarks

In this article, we have presented the results of a practical teaching unit developed in collaboration with DGT, aimed at engaging 32 first-year master’s students in the human evaluation of MT quality at text level through error annotation. Our study shows that students often encounter challenges in accurately detecting and categorizing errors within machine-translated texts. This observation provides empirical support for the widely held belief that students must be continuously trained not only in post-editing but also in MT error detection in order to be fully prepared for careers in the language services industry. The complexity of these tasks highlights the need for ongoing, hands-on practice to develop the necessary skills for effective MT evaluation and post-editing work. However, it is important to stress that if we had conducted this study with second-year master’s students, different trends might have emerged, potentially reflecting their more advanced skills and deeper familiarity with MT error detection and post-editing.

For future research, it would be valuable to examine the quality of students’ post-edited texts to ensure alignment with the MT errors they identified. This could help pinpoint students’ weaknesses in post-editing, enabling the development of targeted exercises or dedicated sessions within post-editing courses to address these challenges. Such practical initiatives would make meaningful contributions to translator education by refining students’ skills in post-editing. Additionally, tracking students’ progress over the course of their master’s program would provide valuable insights into their evolving skills, allowing educators to adapt training methods to better support their development in MT error detection and post-editing. Finally, it would be important to replicate the experiment with the updated version of the eTranslation General Text engine so as to engage students with MT outputs of LLM-based technologies.

References

Banerjee, Satanjeev, and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, MI, pages 65-72.

Bodart, Romane, Justine Piette, and Marie-Aude Lefer. 2024. The Machine Translation Post-Editing Annotation System (MTPEAS), A standardized and user-friendly taxonomy for student post-editing quality assessment. *Translation Spaces*, Online First. https://doi.org/10.1075/ts.24002.bod

Bojar, Ondrej. 2011. Analyzing error types in English-Czech machine translation. *The Prague Bulletin of Mathematical Linguistic*, (95): 63–76.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio Miceli-Barone, and Maria Gialama. 2017. A comparative Quality Evaluation of PBSMT and NMT using Professional Translators. In *Proceedings of Machine Translation Summit XVI: Research Track*, Nagoya Japan, pages 116–131.

Comelles, Elisabet, Jordi Atserias, Victoria Arranz, and Irene Castellón. 2012. VERTa: Linguistic features in MT evaluation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, Istanbul, Turkey. European Language Resources Association (ELRA), pages 3944–3950.

DGT (Directorate B – Translation, ABCD Quality Managers). 2024. Study to evaluate the quality of eTranslation output from multilingual platforms. Report summary.

DGT. 31 March 2020. DGT Guidelines for Evaluation of Outsourced Translations (TRAD19). Key aspects of evaluation under TRAD19 – Quick Reference.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co- occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, San Diego, California, USA, pages 138-145.

Doherty, Stephen, and Dorothy Kenny. 2014. The design and evaluation of a Statistical Machine Translation syllabus for translation students. *The Interpreter and Translator Trainer*, vol. 8(2): 295-315.

Doherty, Stephen, Joss Moorkens, Federico Gaspari, and Sheila Castilho. 2018. On education and training in translation quality assessment. In *Translation Quality Assessment*, edited by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty. Cham, Springer, pages 95–106.

EMT Expert Group. 2022. Competence Framework. https://commission.europa.eu/system/files/2022-11/emt_competence_fwk_2022_en.pdf [last accessed October 6, 2024].

Farrús, Mireia, Marta R. Costa-jussà, José B. Mariño, and José A. R. Fonollosa. 2010. Linguistic-based Evaluation Criteria to identify Statistical Machine Translation Errors. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*, Saint Raphaël, France, European Association for Machine Translation, pages 167–173.

Federico, Marcello, Matteo Negri, Luisa Bentivogli, and Marco Turchi. 2014. Assessing the Impact of Translation Errors on Machine Translation Quality with Mixed-effect Models. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar. Association for Computational Linguistics, pages 1643–1653.

Ginovart Cid, Clara, and Carme Colominas Ventura. 2021. The MT post-editing skill set: Course descriptions and educators' thoughts. In *Translation Revision and Post-editing: Industry Practices and Cognitive Processes*, edited by Maarit Koponen, Brian S. Mossop, Isabelle Robert, and Giovanna Schocchera. London/New York, Routledge, pages 226–246.

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

Kirchhoff, Katrin, Daniel Capurro, and Anne Turner. 2012. Evaluating User Preferences in Machine Translation Using Conjoint Analysis. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, Trento, Italy, European Association for Machine Translation, pages 119–126.

Koponen, Maarit. 2015. How to teach machine translation post-editing? Experiences from a post-editing course. In *Proceedings of 4th Workshop on post-editing technology and practice*, Miami, USA, pages 2–15.

Koponen, Maarit, and Leena Salmi. 2017. Post-editing quality: Analysing the correctness and necessity of post-editor corrections. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 16: 137–148.

Kübler, Natalie, Alexandra Mestivier, and Mojca Pecman. 2022. Using comparable corpora for translating and post-editing complex noun phrases in specialized texts. Insights from English-to-French specialized translation. In Extending the Scope of Corpus-Based Translation Studies, edited by Sylviane Granger and Marie-Aude Lefer. London: Bloomsbury, pages 237–266.

Mellinger, Christophe D. 2017. Translators and machine translation: knowledge and skills gaps in translator pedagogy. *The Interpreter and Translator Trainer*, vol. 11(4): 280-293.

Moorkens, Joss. 2018. What to expect from Neural Machine Translation: a practical in-class translation evaluation exercise. *The Interpreter and Translator Trainer*, 12(4): 375-387.

O’Brien, Sharon. 2002. Teaching post-editing: A proposal for course content. In *6th EAMT Workshop Teaching Machine Translation*, pages 99-106.

Office des publications de l’Union européenne, 2022. *Code de rédaction interinstitutionnel*. Office des publications de l’Union européenne : https://data.europa.eu/doi/10.2830/445722

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, Association for Computational Linguistics, pages 311–318.

Pavlovic, Natasa, and Goranka Antunovic. 2021. Towards acquiring post-editing abilities through research-informed practical tasks. *Strani jzici: Strani jezici: časopis za primijenjenu lingvistiku*, vol. 50(2): 185-205.

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

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

Snover Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, Cambridge, Massachusetts, USA, Association for Machine Translation in the Americas, pages 223–231.

Machine Translation Literacy in the age of GenAI

Aletta G. Dorst
Leiden University
a.g.dorst@hum.leidenuniv.nl

Leena Salmi
University of Turku
leena.salmi@utu.fi

Joke Daems
Ghent University
Joke.Daems@UGent.be

Maarit Koponen
University of Eastern Finland
maarit.koponen@uef.fi

Abstract

One widely used application of artificial intelligence (AI) in today's globalized world is machine translation (MT). Studies show a growing need for an understanding of how to use MT critically, or MT literacy, amongst not only translation and language students but all users. Given the current interest in using generative large language models (LLM) for translation-related tasks, the question arises to what extent MT literacy now also entails knowing how LLMs and generative AI (GenAI) work. Our paper explores how university students enrolled in translation, language and AI courses in Finland, Belgium and the Netherlands understand how MT works and what its defining characteristics are as compared to human translation (HT). We find that, overall, students consider MT distinct from HT, although many also perceive important similarities. However, some of these similarities are based on misconceptions and a tendency to humanize the technology. We argue for a need to (re)define more clearly what MT Literacy entails to empower both professional and informal users to use GenAI for translation effectively and critically.

1 Introduction

Estimated at one billion users, machine translation (MT) is now widely adopted by a variety of users, both informal and professional, for a wide range of purposes (Nurminen, 2021) in anything from low-stakes communication to high-stakes legal and medical contexts (Vieira, O'Hagan & O'Sullivan, 2020). This raises important questions about the consequences of this use: to what extent do the people that use the technology actually understand how it works? Previous work showed that, even in the context of translation studies specifically, students have misconceptions about MT and do not fully understand how it works (Salmi *et al.*, 2023). Misconceptions about technology are far from harmless. By focusing exclusively on the productivity gains made by using MT, for example, the work of translators is increasingly being devalued (do Carmo, 2020). The same is true for the way MT quality results are often presented without the necessary nuance, "creating an unrealistic and uncritical perception of MT among the general public" (Moorkens, 2022:129).

Since the launch of ChatGPT, and the ensuing hype in using large language models (LLMs), the technological landscape has only become more complex, with LLMs and generative AI (GenAI) now actively being used in translation (Kornacki & Pietrzak, 2025), raising even more questions about the need for (nuanced) understanding of technology. To verify to what extent students' conceptualisations of MT also encompass conceptualisations of GenAI, the present paper builds on Salmi *et al.* (2023), expanding the work both in number and recency (data collected until May 2024). In the following sections, we first discuss the importance of MT literacy, relating it to data and AI literacy, and highlight some of the terminological complexities related to (de)humanisation of technology and using computational metaphors to represent human cognition. We then present the methodology and discuss what our results mean for MT literacy going forward.

2 Related research

2.1 MT literacy

Building on earlier work, the concept of MT literacy was refined by Nurminen (2021:44) to encompass a user's ability to:

1. Comprehend the basics of how machine translation systems process texts
2. Understand machine translation systems' strengths and weaknesses
3. Understand how machine translation systems are or can be used for purposes that are important to the user
4. Appreciate the wider implications associated with the use of MT
5. Assimilate information from raw machine-translated texts
6. Evaluate how machine translation-friendly a text is
7. Create or modify a text so that it can be translated more easily by an MT system
8. Modify the output of an MT system to improve its accuracy and readability

While a growing number of studies has focused on MT literacy for students in translation and/or language degrees (Bindels & Pluymaekers, 2022; Dorst *et al.*, 2022; Loock & Léchauguette, 2021), little is known about the need for MT literacy among students in other disciplines (cf. Bowker, 2020). Similarly, there is little research on how professionals in different fields employ MT in their daily workflow (cf. Anazawa *et al.*, 2013; Nurminen, 2019), while the media increasingly report on incidents and malpractices resulting from imprudent use of MT. Today, MT literacy critically interacts and overlaps with information literacy (e.g. Bowker, 2021), data literacy (e.g. Krüger, 2022) and AI literacy (e.g. Ng *et al.*, 2021). Krüger argues that in parallel with the "increasing relevance of MT literacy in the professional translation process, the rise in prominence of another digital literacy can be observed, i.e. *data literacy*" (2022:248, italics original). The same holds for AI literacy: society's widespread adoption of AI tools urgently calls for a definition of AI literacy and a clearer understanding of what people need to know about how LLMs and GenAI work. Ng *et al.* (2021:2) argue that "AI literacy means having the essential abilities that people need to live, learn and work in our digital world through AI-driven technologies", and people need to be taught explicitly "how to use AI technologies judiciously, as well as to discriminate between ethical and unethical practices" (2021:2).

One existing framework for MT literacy training includes technical, linguistic, economic, societal and cognitive competences and highlights the importance of data literacy (Krüger & Hackenbuchner, 2022). In the context of AI literacy, Long & Magerko (2020) suggest no fewer than 17 competencies, including the need to recognise AI, understanding 'intelligence', being aware of AI's strengths and weaknesses, understanding machine learning, and learning from data. All of these definitions and frameworks show a hierarchy of complexity that aligns with Bloom's taxonomy (1956) for learning objectives: Know > Understand > Apply > Analyse > Evaluate > Create. Similarly, the European Master's in Translation Competence Framework

(2022:7) acknowledges that “MT literacy and awareness of MT´s possibilities and limitations is an integral part of professional translation competence” and translators should have a “basic knowledge of machine translation technologies”. However, as argued in Salmi *et al.* (2023), what this ‘basic knowledge’ or ‘knowing how MT works’ entails is not clearly defined. The EMT Competence Framework was also published before the rise of AI tools, making the present work especially timely, as it is “important to understand existing [...] conceptions of AI in order to develop effective AI literacy interventions” (Long & Magerko, 2020:7).

2.2 (De)humanising technology and human cognition as computer processing

Salmi *et al.* (2023:301) found that students had a “tendency to humanise MT when explaining how it works”, although it was impossible to know for sure whether this was a misconception or a lack of terminology to accurately express differences between humans and machines. On the other hand, work on anthropomorphism in technology suggests that “the ascription of human qualities onto non-human entities” is quite a widespread phenomenon, and that it is not entirely harmless (Placani, 2024:692). In the context of AI specifically, humanising the technology leads to overestimations of its performance, as well as distortions of people’s moral judgments, such as increased (misplaced) trust in the technology (Placani, 2024). The tendency to humanise AI can be linked to general narratives where technology is characterised by words related to intelligence and emotions as well as social or ethical aspects (Ekbia & Nardi, 2017:2). In the case of MT, common narratives emphasise its human-likeness, sometimes ascribing it with nearly magical qualities (Vieira 2020:109). Conversely, depending on the translation situation, sometimes people specifically do *not* view MT in humanised terms. In a UK survey examining the perceptions of MT users, for example, the benefits of MT not being human were summarised as follows: MT is “not judgmental”, there is no need to be embarrassed, and MT allows them to do things on their own without relying on others (Vieira *et al.*, 2022:905-906).

Interestingly, early models of human cognition were actually inspired by computer processing leading to the so-called ‘Computational Theory of Mind’ (Horst, 1999). Building on these concepts, translation could then be seen as a problem-solving activity, with translators applying rules and strategies to reach certain goals, in parallel with computers using algorithms to process data and produce output (Risku & Rogl, 2020). This shows that, even in the literature, the conceptualisation of what makes humans human and machines machines is not always clearly delineated, and they have each been conceptualised through the other. In fact, Baria & Cross (2021) warn that while the Computational Metaphor (‘the brain is a computer’) is “the most prominent metaphor in neuroscience and artificial intelligence”, its appropriateness is highly debatable, both in terms of “whether it is useful for the advancement of science and technology” and more particularly “how it may shape society’s interactions with AI” (n.p.). In many disciplines, these computational models of cognition have made way for the notion of “situated, embodied, distributed, embedded and extended cognition” (Risku & Rogl, 2020:481).

3 Methodology

In the present study we focus on the question to what extent MT literacy also entails data literacy and AI literacy by examining how university students enrolled in translation, language and AI courses in Finland, Belgium and the Netherlands understand how machine translation works and what its defining characteristics are as compared to human translation. The following subsections describe the design, methods and participants of the study.

3.1 Questionnaire

In total, 173 students agreed for their data to be used in the study, 47 from University of Turku (Finland), 82 from Leiden University (Netherlands), 15 from University of Eastern Finland

(Finland), and 29 from Ghent University (Belgium). Data was gathered using an online questionnaire that the students filled out in class or as homework. The questionnaire was made available via Webropol (https://webropol.com/) and was offered in Finnish, Dutch and English. The English version was provided because we knew that not all students were (native) speakers of Finnish or Dutch.

The questionnaire opened with a description of the study, including aims and means of data collection and management, as well as contact information on the researchers involved. The students were informed of the purpose of the study, data collection and processing and asked for consent. In the questionnaire, students were first asked to reflect on their understanding of how MT engines work and how humans translate. They were asked to consider what human translators do when they translate and which steps or activities are involved. Then they were asked to briefly answer the following questions: "Do humans translate in the same way machines do? If yes, what is similar about the way they translate? If not, in what way is a human translator different from a machine?" It was stated explicitly that there was no word limit and that they should take approximately 10 minutes for their answer. Finally, students were asked to specify their native language, age, university, course for which they completed the questionnaire, degree, and the start date of their degree. Unfortunately, we did not ask for more specific information regarding their previous experience in translating or using translation technology. Future research could explore in more detail to what degree previous experience influences students' conceptualisations of MT and AI.

3.2 Methods

In total, 173 reflections were analysed, of which 55 were written in Dutch, 62 in Finnish and 59 in English. The reflections were analysed in terms of (a) their answers to the overall question whether humans and machines translate differently or in the same way, and (b) the characteristics they mentioned to explain their views.

Each answer was coded for sameness vs difference and for the characteristics mentioned, linking each characteristic to the human, the machine or both. Reflections often contained statements belonging to different categories, each of which was coded separately. It should be noted that these characteristics represent themes: students did not need to use the exact words of the category label. For example, both "MT produces instant translations" and "machines can go through millions of texts instantaneously" were coded as "Is fast". DeepL was used to translate the Finnish and Dutch answers into English. However, the main analysis was conducted using the original language of the reflections by authors who are speakers of the language in question. The coding for Turku and UEF students was first done by Salmi and checked by Koponen; the coding for Leiden and Ghent students was first done by Dorst and checked by Daems. All unclear, ambiguous and problematic cases were first discussed by the language team and then amongst all authors to reach full consensus.

The coding approach used was inductive thematic analysis. As a starting point, we used a list of data-driven characteristics that had emerged in an unpublished pilot study involving a similar reflection task with students from the Universities of Turku and Eastern Finland (Salmi and Koponen, 2022). These categories were further refined inductively based on the data after the first round of data collection in 2022 (see Salmi *et al.* 2023). After the second round of data collection in 2023-24 a final list of categories was established, including a number of new categories that emerged from the new data. The previously analysed part of the dataset was also rechecked against these new categories to verify whether similar issues could be found. The final list of characteristics, in alphabetical order, can be found in Table 2.

3.3 Participants

University of Turku (Finland): 47 students participated, 39 B.A. and 8 M.A. The first group (n=10) filled out the questionnaire in October 2022 during a 5-ECTS course on intercultural communication, compulsory for the major and minor students of French. The second group (n=15) filled out the questionnaire in October 2022 as part of a 5-ECTS elective course on translation practice, open to all language students, and the third group (n=22) as part of the same course in October 2023. The first group were first-year bachelor's students majoring in French, except one second-year student who had Spanish as their major. The students in the second and third group were majoring in various subjects, most of them in English or other languages. Some of them had translation studies as a minor, and may have completed or been enrolled in a translation technology course at the time of the survey.

Leiden University (Netherlands): 82 students participated, from three different cohorts: 10 B.A. students in October 2022 during a 5-ECTS elective minor course on multilingual translation; 23 M.A. students in November 2022 during a 5-ECTS compulsory course on translation technology; and 4 M.A. and 45 B.A. students in April 2024 during a 5-ECTS elective minor course on AI and the Humanities. The first group was predominantly enrolled in language degrees (especially English, Japanese and Korean) without any prior courses on translation or translation technology; the second group were all enrolled in the 1-year M.A. in translation and had completed a 30-ECTS Minor in Translation during their B.A.; and the third group came from a wide range of degrees, including several exchange students from abroad. This latter group were all enrolled in a Minor in Digital Humanities but we do not know if they had previously taken any courses on translation or translation technology.

University of Eastern Finland (Finland): 15 students participated, 14 B.A. and 1 M.A. The students filled out the questionnaire in April 2024 during a 3-ECTS compulsory course for translation students, elective for other students at UEF, on translation studies. Most were students in the English or Russian translation degrees or the Swedish language degree, with some students from other subjects. Most of them had completed at least one translation course and were enrolled in a translation technology course at the time of the survey.

Ghent University (Belgium): 30 students participated. The first group, consisting of 5 postgraduate students enrolled in a programme on Computer-Assisted Language Mediation and 5 M.A. students enrolled in the European Master's in Technology for Translation and Interpreting, filled out the questionnaire in October 2023 during a 5-ECTS elective course on machine translation and post-editing. Five of the students had a background in translation, but only two indicated they used MT during their studies. Some of the students could have been enrolled in an elective course on terminology and translation technology at the same time, but since they participated in this survey during the first MTPE class, their exposure to translation technology would have been limited. The second group, consisting of 20 M.A. students enrolled in the 1-year Master's in translation, filled out the questionnaire in November 2023 during a 3-ECTS obligatory course on terminology and translation technology. They could have encountered translation technology during an introductory course in the 2nd B.A. year, and - depending on their language combination – have been allowed to use translation technology during their translation work for other courses, but they had no experience post-editing.

4 Results

Table 1 shows the results for the question whether humans and machines translate the same way or differently. "Both" indicates responses saying that there are both similarities and differences between HT and MT. "Unclear" indicates that the student's text did not directly answer the question in a way that could be interpreted as belonging to any of the categories.

One student only wrote some general remarks about how humans translate but did not mention MT at all, while the other mainly reflected on their own experiences when translating.

	Turku	**UEF**	**Leiden**	**Ghent**	**All**
Same	2 (4.3%)	0 (0%)	6 (7.4%)	0 (0%)	8 (4.6%)
Different	27 (57.4%)	10 (66.6%)	53 (64.6%)	20 (69%)	110 (63.6%)
Both	16 (34%)	5 (33.3%)	23 (28%)	9 (31%)	53 (30.6%)
Unclear	2 (4.3%)	0 (0%)	0 (0%)	0 (0%)	2 (1.2%)
Total	47 (100%)	15 (100%)	82 (100%)	29 (100%)	173 (100%)

Table 1. Students' views on whether humans and machines translate in a different or in a similar way.

Table 1 shows that the majority of respondents (63.6%) answered that humans and machines translate in a different way, and nearly all of the remaining (30.6%) saw both similarities and differences. Of the whole group, only a few stated that humans and machines translate the same way. While the precise percentages vary to some extent, the same pattern can be seen across all four universities.

Table 2 shows the characteristics (in alphabetical order) students associated with either humans or machines, or both.

Characteristic	**Human**	**Machine**	**Both**
Considers context and the whole text	74	4	7
Considers target audience and situation	50	0	0
Has a corpus / data / database	1	35	3
Has emotions, cognition, personality	33	0	0
Has experience	20	0	2
Has language skills	24	1	2
Has world knowledge	55	0	2
Is a language learner / non-native speaker	0	6	0
Is creative	27	0	0
Is fast	1	27	0
Learns from prior material	0	6	10
Makes mistakes	3	19	3
Operates mechanically	0	16	4
Searches for information	9	0	3
Translates directly ("word for word")	0	30	3
Translates the same way every time	0	3	0
Understands meaning	46	0	1
Understands nuances	40	0	0
Uses logic / reasoning	6	3	1
Uses predefined knowledge	0	11	9
Uses probabilities / algorithms / statistics	0	34	1
Uses rules	2	14	11
Uses vocabularies / dictionaries	3	4	8

Table 2. Characteristics associated with humans, machines, or both. New or updated categories marked in boldface.

When contrasting Table 2 with the findings from Salmi *et al.* (2023), the same overall pattern emerges but most contrasts between MT and HT are now much clearer, in part because some categories were updated to make more fine-grained distinctions, including the difference between having ´knowledge´ vs ´information´ discussed below. As a result, certain characteristics that were previously classified as belonging to both HT and MT due to a relatively low number of instances are now MT-specific – i.e. 'Is fast', 'Makes mistakes' and 'Operates mechanically' – while 'experience' and 'knowledge' are now clearly associated with humans. The new category 'Understands nuances' is also quintessentially human, as are 'Considers context and whole text', 'Considers target audience and situation', Has emotions, cognition, personality', 'Has language skills', 'Has world knowledge', 'Is creative', and 'Understands meaning'. Characteristics typically associated with machines are 'Has a corpus, database, a lot of data', 'Is fast', 'Makes mistakes', 'Operates mechanically' and 'Translates directly'. In addition, 'Learns from prior materials', Uses predefined knowledge, 'Uses rules' and 'Uses vocabularies and dictionaries' are considered either MT-only or both MT and HT, though the low number of mentions indicates these associations are weak. The counts also show that participants have a much clearer idea of what sets HT apart from MT than the other way around.

While Salmi *et al.* (2023) focused on what these characteristics reveal about common misconceptions in students' conceptualizations of MT, the current paper will focus on the issue of humanisation of technology, relating it to human embodied cognition and the terminology that is central to defining artificial intelligence and machine translation.

5 Discussion

Table 1 showed that across all universities roughly two-thirds consider MT and HT to be different and roughly a third consider them to exhibit both similarities and differences. If we divide the answers according to time period – 2022 (Turku and Leiden), 2023 (Ghent, Turku, UEF) and 2024 (Leiden) – the distribution remains the same, suggesting that the increased visibility of GenAI has not radically changed the way students understand how machine translation works. However, some of the answers do indicate that students make a distinction between traditional MT and LLMs. More importantly though, the presence of 'bias' and 'hallucinations' in LLMs is sometimes explicitly related to LLMs behaving more like humans than traditional MT:

> *L60, English original:* However, like humans, machines can and do make mistakes when translating. Especially LLMs such as GPT who are tasked with translation can produce hallucinations, i.e. outputs that are significantly lesser in quality than desired. This can be compared to how humans may alter sentence structure or grammar rules when translating from one language to another without meaning to, a common occurrence with people who are multilingual. Furthermore, bias exists in both human and machine translation alike.

One issue in teaching MT literacy has been whether it is necessary for users to understand the difference between rule-based, statistical and neural MT. We would argue that for MT in the age of GenAI there is a much greater need for students to understand how data, LLMs and GenAI work and how the processes involved are distinct from human behaviour and cognition.

Looking at Table 2, the categories 'Has emotions, cognition, personality', 'Has experience' and 'Has world knowledge' account for 108 of participants' comments about human translation, while they are never labelled as typical of MT. Only very rarely are they seen as typical of both (twice for 'Has experience' and twice for 'Has world knowledge'). This offers support for the situated, embodied approaches to cognition (Risku & Rogl, 2020). Most comments suggest that there is something typical about ‘the human experience’ and of ‘being in the world’ that sets human translation apart from MT:

L70, English original: I feel as though there is something we acquire from the human experience that allows us to sense the nuances in meaning and feeling when it comes to expressing things that machines simply cannot do.

L81, English original: To me, the crucial difference is that humans have a lived experience of being in this world. This gives them a unique experience of the meaning of certain words and phrases, which is informed by their culture, upbringing, and day to day use of the language.

It is this lived experience that gives people the ‘world knowledge’ that many of the participants deem essential to human translation. This world knowledge covers our consideration of the target audience and purpose of the translation, but also our understanding of nuances as well as the tone and emotional value of the text. It is what enables humans to understand cultural references, humor, and implicit meaning. The answers provided by the students show that in their conceptualisation of human translation context, culture, emotions, nuance and creativity are inextricably linked, and this complex interaction is exactly why machines cannot translate the way humans translate:

L52, English original: human translators can incorporate emotional nuances, humor, and cultural sensitivity into their translations. [...] Algorithm-driven machines, on the other hand, have a hard time dealing with these subjective and emotional factors.

L43, English original: Humans possess cultural and emotional intelligence, allowing them to understand nuances, idioms, and context, which profoundly influence translation accuracy and quality. Emotions imbue human translations with empathy, tone, and cultural sensitivity, making them more adept at capturing the subtleties of language. In contrast, machines rely on algorithms and statistical models to translate, lacking emotional comprehension and cultural awareness.

T05, translated from Finnish: The machine cannot understand intonation, emotion and nuance in language. These, however, affect the translation considerably.

T18, translated from Finnish: The human brings a human tone to the translation; the human understands emotions better than the machine and can play around with words creating a possibly more flowing text.

On the other hand, the concept of ‘experience’ is sometimes interpreted more as ‘gaining experience through practice’ in a way that suggests students are using the Computational Metaphor to draw parallels between humans and machines and the way they ‘learn’ and ‘search for information’:

L19, translated from Dutch: A neural network formed on a large bilingual corpus is to some extent comparable to how a human translator gains ‘experience’ with the two languages over a lifetime.

G10, translated from Dutch: The language data from a translation machine can also be compared with people's language experiences and knowledge, all of which play a role in their translations.

T11, translated from Finnish: The brain has a word storage where one searches for information, which word would fit the translation. Previous experiences can also tell how some term/word has been translated previously and in which context.

T12, translated from Finnish: The human and machine, such as translation memories, both look into their previous knowledge and try to find in their memory correct equivalents for target [sic] language words. They are connected by a relatively mechanical search for equivalents.

A similar example of the Computational Metaphor can be seen in the comment made by T27, who connects databases to both humans and machines, concluding that their translation processes are similar:

T27, translated from Finnish, emphasis added: Both modern MT and a human translator translate based on translation memory, searching **their database** for matches for the specific situation. At this level, the translation process is similar, but the translation memory of a human does not consist of such an extensive repository of options as that of a machine translator.

Here, one problem lies in the fact that many of the central concepts in machine translation, machine learning and artificial intelligence draw on the Computational Metaphor and the

Computational Theory of Mind. The terminology of these fields is fundamentally metaphorical in nature, obscuring the difference between what is essentially human and what typical of machines. In Finnish, this blurring of the two domains is particularly noticeable: both ‘knowledge’ and ‘information’ translate as ‘tieto’, and the word ‘tieto’ also appears in the Finnish equivalents of ‘computer’ (‘tietokone’ – knowledge/information machine) and ‘database’ (‘tietokanta’). Although the majority of the respondents in Finland do think that machines and humans translate differently (Table 1), this may influence some of the students’ thinking and reasoning, when comparing humans and machines. For example:

T12, translated from Finnish, emphasis added: The human and machine translation processes have some similarities, but they are not entirely identical. For example, humans and translation memories, **both explore their prior knowledge** and try to find the correct equivalents of words in the target language. What they have in common, therefore, is a relatively mechanical search for equivalents.

Even though the respondents make a difference between how MT and humans work, they still express themselves in ways that do not make a difference. In the following example from EF03, we have emphasized the words where the original in Finnish has the word ‘tieto’:

EF03, translated from Finnish: One difference between a human translator and a machine is how creatively and extensively humans can use their **knowledge/information** and **information**-seeking skills to find useful sources to come up with a translation. The **knowledge/information** a machine has depends on the texts it is given and how it is programmed to use those texts. Humans, on the other hand, can search more freely for **knowledge/information** from sources that are not necessarily directly related to the subject but can help in the translation process.

The example shows that it has not always been possible to distinguish between “information” and “knowledge” when analysing the responses. Understanding the degree to which a person’s native language and its inherent ambiguities influence their conceptualisations and an analysis on the differences between groups of respondents writing in Finnish, Dutch and English is out of scope for this paper but could be a subject of further research.

6 Concluding remarks

The present study explored how university students enrolled in translation, language and AI courses in Finland, Belgium and the Netherlands understand machine translation and its defining characteristics as compared to human translation, with a focus on what their conceptualisations reveal about the need for more data literacy and AI literacy as part of MT literacy. Building on Salmi *et al.* (2023), the additional data collected showed that across the four universities and three countries approximately 60% of the students considered human and machine translation to be distinct, and around 30% identify both similarities and differences. The characteristics that are assigned to either humans or machines, or both, reveal that students have a relatively clear idea of what sets human and machine translation apart. Most of the quintessentially human characteristics identified can be related to human beings having embodied cognition and world knowledge that encompasses emotions, lived experience, as well as cultural and social awareness.

On the other hand, some reflections revealed a tendency to follow the computational theory of mind in comparing the way humans use their experience to the way a machine’s database and ‘memory’ operate, indications that people may be either humanising the technology or dehumanising people. There were also indications that this tendency may be even stronger for our understanding of LLMs and GenAI, as seen in references to machine ‘hallucinations’ and ‘bias’. It is the terminology itself – using human terms to define machine concepts – that leads to a possibly misleading and potentially dangerous humanisation of MT and AI more generally. There is a clear need to conduct more research on this topic, including a more extensive exploration of how different languages draw the boundaries between humans and machines, i.e. whether they distinguish between knowledge and information, memories and databases. As for pedagogical implications, there is a clear need to include data and AI literacy more systematically in modules where MT is used or taught. Such courses should also take into account informal uses of the technology and explicitly address students' previously formed conceptualisations about translators, both human and machine. MT literacy should help students understand how the technology works without obscuring the differences between humans and machines. At best, this blurring is only misleading and fosters misconceptions. At worst, it lures people into overestimating the machine's capabilities and putting too much trust in its “good intentions”.

References

Anazawa, Ryoko, Hirono Ishikawa, and Kiuchi Takahiro. 2013. Use of Online Machine Translation for Nursing Literature: A Questionnaire-Based Survey. The Open Nursing Journal 7: 22–28. https://doi.org/10.2174/1874434601307010022.

Baria, Alexis T., and Keith Cross. 2021. The Brain Is a Computer Is a Brain: Neuroscience’s Internal Debate and the Social Significance of the Computational Metaphor. arXiv. https://doi.org/10.48550/ARXIV.2107.14042.

Bindels, Joop, and Mark Pluymaekers. 2022. The Use of MT by Undergraduate Translation Students for Different Learning Tasks. *Journal of Data Mining &* Digital Humanities Towards robotic translation? (II. Pedagogical practices): 9019. https://doi.org/10.46298/jdmdh.9019.

Bowker, Lynne. 2020. Machine Translation Literacy Instruction for International Business Students and Business English Instructors. Journal of Business and Finance Librarianship, 25 (1–2): 25–43. https://doi.org/10.1080/08963568.2020.1794739.

Bowker, Lynne. 2021. Machine Translation Use Outside the Language Industries: A Comparison of Five Delivery Formats for Machine Translation Literacy Instruction. In TRITON 2021: Proceedings of the Conference TRanslation and Interpreting Technology ONline, pages 25–36. https://doi.org/10.20381/RUOR-26820.

do Carmo, Félix. 2020. ‘Time Is Money’ and the Value of Translation. Translation Spaces, 9 (1): 35–57. https://doi.org/10.1075/ts.00020.car.

Dorst, Aletta G., Susana Valdez, and Heather Bouman. 2022. Machine Translation in the Multilingual Classroom: How, When and Why Do Humanities Students at a Dutch University Use Machine Translation? Translation and Translanguaging in Multilingual Contexts, 8 (1): 49–66. https://doi.org/10.1075/ttmc.00080.dor.

Ekbia, Hamid, and Bonnie Nardi. 2017. Heteromation and Other Stories of Computing and Capitalism. Cambridge: MIT Press.

Horst, Steven. 1999. Symbols and Computation: A Critique of the Computational Theory of Mind. Minds and Machines, 9 (3): 347–381. https://doi.org/10.1023/A:1008351818306.

Kornacki, Michał, and Paulina Pietrzak. 2025. Hybrid Workflows in Translation: Integrating GenAI into Translator Training. New York, NY: Routledge.

Krüger, Ralph. 2022. Integrating Professional Machine Translation Literacy and Data Literacy. Lebende Sprachen, 67 (2): 247–282. https://doi.org/10.1515/les-2022-1022.

Krüger, Ralph, and Janiça Hackenbuchner. 2022. Outline of a Didactic Framework for Combined Data Literacy and Machine Translation Literacy Teaching. Current Trends in Translation Teaching and Learning E: 375–432. https://doi.org/10.51287/cttl202211.

Läubli, Samuel, and David Orrego-Carmona. 2017. When Google Translate Is Better than Some Human Colleagues, Those People Are No Longer Colleagues. In Proceedings of the 39th Conference Translating and the Computer, pages 59–69. London, UK: ASLING.

Long, Duri, and Brian Magerko. 2020. What Is AI Literacy? Competencies and Design Considerations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–16. Honolulu HI USA: ACM. https://doi.org/10.1145/3313831.3376727.

Loock, Rudy, and Sophie Léchauguette. 2021. Machine Translation Literacy and Undergraduate Students in Applied Languages: Report on an Exploratory Study. Revista Tradumatica, 19: 204–225. https://doi.org/10.5565/rev/tradumatica.281.

Moorkens, Joss. 2022. Ethics and Machine Translation. In Dorothy Kenny, editor, Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence, pages 121–40. Translation and Multilingual Natural Language Processing 18. Berlin: Language Science Press.

Ng, Davy Tsz Kit, Jac Ka Lok Leung, Samuel Kai Wah Chu, and Maggie Shen Qiao. 2021. Conceptualizing AI Literacy: An Exploratory Review. Computers and Education: Artificial Intelligence, 2:100041. https://doi.org/10.1016/j.caeai.2021.100041.

Nurminen, Mary. 2019. Decision-Making, Risk, and Gist Machine Translation in the Work of Patent Professionals. In Proceedings of the 8th Workshop on Patent and Scientific Literature Translation, pages 32–42. Dublin, Ireland.

Nurminen, Mary. 2021. Investigating the Influence of Context in the Use and Reception of Raw Machine Translation. Doctoral dissertation, Finland: Tampere University. https://urn.fi/URN:ISBN:978-952-03-2199-4.

Placani, Adriana. 2024. Anthropomorphism in AI: Hype and Fallacy. AI and Ethics, 4 (3): 691–698. https://doi.org/10.1007/s43681-024-00419-4.

Risku, Hanna, and Regina Rogl. 2020. Translation and Situated, Embodied, Distributed, Embedded and Extended Cognition. In Fábio Alves and Arnt Lykke Jakobsen, editors, The Routledge Handbook of Translation and Cognition, pages 478–499. London: Routledge.

Salmi, Leena, Aletta G. Dorst, Maarit Koponen, and Katinka Zeven. 2023. Do Humans Translate Like Machines? Students' Conceptualisations of Human and Machine Translation. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation, pages 295–304. Tampere, Finland: European Association for Machine Translation. https://aclanthology.org/2023.eamt-1.29/.

Salmi, Leena, and Maarit Koponen. 2022. Do humans translate like machines? Translator students' conceptualisations of human and machine translation processes. Paper presented at the New Trends in Translation and Technology Conference, Rhodes.

Vieira, Lucas Nunes. 2020. Machine Translation in the News: A Framing Analysis of the Written Press. Translation Spaces, 9 (1): 98–122. https://doi.org/10.1075/ts.00023.nun.

Vieira, Lucas Nunes, Minako O'Hagan, and Carol O'Sullivan. 2020. Understanding the Societal Impacts of Machine Translation: A Critical Review of the Literature on Medical and Legal Use Cases. Information Communication and Society, 24 (11): 1515–1532. https://doi.org/10.1080/1369118X.2020.1776370.

Vieira, Lucas Nunes, Carol O'Sullivan, Xiaochun Zhang, and Minako O'Hagan. 2022. Machine Translation in Society: Insights from UK Users. Language Resources and Evaluation, 57: 893–914. https://doi.org/10.1007/s10579-022-09589-1.

.

DeepL abbreviation errors for clinical trial protocols: quality estimation, error classification and health implications

Alicia Picazo-Izquierdo

Universidad de Alicante, Spain

alicia.picazo@ua.es

Abstract

This study explored the use of DeepL for translating clinical trial protocols from English to Spanish and evaluated its quality in translating abbreviations. A corpus of thirty-five clinical trial protocols dealing with diverse health problems was collected from the US National Health Institute database. The DeepL Pro machine translation was used to translate the corpus from English to Spanish. Then, the parallel corpus was aligned and further processed in Sketch Engine to extract abbreviations in context alongside their matching translations. After filtering and deleting abbreviations referring to proper names, trademarks, official names, and titles to analyse only the concept-defining ones, a total of 379 abbreviations were obtained. Results show that less than forty per cent had been translated. More than fifty per cent of translated abbreviations generated errors. With the quantitative data obtained, errors were divided into three categories: inconsistency due to original and translated form, inconsistency caused by several translations, and inconsistency originated by the abbreviation being used for different concepts. Finally, other error categories involved in abbreviation translation are listed.

1 Introduction

Clinical trial protocols are documents that describe how a clinical trial is conducted. Protocols include information about the objectives, design, methodology, statistics, security, and privacy of the study (Chan et al., 2015). Clinical trials are sometimes conducted in several countries, which means that translation into different languages is needed, and machine translation post-editing (MPTE) is often used (Trujillos-Yébenes & Muñoz-Miquel, 2022). The reason for using MTPE is that protocols contain highly specialised language, syntactical structures, formed to be objective, and concept descriptions that elude figurative language. Clinical language also makes use of abbreviations and acronyms (Navarro, 2008) (hereafter, "abbreviations") to represent clinical concepts related to procedures, devices, or conditions. However, unreferenced abbreviations can be difficult to understand, as they may be globally accepted concepts (i.e., BMI for "body mass index") or concepts of a particular domain or study (i.e., PM for "particulate matter" in a technical text or for "project manager"). Abbreviations can be defined, but often are not defined, and often are made up ad hoc by individual physicians (Cohen, 2022). This generates ambiguity, since one abbreviation can have several meanings and one concept may be represented by different abbreviations. It can also lead to the physician misinterpreting the content (Soto-Arnaez et al., 2019; Jayatilake & Oyibo, 2023) and medication errors (PA-PSRS, 2005).

Natural Language Processing (NLP) systems have been developed to extract information, which can be used for different applications, such as decision support systems (Xu et al., 2007). Abbreviations are one of the main NLP challenges, both in terms of segmentation and meaning extraction. When translating abbreviations, automatic translation can generate errors and face some obstacles. Firstly, extracting the meaning for unreferenced abbreviations can be difficult in monolingual and bilingual texts. Secondly, abbreviation disambiguation (Pakhomov et al., 2005) is a challenge for medical NLP applications (Pesaranghader et al., 2019; Jin et al., 2019) and therefore, machine translation.

This paper is structured as follows. Section 2 discusses related work on abbreviation processing and translation. Section 3 presents the methodology used for the dataset creation and for abbreviation detection and extraction. In Section 4, the results are presented in lists of abbreviations, and error categories are discussed. Finally, Section 5 introduces some conclusions and potential future work.

2 Related work

Computational linguistics has irrupted in many professional areas, and the clinical domain is no exception. NLP provides new resources for clinical information retrieval in several applications, such as clinical notes, biomedical literature, and pharmacovigilance. Whereas clinical data can be structured data, such as demographics, diagnosis, and personal information, nearly 80% are unstructured data according to Li (2019) that can be extracted thanks to NLP technologies. This data consists of clinical notes, patient-provided information, health reports and the use of abbreviations, among others. With NLP, clinical data can be processed and used to identify new information and make predictions (Liu et al., 2018; Huang et al., 2020; Murff et al., 2011; Yim et al., 2016; Tsui et al., 2021). However, Ulitkin et al. (2020) point out some difficulties in the interlanguage adaptation of abbreviations since (i) abbreviations are associated with the presence of a common terminological base and personal experience; (ii) polysemy of abbreviations creates difficulties in recognising their semantic content; (iii) abbreviations can be indicative of various parts of speech and express different syntactic functions; (iv) the variability of corresponding translations can conflict with the requirement of a single meaning and unambiguous equivalence in the target language; and (v) syncretism and homonymy play the role of destructive interference at the level of word formation.

NLP deals with abbreviations in different ways. Many efforts are focused on abbreviation disambiguation. Firstly, abbreviation databases are created to train models (Zhou et al., 2006; Grossman et al., 2021; UMLS Reference Manual, 2016). These databases contain abbreviation definitions and act as a glossary, but the fast development of abbreviations or spontaneously abbreviated terms (where the abbreviation cannot be recognised as it is created at that moment) makes it difficult to keep databases updated. There is also a tendency to create models to extract abbreviation definitions through automatic identification and expansion (Jin et al., 2019; Wu et al., 2016; Kreuzthaler et al., 2016). If abbreviation expansion can be achieved at a high level of accuracy, there would be no problem in finding the right meaning, so information extraction and translation could be improved. In this direction, studies focus on technologies to achieve abbreviation sense disambiguation (Moon et al., 2015; Li et al., 2019; Wu et al., 2015; Jaber & Martinez, 2021; Choi & Taghva, 2023).

Machine translation of abbreviations relies on how an MT system deals with disambiguation and will perform according to the available data and models. Accurate and unambiguous abbreviation is important in any field of science and especially in medicine (Shalajeva, 2016). Proposals for clinical abbreviation translation are focused on locating the branch of medicine and reference terminological dictionaries (Kuzmina, Fominykh & Abrosimova, 2015), but finding their unabridged form is often obscure and not easy (Kasprowicz, 2010). Moreover, neural machine translation systems (NMT) can generate inconsistencies (Nitzke, 2019; Forcada, 2017) due to their technical features. For this reason, it is essential to highlight the abbreviation problem in the process of post-editing clinical trial protocols.

3 Methodology

This project aims at exploring how NMT DeepL engine performs when translating abbreviations in clinical trial protocols from English to Spanish. The focus is to check consistency and to detect translation errors. Abbreviation sense is also extracted.

For this purpose, the dataset consisted of clinical trial protocols gathered in an English monolingual corpus, which was translated into Spanish to create a bilingual corpus. The English clinical trial protocols were downloaded from the US National Health Institute to obtain a balanced, representative corpus of 35 complete documents (over 170,000 thousand words). The topics were both technical and clinical, and dealt with different investigational products, devices, techniques, and conditions. Provided with this topic diversity, it is possible to observe abbreviations in the context of different technical sublanguages.

English texts were converted into .txt format and translated automatically with the NMT tool DeepL Pro. They were then aligned automatically in LF Aligner and merged to create an .xml bilingual file. Once the bilingual corpus .xml file was prepared, it was uploaded into Sketch Engine, a corpus management tool that annotates texts automatically. To find abbreviations, a query was created in the Parallel Concordance tool. A simple regular expression[1] was used to look for abbreviations with the Corpus Query Language query type.

After extracting the abbreviations, a manual filtering was performed to remove false positives. The aim of this filtering was to exclude global abbreviations that must not be translated because they represent officially accepted concepts. Since the goal of this paper was to discover inconsistency in abbreviation translation and to find the reasons for this inconsistency, the criteria for excluding abbreviations were as follows: abbreviations of proper names of entities, countries, regions, or states, such as ANOVA, USA, or FL; trademarks, i.e., ASUS; sequences of letters that indicate order or procedure (AB or BA); professional titles, i.e., MD, BA, PhD; clinical studies official names, such as BEUTI; and measurement scales, i.e., MG-ADL, MIRS.

The remaining abbreviations were then classified in terms of translation, occurrence, and consistency, and were presented with the sense (or senses, if necessary).

4 Results and discussion

A total of 379 abbreviations were obtained and subsequently analysed in Sketch Engine. A list of filtered abbreviations was extracted and manually divided into error categories. The first data obtained were relative to the general translation consistency. Some abbreviations were

[1] Query:[word="[[:upper:]]*"]

translated, and others were expressed as in the source language. The following table contains the absolute frequency of consistency in all abbreviations, both translated and untranslated.

		AF	% AF
Repeated abbreviations	Consistency	126	33.24%
	Inconsistency	102	26.92%
	DNT	29	7.65%
Unique abbreviations (only one occurrence)		122	32.19%

Figure 1. abbreviation translation consistency.

When comparing translated to untranslated abbreviations, only 36.67% had been translated (139 out of 379). The 51.72% (196 out of 379) were untranslated, omitted or described (hereafter “untranslated”). Then, the remaining 11.61% (44 out of 379) were abbreviations with the same form both in English and Spanish and, therefore, were Do Not Translate (DNT) words.

Abbreviation translation	AF	% AF
Translated abbreviations	139	36.67%
Untranslated abbreviations	196	51.72%
Same abbreviation in both languages	44	11.61%

Figure 2. abbreviation translation

To obtain the consistency of abbreviations that had been translated, abbreviations that only occurred once in the corpus were removed. After this removal, a total of 116 abbreviations were translated and found in the corpus more than once. This meant that the remaining 23 abbreviations that had been translated and only occurred once in the corpus were not of interest to this project. On account of that, out of the 196 untranslated abbreviations, 84 only appeared once, and 112 more than once in the corpus. The following figure shows the number of abbreviations in each category based on occurrence, translation, and consistency.

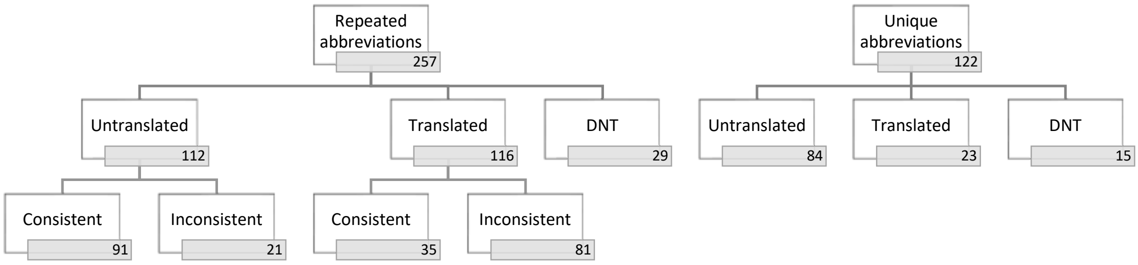


Figure 3. diagram of abbreviations based on occurrence, translation, and consistency

The most outstanding data are those related to the inconsistency of translated abbreviations. Clearly, there is no consistency when translating abbreviations (81 out of 116 abbreviations were inconsistent, which is a 69.82%). It is also notable that consistency is higher in untranslated abbreviations. This may be because there is no sense defined in context, so the NMT engine decides not to translate. Additionally, it is worth noting that almost 50% of the abbreviations were untranslated (excluding DNTs).

A. *Inconsistency in abbreviations with translated and untranslated forms*

The first error category is inconsistency due to two translation proposals. One abbreviation has both translated and untranslated forms in different segments. The following table illustrates some examples of abbreviations and its sense. The first column covers the abbreviation of the original text and the second one defines its sense in English. The third column presents different forms of the abbreviation in translated segments alongside its sense in the fourth column. Finally, the fifth shows the number of times of each abbreviation in the translated text.

OT	**SENSE**	**TT**	**SENSE**	**#**
AI	artificial intelligence	AI	inteligencia artificial	5
		IA		2
ALI	acute limb ischemia	IAM	isquemia aguda de las extremidades	1
		ALI		1
AMS	anxiety management strategies	AMS	estrategias de manejo de la ansiedad	1
		EMA		2
AP	artificial pancreas	PA	páncreas artificial	7
		AP		8
ATD	antithyroid drug	ATD	fármaco antitiroideo	7
		TCA		2
BE	behavioral economics	BE	economía conductual	3
		HCE		2
CA	clavulanic acid	AC	ácido clavulánico	81
		CA		7
CBC	complete blood count	CBC	Hemograma	1
		-		1
CBG	capillary blood glucose	CBG	glucemia capilar	11
		GSC		1
CGM	continuous glucose monitoring	MCG	monitoreo continuo de glucosa	5
		CGM		9
CIP	clinical investigation plan	CIP	plan de investigación clínica	2
		PIC		7
CRPS	complex regional pain syndrome	SDRC	síndrome del dolor regional complejo	1
		CRPS		1
CT	computed tomography	TC	tomografía computarizada	16
		CT		15
DED	dry eye disease	DED	enfermedad del ojo seco	9
		EOS		2
DFA	discriminant function analysis	DFA	análisis de función discriminante	3
		AFD		2
EBA	early bactericidal activity	ABE	actividad bactericida temprana	1
		EBA		21
EDS	Ehlers-Danlos syndrome	SED	síndrome de Ehlers-Danlos	1
		EDS		1
EHR	electronic health record	HCE	historia clínica electrónica	14
		EHR		4
FSS	fibromyalgia severity score	FSS	puntuación de la gravedad de la fibromialgia	3
		SFS		1

FVC	forced vital capacity	FVC	capacidad vital forzada	6
		CVF		1
GI	gingival index	GI	índice gingival	1
		IG		3
HA	healthy adult	AS	adulto sano	2
		HA		5
ICC	intra-class correlation coefficient	ICC	coeficiente de correlación intraclase	2
		CCI		1
ICF	informed consent form	FCI	formulario de consentimiento informado	1
		ICF		2
ICG	indocyanine green	ICG	verde de indocianina	38
		GCI		1
ICS	inhaled corticosteroid	ICS	corticosteroides inhalados	1
		CSI		1
IP	investigational product	IP	producto en investigación	28
		PI		23
ISC	independent scientific group	ISC	grupo científico independiente	3
		omitted		1
LV	last visit	VI	última visita	1
		LV		2
MAE	mean absolute error	MAE	error medio absoluto	5
		EAM		2
MEP	motor evoked potentials	PEM	potenciales evocados motores	1
		MEP		2
MGI	modified gingival index	MGI	índice gingival modificado	7
		IGM		1
MPC	model predictive control	CPM	control predictivo de modelos	1
		MPC		20
NPC	new patient coordinator	NPC	coordinador de pacientes nuevos	1
		CPN		1
NSR	neurostimulation registry	NSR	estimulación y registro neurales	1
		omitted		1
NTE	non-tailpipe emissions	NTE	emisiones no procedentes del tubo de escape	3
		ENT		2
OLS	ordinary least squares	OLS	mínimos cuadrados ordinarios	4
		MCO		2
PEMF	pulsed electromagnetic field	PEMF	campo electromagnético pulsado	4
		BEMER		1
PET	positron emission tomography	PET	tomografía por emisión de positrones	29
		TEP		2
PHI	protected health information	PHI	información sanitaria personal	12
		IPS		1
PI	principal investigator	IP	investigador principal	78
		PI		10
PK	pharmacokinetics	PK	Farmacocinética	14
		FC		1
		PMCF		13

PMCF	post-market clinical follow-up	-	estudio de seguimiento clínico poscomercialización	1
PMG	project management group	PMG	grupo de gestión del proyecto	3
		-		1
PPI	public and patient involvement	PPI	participación pública y de pacientes	6
		IPP		4
QA	quality assurance	QA	garantía de calidad	2
		-		1
QC	quality control	QC	control de calidad	2
		-		2
R&D	research and development	I+D	investigación y desarrollo	2
		R&D		1
RN	registered nurse	RN	Enfermero	1
		-		1
RP	renal profile	PR	perfil renal	1
		RP		1
SCS	spinal cord simulation	EME	estimulación de la médula espinal	3
		SCS		6
SD	standard deviation	SD	desviación estándar	3
		DE		8
SE	standard error	SE	error estándar	2
		EE		2
SLE	systemic lupus erythematosus	LES	lupus eritematoso sistémico	53
		SLE		8
SOP	standard operation procedures	PNT	procedimientos normalizados de trabajo	18
		SOP		5
SSC	surviving sepsis campaign	SSC	campaña sobrevivir a la sepsis	1
		CDC		1
TFT	thyroid function test	TFT	pruebas de función hepática	2
		-		1
TLC	total lung capacity	TLC	capacidad pulmonar total	2
		CPT		1
TMIC	time over minimum inhibitory concentration	TMIC	tiempo sobre la concentración inhibitoria mínima	3
		CIM		1
ULN	upper limit of normal	LSN	límite superior de la normalidad	1
		ULN		2

Figure 4. inconsistency in abbreviations with translated and untranslated forms

Attention should be drawn to some abbreviations that clearly illustrate the inconsistency problem. CGM (continuous glucose monitoring), is translated five times as MCG (the Spanish adaptation that means “continuous glucose monitoring”) and the English form has been kept nine times. In the same line is CT, computed tomography, which appears sixteen times as TC (Spanish adapted) and fifteen times in its original form, CT. Another notable example of this first case of inconsistency is IP (investigation product). In twenty-eight instances the English abbreviation IP has been used, while in twenty-three the translation PI has been used.

Aside from the inconsistency in translation decision-making, there are some errors that have been produced by the wrong sense choice. The context of abbreviations with different meanings can be misunderstood, which, as a consequence, generates wrong translations. For instance, ATD (antithyroid drug) has been translated as TCA twice, while this means “eating disorder” in Spanish. Likewise, the abbreviation that stands for “behavioral economics” (BE), is untranslated and translated as HCE. Both options are mistranslations, and the meaning could not be extracted without the original text. The abbreviations FFS and SSC are also mistranslated.

B. Abbreviations with more than one sense

The second category is related to abbreviations with more than one meaning. These abbreviations were used in the original corpus to represent different concepts. Therefore, they can cause inaccuracies both in the original text and in the translated text and are the root cause for mistranslations. Abbreviations that are already commonly used in one domain to designate a concept are reused for other concepts of the same domain or otherwise.

To illustrate this, the abbreviation CI (confidence interval) is widely accepted in statistics. However, the same abbreviation is used for “chief investigator”, which causes variation. In addition, in the Spanish translation this representation has not been detected, but both translation possibilities (CI and IC) correspond to the concept “confidence interval”. Another example that represents this problem is the acronym SOC. In medical language and, in particular, clinical trial protocols, a SOC is a standard treatment, which is usually translated as TE in Spanish. Instead, this same acronym has been used for “system organ class”. Although it only appeared once, not translating the acronym SOC and not detecting this new denomination is a translation error. Some other examples are shown in the table below.

OT	**SENSE**	**TT**	**SENSE**	**#**
BAL	benralizumab on the airway	BAL	benralizumab en la vía aérea	1
		LBA		1
	bronchoalveolar lavage	BAL	lavado broncoalveolar	6
		LBA		1
CI	chief investigator	IC	investigador jefe	7
		CI		2
	confidence interval	IC	intervalo de confianza	19
MI	myocardial infarction	IM	infarto de miocardio	1
	mental imagery	MI	imágenes mentales	7
		IM		1
PD	Parkinson’s disease	EP	enfermedad de Parkinson	28
		PD		5
	probing depths	PD	profundidades de sondaje	4
PM	particulate matter	PM	materia particulada	11
	project manager	PM	gestor de proyectos	2
POC	proof of concept	POC	prueba de concepto	2
	point of care		punto de atención	4
SOC	standard of care	SOC	tratamiento estándar	11

	system organ class		clase de órgano del sistema	1

Figure 5. abbreviations with more than one sense

C. *Inconsistency due to translation variation*

The third category refers to abbreviations with variation in the translation. This variation can be produced by the translation strategy (amplification or explanation) or by the lexical choice in Spanish. It is easily illustrated with the case of "evento adverso" or "acontecimiento adverso". The abbreviation AE (adverse event) can be translated as AA or EA, and it would have exactly the same meaning. There are also other examples of abbreviations untranslated and two or more translations, as in CRF. Another notable case is the use of two different abbreviations in Spanish, such as EDD (estimated date of delivery), which has been translated as FPE and FEP, meaning "fecha del parto estimada" and "fecha estimada del parto", respectively. The following table shows more examples.

OT	**SENSE**	**TT**	**SENSE**	**#**
AE	adverse event	EA	evento adverso	62
		AA	acontecimiento adverso	2
		EAs	eventos adversos	1
		acontecimientos adversos	-	2
		AE	-	1
AKI	acute kidney injury	LRA	lesión renal aguda	9
		IRA	inflamación renal aguda	3
CRF	case report form	CRF	-	20
		FCI	formulario de casos informados	1
		FRC	formulario de recogida de casos	2
CXR	chest X ray	RxC	rayos X de tórax	1
		CXR		1
		RXC		1
DTA	descending thoracic aorta	DTA	aorta torácica descendente	1
		aorta torácica descendente	-	2
EBC	exhaled breath condensate	CPE	condensado espiratorio exhalado	1
		CBE	-	1
		EBC	-	21
EDD	estimated date of delivery	FPE	fecha del parto estimada	2
		FEP	fecha estimada del parto	1
EMG	electromyography	EMG	electromiografía	22
		electromiografía	-	1
EMR	electronic medical record	EMR	-	12
		HCE	historia clínica electrónica	1

GA	gestational age	EG	edad gestacional	18
		GA	-	5
		AG	-	9
GCP	good clinical practice	GCP	-	2
		BPC	buenas prácticas clínicas	15
		buenas prácticas clínicas	-	1
HEI	high-education institute	IES	instituto de educación superior	1
		instituto de educación superior	-	1
HFJV	high-frequency jet ventilation	VChAF	ventilación por chorro de alta frecuencia	1
		VHJF	-	8
		HFJV	-	7
		VNFH	-	2
ID	identification	ID	Identificación	7
		identificación	-	10
		identificador	-	3
		número de identificación	-	8
		identificación numérica	-	1
IEC	independent ethics committee	CEI	comité de ética independiente	4
		CEIC	comité de ética independiente del centro	1
IRB	institutional review board	CEI	comité de ética independiente	42
		CRI	comité de revisión independiente	2
		JRI	junta de revisión independiente	2
		CIR	comité independiente de revisión	1
		IRB	-	34
IT	information technologies	servicios informáticos	-	1
		entorno informático	-	1
		sistema informático	-	1
LAR	legal authorized representative	LAR	-	14
		representante legal	-	9
		cuidador	-	2

		cuidador o tutor	-	2
LMP	last menstrual period	FUM	fecha última menstruación	10
		FUR	fecha última regla	2
		LMP	-	2
MDR	multi-drug resistant	MDR	tuberculosis multirresistente	6
		TB-MDR		4
		MDR-TB		1
MGD	Meibomian gland dysfunction	DGM	disfunción de las glándulas de Meibomio	6
		MGD	-	6
		DMG	-	1
MIC	minimum inhibitory concentration	CIM	concentración inhibitoria mínima	6
		MIC	-	1
		CMI	concentración mínima inhibitoria	7
MRI	magnetic resonance imaging	RM	resonancia magnética	3
		resonancia magnética	-	3
NICU	neonatal intensive care unit	UCIN	unidad de cuidados intensivos neonatal	2
		UCI neonatal	-	1
NSCLC	non-small cell lung cancer	NSCLC	-	1
		CPNM	cáncer de pulmón no microcítico	17
		CPCNP	cáncer de pulmón de células no pequeñas	5
OS	operating system	SO	sistema operativo	1
		sistema operativo	-	2
PAD	peripheral arterial diseases	EAP	arteriopatía periférica	7
		arteriopatía periférica	-	2
PDN	Parkinson’s disease without cognitive impairment	PDN	-	3
		NDP	-	2
		EPN	enfermedad de Parkinson sin deterioro cognitivo	1
		EP-ND	enfermedad de Parkinson no deterioro	1
QCA	qualitative comparative analysis	QCA	-	2
		ACC	análisis cualitativo comparativo	2
		ACQ	-	3
RAI	radioactive iodine	IRA	yodo radiactivo	13
		RAI	-	26
		IAR	-	2

		yodo radiactivo	-	2
REC	research ethics committee	CEI	comité de ética de investigación	6
		CEIC	comité de ética de investigación del centro	6
		REC	-	1
RRT	renal replacement therapy	TRS	tratamiento renal sustitutivo	1
		TRR	terapia de reemplazo renal	4
SAE	serious adverse event	SAE	acontecimiento/evento adverso grave/serio	8
		EFG		1
		EAE		16
		-		2
		EAS		4
		EAG		2
		AAG		1
		omitted		1
SS	Strava Sun	SS	Strava Sun	62
		protección solar	-	1
		seguridad solar	-	2
TB	tuberculosis	TB	Tuberculosis	53
		-		36
		omitted		1
		antituberculosos		5
		tuberculosa		2
		TBC		1
TED	thyroid eye disease	EOT	enfermedad ocular tiroidea	1
		enfermedad ocular tiroidea	-	1
US	ultrasound	ecografía	-	1
		omitted		1

Figure 6. inconsistency due to translation variation

Concordance is also noteworthy, because some errors are generated by the omission or misuse of an essential article in Spanish. For instance, one abbreviation that is not present in the previous tables is BOP, whose meaning is “bleeding upon probing”. In context, “reductions in BOP” should be translated as “reducciones en el BOP”. However, the article “el” has been omitted, probably because concordance could not be solved due to the unknown sense of this abbreviation. Other concordance errors can be produced by ambiguity. For example, the abbreviation BOE, meaning “best obstetric estimate”, was not translated, but left in English. In Spanish, BOE is widely known as “Boletín Oficial del Estado” (which is masculine), so the chosen article was masculine. The correct translation would be “mejor estimación obstétrica”, which is feminine.

As a result of this analysis, some trends related to the abbreviation translation can be observed. Firstly, abbreviation translation is more accurate when a description of the abbreviation is provided. Instances in which the abbreviation is followed by its sense (whether in brackets or not) show more accuracy than those without the sense. Nevertheless, the abbreviation that is correctly translated in that case can be mistranslated afterwards, even within the same text. To illustrate this, the abbreviation CRPS, which means "Complex Regional Pain Syndrome", is correctly translated in its first occurrence with the sense in brackets (SDRC in Spanish for "síndrome de dolor regional complejo"). However, the second occurrence in the same text is not translated.

It has also been found that consistency is higher when abbreviations are untranslated. But this generates translation errors when untranslated abbreviations can be misunderstood in the target language, or there is already an official abbreviation in that language. There should be instructions to abbreviation translation based on the nature of the project, the target audience, and the scope of the text. Finally, it has been observed that if there is no possible translation for any abbreviation, the NMT engine follows some reorganization rules based on word order, as in QCA – ACQ; PDN – NDP).

5 Conclusions and future work

In this paper, we analysed the problem of DeepL inconsistency when translating abbreviations from English to Spanish. Different translation errors generated by incorrect senses, omissions or mistranslations were detected. As a conclusion, this paper contributes to the well-known problem of abbreviation translation for NMT systems, and sheds light on how errors are generated, as well as the processes that result in the wrong translation decisions. These results are also useful for NMT post-editing purposes. It shows that abbreviations are a potential problem due to inconsistencies and mistranslations. Some strategies for abbreviation error mitigation are the creation of task-specific glossaries that can be applied to the NMT tool.

Abbreviation mistranslation is a major concern in clinical applications as it can be the root cause of health risks or inaccurate patient care. Some examples described in this article show that abbreviations provide incorrect information about diseases, treatments, devices, or care methods. For instance, if TCA is used for "antithyroid drug" in a Spanish text, the most possible option for interpretation would be "eating disorder", which could lead to major care decisions.

Although abbreviation processing (and therefore translation) remains a challenge, NLP efforts are focused on disambiguation of abbreviations and sense detection in order to address this issue. For future work, a more exhaustive analysis of abbreviation translation in clinical trial protocols will be carried out with the purpose of obtaining a higher sample of the problem. It would also be beneficial to collect a larger sample of abbreviations in context to describe the process of abbreviation creation and make it available for NLP applications.

References

Chan, An-Wen; Jennifer Tetzlaff, Douglas Altman et al. (2015) "Declaración SPIRIT 2013: definición de los elementos estándares del protocolo de un ensayo clínico". *Revista Panamericana de Salud Pública* 38:6, pp. 506-514. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5114122/pdf/nihms6261.pdf

Choi, Vice Sing & Kazem Taghva. (2023) "Abbreviation Disambiguation: A Review of Modern Techniques to Improve Machine Reading Comprehension". In: Arai, K. (ed.) 2023. *Intelligent Computing. SAI 2023. Lecture*

Notes in Networks and Systems 711, pp. 250-261. https://link.springer.com/chapter/10.1007/978-3-031-37717-4_17#citeas

Cohen, Kevin B. (2022) "Natural Language Processing for Biomedical Texts". In: Mitkov, Ruslan (ed.) 2022. *The Oxford Handbook of Computational Linguistics.* New York: Oxford University Press, pp. 1133-1164.

Forcada, Mikel. (2017) "Making sense of neural machine translation". *Translation Spaces* 6:2, pp. 291-309. https://www.dlsi.ua.es/~mlf/docum/forcada17j2.pdf

Grossman, Lisa et al. (2021) "A Deep database of medical abbreviations and acronyms for natural language processing". *Scientific Data* 8:149, pp. 1-9. https://pubmed.ncbi.nlm.nih.gov/34078918/

Huang, Kexin; Jaan Altosaar & Rajesh Ranganath. (2020) "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission". Preprint at https://arxiv.org/abs/1904.05342

Jaber, Areej & Paloma Martinez. (2021) "Disambiguating Clinical Abbreviations using Pre-trained Word Embeddings". *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies HEALTHINF (BIOSTEC)* 5, pp. 501-508. https://www.scitepress.org/PublishedPapers/2021/102561/102561.pdf

Jayatilake, Dineth & Samson O. Oyibo. (2023) "Interpretation and Misinterpretation of Medical Abbreviations Found in Patient Medical Records: A Cross-Sectional Survey". *Cureus* 15:9, pp. 1-8. https://pmc.ncbi.nlm.nih.gov/articles/PMC10479966/pdf/cureus-0015-00000044735.pdf

Jin, Qiao; Jinling Liu & Xinghua Lu. (2019) "Deep Contextualized Biomedical Abbreviation Expansion". In: Demner-Fushman et al., (eds.) 2019. *Proceedings of the 18th BioNLP Workshop and Shared Task.* Florence: Association for Computational Linguistics, pp. 88-96. https://aclanthology.org/W19-5010/

Kasprowicz, Malgorzata. (2010) "Handling Abbreviations and Acronyms in Medical Translation". *Medical Translations* 14:2. http://www.translationjournal.net/journal/52abbreviations.htm

Kreuzthaler, Markus; Michel Oleynik, Alexander Avian & Stefan Schulz. (2016) "Unsupervised Abbreviation Detection in Clinical Narratives". *Proceedings of the Clinical Natural Language Processing Workshop*, pp. 11-17. https://aclanthology.org/W16-4213/

Kuzmina, Olga D.; Anna D. Fominykh & Natalia A. Abrosimova. (2015) "Problems of the English abbreviations in medical translation". *Procedia – Social and Behavioral Sciences* 199:3, pp. 548-554. https://www.sciencedirect.com/science/article/pii/S1877042815045565?ref=cra_js_challenge&fr=RR-1

Li, Irene et al. (2019) "A Neural Topic-Attention Model for Medical Term Abbreviation Diambiguation". *Proceedings of the Machine Learning of Health Workshop*, pp. 1-9. https://arxiv.org/abs/1910.14076

Li, Rumeng. (2020) *Overview of NLP in Clinical Domain. Applications of Natural Language Processing.* Amherst: University of Massachusetts. https://people.cs.umass.edu/~brenocon/cs490a_f20/lectures/20-clinical-nlp-li.pdf

Liu, Jingshu; Zachariah Zhang & Narges Razavian. (2018) "Deep EHR: Chronic Disease Prediction Using Medical Notes". *Proceedings of the 3rd Machine Learning for Healthcare Conference* 85, pp. 440-464. https://arxiv.org/pdf/1808.04928

Moon, Sungrim; Bridget McInnes & Genevieve B. Melton. (2015) "Challenges and Practical Approaches with Word Sense Disambiguation of Acronyms and Abbreviations in the Clinical Domain". *Healthc. Inform. Res.* 21:1, pp. 35-42. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4330198/

Murff, Harvey J. et al. (2011) "Automated identification of postoperative complications within an electronic medical record using natural language processing". *JAMA* 206:8, pp. 848-855. https://pubmed.ncbi.nlm.nih.gov/21862746/

Navarro, Fernando. (2008) "Repertorio de siglas, acrónimos, abreviaturas y símbolos utilizados en los textos medicos en español". *Panacea* 9:27, pp. 55-59. https://www.tremedica.org/wp-content/uploads/n27_tradyterm-navarro.pdf

Nitzke, Jean. (2019) *Problem solving activities in post-editing and translation from scratch.* Berlin: Language Science Press. https://langsci-press.org/catalog/book/196

PA-PSRS, Pennsylvania Patient Safety Reporting System. (2005) "Abbreviations: a shortcut to medication errors". Patient Safety Advisory 2:1, pp. 19-21. https://patientsafety.pa.gov/ADVISORIES/Pages/200503_19.aspx

Pakhomov, Serguei; Ted Pedersen & Christopher G. Chute. (2005) "Abbreviation and Acronym Disambiguation in Clinical Discourse". *AMIA Symposium Proceedings 2005*, pp. 589-593. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1560669/#b4-amia2005_0589

Pesaranghader, Ahmed; Stan Matwin, Marina Sokolova & Ali Pesaranghader. (2019) "deepBioWSD: effective deep neural word sense disambiguation of biomedical text data". *J. Am. Med. Informatics Assoc.* 26:5, pp. 438–446. https://pubmed.ncbi.nlm.nih.gov/30811548/

Shalajeva, A. V. (2016) “Translation of abbreviations in medical texts. The main ítems to be pointed out”. *Науковий вісник Міжнародного гуманітарного університету. Серія «Філологія»* 25:2, pp. 204-207. https://europub.co.uk/articles/translation-of-abbreviations-in-medical-texts-the-main-items-to-be-pointed-out-A-459798

Soto-Arnaez, Francisco et al. (2019) “A descriptive study of the knowledge of nurses and doctors of clinical abbreviations in hospital discharge reports”. *Enfermería Clínica* 29:5, pp. 302-307. https://www.sciencedirect.com/science/article/pii/S1130862118302547?via%3Dihub

Trujillos-Yébenes, Lorena & Ana Muñoz-Miquel. (2022) “La traducción automática y la posedición en el ámbito médico”. *Tradumàtica* 20, pp. 57-76. https://ddd.uab.cat/pub/tradumatica/tradumatica_a2022n20/tradumatica_a2022n20p57.pdf

Tsui, Fuchiang R. et al. (2021) “Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts”. *JAMIA Open* 4:1, pp. 1-13. https://academic.oup.com/jamiaopen/article/4/1/ooab011/6174413

Ulitkin, Ilya; Irina Filipova, Natalia Ivanova & Yuriy Babaev. (2020) “Use and translation of abbreviations and acronyms in scientific texts”. *E3S Web of Conferences* 210, pp. 1-12. https://www.e3s-conferences.org/articles/e3sconf/pdf/2020/70/e3sconf_itse2020_21006.pdf

UMLS. (2021) *Unified Medical Language System Reference Manual*. Bethesda: National Library of Medicine. https://www.ncbi.nlm.nih.gov/books/NBK9676/pdf/Bookshelf_NBK9676.pdf

Wu, Yonghui et al. (2016) “A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (CARD)”. *Journal of the American Medical Informatics Association* 24:1, pp. 79-86. https://academic.oup.com/jamia/article/24/e1/e79/2631496

Wu, Yonghui; Jun Xu, Yaoyun Zhang & Hua Xu. (2015) “Clinical Abbreviation Disambiguation Using Neural Word Embeddings”. *Proceedings of BioNLP* 15, pp. 171-176. https://aclanthology.org/W15-3822/

Xu, Hua; Peter D. Stetson & Carol Friedman. (2007) “A Study of Abbreviations in Clinical Notes”. *AMIA Symposium Proceedings 2007*, pp. 821-825. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655910/

Yim, Wen-Wai; Melilha Yetisgen, William P. Harris & Sharon W. Kwan. (2016) “Natural Language Processing in Oncology: A Review”. *JAMA Oncology* 2:6, pp. 797-804. https://pubmed.ncbi.nlm.nih.gov/27124593/

Zhou, Wei; Vetle I. Torvik & Neil R. Smalheiser. (2006) “ADAM: Another database of abbreviations in MEDLINE”. *Bioinformatics* 22, pp. 2813–2818. https://pubmed.ncbi.nlm.nih.gov/16982707/

AREA: Automating Repetitive Editing Actions in Post-Editing

Silvia Terribile
University of Manchester
Oxford Road, Manchester, M13 9PL, UK
silvia.terribile.research@gmail.com

Abstract

An analysis of edits in a small corpus of real-world English-to-Italian post-editing (PE) tasks from Toppan Digital Language revealed that numerous edits – 13% of all edits in PE tasks requiring high PE effort, and 21% in tasks requiring low/medium effort – were repetitions of corrections previously made in the same PE task (Terribile, 2024). Translators had to manually carry out these edits repeatedly, because computer-assisted translation (CAT) tools auto-propagate edits only in full translation memory matches. Repetitive work may considerably increase PE effort, and previous research (e.g. Guerberof Arenas, 2013; Alvarez-Vidal, Oliver and Badia, 2020) has reported that it also has a high negative impact over post-editors' job satisfaction. This paper proposes the AREA algorithm (Automating Repetitive Editing Actions) as a possible solution to this problem. AREA sets the minimum similarity conditions between two segments in a machine translation output to automate edits when repeated. It can be implemented in any CAT tool and could automate up to 46% of repetitive corrections in the English-to-Italian language combination.

1 Introduction

Numerous studies have highlighted that post-editing (PE) can frequently be highly repetitive (e.g. Lagarda et al., 2015; Chatterjee, 2019; Alvarez-Vidal, Oliver and Badia, 2020). In Terribile (2024), I conducted a mixed-methods analysis of the types of edits made in a small corpus of 30 real-world English-to-Italian PE tasks from the language service provider (LSP) Toppan Digital Language. My findings were consistent with previous research, as a large number of edits – 13% of all edits in PE tasks requiring high PE effort, and 21% of edits in tasks requiring low/medium effort – were repetitions of changes previously carried out in the same PE job. Linguists had to manually carry out these edits repeatedly, because computer-assisted translation (CAT) tools only auto-propagate edits in full translation memory (TM) matches.[1] These repetitive actions considerably increase PE effort, and much previous research (e.g. Guerberof Arenas, 2013; Moorkens and O'Brien, 2015) has reported that they also have a high negative impact on post-editors' satisfaction and enjoyment of their work.

This paper proposes the AREA algorithm (Automating Repetitive Editing Actions) as a possible solution to this problem. AREA sets the minimum similarity conditions between two segments in a machine translation (MT) output to automate edits when repeated. The acronym also highlights its core principle, as it considers the area (or context) around an edit, to define the conditions for automating that edit if repeated. AREA can be implemented in any CAT tool and could potentially automate up to 46% of repetitive edits in the English-to-Italian language combination, corresponding to 6% of all corrections made during PE. However, the implementation of this algorithm in a CAT tool and carrying out quantitative testing lie beyond

[1] While the 'fuzzy match repair' feature available in various CAT tools (e.g. the 'upLIFT' feature in Trados) can automatically adapt fuzzy matches (RWS Support Gateway, 2021), it is not designed for automating repeated edits in partially repeated segments.

the scope of this work and will be presented in a future study. This paper presents this algorithm as a suggestion for localisation researchers, LSPs and CAT tool developers, who may be interested in implementing AREA in their CAT tool and testing it against text domains and language pairs of their choice.

2 Related work

Automating certain PE tasks can considerably enhance translators' efficiency – a concept discussed by Bar-Hillel in his seminal 1960 paper, where he advocated for a "*machine-post-editor partnership*" (1960: 94). Automatic post-editing (APE) has long aimed to "emulate what the human is doing" (Knight and Chander, 1994: 779), and it has been effective in automatically correcting systematic errors in MT outputs before they are post-edited by a human translator (Chatterjee, 2019). APE models need to be trained on large quantities of high-quality data. They typically learn to identify and correct MT errors by training on source texts, MT outputs, and their post-edited versions (do Carmo et al., 2021). Model training is frequently enhanced through additional data, such as parallel corpora of source texts and human reference translations, error annotations, and human evaluation scores, among others. APE researchers have fruitfully utilised rule-based, statistical, hybrid, and neural approaches (e.g. Porro et al., 2014; Chatterjee, 2019). However, APE models tend to struggle to correct complex errors and sometimes make unnecessary or incorrect changes.

A major limitation of APE is that it is useful only for correcting systematic errors. My analysis of the types of corrections that post-editors made to neural MT outputs showed that they frequently had to correct the same errors repeatedly within a PE task; however, these errors were less common in other PE tasks from the same domain, suggesting that they may not be systematic (Terribile, 2024). Microsoft researchers have recently demonstrated that GPT-4 can be successfully used to automate PE, without limiting its focus to systematic errors (Raunak et al., 2023). However, they also reported that GPT-4 produced some "hallucinated edits, thereby urging caution in its use as an expert translation post-editor" (ibid.: 1).

This paper presents AREA, a rule-based algorithm that can be implemented in any CAT tool to automate corrections to MT errors, that are not necessarily systematic, when repeated within an individual PE task. While AREA differs significantly from rule-based APE approaches, it also shares certain similarities, particularly with the work of Simard and Foster (2013) and Lagarda et al. (2015). Both studies proposed using online learning techniques to enable an APE system to learn from post-editors' corrections in real time. The AREA algorithm also refers to the corrections made by a post-editor in real time, but it does not learn the type of correction required. Rather, it relies on a fixed set of rules to identify the area of a segment that requires the same correction made in a previous segment, and it automates the repeated edit in a mechanical manner. As such, AREA does not require any learning techniques. The key advantages of this approach are that, unlike rule-based APE models, AREA (1) can automate any type of repeated edit, rather than focusing solely on specific types of corrections to systematic errors; and (2) does not require training on corpora.

3 Development and features of AREA

In Terribile (2024), I carried out a mixed-methods, linguistic analysis of the types of edits made in a small corpus of 30 real-world English-to-Italian PE tasks in the marketing and technical marketing domain (i.e. marketing texts with a high percentage of technical terminology). This research examined edits as “Post-Editing Actions” (PEAs), i.e. “a set of minimal and logical edits”, which may affect one or more words; they are described as ‘logical’ because they “linguistically make sense”, and ‘minimal’ because they are the smallest possible “independent edit[s]” (Blain et al., 2011: 165). All edits presented in the current paper are examples of PEAs.

My analysis aimed to understand which edit types are typically implemented in tasks requiring high PE effort (Terribile, 2024). To this end, it considered edits in 15 PE tasks requiring high levels of temporal and technical PE effort (total: 8,620 words; 2,108 PEAs), and in a corpus of similar size composed of texts requiring low to medium effort (8,833 words; 586 PEAs). The LSP’s CAT tool automatically recorded both temporal and technical effort (Terribile, 2023). For the former, it tracked words per hour, representing the source word count divided by the time a linguist spent in the CAT environment. For the latter, it measured edit distance values using the Levenshtein algorithm, which “calculates the minimum number of character edits that are necessary to transform one string into another string” (Kosmaczewska and Train 2019, 170). As previously mentioned, many repetitive editing actions were identified in both datasets[2]. Thus, I decided to further investigate repetitive edits in these PE tasks to understand whether they could be automated, at least to some extent.

The development of AREA started with a qualitative analysis of the tokens around the edited token(s) in the MT output. A ‘token’ refers to a word, punctuation, or digit (Sketch Engine, 2016), and ‘edited token(s)’ here indicate tokens present in at least two segments and edited in the same way within a PE task. For clarity, I refer to the relevant segments as segments 1 and 2, although they may occur anywhere in a text, and there may be more than two segments requiring the same correction in the text. I also refer to segment 2 as a ‘partially repeated segment’, because it includes some content that is repeated from segment 1, but it is not a full TM match. This investigation started from the classification of partially repeated segments into (1) those where only the edited token(s) are repeated, and (2) those where other tokens immediately before and/or immediately after the edited token(s) are also repeated. Examples are shown in Table 1.

[2] A wide range of edit types were repeated: an in-depth analysis is available in Terribile (2024).

MT output vs post-edited text	
Segment 1	**Segment 2**
~~Se si verifica~~ In caso di rottura dell'articolo gettarlo immediatamente! **Back translation (BT)**: ~~If occurs~~ In case of breakage of the article throw it immediately!	~~Se si verifica~~ In caso di irritazione, consultare un medico. **BT**: ~~If occurs~~ In case of irritation, consult a doctor.
~~Segui~~ Resta al passo con le tendenze di questa stagione e indossa il nostro top corto [PRODUCT NAME 1] […] **BT**: ~~Follow~~ Keep up with the trends of this season and wear our [PRODUCT NAME 1] crop top […]	~~Segui~~ Resta al passo con le tendenze di questa stagione indossando il top corto [PRODUCT NAME 2] […] **BT**: ~~Follow~~ Keep up with the trends of this season wearing the [PRODUCT NAME 2] crop top […]

Table 1. Examples of repeated edits without/with identical token(s) around the edited token(s)[3]

I hypothesised that, in partially repeated segments where some tokens around the edited token(s) were repeated from segment 1, the identical context could be used to determine when automating these edits would be appropriate. I aimed to identify patterns among these repeated edits, to determine what minimum identical context would enable us to say, with a reasonable level of confidence, that if an edit is made in segment 1, it is also needed in segment 2.

This exploration started by considering the number of words in the MT output immediately before and immediately after the edited token(s) that were identical in both segments. However, this number was typically very small, often limited to 1 or 2 words, which were often high-frequency words. Thus, this parameter would not provide a reliable indication of the specific context in which edits were made. I then considered relying on phrases – i.e. one or more words constituting a grammatical unit (Finch, 2000). I evaluated whether it would be possible to argue that if an entire phrase is identical in segments 1 and 2, any edits made in segment 1 within that phrase would also be correct in segment 2. However, this parameter was also insufficient, as phrases can be as short as one word.

Finally, I drew on the distinction between content words, which convey meaning independently, and function words, which play a grammatical role but have limited independent meaning (Segalowitz and Lane, 2000). This enabled me to identify a prominent pattern in the analysed data: the same edit tended to be made in both segments when the first content word immediately before and/or immediately after the edited token(s) was identical in both segments, whereas the same token(s) were typically edited differently when next to at least one different content word, as in Table 2.

[3] In all tables in this paper except for Tables 6 and 7, source texts have been omitted for concision, as they are not needed to understand the examples presented. All back translations follow the Italian word order, even when this is incorrect in English, to enable all readers to understand which tokens are present around the edited token(s) in the MT output.

MT output vs post-edited text	
Segment 1	**Segment 2**
In caso di contatto con la pelle lavare con acqua ~~calda~~tiepida. **BT**: In case of contact with the skin wash with water ~~hot~~ warm.	In caso di contatto con gli occhi lavare con abbondante acqua ~~calda~~ tiepida tenendo l'occhio aperto. **BT**: In case of contact with the eyes wash with plenty of water ~~hot~~ warm keeping the eye open.
Il tuo ~~report~~ rapporto sui dati è stato salvato! **BT**: Your ~~report~~ report on data has been saved!	Il salvataggio del ~~report~~ rapporto è stato annullato. **BT**: The saving of the ~~report~~ report has been cancelled.

Table 2. Examples of partially repeated segments with/without an identical content word immediately before and/or immediately after the edited token(s)

These observations led me to hypothesise that the presence of even just one identical content word immediately before and/or immediately after the edited token(s) could be used as a key parameter to determine whether repeated edits could be automated. I then considered that function words and/or punctuation can significantly affect the meaning of a segment. For example, a comma can make a great difference in meaning, as in “Let’s eat, grandma!” versus “Let’s eat grandma!” (Digital Synopsis, 2023). As such, I evaluated that, if any function words and/or punctuation are present between the edited token(s) and the identical content word, they would also need to be identical, for the automation of edits in partially repeated segments to be usually correct.

Finally, I considered whether repeated edits could be automated, if the edited token(s) are at the beginning or end of a segment, and therefore there cannot be an identical content word in segments 1 and 2 both before and after them. In these cases, I decided that the position of the edited token(s) could be used to delimit the area where the automation would take place. Additionally, in the analysed texts, the presence of an identical content word on the other side of the segment appeared to make the context around the edited token(s) specific enough for the automation of repeated edits to be usually correct (see Table 1, example 2). These considerations led to the development of AREA, presented in plain language in Figure 1.

The AREA (**A**utomating **R**epetitive **E**diting **A**ctions) **algorithm**

Minimum conditions in the MT output, to enable the automation of an edited token, if repeated within the same text:

No content words before the edited token(s) in segment 1 or in segments 1 and 2, and identical function words and punctuation (if any exist) before the edited token(s) OR minimum 1 identical content word immediately before the edited token(s), plus identical function words and punctuation (if any exist) between the content word and the edited token(s)

AND

no content words after the edited token(s) in segment 1 or in segments 1 and 2, and identical function words (if any exist) after the edited token(s) OR minimum 1 identical content word immediately after the edited token(s), plus identical function words and punctuation (if any exist) between the edited token(s) and the content word.

Notes:

1. ‘Edited token(s)’ indicate tokens present in segments 1 and 2, in which the edits carried out by a post-editor in segment 1 would be automatically implemented in segment 2 by implementing this algorithm. Edited token(s) do not include empty tokens.
2. This algorithm works at the level of individual tokens: edit(s) to each token need to fulfil the criteria above to be automated. Nevertheless, an edited token may be replaced with multiple tokens and vice versa. AREA simply requires that the context immediately before and immediately after the edited token(s) meets its criteria, to enable the automation of any edits within that context.
3. AREA is not case sensitive.
4. Punctuation after the edited token(s) may differ in segments 1 and 2, because there may be punctuation to end segment 1, that may be missing in segment 2, if the repeated edited token(s) occupy a different position in segment 2.

Figure 1. The AREA algorithm

To summarise the basic principle of AREA in one sentence, to automate an edit in the MT output of segment 2, there needs to be either no content words or minimum 1 content word identical to a corresponding content word in segment 1 both immediately before and immediately after the edited token(s), and any function words and/or punctuation between the edited token(s) and the identical/no content words also need to be identical.

Table 3 displays examples of repeated edits that were manually made in the PE tasks analysed in Terribile (2024), but they could be automated, as they meet the criteria of AREA.

MT output vs post-edited text		
Segment 1	**Segment 2**	**Fulfilled criteria**
Realizzat~~e~~o con la struttura [BRAND NAME] [PRODUCT NAME], questo modello […] **BT**: Realised [~~Feminine, plural~~ past participle of verb] [Masculine, singular past participle of verb] with the structure [BRAND NAME] [PRODUCT NAME], this model […]	Realizzat~~e~~o con la struttura [BRAND NAME] [PRODUCT NAME], questo classico modello […] **BT**: Realised [~~Feminine, plural~~ past participle of verb] [Masculine, singular past participle of verb] with the structure [BRAND NAME] [PRODUCT NAME], this classic model […]	• No content words before the edited token in segments 1 and 2; • 1 identical content word immediately after the edited token.
Accessorio con chiusura ~~a~~ con zip, […] **BT**: Accessory with fastening ~~at~~ with zip	Chiusura ~~a~~ con zip **BT**: Fastening ~~at~~ with zip	• 1 identical content word immediately before and 1 identical content word immediately after the edited token.
~~Regalo~~ Omaggio gratuito **BT**: Free ~~gift~~ freebie	Ottieni il tuo ~~regalo~~ omaggio gratuito **BT**: Get the your free ~~gift~~ freebie	• No content words before the edited token in segment 1; • 1 identical content word immediately after the edited token.
In caso di contatto con la pelle lavare con acqua ~~calda~~tiepida. **BT**: in case of contact with the skin wash with water ~~hot~~ warm.	In caso di contatto con gli occhi lavare con abbondante acqua ~~calda~~ tiepida tenendo l'occhio aperto. **BT**: in case of contact with the eyes wash with plenty of water ~~hot~~ warm keeping the eye open.	• 1 identical content word immediately before the edited token; • No content words after the edited token in segment 1.

Table 3. Examples of repeated edits meeting the criteria of the AREA algorithm

Figure 2 displays the percentages of repetitions that were not auto-propagated in the PE tasks analysed in Terribile (2024), and that (1) do not meet the criteria of AREA; or (2) meet the criteria of AREA, and therefore could potentially be automated.

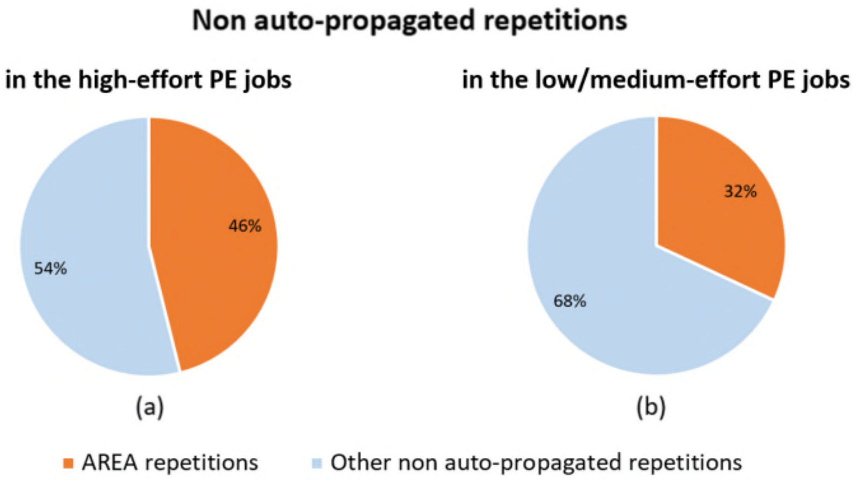


Figure 2. Non auto-propagated repetitive edits

According to these results, implementing this algorithm would have the potential of reducing manual repetitive actions to a great extent, i.e. by 46% in PE tasks requiring high PE effort, and by 32% in those involving lower levels of effort – which correspond, respectively, to 6% and 7% of all PEAs.

Let us now discuss the features of this algorithm. AREA can be implemented in any CAT tool through programming. A language model is not required: providing the CAT tool with a list of function words in any relevant language and programming it to classify all other words as content words will suffice. AREA is intended to be applied to words in the MT output, and it does not consider the source text. For this reason, I recommend disabling AREA to automatically execute edits in fuzzy TM matches. Indeed, whereas content in the source text typically corresponds to content in the MT output, this is frequently not the case in fuzzy matches.

As CAT tools typically auto-propagate edits in full TM matches at the level of individual documents, I propose to do the same for automations enabled by AREA to reduce the likelihood of incorrect suggestions. Whereas it is possible – and perhaps not unlikely – that a sequence of repeated words originated by MT fulfils the criteria of AREA, yet the two MT outputs do not require the same changes at the PE stage due to words being used differently in diverse contexts, this appears to be more unlikely within the same text (see Section 5). It would also be useful to add a setting that allows users to enable AREA to automate edits across different documents within the same project, since these texts may contain similar content, as in the existing option for auto-propagating edits in full TM matches.

Furthermore, this algorithm works at the level of individual tokens: each edited token needs to meet the criteria mentioned in Figure 1 to be automated. As such, it is possible that not all edits belonging to an individual PEA would be automated, but only those to some tokens. Nevertheless, these cases were extremely rare in the data analysed in this research (only 2 out of 165 cases), and applying this algorithm would enable PE effort and repetitive actions to be reduced even in these cases.

Additionally, AREA requires that there are either identical tokens or no tokens in the segment immediately before and immediately after the edited token(s), to be able to correctly locate where the edits need to be implemented in segment 2. Consequently, when more than one contiguous token is edited, if the first or last of the identical tokens are edited and they are not positioned either at the beginning or at the end of segment 1, they do not fulfil the criteria of AREA and generate a domino effect. In particular, editing the first or last identical tokens causes the following/previous tokens to lack a specific context for the algorithm to locate where to implement an edit and so on, preventing edits to all other identical tokens from being automated. Although this may reduce the number of automations, instances of this kind do not appear to be very frequent, as no cases were found in the data analysed in this research.

This constraint cannot be removed because it is fundamental to avoid incorrect automations. Indeed, without some boundaries to delimit the location and extent of an automation, if some tokens that are contiguous to the relevant edited tokens are also edited but they are not repeated in the partially repeated segment, AREA would not be able to identify where edits to the repeated tokens start/finish. This is why the algorithm mentions that, if there are no tokens in the segment immediately before or immediately after the edited token(s), this must be the case in segment 1 (and optionally in segment 2). Indeed, as can be seen in the example presented in Table 4, if there are no tokens before the edited token(s) in segment 2 but not in segment 1, AREA may not be able to identify which tokens need to be included in the automation.

MT output vs post-edited text	
Segment 1	**Segment 2**
[…] ~~ed etichetta~~ e un'applicazione con logo [BRAND NAME] in vita. **BT**: […] ~~and label~~ and a patch with logo [BRAND NAME] on waist.	~~Etichetta~~ Applicazione con logo **BT**: ~~Label~~ Patch with logo

Table 4. Example of a case that does not meet the criteria of AREA, because there are no tokens immediately before the edited token in segment 2, but not in segment 1

Conversely, if there are no tokens immediately before or immediately after the edited token(s) in segment 1, but not in segment 2, AREA would be able to use the first or last identical token in segment 2 as a reference for the location of the automation, as in Table 5.

MT output vs post-edited text	
Segment 1	**Segment 2**
~~Regalo~~ Omaggio gratuito **BT**: ~~Gift~~ Freebie free	Ottieni il tuo ~~regalo~~ omaggio gratuito **BT**: Get the your ~~gift~~ freebie free

Table 5. Example of a case that meets the criteria of AREA, with no tokens immediately before the edited token in segment 1, and not in segment 2

Moreover, edits to empty tokens are not accounted for in AREA, to prevent incorrect automations of certain types of insertions or reorderings. Indeed, insertions or reorderings in segment 1 may involve substituting an empty token with one or more tokens that are not repeated in segment 2. Table 6 presents an example where an incorrect automation would take place, if edits to empty tokens were accounted for in AREA.

	Source text	MT output vs post-edited text
Segment 1	Teal Textured Square Cushion	Cuscino ottanio quadrato ~~ottanio~~ strutturato **BT**: Cushion teal square ~~teal~~ structured
Segment 2	Ivory All Over Print Square Cushion	Cuscino quadrato avorio con stampa all-over **BT**: Cushion square ivory with print all-over

Table 6. Example demonstrating the potential for incorrect automations if AREA considered edits to empty tokens

If AREA allowed for the automation of edits to empty tokens, the space between the words "Cuscino" [Cushion] and "quadrato" [square] in segment 2 would be incorrectly substituted with the word "ottanio" [teal].

4 Limitations

There is potential for incorrect automations of insertions or reorderings, where there are simultaneously (1) edited token(s) that fulfil the criteria of the AREA algorithm, for which a correct automation would be made, and (2) one or more tokens inserted within the edited token(s) as part of a separate insertion or reordering change. In these cases, AREA would not be able to distinguish the insertion/reordering change from the edits that would need to be implemented in segment 2, and it would also execute the insertion/reordering. An example is presented in Table 7.

	Source text	MT output vs post-edited text
Segment 1	White Embossed Tall Grass Faux Plant And Pot	Pianta ~~e~~ artificiale con vaso bianco con erba alta lavorato **BT**: Plant ~~and~~ artificial with pot white with tall grass finished
Segment 2	White Embossed Faux Plant And Pot	Pianta ~~e~~ artificiale con vaso ~~finti~~ bianco lavorato **BT**: Plant ~~and~~ artificial with pot ~~fake~~ white finished

Table 7. Example displaying the potential for incorrect automations of insertions or reorderings

Here, a post-editor substituted "e" [and] with "con" [with], a change that AREA would correctly implement in segment 2. However, the word "artificiale" [artificial] was also inserted as it was missing in the MT output. AREA would not have been able to distinguish between these changes, and it would have implemented both in segment 2. Both edits would have happened to be correct (because "Faux" was mistranslated as "finti" [fake] in segment 2; "finti" would still need to be manually deleted), but this is purely coincidental. Nonetheless, instances of this kind appear to be very rare, as this was the only instance (out of 165 cases) in the data analysed in this research.

There is also potential for AREA automations to be partially incorrect, especially when the edited token(s) in segments 1 and 2 are supposed to differ in terms of number and gender, as in Table 8.

MT output vs post-edited text	
Segment 1	**Segment 2**
Shorts blu con stampa floreale in raso ~~con laccetti~~ allacciati in vita **BT**: Blue shorts with floral print in satin ~~with laces~~ laced [masculine, plural adjective] on waist	Vestaglia rosa cipria in raso ~~con laccetti~~ allacciata in vita **BT**: Blush pink nightgown in satin ~~with laces~~ laced [feminine, singular adjective] on waist

Table 8. Example displaying the potential for partially incorrect automations due to differences in number and/or gender

Here, the edit made in segment 1 would be correctly implemented in segment 2. However, since the prepositional phrase "con laccetti" [with laces] was replaced with an adjectival phrase "allacciati" [laced] during PE, this automation would be partially incorrect, as the adjective would refer to a masculine, plural noun in segment 1 ("shorts"), but to a feminine, singular noun in segment 2 ("vestaglia" [nightgown]). However, such instances appear to be extremely rare (in these data, only this instance out of 165 cases), because they only occur when a phrase that does not specify number and gender is substituted with one that does. Moreover, I hypothesise that automating edits may reduce PE effort even in these cases, as the post-editor would only need to correct the gender and/or number of the edited token(s).

To mitigate the negative impact of any incorrect automations, I recommend presenting AREA suggestions in CAT tools in the right sidebar used to display fuzzy matches, rather than auto-propagating them in the target segments. Different colours could be used to distinguish AREA suggestions from fuzzy matches. It would also be useful to add a setting that allows users to choose whether AREA suggestions are displayed in the right sidebar or auto-propagated directly in the target segments, if they find that they usually accept these suggestions.

5 Preliminary, qualitative testing

A preliminary, qualitative testing of AREA was conducted to get a general understanding of whether and to what extent it would potentially generate any other types of incorrect automations. The first stage of testing involved manually checking all edits that fulfil the algorithm's criteria in all texts analysed in this research, to spot any incorrect automations that AREA could possibly make. No potential errors except for the ones mentioned in the previous section were identified. Secondly, I considered cases meeting the criteria of AREA, including unedited repetitions that could potentially have been edited. In particular, I compared the meaning and grammatical function of relevant repeated tokens in segments 1 and 2. I hypothesised that, if these tokens had the same meaning and function in both segments, it would be possible to say, with a reasonable level of confidence, that if an edit was carried out in the relevant area of segment 1, it would usually be correct also in segment 2. To compare the meaning and grammatical function of these tokens, I have:

1) identified all cases meeting the criteria of AREA (including unedited repetitions) in the MT output of a PE task, by identifying n-grams – i.e. sequences of repeated token(s) (Sketch Engine, 2023a) – through Sketch Engine (2023b) and manually checking which would fulfil the algorithm's criteria;
2) manually compared the meaning of the relevant tokens in segments 1 and 2;

3) compared their parts of speech (POS) in segments 1 and 2, by utilising LancsBox's automatic POS tagging (Brezina and Platt, 2023), and manually correcting any incorrect tags.

Table 9 presents an example of these semantic and part-of-speech comparisons.

MT output of segment 1	MT output of segment 2
T-shirt corta a maniche lunghe [BRAND NAME PRODUCT NAME1] **BT**: T-shirt cropped long sleeve [BRAND NAME PRODUCT NAME1] **POS**: T-shirt-NOM corta-ADJ a-PRE maniche-NOM lunghe-ADJ […][4]	La T-shirt corta a maniche lunghe [PRODUCT NAME 2] è realizzata in 100% jersey di cotone con grafiche sul retro che mettono in risalto tutti e tre i loghi per un'atmosfera costiera unica. **BT**: The T-shirt cropped long sleeve [PRODUCT NAME 2] is made in 100% cotton jersey with back graphics that highlight all three logos for a unique coastal vibe. **POS**: La-DET:def T-shirt-NOM corta-ADJ a-PRE maniche-NOM lunghe-ADJ […]
Green: Tokens where any edits to segment 1 could potentially be automated in segment 2. Blue: Identical tokens immediately before or immediately after them.	

Table 9. Example of semantic and part-of-speech comparisons of tokens meeting the criteria of AREA

Here, if a post-editor was to make any edits to "corta a maniche" in segment 1, AREA would have automatically implemented these edits in segment 2. As can be observed by looking at the back translation and at the part-of-speech tagging, the words "corta a maniche" have the same meaning and grammatical function in both segments. As such, I hypothesise that if "corta a maniche" was edited in segment 1, the same edits would usually be correct in segment 2. Since this second phase of qualitative testing is extremely time-consuming, it was conducted only on one text requiring high PE effort[5]. In this text, 21 repetitions met the criteria of AREA, of which only one was edited. In all cases, all relevant tokens presented the same meaning and POS in both segments.

6 Conclusion and future work

This paper has presented a research-informed algorithm that could be implemented in any CAT tool through programming, to automate up to 46% of repetitive edits in partially repeated segments in the English-to-Italian language combination, corresponding to 6% of all corrections made during PE. Based on the preliminary, qualitative testing carried out as part of this project, incorrect automations enabled by AREA appear to be extremely rare (2/165 cases). Although this research has only considered Italian MT outputs, I hypothesise that AREA would work in the same (or in a similar) way in languages presenting similar language structures, such as English, French, Spanish, etc. Conversely, I hypothesise that AREA would be less useful for

[4] Italian POS tags used in LancsBox (Stein, no date). For the sake of concision, only relevant POS tags are included in this example. ADJ = adjective; DET:def = definite article; NOM = noun; PRE = preposition.

[5] The full semantic and part-of-speech comparison of tokens fulfilling the criteria of AREA in this text is not presented here due to spatial constraints. It is available in Terribile (2024: 296-303).

languages with case systems (e.g. German, Russian, etc.), or languages presenting very different structures from Italian (e.g. Chinese, Japanese, etc.). Nevertheless, it is possible that automations would still be useful in these target languages, as editing them might require lower PE effort than making the same edit repeatedly. Thus, I recommend that LSPs and CAT developers conduct quantitative testing of AREA, considering text domains and language pairs of their choice.

A future study is planned to implement AREA in a CAT tool and quantitatively test its accuracy for the English-to-Italian language pair. This research will also include a user study where professional translators perform PE tasks with or without using AREA, to (1) compare their measured and perceived PE effort in the temporal, technical and cognitive dimensions; and (2) assess the tool's usability and utility through questionnaires and semi-structured interviews. Future research could also explore the possibility of developing a hybrid model that combines rule-based and machine learning (ML) approaches. For example, it would be useful to investigate how ML-driven methods could augment AREA and address challenges related to number, gender, and/or rich morphology.

Acknowledgements

This research was funded by the Arts and Humanities Research Council (AHRC) of UK Research and Innovation (UKRI), with grant number 2498533. I would like to express my gratitude to Prof Maeve Olohan for her invaluable guidance and feedback on my research. I would like to extend my thanks to Toppan Digital Language for their support and for enabling me to analyse their translation projects. I would also like to thank Dr Rebecca Tipton and Dr Henry Jones for their constructive advice.

References

Alvarez-Vidal, Sergi, Antoni Oliver, and Toni Badia. 2020. Post-Editing for Professional Translators: Cheer or Fear?, *Tradumàtica*, 18, pages 49–69. https://doi.org/10.5565/rev/tradumatica.275.

Bar-Hillel, Yehoshua. 1960. The Present Status of Automatic Translation of Languages. In Franz L. Alt (ed.) *Advances in Computers*, 1. Academic Press, New York and London, pages 91–163. https://aclanthology.org/www.mt-archive.info/50/Bar-Hillel-1960.pdf. (Accessed: 29 October 2024).

Blain, Frédéric, Jean Senellart, Holger Schwenk, Mirko Plitt, and Johann Roturier. 2011. Qualitative Analysis of Post-Editing for High Quality Machine Translation. In *Proceedings of the Machine Translation Summit XIII*, Asia-Pacific Association for Machine Translation (AAMT), Xiamen, pages 164–171. https://aclanthology.org/2011.mtsummit-papers.17.pdf (Accessed: 24 May 2021).

Brezina, Vaclav, and William Platt. 2023. *#LancsBox X* [software], Lancaster University. http://lancsbox.lancs.ac.uk (Accessed: 13 December 2023).

Chatterjee, Rajen. 2019. *Automatic Post-Editing for Machine Translation*. (Doctoral thesis, University of Trento). https://doi.org/10.48550/arxiv.1910.08592. (Accessed: 15 May 2024).

Digital Synopsis. 2023. *10 Hilarious Examples of How Punctuation Makes a Big Difference*. https://digitalsynopsis.com/tools/punctuation-marks-importance-rules-usage/ (Accessed: 13 December 2023).

do Carmo, Félix, Dimitar Shterionov, Joss Moorkens, Joachim Wagner, Murhaf Hossari, Eric Paquin, Dag Schmidtke, Declan Groves, and Andy Way. 2021. A Review of the State-of-the-Art in Automatic Post-Editing. *Machine Translation*, 35(2), pages 101–143. https://doi.org/10.1007/s10590-020-09252-y.

Finch, Geoffrey. 2000. *Linguistic Terms and Concepts*. Basingstoke: Macmillan.

Guerberof Arenas, Ana. 2013. What Do Professional Translators Think About Post-Editing, *The Journal of Specialised Translation*, 19, pages 75–95. https://www.jostrans.org/issue19/art_guerberof.pdf (Accessed: 12 May 2023).

Knight, Kevin, and Ishwar Chander. 1994. Automated Post-Editing of Documents. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI),* Seattle, Washington, USA, pages 779–784. https://cdn.aaai.org/AAAI/1994/AAAI94-119.pdf. (Accessed: 29 October 2024).

Kosmaczewska, Kasia, and Matt Train. 2019. Application of Post-Edited Machine Translation in Fashion eCommerce. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks,* Dublin, Ireland, pages 167–173. https://www.aclweb.org/anthology/W19-6730. (Accessed: 2 October 2020).

Lagarda, Antonio L., Daniel Ortiz-Martínez, Vicent Alabau, and Francisco Casacuberta. 2015. Translating Without In-Domain Corpus: Machine Translation Post-Editing with Online Learning Techniques, *Computer Speech & Language*, 32(1), pages 109–134. https://doi.org/10.1016/j.csl.2014.10.004.

Moorkens, Joss, and Sharon O'Brien. 2015. Post-Editing Evaluations: Trade-Offs Between Novice and Professional Participants. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 75–81. https://aclanthology.org/W15-4910 (Accessed: 18 November 2022).

Porro, Victoria, Johanna Gerlach, Pierrette Bouillon, and Violeta Seretan. 2014. Rule-Based Automatic Post-Processing of SMT Output to Reduce Human Post-Editing Effort. In *Proceedings of Translating and the Computer* (36). https://aclanthology.org/2014.tc-1.8/ (Accessed: 18 November 2022).

Raunak, Vikas, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. Leveraging GPT-4 for Automatic Translation Post-Editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore. Association for Computational Linguistics, pages 12009–12024. https://doi.org/10.18653/v1/2023.findings-emnlp.804.

RWS Support Gateway. 2021. *About upLIFT*. https://gateway.sdl.com/apex/communityknowledge?articleName=000002847#:~:text=parts%20of%20TUs.-,Fuzzy%20Match%20Repair,Punctuation%20changes%20can%20also%20occur. (Accessed: 29 October 2024).

Segalowitz, Sidney J., and Korri C. Lane. 2000. Lexical Access of Function Versus Content Words, *Brain and Language*, 75(3), pages 376–389. https://doi.org/10.1006/brln.2000.2361.

Simard, Michel, and George Foster. 2013. PEPr: Post-Edit Propagation Using Phrase-Based Statistical Machine Translation. In *Proceedings of the XIV Machine Translation Summit*, Nice, France, pages 191–198. https://aclanthology.org/2013.mtsummit-papers.24.pdf. (Accessed: 29 October 2024).

Sketch Engine. 2016. *Token*. https://www.sketchengine.eu/my_keywords/token/ (Accessed: 05 December 2022).

Sketch Engine. 2023a. *N-gram*. https://www.sketchengine.eu/my_keywords/n-gram/ (Accessed: 05 December 2023).

Sketch Engine. 2023b. *What is Sketch Engine?* https://www.sketchengine.eu/ (Accessed: 05 December 2023).

Stein, Achim. No date. Italian Tagset Used in the TreeTagger Parameter File, *Centrum für Informations- und Sprachverarbeitung - LMU Munich*. https://www.cis.lmu.de/~schmid/tools/TreeTagger/data/italian-tagset.txt. (Accessed: 13 March 2023).

Terribile, Silvia. 2023. Is Post-Editing Really Faster Than Human Translation?, *Translation Spaces*. https://doi.org/10.1075/ts.22044.ter.

Terribile, Silvia. 2024. *Productivity in the Post-Editing of Neural Machine Translation: a Mixed-Methods Analysis of Speed and Edits at Toppan Digital Language*. (Doctoral thesis, University of Manchester). https://research.manchester.ac.uk/en/studentTheses/productivity-in-the-post-editing-of-neural-machine-translation-a- (Accessed: 20 May 2024).

Feeding the multilingual terminological knowledge base EcoLexicon with metaphor-based names of flowers and plants

Amal Haddad Haddad

University of Granada

amalhaddad@ugr.es

Abstract

EcoLexicon is a multilingual terminological resource developed by the LexiCon Research Group at the University of Granada. It represents the conceptual structure of the specialised domain of the Environment in the form of a visual thesaurus following the premises of Frame Based-Terminology. In this paper, we explain the results of the project "Feeding the multilingual terminological knowledge base EcoLexicon with metaphor-based names of flowers and plants" and we show how the concepts and terms are being inserted in EcoLexicon, how the conceptual systems are being constructed with a culture-sensitive approach, and how the terms are being inserted in different languages. Finally, we draw conclusions on the use of this tool.

1 Introduction

This paper explains the results of the project "Feeding the multilingual terminological knowledge base EcoLexicon with metaphor-based names of flowers and plants"[1]. The main aim of this project is to add metaphor-based names of flowers and plants to the online multilingual Terminological Knowledge Base (TKB) EcoLexicon[2], to make it accessible online to any interested groups or individuals, especially translators, terminologists and environmental experts. It is important to highlight that in this research we focus on metaphoric names with the objectives of providing results that are useful for the advancement of metaphor research and applications, however, in EcoLexicon, we also work parallelly on non-metaphoric names as part of the EE.

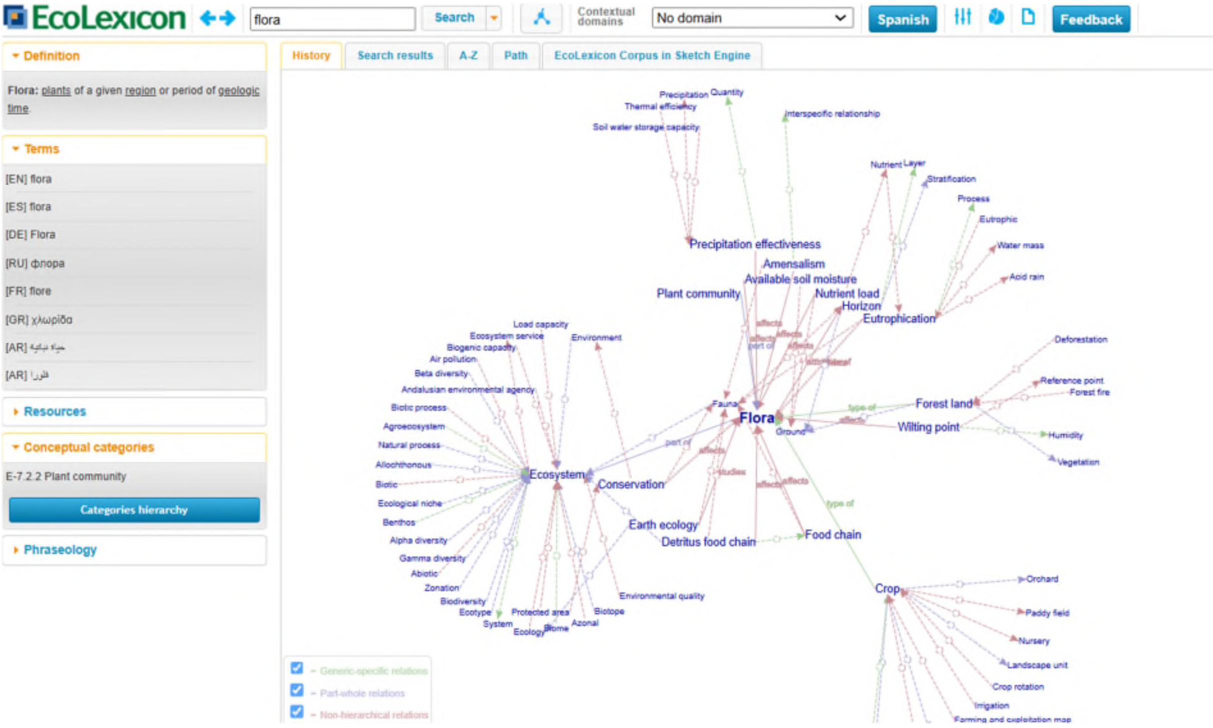
EcoLexicon represents the conceptual structure of the specialised domain of the Environment in the form of a visual thesaurus. This thesaurus has been elaborated according to the theoretical premises of Frame-Based Terminology (FBT) (Faber 2012). Consequently, each concept appears in the context of a specialised frame that highlights its relation to other concepts and makes explicit its designations in different languages.

The representation of knowledge in EcoLexicon takes into account the conceptual organisation, the multidimensional nature of specialised knowledge, and the extraction of semantic and syntactic information using multilingual corpora (León-Araúz, et al. 2019: 224) in a way that helps in knowledge representation and acquisition of the specialised environmental concepts. All related entities and processes in the domain of the environment are

[1] The project is part of the Researching and Applying Metaphor (RaAM) Building Bridges Fund 2024, aiming at supporting activities that will lead to the sharing of existing research on all types of figurative language carried out by RaAM members with an emphasis on its scientific, social and economic impact.

[2] EcoLexicon: https://ecolexicon.ugr.es/en/index.htm

delimited within a general event-frame called the ENVIRONMENTAL EVENT (EE). This event is sensitive to multidimensionality so that it can absorb concepts from other domains when they form part, directly or indirectly, of the EE. At a conceptual level, the EE is conceived as PROCESSES initiated by an AGENT (natural or human) which affect other entities with the function of PATIENT and produce specific RESULTS. These conceptual macro-categories and the relations linking them help articulate the other concepts of the domain. In the case of flowers and plants, those entities that form an important part of the EE, being part of the ecosystem. Figure 1 shows how the concept of FLORA is connected to the EE through other connected subevents. It also shows that the definition of *flora* as extracted from the conceptual relations is: “plants of a given region or period of geologic time.”


Flora: plants of a given region or period of geologic time.

[EN] flora
[ES] flora
[DE] Flora
[RU] флора
[FR] flore
[GR] χλωρίδα
[AR] حياة نباتية
[AR] فلورا

E-7.2.2 Plant community

Figure 1. EcoLexicon main view: entry for *flora*

After having carried out research on the automatic extraction of metaphor-based names of flowers and plants, by applying natural language processing (NLP) techniques (Haddad et al 2023, Premasiri et al. 2023), the objective is to feed EcoLexicon with the annotated names in English and Spanish as well as adding the cultural terminological elements characterising each term in accordance with the principles of cultural adaptation of EcoLexicon within the approach of FBT (León-Araúz and Faber, 2024). The inclusion of this cultural component would help to transform linguistic resources into inclusive knowledge bases (León-Araúz and Faber, 2024).

2 Methodology

The project consists of three phases: first, we add the concepts and their conceptual relations as well as their ontological categories within the corresponding environmental subframes. Then we add the English and Spanish terms associated with them. Afterwards, we add the definitions of each term in accordance with the definitional templates and through the extraction of semantic relations from corpus, following bottom up and top-down approaches. Finally, we add the equivalent terminological variants in the same languages, as well as other possible equivalents in other languages, such as Arabic. The focus of this project is the metaphoric names of flowers and plants, for this reason, this process starts by defining the families and genres of those plants to be associated with the EE, in order to be able to link these plants, through conceptual relations, to the whole system. In the section dedicated to 'notes' of each metaphoric name, we add the reason why it can be considered as metaphoric, and we highlight the cultural elements within the TKB structure.

3 Example

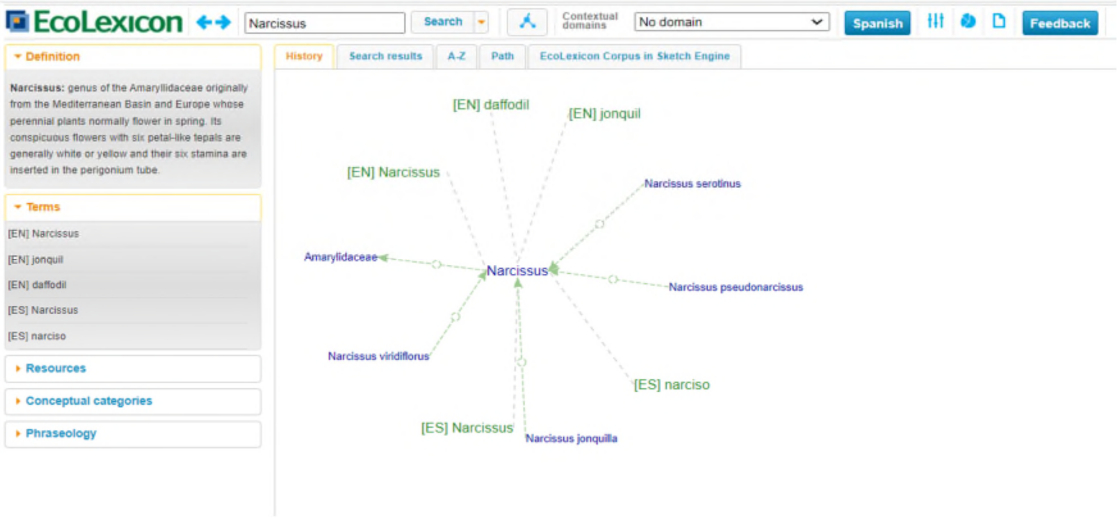
Regarding the definition of a metaphor, it is a figure of speech in which two domains are linked by mapping attributes from one onto the other (Humar 2021). In the case of the names of flowers and plants, the name is considered metaphoric if it contains a lexical unit, a prefix, root and/or suffix with a semantic meaning that does not belong to the domain of plants. This can be either a whole word forming the name as for example *Moon light*, being both *moon* and *light* a metaphoric lexical unit; or the metaphoric part would be only part of the word as in the case of the scientific name of the flower *Leontopodium alpinum*. This name is metaphoric as the first part of the name *Leontopodium* combines the Greek word *léōn* ('lion') and *pódion* ('foot'), based on information extracted from etymological dictionaries. Another example is the name *Edelweiss*, which combines the prefix *edel* meaning noble in German, and the suffix *weiss*, meaning white. The first two examples are considered multiword expressions, while the third one is a one-word name. Those names may be challenging when it comes to a translation task and require specialised knowledge in order to transfer the meaning correctly.

As a practical example of this tool, different metaphor-based names and their conceptual relations will be explained. For instance, the name *Narcissus* has been annotated as metaphoric, as its etymology comes from the Ancient Greek *narkissos* meaning a *narcissist*. In order to insert this name in EcoLexicon[3], first of all, the whole conceptual structure of this term was defined, above all, the family of plants it belongs to, its genre, and the conceptual relations that link it within the EE. This information was extracted by means of a top-down and bottom-up approach, extracting information from specialised resources as well as from corpus. The main tool for corpus compilation and analysis is Sketch Engine (Kilgarriff et al. 2014).

The concept NARCISSUS IS defined in EcoLexicon as "genus of the Amaryllidaceae originally from the Mediterranean Basin and Europe whose perennial plants normally flower in spring. Its conspicuous flowers with six petal-like tepals are generally white or yellow and their six stamina are inserted in the perigonium tube." After adding the definition of the concept in English and Spanish, the terms associated with it as well as its variants are inserted, clarifying whether the terms added are the main terms or variants of the names, or whether it is a diaphasic variant, a synonym, acronym, etc. The conceptual relations of the concept, its ontological category as well as any other information is added to its internal annotated structure. The result of this annotation can be seen in EcoLexicon when searching for the term *Narcissus* as can be seen in Figure 2, which shows the entry for the concept, its conceptual relations, for instance the hyponym relation *type of,* as well as the other concepts associated with it, such as *Narcissus*

[3] This concept was created in EcoLexicon by Arianne Reimerink.

viridiflorus, *Narcissus serotinus*, etc. Clicking on each concept would expand the whole conceptual structure of the concept, showing its relation to the environment.



Narcissus

Contextual domains: No domain

Narcissus: genus of the Amaryllidaceae originally from the Mediterranean Basin and Europe whose perennial plants normally flower in spring. Its conspicuous flowers with six petal-like tepals are generally white or yellow and their six stamina are inserted in the perigonium tube.

Terms

[EN] Narcissus

[EN] jonquil

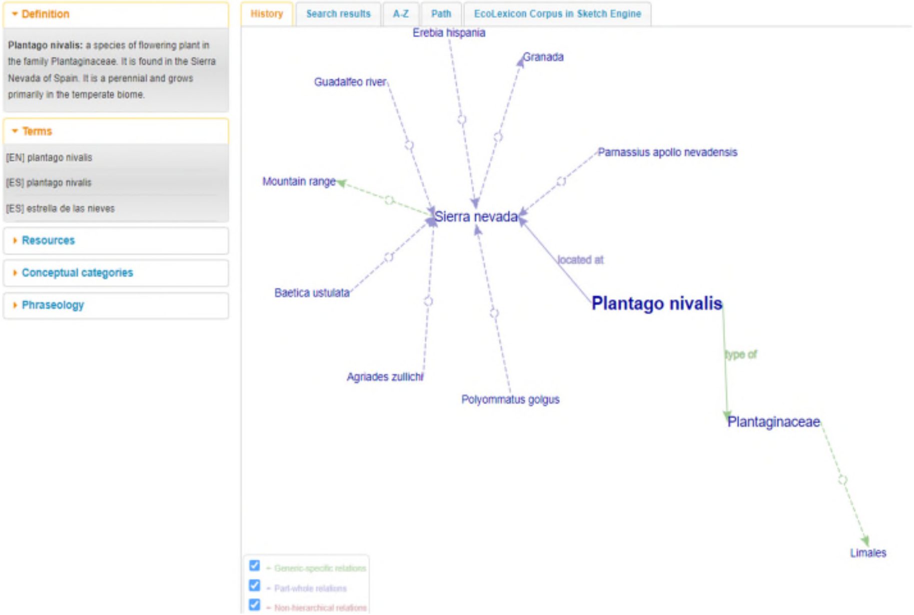
[EN] daffodil

[ES] Narcissus

[ES] narciso

Figure 2. EcoLexicon main view: entry for *Narcissus*

Another example is the concept PLANTAGO NIVALIS (Figure 3), considered metaphoric as the etymology of the lexical unit *nivalis* indicates the meaning of *snow* coming from the oblique stem *niv-* of nix ("snow") + *-ālis* ("-al", adjective-forming derivational suffix). The conceptual relations reflect the relation *type of* to indicate its plant family, and the genre, as well as the relation *located in*, considered as a terminological cultural marker to highlight that it is native to Sierra Nevada in Granada (Spain).



Plantago nivalis: a species of flowering plant in the family Plantaginaceae. It is found in the Sierra Nevada of Spain. It is a perennial and grows primarily in the temperate biome.

Terms

[EN] plantago nivalis

[ES] plantago nivalis

[ES] estrella de las nieves

☑ – Generic-specific relations

☑ – Part-whole relations

☑ – Non-hierarchical relations

Figure 3. EcoLexicon main view: entry for *Plantago Nivalis*

4 Conclusions

The aim of this project is to feed the multilingual terminological knowledge base EcoLexicon with metaphor-based names of flowers and plants annotated in previous research (Haddad et al.

2023; Premasisi et al. 2023). To achieve this result, it is necessary to feed EcoLexicon with the whole conceptual structural relations and concepts associated with those metaphoric names. Adding the metaphor-based names to the TKB also provides a terminological resource for translators, terminologists and environmental experts in the acquisition of specialised knowledge in the domain of Botany. Secondly, it serves as dynamic dictionary with definitions and synonyms, distinguishing between scientific names and vernacular names of plants. Moreover, adding these names to EcoLexicon following a cultural-sensitive approach, helps to convert this tool into inclusive knowledge bases (León-Araúz and Faber, 2024). Furthermore, this resource can be useful for the coinage of names in languages that lack the names of certain plants or to coin new names for newly discovered plants, as the conceptual systems lying behind the names are visible and illustrative, and the notes related to the dimension of metaphoricity of each name is explained within the system.

Finally, this research is a step forward in figurative language research, aimed at supporting activities that will lead to the sharing of existing research on all types of figurative language. For instance, the information available in the database EcoLexicon may be helpful in the advancement of machine translation related to metaphoric multiword expressions in the domain of Botany and in resources related to knowledge representation.

In future applications of this research, we aim to include annotated images of flowers and plants in EcoLexicon.

Acknowledgements

This research was carried out as part of the projects PID2020-118369GB-I00 and A-HUM-600-UGR20, funded by the Spanish Ministry of Science and Innovation and the Regional Government of Andalusia.

References

Faber, Pamela. 2012. *A Cognitive Linguistics View of Terminology and Specialized Language*. De Gruyter Mouton. https://www.degruyter.com/document/doi/10.1515/9783110277203/html

Haddad Haddad, Amal, Damith Premasiri, Tharindu Ranasinghe and Ruslan Mitkov. 2023. Deep Learning Methods for Extracting Metaphorical Names of Flowers and Plants. *Procesamiento del Lenguaje Natural*, 71(0).

Humar, Marcel. 2021. Metaphors as models: Towards a typology of metaphor in ancient science. *HPLS* 43, 101 . https://doi.org/10.1007/s40656-021-00450-2

Kilgarriff, Adam, Vit Baisa, Jan Bušta, Milos Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vit Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7-36. https://doi.org/10.1007/s40607-014-0009-9

León-Araúz, Pilar and Pamela Faber. 2024. Including the Cultural Dimension of Terminology in a Frame-Based Resource. In Silvia Molina-Plaza and Nava Maroto. editor-in-chief, *Aspects of Cognitive Terminology Studies*, 39–72. De Gruyter. https://doi.org/10.1515/9783111073149-003.

León Araúz, Pilar, Arianne Reimerink and Pamela Faber .2019. EcoLexicon and by-products: integrating and reusing terminological resources. In Alcina, A., Costa, R. & Roche, C. editor-in-chief, *Terminology*. Special issue of Terminology and e-dictionaries, 25(2):222-258. John Benjamins Publishing Company. doi:doi.org/10.1075/term.00037.leo.

Premasiri, Damith, Amal Haddad Haddad, Tharindu Ranasinghe and Mitkov, Ruslan. 2023. Deep Learning Methods for Identification of Multiword Flower and Plant Names. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 879-887.

On the Conversion and Linking of Terminological Resources in the Context of Language Data Spaces

Patricia Martín Chozas
Universidad Politécnica de Madrid
patricia.martin@upm.es

Paula Diez Ibarbia
Universidad Politécnica de Madrid
paula.diez@upm.es

Lucía Palacios Palacios
Universidad Politécnica de Madrid
lucia.palacios@upm.es

Elena Montiel Ponsoda
Universidad Politécnica de Madrid
elena.montiel@upm.es

Abstract

In this paper, we present the progress towards an automatic service for the conversion and linking of interoperable terminological resources. On the one hand, we present our efforts towards the conversion of authoritative national and European resources as to the Ontolex-lemon model, the *de facto* standard for language resources. Specifically, we describe the transformation of glossaries from the Catalan Terminology Centre TERMCAT, a highly relevant national resource, and IATE, the most relevant European terminological database. On the other hand, we present the progress towards the development of a Concept Matching algorithm with the aim of unambiguously linking the converted resources mentioned above with other relevant ones, such as EuroVoc.

1 Introduction

The proliferation of Natural Language Processing (NLP) pipelines driven by Artificial Intelligence (AI) has led to an unprecedented increase in the generation of language data. While AI systems are typically trained on vast amounts of general unstructured data, there is a growing demand for domain-specific applications which greatly benefit from domain-specific structured language datasets.

Considering these needs, we are currently involved in the INESData project,[1] a project funded by the Spanish "Ministry for the Digital Transformation and Civil Service" and the European Commission (NextGenerationEU/PRTR) to create a prototype for data spaces. Data spaces are defined as "federated, open infrastructures for sovereign data sharing based on common policies, rules and standards".[2] Amongst other objectives, we aim to develop a Language Data Space in which language resources and language technologies are to be shared for use by third parties (companies, public administrations, etc.). We are particularly interested in the structuring and sharing of language resources according to the standardised and

[1] https://inesdata-project.eu/
[2] https://gaia-x-hub.de/wp-content/uploads/2022/10/White_Paper_Definition_Dataspace_EN.pdf

interoperable formats proposed by the World Wide Web Consortium (W3C).[3] To this end, we are currently working on two strands that we will detail below.

First, we present our efforts towards the conversion of authoritative resources for Spanish terminology to the Ontolex-lemon model,[4] the de facto standard for language resources in the Semantic Web. Specifically, we present our progress towards the conversion of glossaries from the Catalan Terminology Centre TERMCAT,[5] a highly relevant national resource, and IATE,[6] the most relevant terminological database at a European level.

Second, we present the progress towards the development of a Concept Matching algorithm to automatically connect the resources once they are converted. The approach is based on the integration of contextual methods (embeddings) with Language Models and Large Language Models (LLM) to provide contexts that could act as sense indicators.

2 Literature Review

2.1 Related Work on the Conversion of Language Resources

Although interoperable formats for lexicographic and terminological data are limited, some efforts have been made to tackle this challenge. In the lexicographic domain, notable examples include the English lexicon WordNet (McCrae et al., 2014) and the Apertium dictionaries (Gracia et al., 2018), both of which have been converted into Resource Description Frameworks (RDF) using the lemon model. KDictionaries were similarly transformed (Bosque-Gil et al., 2019) based on the Ontolex-lemon model, the evolution of the lemon model.

In the terminological domain, several terminologies from TERMCAT, as well as others like Terminesp,[7] have also been transformed according to the Ontolex-lemon model (Bosque-Gil et al., 2016). Several tools have been developed to assist with these transformations. TBX2RDF[8] and Terme-à-LLOD (Di Buono et al., 2020) are designed to convert TBX files into RDF, while EasySKOS[9] converts CSV or XLSX spreadsheet data into Simple Knowledge Organization Systems (SKOS).[10]

2.2 Related Work on Entity Matching

To interlink specialised terminologies in Semantic Web formats, we adopted Entity or Concept Matching techniques, focused on finding which entries across two knowledge bases refer to the same entity.

The most widespread techniques are those focused on Representation Learning (RL), enabling models to learn low-dimensional vector representations of entities, commonly known as Knowledge Graph Embeddings, such as the TransE model (Bordes *et al.*, 2013). When

[3] https://www.w3.org/
[4] https://www.w3.org/2016/05/ontolex/
[5] https://www.termcat.cat/ca/terminologia-oberta
[6] https://iate.europa.eu/
[7] https://aeter.org/terminesp/
[8] http://tbx2rdf.lider-project.eu/converter
[9] https://terminoteca.linkeddata.es/converter.html
[10] https://www.w3.org/TR/skos-reference/

applied to ontology matching tasks, these approaches can use various types of information, including lexical, structural, semantic, and external sources.

In these experiments, we use lexical information to build sense indicators. Leading-edge systems that leverage lexical information often use text similarity methods and dictionary-based similarity techniques (Liu *et al*., 2021). The former involves analysing textual information from entities and performing string matching to determine similarity (Li *et al*., 2009), while the latter employs NLP methods to analyse the labels and comments of entities. These methods then use resources like dictionaries and thesauri to improve the matching process (Fürst *et al*., 2023).

3 Terminology Conversion

The objective of this terminology conversion task is to provide an automatic service to transform terminological resources in heterogeneous data formats into Linked Data following the Ontolex model. For this purpose, we analysed two resources: TERMCAT and IATE. These resources contain various types of data and cover a great number of domains. However, they are created in different formats, which hinders access and reusability: while TERMCAT terminologies are available in XML, IATE is provided in JSON. To standardise the structure, we propose a four-step pipeline, shown in Figure 1: i) the analysis of the structure and data of each of the resources, to design the modelling; ii) the cleaning and preprocessing of this data; iii) the creation of the mapping rules using mapping tools such as Mapeathor;[11] and iv) the transformation of the resources using tools such as RMLMapper[12] or Morph-KGC (Arenas-Guerrero, 2024) , well-known services for knowledge graph construction.

[11] https://morph.oeg.fi.upm.es/tool/mapeathor

[12] https://rml.io/tools/

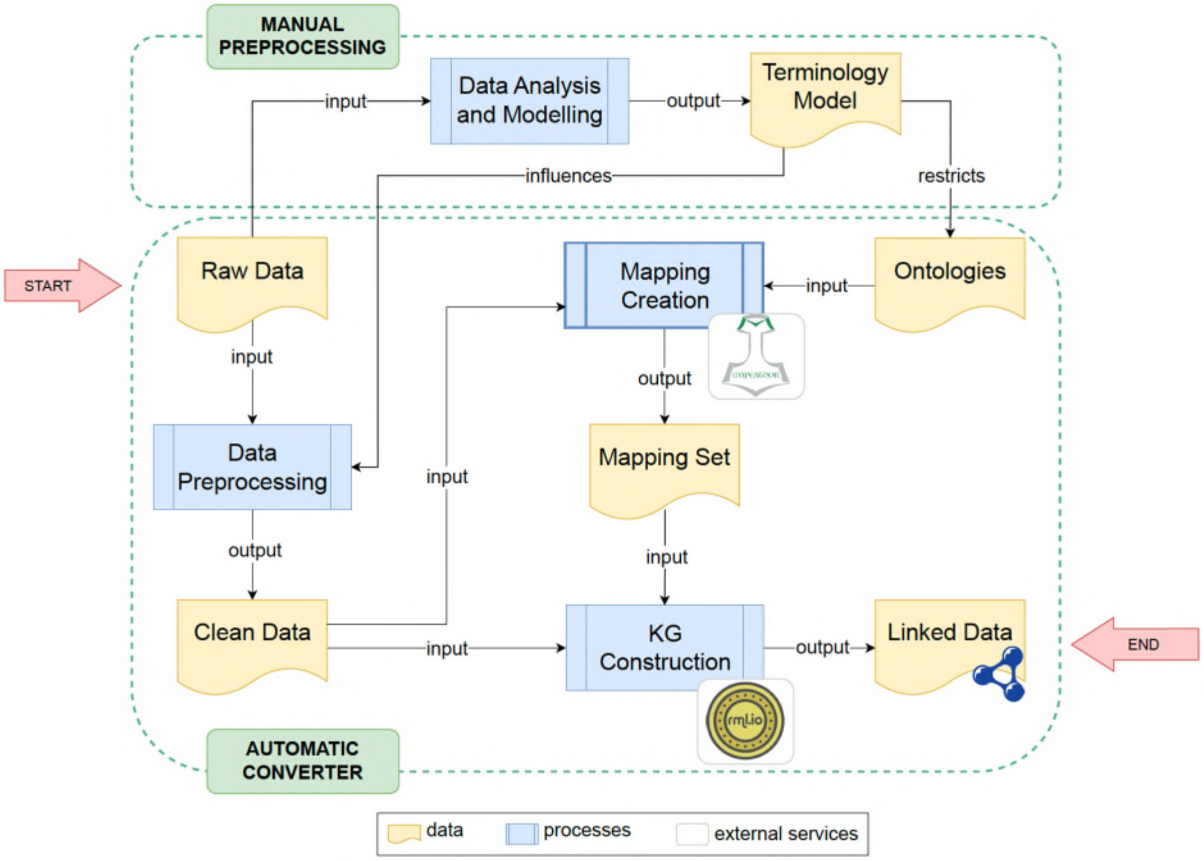


Figure 4. Four-step Conversion Pipeline

The challenges found during the design and implementation of this pipeline are varied. We find especially interesting those related to the analysis of the original resources and the design of the standardised models. The resources transformed include different types of data (see Table 1), which lead to different representations. For instance, while TERMCAT requires a way of modelling terms in sign language (which consists of video URLs), IATE needs a way to represent the sources of terms, definitions and notes. This results in the use of different ontologies (knowledge representation schemas) according to their modelling needs. However, if the resources are to be linked afterwards and follow the same specification, the modelling decisions should be similar (we should not have the same type of data represented in different levels). For this reason, in both cases the representation of the data was based on Ontolex-lemon, the standard model to describe lexicographic data, and both resources follow an onomasiological approach, so we took the modelling decisions accordingly and the final models proposed are complementary.

Some interesting modelling decisions are derived from the types of data in Table 1, such as the use of the EasyTV ontology[13] to model sign language in TERMCAT; the use of the Termlex

[13] https://w3id.org/def/easytv#

proposal[14] to model sources of definitions, notes and usage recommendations in IATE; or the use of OLiA[15] to model transitivity of verbs in TERMCAT.

Types of Information	**TERMCAT**	**IATE**
Languages	17 Official Languages	24 European Official Languages
Language variety information	✓	?
Sign language support	✓	×
Domain of the concept	✓	✓
Notes about the domain of the concept	×	✓
Relations between concepts	×	✓
Concept definitions and notes	✓	✓
Sources of concept definitions and notes	×	✓
Sources of terms	×	✓
Term variants and translations (implicit relations)	✓	✓
Part-of-speech, gender and number	✓	✓
Transitivity and syntactic behaviour of verbs	✓	×
Term-type information	✓	✓
Explicit scientific name information	✓	?
Prefixes and suffixes	✓	×
Usage examples	×	✓
Source of term usage examples	×	✓
Term in context	✓	×
Source of the term in context	×	✓
Domain-specific codes	✓	×
Normative authorization	✓	✓
Term reliability	×	✓

[14] https://termlex.oeg.fi.upm.es/
[15] https://purl.archive.org/olia/olia.owl

Table 4. Different types of information in TERMCAT and IATE.

4 Terminology Linking

Once the resources are converted, we can navigate through the concepts and retrieve information related to the same concept from different resources. As an example, we selected the term *simple majority* that appears both in TERMCAT and IATE. Moreover, this term is also contained in the EuroVoc[16] thesaurus, which is already published as Linked Data.

Therefore, we can now establish links between these three resources and, through a single access point, retrieve different types of data from the three resources:

From TERMCAT,[17] we can retrieve domains, equivalents in other languages, including a translation in Catalan Sign Language,[18] definitions and notes.

From IATE, [19] we can retrieve domains, equivalents in other languages, references, definitions, usage notes and related terms such as relative majority or political majority.

From EuroVoc,[20] we can retrieve domains, equivalents in other languages, and related terms such as majority voting.

4.1 Concept Matching

To automatically establish these links between resources represented as Knowledge Graphs, we are exploring several Concept Matching approaches. At this stage, we are testing different algorithms in a controlled environment with a reduced number of concepts from TERMCAT. We will move to a wider scope (linking IATE and EuroVoc) after getting good performance in current experiments.

At this moment, four experiments have been carried out combining different types of lexical information from the original resources to build contextual embeddings using three multilingual Language Models: Sentence-BERT (Reimers and Gurevych, 2019), RoBERTa (Liu, 2019) and Universal Sentence Encoder.[21]

The different types of information used to build the embeddings are as follows:

Term and translations. For example: *bank account + cuenta bancaria (es) + Bankkonto (de).*

Term and domain. For example: *bank account + economics*.

[16] http://publications.europa.eu/resource/dataset/eurovoc
[17] https://www.termcat.cat/ca/cercaterm/fitxa/NDM2OTQ5OQ%3D%3D
[18] https://youtu.be/ZNyelQiKplQ
[19] https://iate.europa.eu/search/result/1728556981932/1
[20] http://eurovoc.europa.eu/1753
[21] https://www.kaggle.com/models/google/universal-sentence-encoder/tensorFlow2/multilingual-large/2?tfhub-redirect=true

Term, translations and domain. For example: *bank account* + *cuenta bancaria (es)* + *Bankkonto (de)* + *economics*.

Term and fixed context. In this case, it is necessary to build a contextual template for a generic context for all the terms of a domain. In the economics domain, a simple example of the contextual template is: *TERM is related with economics, banks and monetary information*. Therefore, an example of a contextual embedding is: *bank account* + *bank account is related with economics, banks and monetary information.*

4 Conclusions and Future Work

In this paper, we present the progress towards an automatic conversion and linking platform for terminological resources to offer it as a service in a Language Data Space. These experiments are publicly available in a GitHub repository.[22]

The conversion pipeline is already designed and the implementation in the Data Space progresses steadily. However, there is still a manual preprocessing stage that relies heavily on expert knowledge, since the ontologies and models used completely depend on the structure and type of information in the original resource.

The linking service is still in an experimental stage. After analysing the experiments reported, we conclude that, in general, terminological resources are scarce in contextual information. However, using a fixed context is not a good solution either, since the same sentence structure is applied to each term, reducing lexical variety, resulting in homogeneous sentence embeddings. This similarity hinders the identification of exact matches, since the models often generate matches for almost every pair of entries.

To overcome the above-mentioned limitations, the next experiments will include Large Language Models. In the first place, we will feed the models with ontologies, converted resources and mappings, expecting that the models could afterwards propose an automatic RDF design for new resources. In the second place, we will use an LLMs to generate a specific context for each entry that can be used as unique sense indicators for the Concept Matching algorithm.

Acknowledgements

This work was funded by INESData (Infraestructura para la INvestigación de ESpacios de Datos distribuidos en UPM), a UNICO I+D CLOUD project funded by Ministerio para la Transformación Digital y de la Función Pública (PRTR) and the European Union (NextGenerationEU).

References

Arenas-Guerrero, Julián, David Chaves-Fraga, Jhon Toledo, María Pérez Sánchez, and Oscar Corcho. 2024. Morph-KGC: Scalable Knowledge Graph Materialization with Mapping Partitions. *Semantic Web*, (Preprint), 1-20.

[22] https://github.com/oeg-upm/term2LD

Bordes, Antoine, Niculas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. *Advances in Neural Information Processing Systems*, volume 26.

Bosque-Gil, Julia, Elena Montiel-Ponsoda, Jorge Gracia, and Guadalupe Aguado-de Cea. 2016. Terminoteca RDF: A Gathering Point for Multilingual Terminologies in Spain. *In Proceedings of TKE 2016, The 12th International Conference on Terminology and Knowledge Engineering*, pages 136–146.

Bosque-Gil, Julia, Dorielle Lonke, Ilan Kernerman, and Jorge Gracia. 2019. Validating the Ontolex-lemon Lexicography Module with K dictionaries' Multilingual Data. *Proceedings of the eLex Conference*.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. arXiv preprint arXiv:1911.02116.

Di Buono, Maria Pia, Philipp Cimiano, Mohammad Fazleh Elahi, and Frank Grimm. 2020. Terme-a-LLOD: Simplifying the Conversion and Hosting of terminological Resources as Linked Data. *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 28-35.

Fürst, Jonathan, Mauricio Fadel Argerich, and Bin Cheng. 2023. Versamatch: ontology matching with weak supervision. 2023. *In 49th Conference on Very Large Data Bases (VLDB)*. (Vol. 16, No. 6, pp. 1305-1318). Association for Computing Machinery.

Gracia, Jorge, Marta Villegas, Asunción Gómez-Pérez, and Nuria Bel. 2018. The Apertium Bilingual Dictionaries on the Web of Data. *Semantic Web*, *9*(2), 231-240.

Li, Juanzi, Jie Tang, Yi Li, and Qiong Luo. 2008. Rimom: A Dynamic Multistrategy Ontology Alignment Framework. *IEEE Transactions on Knowledge and Data Engineering*, 21(8), 1218-1232.

Liu, Xiulei, Qiang Tong, Xuhong Liu, and Zhihui Qin. 2021. Ontology Matching: State of the Art, Future Challenges, and Thinking based on Utilized Information. *IEEE Access*, *9*, 91235-91243.

McCrae, John, Christiane Fellbaum, and Philipp Cimiano. 2014. Publishing and Linking WordNet using lemon and RDF. *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.

Reimers, Nils, and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Analyzing Interpreting Through NLP and Distributional Semantics — An Automated Interpreter Feedback Framework

Francesco Saina

info@francescosaina.com

Abstract

This work proposes a theoretical and practical automated approach to support the observation and assessment of language interpreting. The analysis of a source speech and its target interpreted rendition (whether pre-recorded or live-streamed audio files or text transcriptions) is performed by computing cosine semantic similarity between the two, derived from a combination of embeddings obtained from cross-lingual sentence encoders and multilingual large language models. This allows for an approximate measure of the transfer of concepts across languages, while enabling the detection of divergences.

Further speech recognition and natural language processing resources are subsequently employed to concentrate on the linguistic elements of the target rendition specifically pertinent to effective interpreting. These aspects include the identification of excessively repeated words (for interpreters to improve their use of synonyms) and the presence of redundant fillers or long pauses (which may be indicative of hesitations). This application could provide interpreting analysis processes with a quantifiable measurement method, beneficial in both professional and educational settings. The system would assist interpreting practitioners in monitoring their performance and honing their skills, and provide interpreter trainers and students with an additional learning aid.

An illustrative prototype tool demonstrating the framework has been developed as an online interface.

1 Introduction

In a language services landscape increasingly characterized by digitalization and a growing need for high-quality interpreting, recent strides in natural language processing (NLP), and notably large language models (LLMs), have led to the exploration of automated approaches to support, enhance, and assess interpreter performance.

While most research on interpreting and technology so far has focused on systems to aid interpreters before and during their activity in the broad sphere of computer-assisted interpreting (CAI), the significantly-enhanced technical capacity of automated resources to process and analyze multilingual data can now also provide interpreters with feedback based on quantifiable metrics obtained from machine learning models.

With automated speech translation systems set to improve their output quality, and thus meet the demand for certain lower-end use cases, interpreters will likely be required to broaden their skillset, advance their expertise, and increasingly offer top-quality services (Saina, 2021). To help meet this need and by leveraging the latest advances in language-related artificial intelligence (AI), this work proposes and introduces an automated framework to analyze the semantic similarity (i.e., the transfer of concepts and meaning) between a source speech and its target interpreted rendition as well as other elements which, as reported in interpreting studies literature, contribute to the perception of quality in interpreting.

This work lays the theoretical foundations (Section 2) for the proposed method of analysis by first situating the framework within the domain of distributional semantics (Subsection 2.1) an approach to representing the meanings expressed by ‘linguistic objects’ in a quantifiable manner, hence enabling their calculation (Bunt and Muskens, 1999). The paper then reviews relevant literature in interpreting studies on criteria for the evaluation of interpreting quality (Subsection 2.2), some of which this framework seeks to address. Section 3 details the technical architecture of the framework and the rationale for it, while Section 4 discusses its potential applications (Subsection 4.1), related work (Subsection 4.2), and limitations (Subsection 4.3), before paving the way for future research and development directions in the Conclusion.

2 Theoretical Background

The success of modern neural models in NLP and language applications of AI, despite their limits, seems to hint at the triumph of distributional, statistical, and probabilistic conceptions of language. As this perspective shifts the focus to the nature of language itself, rather than the cognitive capacities of linguistic agents, it is opposed to the generative approach which has long dominated linguistic research.

While the combinatorial calculus of linguistic units may appear to be only successful at addressing structural aspects of language (Bender and Koller, 2020), the efficacy of modern LLMs still indicates a broader scope for the hypothesis of distributionalism, encompassing semantics as well as different dimensions of language, including phonology and even style.

Current neural language models associate vector representations, or embeddings, to each linguistic unit in a corpus, e.g., each word in the vocabulary of a language. These vectors present dimensions given by the possible linguistic contexts in which those units occur, and are trained as hidden layers in a dedicated neural network, whose task is to predict words out of those surrounding them in a given context. Word embeddings, thus, encode a vast amount of information precisely about word co-occurrence and distribution.

2.1 Distributional Semantics and Its Application in Language Models for NLP and Psycholinguistics

Stemming from structural and corpus linguistics, the distributional semantics approach to language is the theoretical basis for much of the computational work on language, including modern neural networks and deep learning models (Gastaldi and Pellissier, 2021).

Distributional semantics maintains that (any) language is constructed over a distributional structure, based on the occurrence of units (sounds — or, in the domain of modern language processing models, tokens) relative to other units. The distribution environment of such elements is the sum of all its existing co-occurrents in a precise position. The meaning of a word, therefore, can be determined by — or is at least strongly correlated with — the linguistic contexts in which such word occurs (i.e., its statistical distribution). Words appearing in similar circumstances are subsequently assumed to possess analogous or related meanings (Harris, 1954; 1988).

Distributional semantics theories provide methods to represent meaning in natural languages. In NLP and language models, as anticipated above, this is accomplished through vectors encoding the statistical distribution of concepts in linguistic contexts. In addition to being a theoretical model on the expression of meaning for both computational and theoretical linguistics, distributional semantics also provides a diversity of fields with a practical methodology to construct semantic representations, a computational framework for deriving meaning from language data, as well as a cognitive hypothesis on how language usage contributes to shaping meaning (Lenci and Sahlgren, 2023; Gastaldi, 2021).

Over time, the distributional semantics theory has evolved into various computational models enabling a quantification of the semantic distance between units by analyzing their contextual distribution in large corpora. In word embeddings of modern models, dense vector representations of linguistic units capture contextually-rich semantic information by mapping such units into a continuous vector space (Lenci, 2018).

In these spaces, semantic affinity can be measured by cosine similarity, where closer vectors represent semantically related words or concepts (Jatnika et al., 2019). Computing the distance between any pair of such vectors amounts to computing their distributional similarity (the more similar the distribution of two units, the smaller the distance between their vector representations), which turns out to be directly connected with various forms of linguistic relatedness.

Advances in neurolinguistics, psycholinguistics, and cognitive sciences seem to suggest that vectors might effectively represent human concepts computationally (Piantadosi et al., 2024, for a review), confirming the compelling results achieved by recent progress in both LLMs and vector-based symbolic architectures. Theoretical and computational neuroscience is finding that meaning can be appropriately derived from high-dimensional vector spaces through relationships over such concept vectors.

Indeed, concept representations in both cognitive sciences and distributional semantics map similar concepts close to each other in the vector space (Lenci, 2008). This idea also underlies word2vec (Mikolov et al., 2013), BERT (Devlin et al., 2019), and transformer (Vaswani et al., 2017) language models, learning vector-based representations of words to capture features of their usage in context. This suggests that to date, despite all the manifest and undeniable limitations of current models, concepts-as-vectors models appear to hold the largest potential for capturing and encoding meaning in context.

In the framework proposed in this work, the principles of distributional semantics are applied to the observation and analysis of language interpreting. By generating embeddings for sentences in both the source speech and its interpreted rendition, their cosine semantic similarity is calculated as a proxy for conceptual equivalence (Wieting et al., 2019; Zhang et al., 2020). This allows an approximate estimation of the accuracy of the interpreter’s transfer of meaning between languages.

2.2 Research in Interpreting Quality

Assessing the quality of simultaneous interpreting is a complex exercise (Riccardi, 2002), due to the array and nature of strategies commonly adopted by interpreters to convey a message in another language as well as the layered and context-dependent nuances of real-time multilingual communication processes.

While the notion of quality has constantly been explored and debated in interpreting studies, expectations of end users of interpreting services may be very diverse (Kurz, 2001). Solid evaluation methods would, therefore, be helpful for a variety of purposes. Interpreting quality assessment can provide valuable insights for practitioners, trainers, and students, as well as scholars, certification bodies, and even customers. However, like any manual evaluation, it is time-consuming and resource intensive, thus performed in limited scenarios.

Theoretical models of interpreting quality have often emphasized the relevance of a series of aspects, including accuracy, equivalence, fidelity to the source text, fluency, lexical variety, appropriateness, and 'usability' for the audience (Viezzi, 1999; Pöchhacker, 2002). Interpreting fidelity is conceived as the degree to which the meaning of a source speech is accurately conveyed in a target language, despite potential structural or cultural differences. This is closely related to the notion of semantic similarity, making it a suitable focus for NLP-based assessment tools. However, accurate and faithful interpreting does not only refer to the transfer of concepts, but also comprehends the pragmatic dimension, ensuring that the speakers' communicative intents are equally preserved, beyond the transposition of meaning at the word level (Lederer and Seleskovitch, 1984; Setton and Motta, 2007).

The fluency of an interpreted rendition is equally crucial to ensure listener reception and message comprehension. Excessive use of filler words, hesitations, and undue long pauses can undermine an interpreter's credibility or make it difficult for a listener to seamlessly follow the interpretation, thus degrading the perception of quality (Shlesinger, 1997). These elements are typically seen as markers of disfluency and a hindrance to understanding. Lexical variety also contributes to the effectiveness of the communication act, as overuse of certain terms can lead to redundancy, while the usage of a broad, rich, and diverse vocabulary repertoire can contribute to exhibiting a skillful command of language. Word frequency and repetition analysis in the target rendition can provide valuable feedback to interpreters and foster their exploration of synonyms and alternatives.

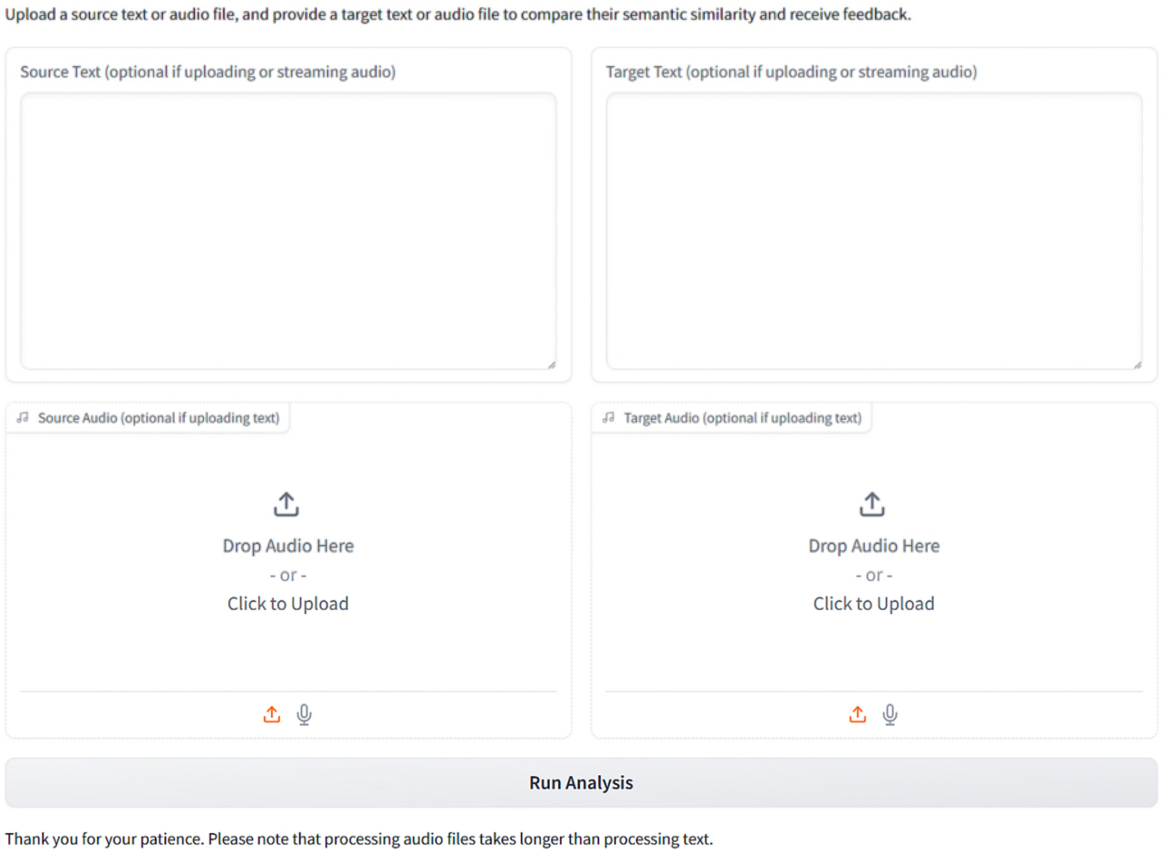
The framework proposed here for an automated measurement system to assist interpreting assessment aims at encompassing at least some of the aspects of interpreting quality evaluation mentioned above. It does so by computing semantic similarity and offering linguistic insights into the delivery of an interpretation. The alignment between source and interpreted texts is calculated as an application of meaning representations derived from distributional semantics, although it is to be acknowledged that semantic similarity measures of this kind may not capture all the nuances of human communication involved in such a process. Likewise, additional linguistic elements reported in interpreting literature as disfluency markers or factors impacting listener reception (e.g., false starts, self-corrections, delivery pace, rhythm, or voice pitch) are not comprised in the framework presented.

3 Methodology and Technical Framework

The framework proposed in this work relies on recent technical advancements in NLP (underpinned by the models of distributional semantics) to assess language interpreting. The evaluation is carried out by analyzing a source text and its target interpreted rendition, provided as either audio files or text transcriptions.

This section outlines the essential stages of the methodology employed, which implements automatic speech recognition (ASR), embedding generation, sentence alignment, semantic similarity computation, and additional linguistic analysis. An illustrative prototype demonstrating the practical application of the framework has been developed as an online interface, currently available as a private Hugging Face Space restricted to research purposes.

The tool allows users to upload source and target speeches as either text transcriptions or audio recordings, and receive an analysis of the interpretations, as outlined in this work. Input languages are detected automatically. The results are returned in an accessible format, including visualizations of the average percentage of semantic similarity, examples of semantically divergent aligned sentences, word frequency and fillers lists, and indications of the longest pauses. This interface serves as a proof of concept, illustrating the potential of the framework in the field of language interpreting.



Upload a source text or audio file, and provide a target text or audio file to compare their semantic similarity and receive feedback.

Source Text (optional if uploading or streaming audio)
Target Text (optional if uploading or streaming audio)
Source Audio (optional if uploading text)
Drop Audio Here
- or -
Click to Upload
Target Audio (optional if uploading text)
Drop Audio Here
- or -
Click to Upload
Run Analysis

Thank you for your patience. Please note that processing audio files takes longer than processing text.

Figure 1. Framework Prototype Input Interface.

In the left boxes, the user can upload either a source text or audio file, and then, in the right boxes, provide a target text or audio file. The analysis compares their semantic similarity and provides feedback.

3.1 Automatic Speech Recognition and Transcription

If the (source or target) input is provided as audio file, whether pre-recorded and uploaded or directly live-streamed onto the framework interface, the first step required is the conversion into text using an ASR model.

In the prototype system developed for the demonstration of the proposed framework, OpenAI's open-source Whisper model (Radford et al., 2023) was used for speech recognition and transcription. It is a state-of-the-art model capable of handling multiple languages (up to

57 with <50% word error rate), a feature needed to process data and examine interpreting performance across several language combinations. The output (a starting point for subsequent processing) is an automatically-generated textual version of the source and target speeches to be analyzed.

3.2 Embedding Generation

After obtaining the transcriptions (or using the pre-provided texts), the next stage is the generation of embeddings encapsulating the meaning at the sentence level. These embeddings are encoded vector representations of both source and target sentences as sequences of numbers in a high-dimensional space. In this space, each component in the vector corresponds to a particular feature of the meaning of the sentence, so that semantically similar sentences are located closer together.

In this framework, the embeddings are produced through a combination of cross-lingual sentence encoders and multilingual LLMs. The pre-trained models used in the framework prototype for the semantic representation of the texts are Multilingual E5 (Wang et al., 2024) and the BLOOM multilingual LLM (BigScience Workshop, 2022). Both were chosen for their wide language coverage (100 and 46 natural languages, respectively) as well as their open nature, allowing for testing and demonstration purposes.

Cross-lingual embeddings map linguistic data from various languages into a shared vector space, enabling direct comparison of content across different languages by placing semantically similar embeddings close to each other. This bypasses the need for intermediate machine translation (MT) and avoids the potential introduction of ‘translationese’ biases. By operating on a higher representation of concepts above the word-for-word level, interference from the source language exhibited in machine-translated texts, e.g., unnatural syntactical patterns or lexical choices (Bizzoni et al., 2020; Vanmassenhove et al., 2021), can be prevented. The over-representation of linguistic properties of the source language in the target language (e.g., in terms of structure or meaning expression) could indeed lead to a less accurate representation of the target language’s natural use. With cross-lingual sentence embeddings, texts may therefore be processed as two independent ‘originals’. However, this approach requires indirect alignment of meanings across languages, which is more challenging and potentially inaccurate.

3.3 Sentence Alignment Using Dynamic Time Warping

Strategies adopted by interpreters to render concepts across languages are varied and articulated. They can include rephrasing, adaptation, expansion, generalization, explanation, addition, or intentional omission to adapt the messages to the shared linguistic and cultural conventions of the target language.

Therefore, an interpretation may not follow the structure of the original speech, and sentences conveying the same meaning may be located in different positions. This raises the need to identify correlated source and target embeddings to allow their effective comparison and analysis.

In response to this challenge, to pair and evaluate the speech transcriptions successfully, a dynamic time warping (DTW) algorithm is introduced into the framework for sentence alignment. Already applied in various speech recognition tasks (Juang, 1984), DTW can compute pairwise cosine distances between the embeddings of source and target sentences, thus enabling the alignment of semantically similar sentences.

3.4 Semantic Similarity Computation

The aligned sentence pairs are employed to calculate the similarity of meaning transferred between the source and target speech transcriptions. Semantic similarity (in this context, meaning transfer) is quantified by using the measure of cosine similarity. Cosine similarity is computed between two vectors in a high-dimensional space. The vectors represent the semantic embeddings of the source and target sentences. The formula is as follows:

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

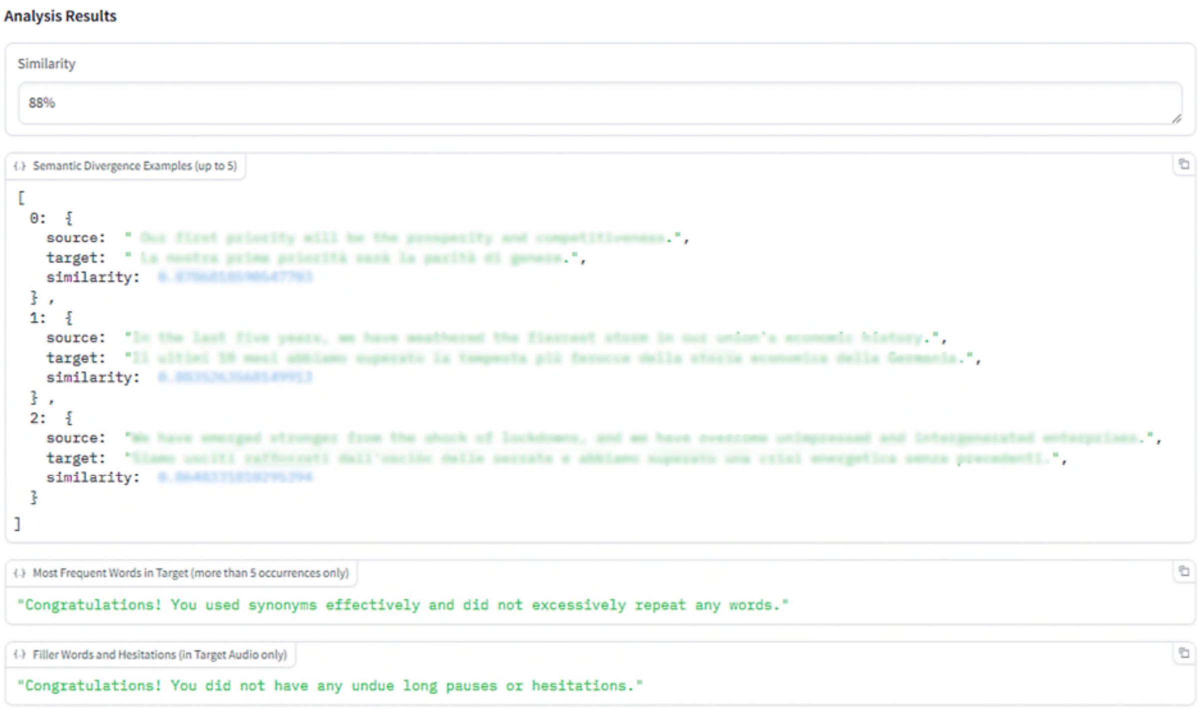
Given two embedding vectors for the source and target sentences (**A** and **B**, respectively), the cosine similarity, **cos(θ)**, is the cosine of the angle between the vectors represented by the dot product of the two, divided by the product of their lengths (‖**A**‖ and ‖**B**‖ are the magnitudes). The resulting similarity is a value ranging from -1 (exactly opposite) to 1 (perfectly the same), with 0 indicating orthogonality, or decorrelation, between the vectors and in-between values indicating intermediate similarity or dissimilarity.

After computing the cosine similarity between the embeddings of each sentence in a pair, all the calculated similarity scores are aggregated to produce an average similarity score for the entire interpretation. This average score (converted into percentage and displayed as such on the prototype interface for ease of consultation) provides a quantitative estimate of how accurately the interpreter has conveyed concepts from the source to the target language. A list of examples of divergence instances, i.e., major discrepancies and sentences where the interpretation may have strayed far from the source speech, is also shown.

3.5 Linguistic Fluency Analysis

In addition to semantic similarity calculation, the framework can also deliver an analysis of the target rendition in text form, focusing on linguistic factors contributing to the fluency and perceived quality of the interpretation. The aspects considered are disfluency markers like excessively repeated words, fillers, and long pauses, all indicative of hesitations or uncertainty. By pointing out these factors, the framework encourages the use of a more varied vocabulary and synonym repertoire, and the mitigation of delivery hesitations. This can provide interpreters with measured, actionable feedback on their performance to improve the overall quality of their interpretation.

To detect instances of these elements in the prototype framework, certain reference thresholds needed to be established: terms repeated more than 5 times, a set of filler expressions (e.g., ‘uh’, ‘ehm’), and pauses longer than 3 seconds. For a more functional and robust framework design, further user perception research would obviously need to be conducted to identify appropriate conditions based on empirical evidence. Such criteria could also be left to customization by users, based on their specific needs, use cases, and application scenarios.



Analysis Results

Similarity

88%

Semantic Divergence Examples (up to 5)

```
[
  0:  {
    source:  " Our first priority will be the prosperity and competitiveness.",
    target:  " La nostra prima priorità sarà la parità di genere.",
    similarity:  [illegible]
  } ,
  1:  {
    source:  "In the last five years, we have weathered the fiercest storm in our union's economic history.",
    target:  "Il ultimi 18 mesi abbiamo superato la tempesta più feroce della storia economica della Germania.",
    similarity:  [illegible]
  } ,
  2:  {
    source:  "We have emerged stronger from the shock of lockdowns, and we have overcome [illegible] and [illegible] enterprises.",
    target:  "Siamo usciti rafforzati dall'uscita delle serrate e abbiamo superato una crisi energetica senza precedenti.",
    similarity:  [illegible]
  }
]
```

Most Frequent Words in Target (more than 5 occurrences only)

```
"Congratulations! You used synonyms effectively and did not excessively repeat any words."
```

Filler Words and Hesitations (in Target Audio only)

```
"Congratulations! You did not have any undue long pauses or hesitations."
```

Figure 2. Framework Prototype Output Interface.

The prototype provides the average semantic similarity score between the two texts (displayed as a percentage value), semantic divergence examples (up to 5), a list of the most frequent words (with more than 5 occurrences), fillers, and hesitations in the target.

4 Discussion

The challenges of automating the evaluation of interpreting stem from the intrinsic nature of interpreted renditions, characterized by the non-linear strategies mentioned above that interpreters adopt to best deliver messages in contextually-appropriate ways. Despite advances in context-aware language models, these still lack the holistic understanding of in-setting and socially-situated quality that only humans possess, and the automatic calculation of meaning alignment disregards essential subtleties of interpersonal communication.

A structured analysis and assessment of spoken interlingual mediation through textual transcriptions only, albeit usable in certain scenarios (Romero-Fresco and Pöchhacker, 2017; Korybski et al., 2022; Alonso-Bacigalupe, 2023), cannot account for the entire set of oral features characterizing spoken language mediated exchanges either. It also leaves out the perspective of user perception in terms of information retention, intelligibility, and ultimately communicative effectiveness.

Nevertheless, the framework aims to represent an advancement in the automated analysis of, and support for, interpreting assessment. By leveraging distributional semantics theories and models as well as state-of-the-art NLP techniques and applications, it can offer interpreters constructive feedback that can help improve their semantic accuracy and delivery fluency.

4.1 Potential Applications

By providing feedback based on quantifiable metrics, this framework could find practical applications spanning both professional and educational settings. For practicing interpreters, this system would offer a resource to monitor performance and refine skills, with the aim of consistently delivering high-quality interpretations. Interpreters may count on such an overview of at least some aspects of their renditions for professional development and continued improvement.

In interpreter training, the framework could serve as a powerful learning aid for trainers and students. The system could support interpreter trainers in the assessment of students' renditions, and, in turn, such a tool would guide trainees in focusing on areas including vocabulary expansion and delivery confidence. Interpreter trainers and students could rely on it as an additional resource to further elaborate on interpreting processes in the classroom.

In consideration of the inherent constraints and only approximate capabilities of an estimation mechanism like this, all review and assessment processes should be attentively overseen to ponder and compensate for the limitations of the fully-automated approach. Indeed, the automatic assessment of an interpreted text using the transcripts only is bound to some limitations. Being on a purely textual level, it cannot consider non-verbal and prosodic features of the delivery (intonation, voice modulation, stress, rhythm), which are essential components of interpreting evaluation.

A comparative assessment uniquely based on source and target texts can only provide an overall quality evaluation that concentrates on a limited set of components. As a comprehensive assessment is therefore out of reach, such a framework is designed to serve as an assisting companion.

4.2 Related Work

NLP- and LLM-based tools and systems designed for automated interpreter quality evaluation and assessment represent a recent trend in interpreting quality research (Ünlü, 2023; Wang and Fantinuoli, 2024). Some of these frameworks also include MT estimation metrics (Stewart et al., 2018; Lu and Han, 2022), while others just offer an overall accuracy percentage as compared to adherence to a pre-loaded reference target text (Corpas Pastor, 2017). All these applications demonstrate a growing interest in similar approaches, with positive correlations between automatic metrics and human scores suggesting potential for automation.

However, often these methods (Ünlü, 2023; Wang and Fantinuoli, 2024) only resort to direct prompting of the language models to assess semantic similarity, with no additional processing. Semantic similarity is frequently also used as the sole measure of overall interpreting quality, while the approach presented in this work aims to go beyond this, by offering more comprehensive feedback on interpreting and linguistic performance. Additionally, the use of cross-lingual sentence embeddings, as in this framework, allows for more accurate comparisons across languages, suffering less from potential biases, errors, or even prompt-induced interference from the models. Nevertheless, at present, none of the current frameworks is capable of addressing dimensions like prosodic features of the delivery.

Still, reported small-scale experiments on the methods mentioned have already confirmed their effectiveness with respect to human assessment and scoring. Additional features, like the generation of concise evaluation reports in natural language (Ünlü, 2023), make them compelling and convenient for the user.

4.3 Technical Limitations and Ethical Considerations

The proposed framework certainly holds promise for the future, although technical limitations exist that cannot be neglected. The accuracy of the analysis depends on several factors, and misinterpretations may occasionally result from errors introduced in previous stages.

Firstly, the quality and precision of the automatically-generated transcription. ASR faults may lead to incorrect texts and, thus, compromise the subsequent steps. Secondly, cross-lingual embeddings and multilingual LLMs have undoubtedly made considerable strides; however, they are not always capable of capturing the whole array of meaning nuances.

Furthermore, DTW assumes correspondences between source and target sentences which might not always hold in less literal or linear interpretations, thus affecting the reliability of the alignment, particularly when the source and target languages have significant structural differences. Likewise, the cosine semantic similarity metric, while useful, cannot account for all traits of human interpretation, leaving out aspects contributing to the construction of meaning like intonation (Tsiamas et al., 2024), emotion, and other non-verbal clues.

In addition to these step-specific shortcomings, all the stages rely on pre-trained language models, which may not perform equally across all languages, in particular low-resource ones. While these technical limitations may be mitigated over time with the ongoing refinement of models, they also constitute inherent constraints in any fully-automated assessment of interpreting that need to be acknowledged.

The use of automated metrics and methods to evaluate interpreting performance also raises relevant ethical considerations about the transparency and fairness of the assessment. While the framework can provide helpful feedback, it should not be viewed as a definitive evaluation system without human oversight. For instance, language industry stakeholders contracting or evaluating interpreting services (service providers, institutions, organizations, accreditation entities) should not regard this as a standalone solution to consistently and objectively measure, examine, or monitor interpreting performance without supervision.

Confidentiality concerns also arise with the handling of data, especially if proprietary commercial models are deployed. Finally, biases of various sorts may be reflected, perpetuated, and propagated in language models, thus impacting in several ways the linguistic analysis and the output of the system.

Yet, these limitations and considerations do not hinder the framework from representing a step forward in the analysis of interpreting, in both research and practice, by integrating linguistic theories with advanced computational techniques.

5 Conclusion

With digital technology applications gaining increasing relevance in interpreting research and practice (Corpas Pastor and Defrancq, 2023), this work proposes a framework offering an innovative and promising approach to interpreting observation and assessment.

Interpreting quality is inherently contextual, as it does not merely depend on measurable elements such as those considered in this work, but also includes factors like the purpose of the interpretation, the expectations of the different parties involved, and the appropriateness for the subjects and settings in which the interpreting takes place. This framework, therefore, aims to offer a foundational assessment that can subsequently be complemented by human judgment in specific and contextually-situated evaluations. These will indeed account for a more comprehensive range of factors which are challenging to discretize and model.

Still, automated frameworks, such as the method presented in this work, aim to stimulate a novel approach to addressing some of these dimensions through the application of theoretical linguistic models and advances in NLP techniques. The integration of innovative systems and methods can offer interpreting research and practice new resources and perspectives.

The framework presented has been implemented as an illustrative, limited-access online interface where users can upload source and target speeches as either text transcriptions, audio files, or live-streamed recordings. The system designed provides an analysis of the interpretation, returning an overall semantic similarity score, aligned sentences with examples of semantic divergences, lists of frequently repeated words and fillers, and instances of long pauses.

The tool could also pursue the integration of additional features, such as the interpreter’s talking speed (the average number of words uttered per minute, when audio files are provided) to assess ‘listenability’, or an option to display the initial transcriptions of both the source and target speeches in their entirety. The latter would enable users to perform an autonomous self-assessment by directly comparing the two texts, one alongside the other. In this way, users would not only count on the automatic metrics, but also engage in a more reflective review of the renditions.

The framework is undergoing empirical experiments with different languages, and preliminary testing has already demonstrated its potential to serve as a helpful resource in interpreting quality observation and assessment for interpreting practitioners, trainers, and students. In the future, if such feedback were provided in real time, interpreters could potentially even make on-the-fly adjustments and enhance their overall performance.

This paper is just the initial stage of a longer research trajectory, the next step being the comparison with and benchmarking against human assessments to gauge and validate the framework’s effectiveness and reliability in practical scenarios. At present, the scarcity of large-scale, multilingual open data and benchmarks for human simultaneous interpreting hampers a rapid, streamlined process for preliminary automated evaluations.

Subsequent directions could then explore an evaluation of the tool’s performance with different models (including smaller ones, run locally), with various levels of model temperature, and, in the longer term, with the integration of solid end-to-end speech-to-speech models to bypass the textual transcription.

This paper lays the theoretical foundations and presents a computational framework for an effective automated approach to observe and analyze language interpreting. While its applied robustness can only be assessed through empirical evaluation, the object of current ongoing efforts, this work commits to the broader advancement and innovation of the field.

References

Alonso-Bacigalupe, Luis. 2023. Joining Forces for Quality Assessment in Simultaneous Interpreting: the NTR Model. *Sendebar. Revista de Traducción e Interpretación*, 34: 198–216.

Bender, Emily M., and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198.

BigScience Workshop. 2022. *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. https://arxiv.org/abs/2211.05100 [last accessed November 4, 2024].

Bizzoni, Yuri, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How Human is Machine Translationese? Comparing Human and Machine Translations of Text and Speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290.

Bunt, Harry, and Reinhard, Muskens. 1999. Computational Semantics. In Harry Bunt, and Reinhard Muskens (eds.), *Computing Meaning*. Dordrecht: Springer, pages 1–32.

Corpas Pastor, Gloria. 2017. VIP: Voice-Text Integrated System for Interpreters. In *Proceedings of the 39th Conference Translating and the Computer*, pages 7–10.

Corpas Pastor, Gloria, and Bart Defrancq (eds.). 2023. *Interpreting Technologies – Current and Future Trends*. John Benjamins.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Gastaldi, Juan Luis. 2021. Why Can Computers Understand Natural Language? The Structuralist Image of Language Behind Word Embeddings. *Philosophy & Technology*, 34(1): 149–214.

Gastaldi, Juan Luis, and Luc Pellissier. 2021. The calculus of language: explicit representation of emergent linguistic structure through type-theoretical paradigms. *Interdisciplinary Science Reviews*, 46(4): 569–590.

Harris, Zellig Sabbettai. 1954. Distributional Structure. *WORD*, 10(2–3): 146–162.

Harris, Zellig Sabbettai. 1988. *Language and Information*. Columbia University Press.

Jatnika, Derry, Moch Arif Bijaksana, and Arie Ardiyanti Suryani. 2019. Word2Vec Model Analysis for Semantic Similarities in English Words. *Procedia Computer Science*, 157: 160–167.

Juang, Biing-Hwang. 1984. On the hidden Markov model and dynamic time warping for speech recognition — A unified view. *AT&T Bell Laboratories Technical Journal*, 63(7): 1213–1243.

Korybski, Tomasz, Elena Davitti, Constantin Orăsan, and Sabine Braun. 2022. A Semi-Automated Live Interlingual Communication Workflow Featuring Intralingual Respeaking: Evaluation and Benchmarking. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 4405–4413.

Kurz, Ingrid. 2001. Conference Interpreting: Quality in the Ears of the User. *Meta*, 46(2): 394–409.

Lederer, Marianne, and Danica Seleskovitch. 1984. *Interpréter pour traduire*. Paris: Didier Érudition.

Lenci, Alessandro. 2008. Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, 1(20): 1–31.

Lenci, Alessandro. 2018. Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4(1): 151–171.

Lenci, Alessandro, and Magnus Sahlgren. 2023. *Distributional Semantics*. Cambridge: Cambridge University Press.

Lu, Xiaolei, and Chao Han. 2022. Automatic assessment of spoken-language interpreting based on machine-translation evaluation metrics: A multi-scenario exploratory study. *Interpreting*, 25(1): 109–143.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. *Efficient Estimation of Word Representations in Vector Space*. https://arxiv.org/abs/1301.3781 [last accessed November 4, 2024].

Piantadosi, Steven T., Dyana C.Y. Muller, Joshua S. Rule, Karthikeya Kaushik, Mark Gorenstein, Elena R. Leib, and Emily Sanford. 2024. Why concepts are (probably) vectors. *Trends in Cognitive Sciences*, 28(9): 844–856.

Pöchhacker, Franz. 2002. Researching interpreting quality: Models and methods. In Giuliana Garzone, and Maurizio Viezzi (eds.), *Interpreting in the 21st Century*. John Benjamins, pages 95–106.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518.

Riccardi, Alessandra. 2002. Evaluation in interpretation: Macrocriteria and microcriteria. In Eva Hung (ed.), *Teaching Translation and Interpreting 4*. John Benjamins, pages 115–126.

Romero-Fresco, Pablo, and Franz Pöchhacker. 2017. Quality assessment in interlingual live subtitling: The NTR Model. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 16: 149–167.

Saina, Francesco. 2021. Technology-Augmented Multilingual Communication Models: New Interaction Paradigms, Shifts in the Language Services Industry, and Implications for Training Programs. In *Proceedings of the 1st Workshop on Automatic Spoken Language Translation in Real-World Settings*, pages 49–59.

Setton, Robin, and Manuela Motta. 2007. Syntacrobatics: Quality and reformulation in simultaneous-with-text. *Interpreting*, 9(2): 199–230.

Shlesinger, Miriam. 1997. Quality in simultaneous interpreting. In Yves Gambier, Daniel Gile, and Christopher Taylor (eds.), *Conference Interpreting: Current trends in research*. John Benjamins, pages 123–132.

Stewart, Craig, Nikolai Vogler, Junjie Hu, Jordan Boyd-Graber, and Graham Neubig. 2018. Automatic Estimation of Simultaneous Interpreter Performance. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 662–666.

Tsiamas, Ioannis, Matthias Sperber, Andrew Finch, and Sarthak Garg. 2024. *Speech is More Than Words: Do Speech-to-Text Translation Systems Leverage Prosody?* https://arxiv.org/abs/2410.24019 [last accessed November 4, 2024].

Ünlü, Cihan. 2023. InterpreTutor: Using Large Language Models for Interpreter Assessment. In *Proceedings of the International Conference on Human-Informed Translation and Interpreting Technology*, pages 78–96.

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*, pages 2203–2213.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30: 5998–6008.

Viezzi, Maurizio. 1999. Aspetti della qualità nell'interpretazione. In Caterina Falbo, Mariachiara Russo, and Francesco Straniero Sergio (eds.), *Interpretazione Simultanea e Consecutiva – Problemi Teorici e Metodologie Didattiche*. Milano: Ulrico Hoepli, pages 140–151.

Wang, Liang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. *Multilingual E5 Text Embeddings: A Technical Report*. https://arxiv.org/abs/2402.05672 [last accessed November 4, 2024].

Wang, Xiaoman, and Claudio Fantinuoli. 2024. Exploring the Correlation between Human and Machine Evaluation of Simultaneous Speech Translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation*, pages 327–336.

Wieting, John, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU: Training Neural Machine Translation with Semantic Similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *ICLR 2020*.

Better multilingual AI for Europe: Translators and supercomputers at the European Commission – DG Translation

Mihai Cristian Brașoveanu
European Commission – DG Translation
mihai.brasoveanu@ec.europa.eu

Bhavani Bhaskar
European Commission – DG Translation
bhavani.bhaskar@ext.ec.europa.eu

Ilja Rausch
European Commission – DG Translation
ilja.rausch@ec.europa.eu

Abstract

Translators and other language professionals are not only users of artificial intelligence (AI) but also co-creators of AI. AI is built on extensive textual data, and the quality of that data plays a key role in the performance of AI models trained with it. Translators and other language professionals create large volumes of multilingual text data of the highest quality every single day. Their work can contribute to building better multilingual AI for everybody. The European Commission's DG Translation has embarked on a new pathway towards creating large language models (LLMs) with better multilingual abilities by leveraging the high quality multilingual data from EURAMIS, the EU's interinstitutional translation database, and the power of the EuroHPC network of European supercomputers. Our results show that when trained on high-quality data translated by experts, the model's performance can be improved in tasks such as machine translation and text completion, while we also highlight the challenges of catastrophic forgetting that arise when trained on multiple languages.

1 Introduction

The advent of generative artificial intelligence (AI) with large language models (LLMs) represents a huge breakthrough in model capabilities. Such models, typically trained on extensive textual data, have shown remarkable performance in tasks like text generation, question answering, data retrieval and summarization. They also show promising results for translation from and to high-resource languages (Dubey et al., 2024). However, current LLMs are often hampered by limitations, including incomplete language coverage, which leads to an increased imbalance between languages. This is particularly concerning for EU institutions and more broadly in the EU, where multilingualism is a fundamental aspect of communication.

The commonly used datasets for training LLMs are often open and transparent (Penedo et al., 2024), but this is not always the case, and the use of non-transparent datasets can raise concerns about data quality, bias, and copyright infringement (Ferdaus et al. 2024). Furthermore, the under-representation of low-resource languages in these datasets results in LLMs that are less inclusive and less effective for those languages.

To address these limitations, the European Commission's Directorate-General for Translation (DGT) has started a project to leverage its high-quality multilingual text data to improve the multilingual capabilities of LLMs. The DGT EU LLM project is also expected to produce new insights into how high-quality data created by language professionals affects LLM performance.

2 Translators and text data at DGT

DGT is one of the largest translation services in the world with around 1 400 translators, language technology experts, quality officers, terminologists, and revisers (OP, 2023). DGT works primarily on translating legislation, policy documents and communications with citizens in the EU's 24 official languages (OP, 2023).

DGT has a large volume of text data from translation work, growing every single day. In 2023, DGT translators produced 2.5 million pages of translation across all EU official languages and all EU policy areas (OP, 2023). All the text data from DGT's translation work is kept in a specialised database for multilingual translation memories called EURAMIS (Leick, 1995). The EURAMIS database also includes text data from other translation services in the EU institutions. In total, the EURAMIS database currently contains around 2 billion translation segments, representing about 100 billion tokens for LLM training.

DGT data is of the highest quality. Translators recruited by DGT and other translation services in other EU institutions are required to have perfect command of their mother tongue or main language, hold a university degree of minimum 3 years, and have other key qualifications (OP, 2023). DGT translators' roles include producing translations that meet the quality requirements laid down in DGT's translation quality management framework, but also performing quality control (DGT, 2023).

DGT and the other translation services in other EU institutions also use comprehensive translation quality management systems. In the case of DGT, the internal Translation Quality Management System (TQM) ensures that quality is embedded in the translation process (DGT, 2023). DGT's TQM framework includes key translation quality management principles, guidelines for translation quality and evaluation of translation quality, and other language-specific guidelines. Quality requirements for DGT translations cover accuracy, terminology, style, linguistic norms, and design (DGT, 2023). Quality assurance involves activities before, during, and after production, by different actors including translators and assistants, quality officers and quality managers (DGT, 2023).

Thus, annually, DGT and other translation services in the other EU institutions generate a vast amount of human-translated content of the highest quality covering all 24 official EU languages and encompassing a broad range of EU policy areas, thus offering a rich source of diverse text data. The EURAMIS database, which houses this data, is therefore a comprehensive and multilingual resource. Its value lies not only in its high level of quality but also in its considerable corpus of texts from low-resource languages, including Maltese, Irish, Estonian and more. DGT, in partnership with DG Communications Networks, Content and Technology (CONNECT), initially used the EURAMIS database to build the eTranslation neural machine translation, started in 2018 and serving not only the European Commission and other EU institutions but also a broad range of stakeholders in the EU (EC, 2024).

3 DGT EU LLM project

There is considerable potential value in leveraging DGT's high quality multilingual text data from the translation work of professional EU institutional translators and using it in the process of building LLMs. The DGT EU LLM project explores that value by using DGT's data to perform continual pre-training of an existing open source LLM using the supercomputing infrastructure provided by the EuroHPC Joint Undertaking.

3.1 Methodology

Our project revolves around the use of non-public, high-quality, multilingual data from DGT's EURAMIS database for continual pre-training of LLMs. Since its initiation, the project was divided into two phases: a small-scale phase using the Llama 2 13B model (Touvron et al., 2023) continually pre-trained on two languages (first on Slovenian and, subsequently, on Croatian), followed by a larger-scale phase involving all 24 official languages of the European Union and a larger Mixtral model (at least Mixtral 8x7B with ca. 47B parameters in total) (Jiang et al., 2024). The data used for this project is sourced from the EURAMIS database (Leick, 1995). As the bulk of the EURAMIS data is not public, the continual pre-training is done on largely unseen data, avoiding overfitting.

The compute resources for this project are provided by the EuroHPC infrastructure, which enables the efficient processing of large-scale data (Skordas, 2019). The small-scale phase of the project was executed on the MeluXina supercomputer with a development access to 3000 node hours on 4xA100(40GB) GPU nodes (LuxProvide, 2024). The larger-scale phase is leveraging the Leonardo supercomputer with an access to 50 000 node hours on 4xA100(64GB) GPU nodes (Turisini *et al.*, 2023). The team involved in this project consists of experts with a range of skills and profiles, including linguists, AI engineers and IT specialists.

The new models produced in the first phase were evaluated in three steps (Rausch *et al.*, 2024). The first evaluation step used two standardized benchmarks (ARC and HellaSwag) (Chollet, 2019; Zellers *et al.*, 2029), which were translated into Slovenian and Croatian by the project team using DGT's own eTranslation neural machine translation system (EC, 2024). These benchmarks contain questions and multiple-choice answers. Both benchmarks were evaluated with three configurations: (i) question in English and answer in Slovenian or Croatian (short: en2sl and en2hr), (ii) vice versa (short: sl2en and hr2en) and (iii) both question and answer in Slovenian or Croatian (short: sl2sl and hr2hr).

The second evaluation step involved tests on segment-based translation tasks, in which the models were given English input segments and asked to translate them into Slovenian and Croatian. The results were measured using the SacreBLEU score (Post, 2018), which assesses the similarity between the model's output and the reference translation.

The third evaluation step involved professional translators as evaluators. The models were assessed based on the text completion task in Slovenian and Croatian, where the evaluators were asked to choose their preferred output and rate the relevance and fluency of each completion.

3.2 Results

A more detailed description of our study's methodology, results, and discussion can be found in (Rausch et al., 2024). This paper provides a concise summary of the key findings and implications of the study. We observed mixed results for the impact of DGT’s data on the accuracy of the continually pretrained model. However, the models that underwent continual pre-training with DGT data did outperform the original base model in several tasks in the respective target languages.

The evaluation results based on the ARC and HellaSwag standardized benchmarks showed that the models' performance decreased in accuracy for sl2en and hr2en (Rausch *et al.*, 2024). However, the Slovenian model (sl-model) slightly outperformed the others in its target language (xx2sl). The Croatian model (hr-model) performed best only in the en2hr case, but not in the other configurations. Overall, the results suggest that the models' performance is affected by the language of the question and answer, highlighting the need for high-quality benchmark translations, ideally by professional translators who understand the subtleties and complexities of the questions and answers.

In the tests on segment-based translation tasks, the sl-model and hr-model outperformed even the larger Llama2 70B model for their respective target languages (Rausch *et al.*, 2024). In the English-to-Slovenian translation task, Llama2 achieved the SacreBLEU scores 91.5, 92.4 and 93.1 with its 7B, 13B and 70B versions, respectively. In contrast, our sl-model (with 13B parameters) achieved 93.6 in the same task. Similarly, the hr-sl-model outperformed other models in the English-to-Croatian translation task with a score of 93.2, while Llama2 achieved 91.2, 92.0 and 92.9 with the 7B, 13B and 70B versions, respectively.

The results of the human evaluation involving translators showed that the sl-model performed best in tasks in Slovenian, while the hr-model performed best in tasks in Croatian (Rausch *et al.*, 2024). However, the Llama2 13B model, which was barely trained on Slovenian, outperformed the hr-model in Slovenian tasks, despite the hr-model being extensively trained on Slovenian text before being continually trained on Croatian. This suggests that the models were suffering from "catastrophic forgetting", where the performance of a model drops when it is further trained on a different task or language. The results were consistent across all three evaluations, and the Llama2 13B model always scored in second place after the model whose latest training corresponded to the input language.

Overall, the results highlight the challenges of training LLMs on multiple languages. While the DGT data can contribute to improved LLM performance in low-resource languages, further work is needed to mitigate the effects of catastrophic forgetting and to improve the model's performance in general. This work continues in the second phase of the project, which involves continual pre-training of an LLM at full scale with all DGT data covering all 24 EU official languages and using the more powerful Leonardo supercomputer and a larger open source LLM (DGT, 2024a).

4 Conclusion and outlook

The results from the first phase of the DGT EU LLM project show that high-quality text data created by translators and other language professionals has the potential to improve the performance of LLMs for low-resource languages. The first phase of the project has also highlighted the value of involving translators and other language professionals in the process of evaluating LLMs.

The second phase of the DGT EU LLM, currently ongoing and involving DGT data for all 24 EU official languages and a larger LLM is expected to add new and deeper insights on the extent to which high quality multilingual text created by translators and other language professionals can play a key role in improving the performance of LLMs.

The second phase of the DGT EU LLM project will continue to involve DGT translators in the process of evaluating LLMs. This participation can be extended beyond the actual evaluation of the outputs of LLMs, to the development of evaluation benchmarks and datasets. In particular, given the scarcity of evaluation datasets for languages other than English and a few others, translators can contribute by translating monolingual or limited LLM evaluation datasets into other languages, especially low resource languages, thus creating broadly multilingual gold-standard LLM evaluation datasets.

The participation of translators and linguistic experts in AI projects can help bridge the gap between human expertise and machine learning capabilities, leading to more accurate and reliable results. Their expertise might prove essential for ensuring that LLMs are trained on high-quality data and that the results are accurate and reliable.

Acknowledgements

The authors greatly appreciate the 60 colleagues who volunteered to evaluate the output of our models. The volunteers were translators and other language professionals from DGT and other members of EU staff, almost all of them native speakers of Slovenian or Croatian. The bulk of the model training was performed on the MeluXina supercomputer in Luxembourg, operated by LuxProvide and funded by the EuroHPC Joint Undertaking and by the national government and other entities in Luxembourg. The authors gratefully acknowledge the support received from both the EuroHPC team and the LuxProvide team.

In line with this reuse policy this document is submitted for presentation at the 46th Translating and the Computer Conference (TC46), in Luxembourg, 18-20 November 2024, and its publication on the AsLing website(s) and in the publication, Proceedings of TC46, co-published by AsLing and Editions Tradulex, as well as pre-publications, provided that appropriate credit is given and any changes are indicated.

References

Chollet, F., 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547.*

DGT – Directorate-General for Translation, 2023. DGT translation quality management framework (internal document).

DGT – Directorate-General for Translation, 2024a. Leveraging high quality internal data of the European Institutions at scale to build an EU institutional large language model (LLM). URL: https://eurohpc-ju.europa.eu/leveraging-high-quality-internal-data-european-institutions-scale-build-eu-institutional-large_en [last accessed October 31, 2024].

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A. and Goyal, A., 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783.*

EC – European Commission, 2024. eTranslation – The European Commission's machine translation system. URL: https://commission.europa.eu/resources-partners/etranslation_en [last accessed October 31, 2024].

Ferdaus, M.M., Abdelguerfi, M., Ioup, E., Niles, K.N., Pathak, K. and Sloan, S., 2024. Towards Trustworthy AI: A Review of Ethical and Robust Large Language Models. *arXiv preprint arXiv:2407.13934.*

Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.D.L., Hanna, E.B., Bressand, F. and Lengyel, G., 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088.*

Leick, J.M., 1995. Euramis: Integrated multilingual services for a large multilingual community, In *Proceedings of Machine Translation Summit V*. URL: https://aclanthology.org/1995.mtsummit-1.15.pdf [last accessed October 9, 2024].

LuxProvide, 2024. System overview - MeluXina user documentation. URL: https://docs.lxp.lu/system/overview/. Accessed: July, 2024. [last accessed October 9, 2024].

OP – Publications Office of the European Union, 2023. Translating for Europe – Facts and Figures. URL: https://op.europa.eu/en/publication-detail/-/publication/9a56d8d5-8070-11ef-a67d-01aa75ed71a1 [last accessed October 31, 2024].

Penedo, G., Kydlíček, H., Lozhkov, A., Mitchell, M., Raffel, C., Von Werra, L. and Wolf, T., 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*.

Post, M., 2018. A call for clarity in reporting BLEU scores. *arXiv preprint arXiv:1804.08771.*Skordas, T., 2019.

Toward a European exascale ecosystem: the EuroHPC joint undertaking. *Communications of the ACM,* 62(4), pp.70-70.

Rausch, I., Bhaskar, B., Safont-Andreu, A., Ewetz, H., Kolovratnik, D., Oravecz, C., Runonen, M., 2024. Towards effective continued pre-training of EU institutional LLMs on EuroHPC supercomputers. *Proceedings of the First EuroHPC user day* (in print)

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. and Bikel, D., 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288.*

Turisini, M., Amati, G. and Cestari, M., 2023. Leonardo: A pan-European pre-exascale supercomputer for HPC and AI applications. *arXiv preprint arXiv:2307.16885.*

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A. and Choi, Y., 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*

How does the use of Text-to-Speech technology affect the effort of post-editors working into their L2?

Foteini Kotsi

Ghent University
kotsi.translations@gmail.com

Todor Lazarov

New Bulgarian University

todorlazarov91@abv.bg

Joke Daems

Ghent University
joke.daems@ugent.be

Abstract

The rise of neural machine translation (NMT) coupled with a rise in demand for localisation services has transformed the face of the language service industry (LSI) over the past few years. One of the most up-and-coming services is machine translation post-editing (MTPE), with shorter turnaround times and more affordable prices. With the rising popularity of MTPE, professionals also face an increasing workload in their first foreign working language (L2), and, as a result, they must cope with growing mental and physical fatigue. A potential solution to lower MTPE effort could be the use of text-to-speech (TTS) technology. This study revolves around an experiment to investigate the MTPE effort of professionals working into their L2, from Greek into English, with and without the use of TTS, as well as reporting on their perceived effort during the experiment. According to our results, the use of TTS can increase post-editing quality, while it does not affect the temporal effort negatively in any significant way. On the other hand, it increases technical and cognitive effort as it involves more keystrokes and mouse activity. The perceived effort of the participants seems to be heavily influenced by the actual gains they had while using TTS.

1 Introduction

With the growing technological possibilities of the 21st century, there is an uptake of speech tools such as automatic speech recognition (ASR; speech-to-text) and Text-to-Speech (TTS) by professional translators (ELIA et al., 2022). Some benefits are more prominent than others. For instance, the fact that without quality ASR and TTS technologies blind and visually-impaired individuals would not have access to the profession, and people with physical impairments — such as repetitive strain injury (RSI) — would not have a healthier alternative to carry out traditional translation or machine translation post-editing tasks (Ciobanu, 2016). There is also research focusing on the positive effect that ASR and TTS have specifically on productivity and quality when it comes to translation revision and NMT error annotation tasks (Ciobanu et al., 2019 and Brockmann et al., 2022). According to Ciobanu and Secară (2019) "several areas are now ripe for much more systematic research, such as: translator ergonomics, productivity, as well as the impact of ASR and TTS on the process and output of translation, revision and review as defined in the ISO 17100:2015 standard".

Even though MTPE can be less time-consuming, the increasing demand on the market has increased the professional post-editor's workload and this increase can lead to potential cognitive overload. There seems to be a gap in research concerning the effect of TTS technology on machine translation post-editing effort.

When it comes to the implementation of speech technologies in MTPE, the focus has been mostly on ASR (Dragsted et al., 2011; Ciobanu, 2015; Ciobanu, 2016; Ciobanu and Secară, 2019) but research on TTS implementations is lagging, especially in the target segment of the machine translation post-editing process. Our research will take into consideration the already published findings and aims to fill the gaps concerning how the use of TTS in MTPE can affect the actual and perceived effort of post-editors, by implementing different methods of measuring MTPE effort and both qualitative and quantitative data.

2 Related Research

Even though TTS technology is not recent, it has seen significant improvements over the last few years, becoming more similar to natural human speech (Gottardi et al., 2022). There are real-life scenarios where listening to a text helps linguists spot issues. For example, there are international organisations such as the European Space Agency where, for reasons of speed and efficiency, the traditional Translation – Editing – Proofreading (TEP) model, where translators, revisers, and reviewers all use Track Changes and pass documents among themselves, has been replaced by a Translation + Face-to-face Review model (Ciobanu, 2015). In this process, the translator has their translation read back to them by a colleague. In this way the colleague is doing a monolingual target language review at the same time as the translator is doing a bilingual revision (ibid). Several translators have introduced TTS technology during the revision step of their own translations and this “could help linguists spot both fluency and accuracy errors more easily” (Ciobanu and Secară, 2019). Similarly, translators that have introduced ASR into their workflow, believe TTS technology to be an effective method of catching “speakos” (errors that might have occurred as the linguist spoke their translation through ASR) (Ciobanu, 2015). According to Brockmann et al. (2022), “this intuitively perceived benefit of aurally processing text points to the potential of TTS as an attention-raising technology, may also help post-editors identify subtle NMT errors”.

An experiment conducted by Ciobanu, Ragni, and Secară (2019) introduced TTS technology in the translation revision workflow by reading out the source text to the reviser, allowing them to focus on the target text, and the results were encouraging especially as far as accuracy errors were concerned. Accuracy has been identified as one of the major challenges for NMT (Vardaro et al., 2019; Brockmann et al. 2022). It should be mentioned that the participants were a mix of professional translators and trainees, and they worked in memoQ[1]. The results mostly focused on quality, and they were promising for the use of TTS. Secondly, Brockmann et al. (2022) focused on error annotation with the TTS technology enabled for both the source and target segments. The participants were students, and the experiments were carried out in Microsoft Word 365[2]. Given the nature of the task, the results were given in error over- and under-annotation, and gave good indications of quality while working with TTS. The introduction of

[1] https://www.memoq.com/

[2] https://www.office.com/

TTS technology in the MT error annotation workflow, also shows benefits for content comprehension and error identification, alongside measurable benefits for reducing error under-annotation (Brockmann et al., 2022). Given all the aforementioned, and the fact that the currently widely used neural MT still has some room for improvement, the implementation of TTS in MTPE could help professionals spot fluency and accuracy errors, as well as lower their post-editing effort.

2.1 Research Questions

The purpose of our study is to focus on aspects of MTPE that, to the best of our knowledge remain unexplored, and that could potentially help professional post-editors organize an optimal workflow. The main focus of the research will be machine translation post-editing with and without the use of TTS, since MTPE constitutes a highly relevant service that receives more and more attention with the continuing growth of NMT, while speech technologies have been harnessed across many different industries with great advantages and gains. The chosen service was also MTPE because the two most relevant studies (Ciobanu et al., 2019 and Brockmann et al., 2022) to date have focused on translation revision and error annotation, respectively. The three services do present similar characteristics but are still distinct in many respects. We intend to approach the measurement of machine translation post-editing effort with process- and product-based methods, as well as take into consideration each participant's perceived effort. This will allow us to determine not only whether the use of TTS can help post-editors handle their growing workload but also their perceptions. To cover these research grounds, we formulated the following research questions (RQs):

1. How does the use of TTS influence the actual machine translation post-editing effort when working into L2?

2. How does the use of TTS influence the perceived machine translation post-editing effort when working into L2?.

3 Methodology

To investigate these research questions, an experiment was constructed.

There were specific criteria that the participants had to meet to take part in the experiment. First of all, they had to be native Greek speakers with the additional requirement that their L2 had to be English. Moreover, they had to hold a bachelor's degree in Translation and have between two and ten years of experience in the LSI. Full MTPE will be carried out since most of the participants did not have extensive experience with MTPE and could struggle with the notion of focusing only on critical errors and making minimal changes.

The two texts used during the main experiment referred to the same overarching topic, came from the same news website, were of identical length, and obtained the same readability score in the Text Readability software for Greek developed by the Centre for Greek Language[3].

[3] https://www.greek-language.gr/certification/readability/index.html

To determine which MT engine would be used, we compared the raw output of ModernMT[4], DeepL[5], and eTranslation[6] for both of the texts. The outputs were ranked by the researcher segment by segment. The best (DeepL) and worst performing (ModernMT) engines, were excluded to avoid conditions where too much or too little effort would be required by the participants. As a final step, the eTranslation output for both texts was annotated for errors, using the TAUS DQF-MQM error typology[7] without critical errors being detected, and the major and minor errors were comparable in numbers. The texts were also alternated in order and use to ensure the reliability of the results.

MateCat[8] was selected for this experiment due to its straight-forward interface, its popularity and easy access. MateCat also counts the percentage of text changed at segment level.

The Google Chrome Read Aloud[9] browser extension was selected as the TTS tool. What set Read Aloud apart from other TTS tools for this particular experiment was its variety of display modes, that allowed participants to operate the TTS function just by using the shortcuts without having an extra element on their workbench.

Lastly, the keystroke logging tool Inputlog[10] (Leijten and Van Waes, 2013) was employed, to collect data regarding the MTPE temporal, cognitive, and technical effort.

The experiment consisted of three main stages: pilot experiment, main experiment, and post-experiment questionnaire. The pilot experiment consisted of two different MTPE tasks that were completed with the use of TTS so the participants could get used to the new interface and way of working. During the main experiment, the participants completed one MTPE task with TTS technology and one without it, so we could collect and analyse the data for both ways of working. Lastly, the post-experiment questionnaire served as a way of collecting data on the perceived effort from the participants.

4 Results

In this segment, we will present and analyse the data collected during the main experiment.

4.1 Quality

Even though our research does not focus on the quality of the final product, we had to ensure that the post-edited text met a certain level of quality. To determine the quality of the post-edited text, we used the errors that were annotated on the eTranslation output for each text when we determined which MT engine to be used during the experiment. Having the already annotated text and the Quality Report (QR) tab on MateCat facilitated this procedure. We were able to determine how many of the annotated errors were spotted and corrected by each

[4] https://www.modernmt.com/
[5] https://www.deepl.com/en/translator
[6] https://commission.europa.eu/resources-partners/etranslation_en
[7] https://www.taus.net/data-solutions/dqf-mqm-error-annotation
[8] https://www.matecat.com/
[9] https://chromewebstore.google.com/detail/read-aloud-a-text-to-spee/hdhinadidafjejdhmfkjgnolgimiaplp
[10] https://www.inputlog.net/

participant for each text. These fractions were then turned into percentages to obtain the percentage of the MT errors fixed per task.

In 15 out of the 16 tasks, the percentage of errors fixed reached a level of at least 50%, which is considered satisfactory given the fact that the error annotation was conducted by only one native speaker, and our research does not focus on the quality of the post-edited text. There was one task that fell below the 50% threshold and had 45.45% as a quality level. If we take into account the different conditions under which the participants worked, TTS seemed to benefit all of them, with the exception of one. More specifically, when working without TTS, the participants were 58% accurate in the correction of the examined errors, whereas for their TTS-enabled tasks, they scored 71% on average. The gains for each participant, as highlighted in Figure 1, when using TTS during the machine translation post-editing ranged from 4.55% to 36.4%. It is worth mentioning that the two tasks with 100% errors fixed were carried out using TTS on two different texts. This finding is completely in line with both Ciobanu et al. (2019) and Brockmann et al. (2022), further corroborating that TTS can help with accuracy and error under-annotation.

To conclude, according to our results, post-editors can benefit from the use of TTS in terms of quality while working into their L2. Our results pertaining to quality are also corroborated by the findings of Ciobanu et al. (2019) and Brockmann et al. (2022).

4.2 Product-based evaluation of Machine Translation Post-editing effort

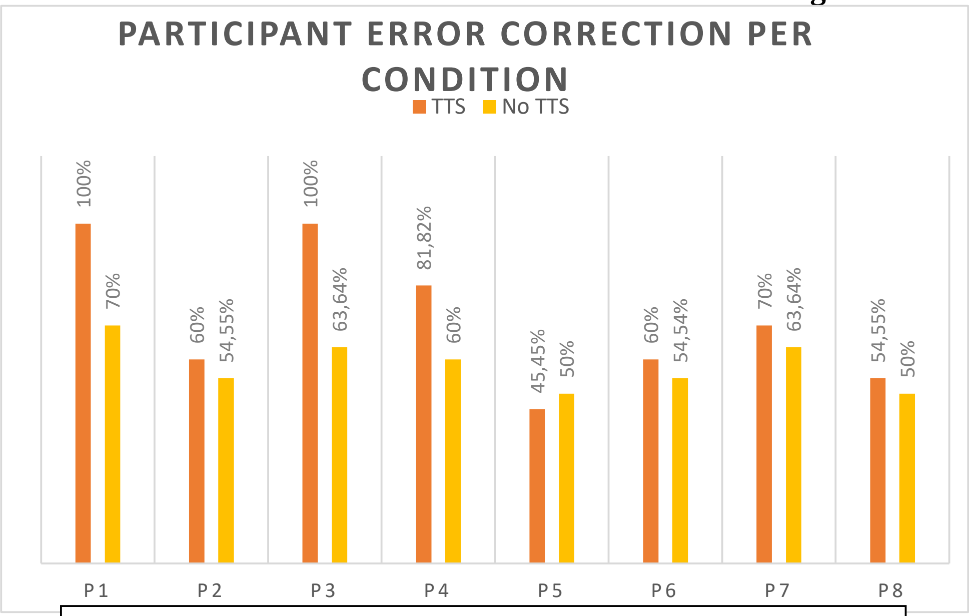


Figure 1 Participant error correction percentage per condition

To measure the machine translation post-editing effort using the product-based approach, we decided to use the Translation Edit Rate (TER), the total score derived from the total edits a human performs to modify the MT output, so it matches the reference translation (Snover et al., 2006). For our study, we retrieved the TER scores of all tasks expressed as percentages from the QR tab on MateCat.

We compared the average TER of the participants when working with TTS and without it. Under this prism, the differences between the two conditions were very slight. More specifically, the average TER for all tasks carried out without TTS was 16%, and for those carried out with TTS, it was 17%.

As with any other automatic metric, TER merely captures the number of edits made from the original MT output to the human post-edited text, and as a direct result, does not necessarily reflect the actual effort put into the final product (Santos, 2023). The time and effort spent on this research cannot be reflected by the final scores of automatic metrics such as TER alone.

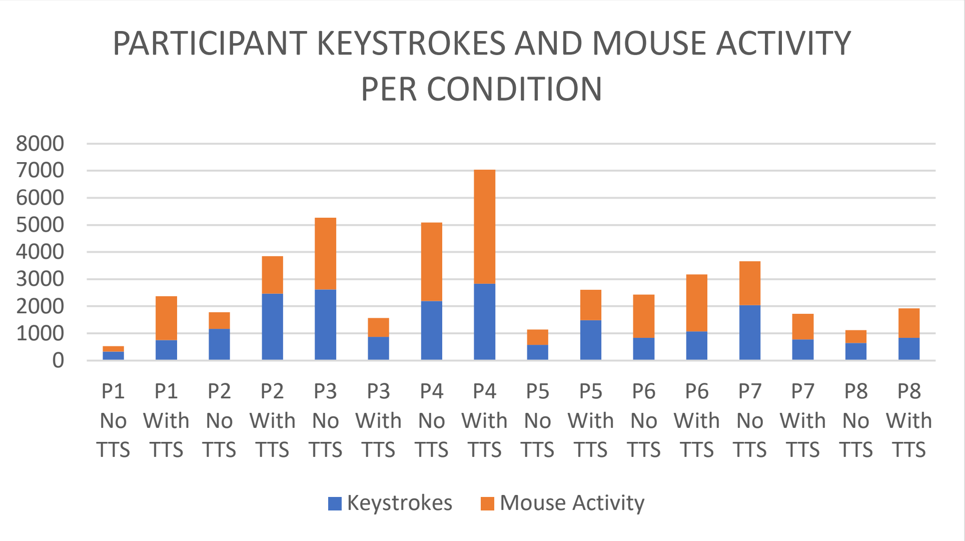
4.3 Process-based evaluation of Machine Translation Post-editing Effort

Contrary to the automatic metrics employed by the product-based approach, the process-based approach aims to analyse the post-editors' activity during their tasks. To achieve this part of the research, we used the data recorded by Inputlog, namely, the number of keystrokes, mouse activity, and the recorded time spent on each task.

4.3.1 Keystrokes and Mouse Movements

The types of events included in the general analysis provided by Inputlog have the following names: Keyboard, Mouse, and Focus. The first two were our means of measuring the cognitive and technical effort.

Looking at the keystrokes and mouse activity, as presented in Figure 2 below, they were generally higher when working with TTS. Only two participants had more keystrokes and higher mouse activity for the tasks carried out without TTS than those carried out with TTS. In general, TTS-enabled MTPE can be expected to have more activity of this sort because more shortcuts are needed for simple functionality such as play/stop/rewind or selecting the specific segment or sub-segment the post-editor would like to listen to with Read Aloud.



As a first approach, the experimental design measured the effort needed to correct the MT output for both tasks and under both conditions by applying the measures introduced by Barrachina et al. (2009):

Figure 2 The keystrokes and mouse activity

of each participant under the different conditions

- **Keystroke Ratio (KSR):** Number of keystrokes divided by the total number of reference characters.

- **Mouse-Action Ratio (MAR):** Number of pointer movements plus one more count per sentence (aimed at stimulating the user action needed to accept the final translation), divided by the total number of reference characters.

- **Keystroke and Mouse-Action Ratio (KSMR):** KSR plus MAR.

These metrics, however, are designed for computer calculations and not corrections made by human post-editors, and as such do not capture the actual effort made by linguists. To better capture the actual effort, Daems and Macken (2019) suggest dividing the number of keystrokes and mouse activity events by the total number of characters in the final version of the post-edited text. Using this method, we arrived at the results presented in Figure 3.

While looking at the KSMR above, it is evident that the machine translation post-editing effort was higher while working with TTS. For functional reasons, an increase in keystrokes and mouse activity was to be expected. In the following segment, we will continue analysing the data retrieved by Inputlog to complete the image of the process-based effort.

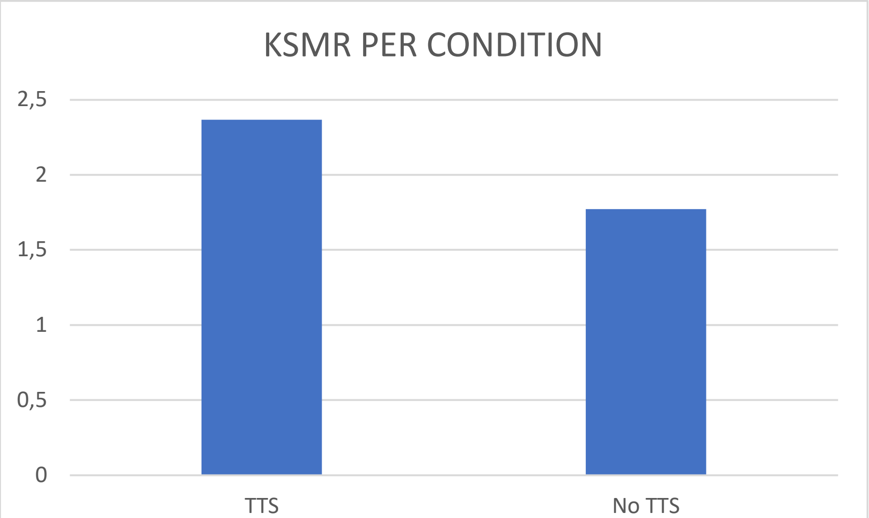


Figure 3 Comparison of the overall KSMR under each condition.

4.3.2 Time Spent

According to the Inputlog data, on average, all parties, regardless of the conditions they were working under, spent about 44 minutes post-editing Text 1 and around 56 minutes post-editing Text 2. Even though there seems to be a gap between the time spent in the two texts, it should be mentioned that the longest sessions, two hours and three minutes and one hour and 38 minutes respectively, were recorded on different texts, and both times the tasks were carried out without the use of TTS.

When looking at individual performance in terms of time, the comparison between conditions differs. As demonstrated in Figure 4, half of the participants spent more time post-editing the TTS-enabled tasks, and the other half took longer to post-edit without it, with both texts appearing in both of these groups.

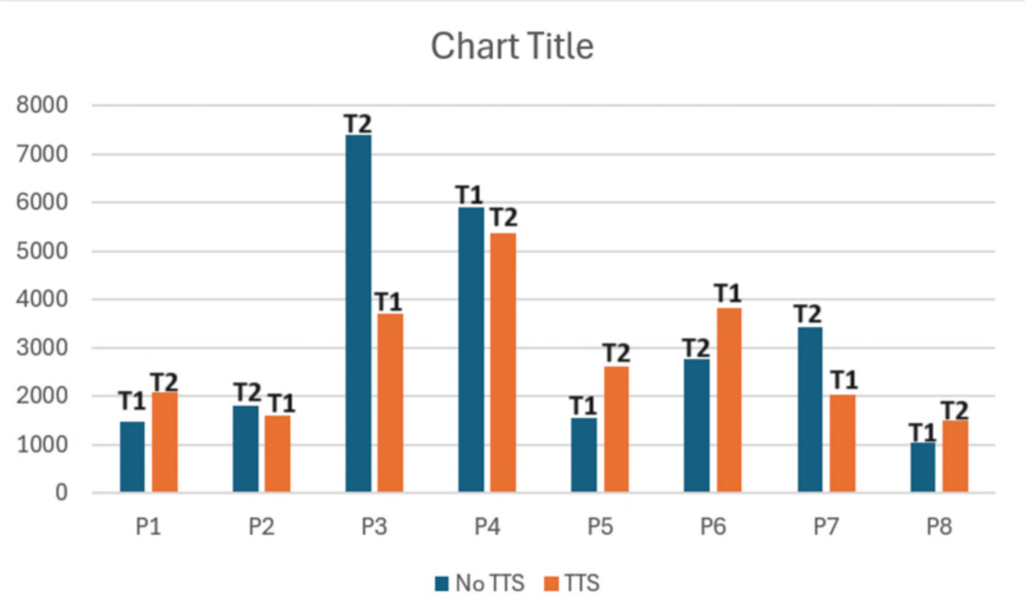


Figure 4 The time each participant spent post-editing under each condition in seconds.

By compiling all of the data and comparing all the TTS-enabled tasks to the ones carried out without it in Figure 5, TTS-enabled tasks were on average less time consuming according to our data. In particular, when the participants were post-editing with the aid of TTS, they worked on the text for an average of 47 minutes and 28 seconds, and when they were working without it, an average of 52 minutes and 48 seconds. Even though the margin is not very wide, it is worth highlighting that this was achieved despite the fact that on multiple occasions the

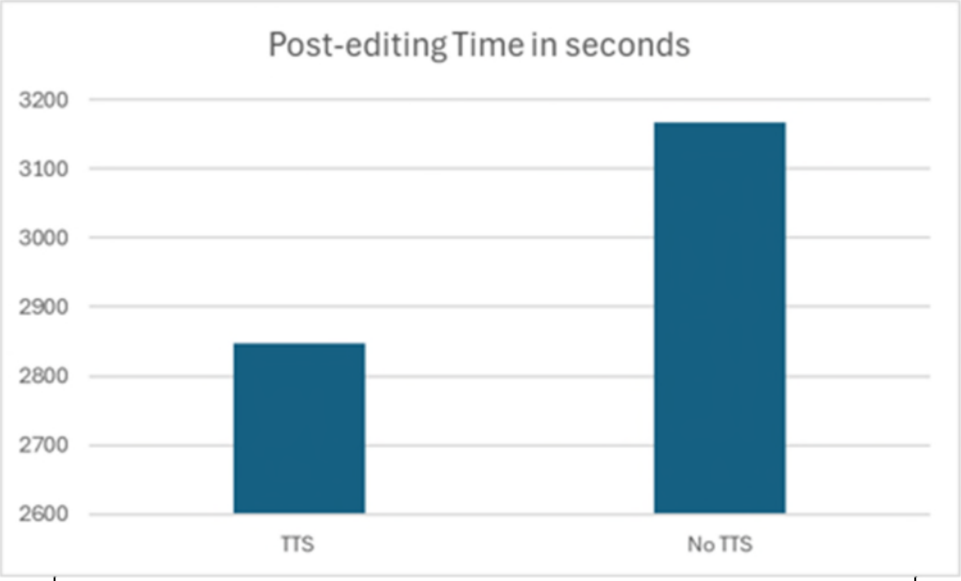


Figure 5 Average time spent post-editing under each condition.

participants chose to listen to the target segments more than once and still, on average, spent less time while working with TTS.

Our findings in this segment point to TTS being able to lower temporal effort in machine translation post-editing. This finding corroborates Brockmann et al.'s (2022) finding that the use of TTS can increase productivity.

Technical and cognitive effort seemed to increase by almost 34% in our experiment when using TTS. However, temporal effort decreased by a little over 11% under the same conditions.

To conclude the analysis of our data, we will focus on the answers given by all the participants in the post-experiment questionnaire upon completion of all tasks. The questions were mainly relevant to how each participant perceived the effort they put in under each condition.

4.4 Perceived Machine Translation Post-editing Effort

We asked the participants two questions directly linked to quality. First of all, whether they believed the use of TTS helped them spot more errors in the MT output. Participants P1, P3, and P4, who were the ones with the highest gains in terms of errors corrected, all answered that it helped them to a small extent. P5, the only participant that was less accurate when working with TTS, answered it did not help them at all. Out of the remaining participants who had smaller increases in their percentage of error correction while working with TTS, three answered that it helped them to a small extent and one that it did not help them at all. Moreover, the group was asked whether their use of TTS contributed to them focusing more on details of the source text. Again, P1, P3, and P4 reported that it helped them to a small extent, while P5 responded that they did not find it at all helpful.

Similarly, to the previous questions, the participants with single-digit gains from the use of TTS had different opinions, with two of them claiming it helped them and the other two that it did not. It should be mentioned that none of the participants selected the third option, "to a big extent," for either of the above questions.

Regarding the number of edits, they had to perform on the MT output to reach their final version, we asked them if they believed they had made more modifications under one condition and which one that was, or if they believed they had made about the same number of corrections. Even though the participants did not seem inclined to choose a specific condition, only P2 was accurate in their assumption when we compared the participants' answers to the recorded TER scores.

Of course, the above question related to the technical effort as well. Even though only one answer out of the eight was accurate, half of the participants felt they had made the same number of modifications. Three out of these four participants had modified from 5 to 7% more of the text when they were working on it with TTS, and one of them, 5% more while working without TTS. Three responders felt they had modified less with TTS. Two of them were right, and the third actually had a lower TER while working without TTS. Overall, half of the participants underestimated how many modifications they made while working with TTS, meaning they considered they made fewer or the same number of modifications using TTS compared to the task without TTS, when the opposite was true, and two others were aware of the fact that the

TTS-enabled tasks involving fewer edits on their part. Only P5 and P3 perceived a greater number of edits with TTS than was actually the case according to their TER scores.

Regarding the technical effort, we aimed to investigate whether, according to the participants, checking external resources becomes easier using TTS since this constitutes a big part of the machine translation post-editing process. While researching external resources, TTS seems to have helped P3 and P7. More specifically, P3 said that TTS helped them to a large extent while taking to external resources for translation research. This participant also seemed to be the one who used all of the shortcuts to play/stop/replay each segment according to their needs, even when not on the MateCat tab. P7 said it only helped them to a small extent while checking external resources.

Regarding cognitive effort, we asked the participants if spotting the errors in the MT output required less effort with TTS. The answers followed a similar pattern to the questions about quality. Participants P1, P3, and P4 answered ‘yes, to a small extent’. P5 responded the effort was not reduced at all when using TTS. Out of the four participants with low gains in error correction, three answered it did not make it at all easier, and one that it did make it easier to a small extent. Once again, the option "yes, to a big extent" was not selected by any of the participants.

To draw conclusions about both cognitive and technical effort, we asked the participants which task they found to be more tiring. 62.5% of the participants found the TTS task to be more tiring. In general, this seems to be in line with the conclusions we drew from the KSMR scores, since all participants except for two had elevated KSMR while working with TTS. However, the most surprising result is that once again, P1, P3 and P4 gave positive answers about TTS use. P4 found the TTS-enabled task less tiring even though their KSMR was significantly higher when working with TTS. P1, whose KSMR more than doubled when working with TTS, said they found both tasks equally tiring. Lastly, even though P3 reported to feel that the tasks were equally tiring, showed a significantly lower KSMR when working with TTS. The exact reverse happened with P7, who, even though reported to have found TTS-enabled machine translation post-editing more tiring, had half the KSMR when working with TTS.

To collect the participants' impressions on the time spent, we asked them which task they considered to be more time-consuming. Even though half of the participants were faster while post-editing with TTS, and by significant margins in two of the cases, seven out of eight participants considered TTS to be more time-consuming and one believed that the tasks took about the same time to complete.

5 Discussion

We designed and carried out this experiment to determine how the use of TTS while post-editing, enabled in the target segments of the text, affects both the actual and perceived effort of professionals working into their L2. To answer this, we employed both quantitative and qualitative methods to obtain a clearer image. More particularly, we worked with automatic metrics for the product-based analysis, we calculated the keystrokes and mouse activity to analyse our results with the process-based approach, and, lastly, we designed questionnaires to be filled in on specific stages of the experiment, so we could gather invaluable data from our participants to assess their perceived effort and to complete the numeric data gathered.

Even though our experiment was not focused on measuring the quality of the post-edited text, we considered it important that the texts delivered fulfilled a certain quality standard. That is because if any of the tasks were completed with too few or no modifications, this would also be reflected in the data collected for the machine translation post-editing effort measurement and it would taint the overall results. According to our analysis of the error correction percentages of the participants, we determined that the use of TTS can indeed aid users spot and correct more NMT output errors since seven out of eight participants corrected more of the annotated errors when using TTS. Our results corroborate those of Ciobanu et al. (2019) and Brockmann et al. (2022). More specifically, Ciobanu et al. (2019) mention that TTS can contribute to users spotting more accuracy errors during translation revision. Similarly, when working with NMT which presents accuracy errors, TTS seemed to help in our experiment. Brockmann et al. (2022) concluded that TTS helps limit under-annotation in MT output which is in line with our findings.

Concerning our Research Questions, we have come to the following conclusions.

1. How does the use of TTS influence the actual machine translation post-editing effort when working into L2?

Regarding the product-based approach, as expected, increased error corrections lead to more edits in the MT output. However, by measuring the effort using the product-based approach we arrived at the conclusion that the use of TTS does not necessarily affect the final TER of the post-edited text in a significant way. More specifically, the TER of the TTS-enable tasks for our experiment was 17% and for the ones carried out without TTS was 16%. Furthermore, if we take into account that error correction increased by 13.2% when using TTS, the low TER margin could even indicate a decrease in preferential changes.

Regarding the process-based analysis, first of all, when measuring technical and cognitive effort with KSMR it became evident that the use of TTS increases, by almost 34%, the measurements for these types of effort. More specifically, depending on the way the user activates and uses the system, it increases the keystrokes and/or mouse activity. Since MateCat does not offer the TTS function, it is possible that a fully integrated TTS function in a CAT tool interface could probably limit the KSMR to some extent. However, with the added functions, namely play/stop/rewind, we cannot expect the use of TTS to have no effect on the technical effort. Moreover, the augmented KSMR can to an extent be due to the increase in edits since more errors were spotted with TTS.

Regarding temporal effort, according to our results, the TTS-enabled tasks were on average shorter by 11.23%. Although this is not a high percentage, it is an indication that the use of TTS does not affect temporal effort negatively in any significant way, especially when we consider the time spent listening to the spoken text and the additional time required to operate the TTS system. These results coupled with the increased error correction could mean better quality texts with potentially shorter turnaround times.

2. How does the use of TTS influence the perceived machine translation post-editing effort when working into L2?

Since this new proposed way of working is targeted mainly at professionals, it is important to analyse how they perceive the whole procedure, which brings us to our second research question, regarding the effect that TTS has on the perceived effort of post-editors working in their L2. A particularly positive finding is that the participants who benefited the most from the use of TTS in general perceived the effort with TTS to be less intense than indicated by the actual data.

It is important to note that most participants did not have an accurate perception concerning the conditions under which they edited the corresponding text, with only two of them naming the correct task.

When asked which task they found more tiring, the participants' answers were in agreement with the KSMR scores, since most of them found the TTS-enabled task more tiring. However, here participants with high error correction and time gains with TTS gave answers more favourable to TTS use than what was indicated by the collected data. This is an indication that those who benefit from this new way of working and that have a way of working compatible with it, find it more pleasant, even when metrics suggest otherwise.

One of the most interesting findings is that in terms of time, most participants believed TTS-enabled machine translation post-editing to be more time-consuming even when that was not the case for their task. A contributing factor to this perception could be that even though the participants had worked on three TTS-enabled tasks by the completion of the experiment, it is unlikely that they were comfortable enough with TTS to be doing something else while the speaking took place, such as searching external resources. As a result, this window must have felt like a "down time," during which they did not feel productive enough.

In conclusion, our findings revealed that TTS can help users spot and correct more MT errors, in-line with previous studies. Interestingly, while increased error correction led to more edits, the use of TTS did not significantly affect the final TER of the post-edited text. However, it did impact various aspects of machine translation post-editing effort differently. The technical and cognitive effort clearly increased when using TTS technology. On average temporal effort decreased when working with TTS. Participants' perceptions of TTS-enabled machine translation post-editing varied, with some finding it less intense while others perceived it as more tiring, highlighting the subjective nature of individual experiences. These insights could form a basis for further research on TTS-enabled MTPE.

Acknowledgements

A special thank you to Ivanna, Ioanna, Anastasia L., Anastasia S., Nikolina, Sophia, Athina, and Katerina for all their support throughout the research.

References

Brockmann, Justus, Claudia Wiesinger, and Dragoș Ciobanu. "Error Annotation in Post-Editing Machine Translation: Investigating the Impact of Text-to-Speech Technology." In Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, 251–59. Ghent, Belgium: European Association for Machine Translation, 2022. https://aclanthology.org/2022.eamt-1.28.

Ciobanu, Dragoș. "Automatic Speech Recognition in the Professional Translation Process:" Translation Spaces 5, no. 1 (October 20, 2016): 124–44. https://doi.org/10.1075/ts.5.1.07cio.

Ciobanu, Dragos. “Of Dragons and Speech Recognition Wizards and Apprentices.” Revista Tradumàtica : Traducció i Tecnologies de La Informació i La Comunicació 12 (January 1, 2015) : 524–38. https://doi.org/10.5565/rev/tradumatica.71.

Ciobanu, Dragoş, and Alina Secară. “Speech Recognition and Synthesis Technologies in the Translation Workflow.” In The Routledge Handbook of Translation and Technology. Routledge, 2019.

Ciobanu, Dragoş, Valentina Ragni, and Alina Secară. “Speech Synthesis in the Translation Revision Process: Evidence from Error Analysis, Questionnaire, and Eye-Tracking.” Informatics 6, no. 4 (November 11, 2019): 51. https://doi.org/10.3390/informatics6040051.

Densmer, L. (2014). Light and Full MT Post-Editing Explained, available at http://info.moravia.com/blog/bid/353532/Light-and-Full-MT-Post-Editing-Explained

Dragsted, Barbara, Inger Mees, and Inge Hansen. “Speaking Your Translation: Students’ First Encounter with Speech Recognition Technology.” Translation and Interpreting 3 (July 28, 2011).

European Language Industry Association (ELIA), EMT, EUATC, FIT Europe, GALA, LIND, Women in Localization. 2022. 2022 European Language Industry Survey. Trends, expectations, and concerns of the European language industry. Available at: https://fiteuroperc.org/wpcontent/uploads/2022/03/ELIS2022_survey_results_final_report.pdf?x77803

Gottardi, William, Janaina Fernanda de Almeida, and Celso Henrique Soufen Tumolo. “Automatic Speech Recognition and Text-to-Speech Technologies for L2 Pronunciation Improvement: Reflections on Their Affordances.” Texto Livre 15 (April 20, 2022). https://doi.org/10.35699/1983-3652.2022.36736.

Hu, Ke, and Patrick Cadwell. “A Comparative Study of Post-Editing Guidelines,” 2016. https://doi.org/10.13140/RG.2.1.2253.1446 346-353.

“ISO 18587:2017(En), Translation Services — Post-Editing of Machine Translation Output — Requirements.” Accessed April 8, 2023. https://www.iso.org/obp/ui/#iso:std:iso:18587:ed-1:v1:en.

Jakobsen, A-L. (1998). Logging Time Delay in Translation, LSP Texts and the Translation Process. Copenhagen Working Papers. 73 − 101.

Koponen, Maarit. “Comparing Human Perceptions of Post-Editing Effort with Post-Editing Operations.” In Proceedings of the Seventh Workshop on Statistical Machine Translation, 181–90. Montréal, Canada: Association for Computational Linguistics, 2012. https://aclanthology.org/W12-3123.

Koponen, Maarit, Leena Salmi, and Markku Nikulin. “A Product and Process Analysis of Post-Editor Corrections on Neural, Statistical and Rule-Based Machine Translation Output.” Machine Translation 33 (June 1, 2019). https://doi.org/10.1007/s10590-019-09228-7.

Krings, Hans P. Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes. Kent State University Press, 2001.

Lacruz, Isabel, Michael Denkowski, and Alon Lavie. “Cognitive Demand and Cognitive Effort in Post-Editing.” In Proceedings of the 11th Conference of the Association for Machine Translation in the Americas, 73–84. Vancouver, Canada: Association for Machine Translation in the Americas, 2014. https://aclanthology.org/2014.amta-wptp.6.

Läubli, Samuel, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk. “Assessing Post-Editing Efficiency in a Realistic Translation Environment.” In Proceedings of the 2nd Workshop on Post-Editing Technology and Practice. Nice, France, 2013. https://aclanthology.org/2013.mtsummit-wptp.10.

Mesa-Lao, Bartolomé. “Speech-Enabled Computer-Aided Translation: A Satisfaction Survey with Post-Editor Trainees.” In Proceedings of the EACL 2014 Workshop on Humans and Computer-Assisted Translation, 99–103. Gothenburg, Sweden: Association for Computational Linguistics, 2014. https://doi.org/10.3115/v1/W14-0315.

O’Brien, Sharon. “Pauses as Indicators of Cognitive Effort in Post-Editing Machine Translation Output:” Across Languages and Cultures 7, no. 1 (June 1, 2006): 1–21. https://doi.org/10.1556/Acr.7.2006.1.1.

Pérez, Celia Rico. ‘Re-Thinking Machine Translation Post-Editing Guidelines’. The Journal of Specialised Translation, no. 41 (30 January 2024): 26–47. https://doi.org/10.26034/cm.jostrans.2024.4696.

Santos, Hitalo. ‘Machine Translation Post-Editing of Automatic Subtitling for Brazilian Portuguese,’ 2023. https://doi.org/10.13140/RG.2.2.20699.34085.

Stasimioti, Maria, Vilelmini Sosoni, and Konstantinos Chatzitheodorou. “Investigating Post-Editing Effort: Does Directionality Play a Role?” Cognitive Linguistic Studies 8, no. 2 (November 22, 2021): 378–403. https://doi.org/10.1075/cogls.00083.sta.

Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. “A Study of Translation Edit Rate with Targeted Human Annotation.” In Proceedings of the 7th Conference of the Association for

Machine Translation in the Americas: Technical Papers, 223–31. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, 2006. https://aclanthology.org/2006.amta-papers.25.

Toledo Báez, M.C. (2018). Machine translation and post-editing: Impact of training and directionality on quality and productivity. Revista Tradumàtica. Technologies de la Traducció, 16, 24–34. https://doi.org/10.5565/rev/tradumatica.215

Vidal, E. 1997. Finite-State Speech-to-Speech Translation. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Munich, Germany

Benchmarking terminology building capabilities of ChatGPT on an English-Russian Fashion Corpus

Anastasiia Bezobrazova
Centre for Translation Studies
University of Surrey, UK
bezobrazovaanastasia@gmail.com

Miriam Seghiri
IUITILM
University of Malaga, Spain
seghiri@uma.es

Constantin Orasan
Centre for Translation Studies
University of Surrey, UK
c.orasan@surrey.ac.uk

Abstract

This paper compares the accuracy of the terms extracted using SketchEngine, TBXTools and ChatGPT. In addition, it evaluates the quality of the definitions produced by ChatGPT for these terms. The research is carried out on a comparable corpus of fashion magazines written in English and Russian collected from the web. A gold standard for the fashion terminology was also developed by identifying web pages that can be harvested automatically and contain definitions of terms from the fashion domain in English and Russian. This gold standard was used to evaluate the quality of the extracted terms and of the definitions produced. Our evaluation shows that TBXTools and SketchEngine, while capable of high recall, suffer from reduced precision as the number of terms increases, which affects their overall performance. Conversely, ChatGPT demonstrates superior performance, maintaining or improving precision as more terms are considered. Analysis of the definitions produced by ChatGPT for 60 commonly used terms in English and Russian shows that ChatGPT maintains a reasonable level of accuracy and fidelity across languages, but sometimes the definitions in both languages miss crucial specifics and include unnecessary deviations. Our research reveals that no single tool excels universally; each has strengths suited to particular aspects of terminology extraction and application.

1 Introduction

The rise of digital communication and the increasing globalisation of industries have underscored the necessity for reliable multilingual dictionaries that accurately reflect the nuances and specific terminology of various specialised fields. Specialised fields such as law, engineering, medicine, or fashion have their own terminologies and vocabularies, requiring precise and comprehensive bilingual lexicons for efficient language processing (Chodkiewicz et al. 2002). Due to the specialist character of this terminology, extracting these lexicons from specialised corpora may pose particular difficulties. Tools like SketchEngine (Kigarriff et al, 2014) and TBXTools (Oliver and Vazquez, 2015) are commonly used to extract terms from domain specific corpora, but they cannot provide definitions for the extracted terms. The recent developments in the field of Generative AI (GenAI) have attracted the attention of terminologists who proposed ways to used Large Language Models (LLMs), and ChatGPT in particular, to support the process of building terminologies (Giguere et al, 2023; Massion, 2024). In contrast to the commonly used tools for terminology extraction, GenAI can extract terms from domain-specific corpora and propose definitions for them. This characteristic was successfully employed by Lew (2023) to generate definitions for dictionary entries.

The aim of this research is to demonstrate the feasibility of compiling a reliable and high-quality corpus of fashion texts in English and Russian which can serve as a valuable source for creating a bilingual glossary that can aid translators in their work. In addition, we compare the accuracy of the terms extracted using SketchEngine, TBXTools and ChatGPT, discussing their strengths and weaknesses. Given the ability of ChatGPT to generate texts on the basis of prompts that it receives, we also evaluate the quality of the definitions produced by ChatGPT. The research presented in this paper focuses on the English-Russian language pair, but the proposed methodology can be easily adapted to other language pairs and applied to other domains.

Whilst ChatGPT has proved its usefulness in numerous applications, there is limited research that systematically assesses its ability to extract terminologies. Apart from Giguere et al. (2023) who compared the terms extracted by GPT4 with a statistical model, to the best of our knowledge, there are no other academic studies that assesses the performance of ChatGPT for terminology extraction. The majority of publications on this topic are focused on how ChatGPT can support translation professionals by providing practical guidance on the use of ChatGPT for this purpose. Good examples for this are (Muegge, 2023) and numerous posts on social media[1] and company blogs[2]. Whilst such publications can offer invaluable practical information on how to use ChatGPT for building terminologies, they do not provide comprehensive evaluations of its performance. This research aims to address this gap by carefully evaluating the performance of SketchEngine, TBXTools and ChatGPT on our corpus.

The rest of the paper is structured as follows: Section 2 presents the process of preparing the data used in this study. The three tools employed in this study are briefly described in Section 3, followed by their evaluation in Section 4. The paper finishes with a discussion and conclusions.

2 Data preparation

In this section we present the corpus compilation process and how the gold standards used in this research were created.

2.1 Corpus compilation

The essence of the corpus-driven terminology extraction methods lies in their ability to highlight relevant terms for the domain represented in the corpus. In the case of this research, this means that provided that the corpus represents the fashion domain accurately, the terminology extracted from the corpus captures the language used within fashion-related

[1] https://www.linkedin.com/pulse/large-language-models-terminology-extraction-yannis-evangelou/

[2] https://www.oneword.de/en/term-extraction-ai/

discourses in Russian and English, (Afzaal et al. 2023). Achieving this representativeness depends very much on how the corpus used in the extraction process was compiled. Our research was carried out on a comparable corpus of magazines and webpages related to fashion written in English and Russian. This section presents the process of corpus compilation and details about the corpus.

The **first step** in corpus building entailed identifying sources of information and locating texts to be included in the corpus. The Google search engine was utilised to identify popular and reputable fashion magazines and websites renowned for their significant impact and readership. This process was carried out by the first author of the paper, who possesses expertise in the fashion domain. By analysing various resources such as '*Top 14 Fashion Magazines In The World*', '*Top 10 Luxury Fashion Magazines*', '*Fashion Magazines: History of the Biggest Magazines - Vogue, ELLE & Co.*', we determined that *Vogue* was the most important magazine globally. Subsequently, other prominent magazines like *Cosmopolitan*, *ELLE*, and *Glamour* were also selected as a source of the texts to be included in the corpus. This ensured a comprehensive and authoritative representation of the fashion domain.

In addition to these magazine texts, the corpus was enriched by the inclusion of content from 52 leading fashion websites. These websites were selected using the same criteria as the magazines, ensuring that they meet high standards of relevance, authority, and influence in the fashion domain. This expansion served to diversify the corpus's virtual nature further, by embracing the dynamic and evolving landscape of online fashion discourse. All the magazines and webpages downloaded were published between 2021 and 2024, thus ensuring that the corpus represents the language currently used in the fashion industry.

The **second step** involved downloading the chosen magazines from their web pages for future storage. This task was accomplished by right-clicking on the desired content and selecting 'Download' in PDF format. Given that our analysis also required the magazines in plain text format (TXT) for the terminology extraction programs, an online converter was employed to convert the downloaded files to the required format. In parallel to the acquisition of magazine content, digital materials sourced from the selected fashion websites were stored in HTML format. This decision was informed by the need to preserve the original formatting and interactive elements inherent to web-based content, which might be useful in future analysis of digital fashion discourse. In addition, the HTML files were converted to text format by saving them in TXT format.

The result of this process was a comparable, virtual, bilingual corpus comprising 24 fashion magazines, with 12 in Russian and 12 in English. In addition, we collected web pages from 52 websites, equally distributed between English and Russian. This composition ensures that the corpus is not only diverse in terms of content source, but also balanced linguistically, making it qualitatively representative of the fashion domain, and providing a comprehensive overview of the industry's discourse. After cleaning the corpus (see Section 2.2) we tested the representativeness of the corpus using the ReCor tool[3]. ReCor relies on lexical density to determine the minimum number of texts and words that should be included in a specialised language corpus in order to be representative for that particular domain (Corpas Pastor and

[3] http://www.lexytrad.es/en/resources/recor-3/

Seghiri 2007). ReCor applied to our corpora confirmed that "the corpora are nearing a state of representativeness".

2.2 Corpus cleaning

The process of creating a corpus from fashion magazines involved downloading the PDF and HTML files and converting them to TXT format. However, the conversion process proved to be a challenging task. Despite testing numerous programs, the conversion often resulted in TXT files that were not presentable due to a plethora of issues, including the presence of symbols, numbers within words, significant gaps between words, and a slew of unnecessary symbols. This was particularly true when we converted the PDF files, due to the fact that they contained many images and employed a very creative layout for pages. The appendix contains two examples of these noisy texts.

Despite experimenting with several programs such as Text Cleaner[4], Text Tools[5], ReText.AI[6], and Code Beautify[7], the outcomes were unsatisfactory. Although these programs boasted user-friendly interfaces, they failed to deliver acceptable results. The resulting TXT files still contained significant levels of noise, or an important part of information was lost. Consequently, the decision was made to explore ChatGPT as an alternative text cleaning tool. To this end, we experimented with a number of prompts where ChatGPT was asked to remove the noise and make the text readable. Given that the corpus was collected from the Internet it is possible that the texts have been ingested by ChatGPT during its training, which made the task easier. Given the tendency of Large Language Models (LLMs) to hallucinate, the output was carefully checked to ensure that it did not contain information which was not present in the original or that no essential information was removed. Even with this manual checking step, the cleaning of the corpus was much faster with the help of ChatGPT, than if it had been done manually.

ChatGPT produced markedly better results than the other tools used, albeit with its own set of challenges. The most effective prompts for cleaning English texts included: "*Clean this text and make it readable.*" and "*Make this text readable, keep it as original as possible, remove the noise, keep all information.*"[8] For Russian texts, the approach was similar, with some prompts in English but annotated with '*in Russian*' for instance, i.e. '*make it readable in Russian.*' Without such a note, the cleaned Russian texts were at times inadvertently translated into English. Additionally, the same prompt could yield different results, and occasionally, ChatGPT would unexpectedly cease providing output with the message, '*I'm sorry, but I can't assist with that request*' despite having functioned moments before. In such situations, the process had to be restarted using a new chat.

One of the challenges encountered during the cleaning process was that at times, prompts intended to clean the text and reduce noise sometimes resulted in a summary rather than the original text, though the prompt '*clean the noise, make this text readable and keep it as original*

[4] https://textcleaner.net/
[5] https://texttools.ru/text-cleaner
[6] https://retext.ai
[7] https://codebeautify.org/text-cleaner
[8] https://chat.openai.com/share/5d50314b-d1de-431c-9aab-19260b84e0a5

as possible’ generally produced good results. Nevertheless, repeated prompting was occasionally necessary to achieve the desired outcome, and even then, the results were sometimes unsatisfactory. Switching to a new chat and repeating the exact same prompt would, however, complete the request as needed.

We also experimented with more complicated prompts that considered the characteristics of a particular source. Whilst the prompt usually worked better for that particular source, it was rarely very useful for other sources that had a completely different layout. As a result, we decided to use these generic prompts that worked across all the sources, and manually check the results. All the experiments presented in this paper were carried out using the web interface of ChatGPT. In future experiments, we plan use the API which provides more flexibility and would allow us to apply a cascade of prompts to clean a text step by step. This may allow us to automate the process more.

For cleaning, we experimented with both ChatGPT-3.5 and ChatGPT-4. In the initial stages of this research, only ChatGPT-3.5 was available, but we switched to ChatGPT-4 as soon as it became available. The majority of text cleaning was performed with ChatGPT-4. Overall, the text cleaning process was time-consuming as it had to be done manually, one text at a time, but this approach was the only one that provided satisfactory results. The resulting corpus contains over 1.8 million words, with 1 million words in English and 800,000 words in Russian.

2.3 Gold standard development

To effectively evaluate the terms extracted by term extraction tools such as SketchEngine, ChatGPT, and TBXTools, it was necessary to establish a robust gold standard.[9] For this reason, we used the Google Search engine to locate pages containing lists of fashion related terms and harvested them using bespoke python scripts. The selected pages were relatively easy to harvest as the terms were listed using clearly structured tables. At times these tables also contained the translation of terms and their definitions.

The gold standard developed comprises a carefully curated list of terms pertinent to the fashion domain. The terminology was primarily sourced through an automated harvesting process from bilingual English-Russian websites, alongside monolingual English sites. The terms harvested from English only websites were automatically translated and rigorously corrected by a native speaker to ensure their accuracy and relevance within both linguistic frameworks.

The gold standard contains a total of 354 terms in English and Russian. Sixty of these terms also had definitions in English. As with the terms, we automatically translated the definitions to Russian and carefully checked their accuracy. The definitions were used to assess the ChatGPT’s ability to extract definitions.

[9] The term *gold standard* is used in the field of Natural Language Processing to refer to a resource that was validated by humans and is used to evaluate automatic processing methods by comparing their output with the gold standard using establish comparison methods.

3 Terminology extraction tools

Once the data was preprocessed, terms were identified in the corpora with the help of SketchEngine, TBXTools, and ChatGPT. These tools use various algorithms and techniques to identify frequent and domain-specific terms, facilitating a comprehensive analysis of the terminology within the fashion industry.

SketchEngine is a widely used tool for terminology extraction in academic circles, owing to its highly customisable features and comprehensive linguistic resources encompassing corpora, dictionaries, and thesauri. Its advanced querying capabilities facilitate precise and efficient extraction, utilising linguistic templates and domain-specific terms. Moreover, the tool incorporates phrase analysis functionality to identify commonly occurring phrases, while its evaluation and validation tools guarantees the quality and reliability of the extracted terms by conducting comparisons with external resources or expert knowledge (Kilgarriff et al. 2014).

TBXTools is a highly capable tool for terminology extraction, offering a range of functionalities. It employs statistical and linguistic methods to extract multiword terms from specialised corpora. The tool supports statistical term extraction using n-grams and stop words, linguistic term extraction using morpho-syntactic patterns and a tagged corpus, detection of translation candidates in parallel corpora, and automatic learning of morphological patterns. Additionally, it offers evaluation capabilities to assess precision and recall based on different frequency thresholds. TBXTools proved its effectiveness in various language processing tasks, including ontology learning, machine translation, computer-assisted translation, thesaurus construction, classification, indexing, information retrieval, and text mining. Its Python-based nature enhances usability, flexibility, and compatibility across platforms and systems, making it a valuable asset for researchers and practitioners in the field of terminology management. The fact that it is open source enables other researchers to extend it and adapt it as needed (Oliver and Vàzquez 2015).

ChatGPT is not specifically designed for terminology extraction, but it can extract terms if prompted correctly. We experimented with a variety of prompts such as:

> *Extract the terms related to fashion, such as all kinds of clothes, shoes, accessories etc from the given text and list them. Can you please extract terms, that can be found ONLY in the given text*[10]

One limitation of ChatGPT is that it demands considerable time to extract an extensive array of terms. Initially, ChatGPT may provide between 20 to 50 highly relevant terms. To elicit further terminology, it is necessary to continue prompting iteratively. The process was repeated till ChatGPT started giving repetitive words, non-existing words, colour plus clothing items, or unrelated words like “shower”. During the process it was noticed that ChatGPT can sometimes deviate from the specific corpus under consideration, beginning to extract domain-specific terminology—such as fashion terms—in a more general context without relying on the text provided.

[10] https://chat.openai.com/share/4e30b4e7-1086-4df2-ac66-8b703e4ae17d

4 Results

4.1 Terminology extraction

In this section we compare the performance of the SketchEngine, TBXTools and ChatGPT at extracting terms from our corpus. The performance is calculated using standard metrics like precision, recall, and f-measure. **Precision** measures the accuracy of the extracted terms by evaluating how many of the extracted terms are in our gold standard. A high precision indicates that fewer irrelevant terms are extracted. **Recall** measures the completeness of the term extraction by assessing how many of the terms from the gold standard were extracted. High recall means that most of the terms from the gold standard were extracted. **F-measure** is the harmonic mean of precision and recall, providing a single metric to balance the two. The results for these metrics for each language are presented in Table 5. The analysis presented in this section will highlight the strengths and weaknesses of each tool, with a particular emphasis on how well they balance precision and recall in the extraction process.

	English corpus			**Russian corpus**		
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
TBXTools	0.022	0.704	0.044	0.055	**0.833**	0.104
SketchEngine	0.009	**0.732**	0.018	0.008	0.629	0.016
ChatGPT	**0.283**	0.360	**0.317**	**0.335**	0.358	**0.346**

Table 5. The accuracy of term extraction using the three tools and measured using precision, recall and f-measure. The values in bold represent the highest values observed for a metric and a corpus

TBXTools extracted approximately 15,000 terms in English and 7,800 terms in Russian, with frequencies ranging from a minimum of 1 to a maximum of 4. After we applied an automatic cleaner, which removed superfluous characters such as dashes, ‘at’, and numerals, standardised plurals and converted the terms to lowercase the total number of terms was reduced to approximately 10,000 in English and 5,000 in Russian. Table 5 presents the accuracy of TBXTools. The results from the TBX Tools term extraction exhibit a high recall, especially for Russian, suggesting the tool’s efficacy at identifying a broad array of terms, including a large proportion of relevant terms from the gold standard.

SketchEngine extracted approximately two to three times as many terms in English and Russian compared to TBXTools, resulting in a higher incidence of noise. In the case of English, the predominant issue was the amalgamation of separate phrases, exemplified by terms such as ‘andjeans.’ For Russian, the errors were twofold: some terms appeared in English rather than Russian, and others where single words were incorrectly split into two words. Following both automatic and manual cleaning processes, the number of terms was reduced to approximately 10,000 for each language. Even after cleaning the lists, the precision remains low for both languages, as can be seen in Table 5. SketchEngine shows a notable ability to achieve high recall in both English and Russian, effectively capturing a large proportion of relevant terms. This is not surprising given the large number of terms it extracts. However, it struggles significantly with precision, with the inclusion of many irrelevant terms leading to a high incidence of noise.

As mentioned above, one of the challenges of using ChatGPT for extracting terms was that it had to be prompted repeatedly in order to produce a longer list of terms as each time the list contained between 20 to 50 terms. To elicit further terminology, it is necessary to continue prompting iteratively. Additionally, ChatGPT can sometimes deviate from the specific corpus under consideration, beginning to extract domain-specific terminology - such as fashion terms - in a more general context without relying on the provided text. No cleaning was applied to the list of terms produced by ChatGPT. As can be seen in Table 5, ChatGPT obtains a significantly higher precision and f-measure scores, but the recall is the lowest. This can be explained by the fact that the number of terms extracted using ChatGPT was in the hundreds, rather than thousands as TBXTools and SketchEngine produce.

A notable advantage of using ChatGPT is its ability to classify terms into distinct categories. We noticed that at times ChatGPT would organise the extracted terms into categories such as ‘clothes’, ‘shoes’, ‘accessories’, and ‘bags’. This capability facilitates a more structured approach to understanding and organising terminology, which can be particularly beneficial for academic and research purposes in specialised fields.

4.2 Evaluation of ranking of terms

TBXTools and SketchEngine return lists of terms ranked according to their term likelihood. Because our gold standard contains only 354 terms, we decided to evaluate how increasing the number of terms we consider up to 354 terms impacts on the accuracy of the terms extracted. This would simulate a scenario where a terminologist uses automatic tools to build a glossary, and they consider the extracted terms in the order returned by the tools.

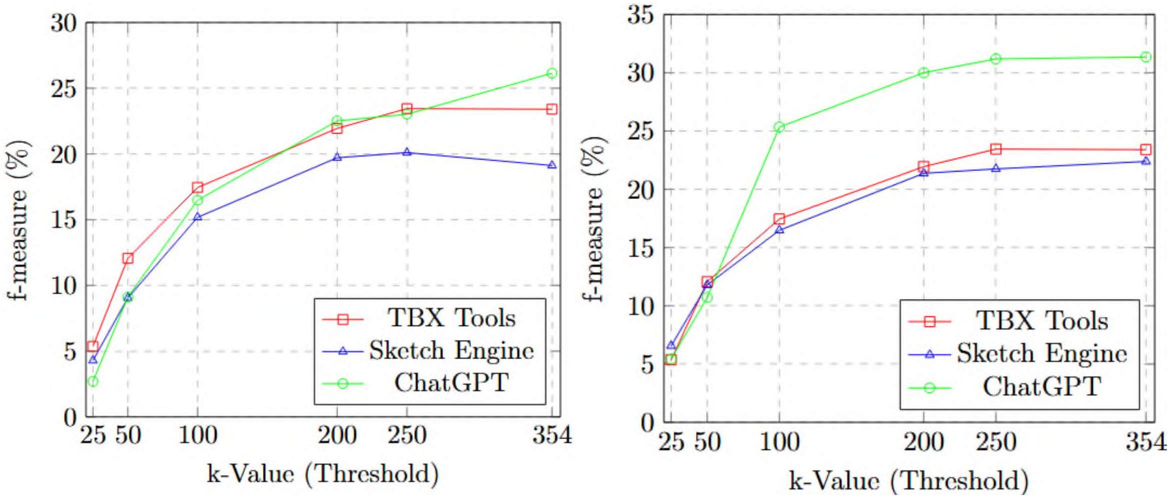


Figure 5. Changes in the f-measure scores as we increase the number of terms considered for the English corpus (left) and Russian corpus (right)

Figure 5 demonstrates how the scores for f-measure change as we increase the number of terms we consider up to the size of the gold standard. Due to space reasons, we do not include the graphs presenting the precision and recall scores. While analysing these scores, we noticed that in most of the cases as the number of terms we consider increases, the recall also increases at the expense of precision. The exception is when we extract terms using ChatGPT. As we increase the number of terms from 25 to 100, both precision and recall increase (precision from 0.20 to 0.40, recall from 0.01 to 0.10). If the number of terms is increased further, the pattern observed with the other term extraction methods is followed i.e. the precision decreases whilst the recall keeps increasing.

Analysis of the f-measure scores across both languages shows that SketchEngine typically has lower scores, reflecting the challenge in maintaining a balance between high recall and lower precision. This is evident in both English and Russian, where the f-measure scores increase with k, but remain modest due to the precision drop. TBXTools consistently achieves slightly better f-measure scores than SketchEngine, though it is still hampered by low precision levels. The f-measure scores improve as more terms are considered, peaking at k=250 before the decline in precision affects the score adversely. ChatGPT achieves higher f-measure scores in both language datasets, indicating a better overall balance between precision and recall. This suggests that ChatGPT might be more effective when managing the extraction scope through specific prompting and adjustments, adapting better to the nuances of each language compared to the more static algorithms of TBXTools and SketchEngine.

4.3 Error analysis of terminology extraction

In order to understand better how good SketchEngine, TBXTools and ChatGPT are at term extraction we conducted a detailed error analysis. For this purpose, we analysed the 50 top terms extracted by each of the tools. These terms were compared against our fashion-specific gold standard to identify deviations and inaccuracies in fashion terminology. The objective being to pinpoint significant discrepancies.

Analysis of the terms extracted by TBXTools and SketchEngine reveals that they extract a large number of words that are not terms. TBXTools favours extraction of words related to the fashion industry, but which are too general to be included in our gold standard. Examples of such words are terms such as ‘collection,’ ‘design,’ ‘colour’, ‘мода’ (fashion), ‘цвет’ (colour), ‘одежда’ (clothing). Most of the terms correctly extracted by these tools, but not included in the gold standard, are names of well-known fashion brands such as Chanel, Gucci, Prada, Versace, and Balenciaga.

The comparison between ChatGPT’s fashion-related terminology and the gold standard highlights ChatGPT’s effectiveness in capturing both contemporary and niche terms not included in traditional terminologies. All the top 50 terms extracted by ChatGPT are fashion-related terms in English and only one is not a term in Russian. However, a large proportion of these terms are not present in our gold standard which shows one of the limitations of employing automatic evaluation metrics.

4.4 Definition extraction using ChatGPT

We also analysed the definitions produced by ChatGPT for 60 commonly used terms in English and Russian. As a reference, we used the definitions we extracted from the online glossaries.

The definitions for the extracted terms were produced by prompting ChatGPT to define a term in the context of the fashion industry and given our corpus.

The effectiveness of ChatGPT in providing definitions is examined by measuring the similarity between the model-generated definitions and the reference definitions using the Levenshtein distance (Levenshtein, 1966). This metric, quantifying the minimum number of edits needed to change one sequence into another, serves as an indicator of how closely the definition by ChatGPT match the expected text. We calculated the distance at the word level, rather than character level. We decided to use Levenshtein distance because it shows the number of word level edits a terminologist needs to make in order to produce the definition from the gold standard.

Significant variations in Levenshtein distances were observed, with values ranging from as low as 1 to as high as 221. The lowest distances indicate cases where ChatGPT's definitions are almost identical to the reference, demonstrating high fidelity in reproducing accepted definitions with minimal alteration. This could suggest that the pages used to extract the reference definitions are included in ChatGPT. At the other end are the highest distances which reflect instances where the definitions have been substantially modified, suggesting that ChatGPT has either added extraneous information or shifted the focus of the definition, potentially leading to deviations from the intended meanings.

For the English definitions, the lowest recorded distance is 1, where ChatGPT made a minor modification by replacing 'usually' with 'typically', showing high fidelity in reproducing the reference definition almost identically. In contrast, the highest distance observed is 221, where ChatGPT expanded significantly on the reference by adding various contextual details, leading to a potential shift away from the original meaning and introducing possible inaccuracies. The average Levenshtein distance between the reference descriptions and definition by ChatGPT is 14.91 indicating that we would need an editor to make around 15 word changes per definition. For the Russian definitions, the Levenshtein distances, range from as low as 0 to as high as 94, with an average distance of 8.69 tokens.

In order to gain a better understanding of the differences between the definitions produced by ChatGPT and the ones in our gold standard, we carried out a detailed analysis and noticed the following phenomena across both languages:

Handling of Core Concepts: ChatGPT generally retains the core concepts of the terms it defines. For instance, regardless of language, the descriptions for items like '*jacket*', '*sweat-shirt*', and '*coat*' maintain essential elements such as their use and basic form (e.g., long sleeves, upper body coverage).

Synonym Replacement and Structural Changes: Changes often involve synonyms or slight structural adjustments without dramatically altering meaning, like changing '*usually*' to '*typically*' which affects the Levenshtein distance, but not the overall meaning.

Elaboration: ChatGPT often elaborates on the reference. For instance, the reference might simply describe an item as '*a jacket made of wind-resistant material*' while ChatGPT might expand this to '*a thin lightweight jacket designed to resist wind chill and light rain*'. This adds context and details which are not present in the reference leading to high Levenshtein distance. The English definitions tend to be more expanded and detailed compared to the Russian ones.

Specification and Detailing: ChatGPT descriptions tend to specify materials, contexts, or uses, such as changing '*a long thick coat worn in cold weather*' to '*a long warm coat worn over other clothing in cold weather*'. At times, this can significantly alter the semantic meaning and applicability of the descriptions.

Omission of Specifics: There are instances where ChatGPT omits specific details that could be crucial for a complete understanding of the term. This occurs in both languages, such as missing the '*strong blue cotton cloth*' for jeans or omitting '*military style*' from the description of a trench coat.

Quality of Additional Information: In both languages, the added details can either enrich the understanding of a term or potentially lead to inaccuracies. The effectiveness of these additions depends on the context in which the definition is used. In educational or technical contexts, such precision and additional context may be valuable, but it could also complicate understanding in more general uses.

In summary, while the definitions produced by ChatGPT maintain a reasonable level of accuracy and fidelity across languages, the English definitions exhibit a greater degree of elaboration and variability. ChatGPT shows proficiency in handling core concepts but could benefit from improvements in consistently including crucial specifics and managing the extent of elaboration to avoid unnecessary deviations. This analysis underscores the importance of fine-tuning and potentially adjusting the model outputs based on the target language and the specificity required by the usage context. Moreover, it is important to recognise that automatic evaluations, such as those using the Levenshtein distance, can sometimes yield lower results even when definitions are grammatically and factually correct. This discrepancy can arise because the definition might use synonyms or include specific details that do not match the reference exactly but are still accurate. Such instances highlight the need for nuanced interpretation of evaluation metrics to appreciate the full accuracy and utility of the definitions provided.

5 Conclusions

This paper has presented the process of building a domain specific corpus and an evaluation of the terminology extraction tools TBXTools, SketchEngine, and ChatGPT on this corpus, providing insights into the capabilities and limitations of each tool. TBXTools showed a high recall rate, capturing a broad array of relevant terms, yet struggled with precision due to the inclusion of many irrelevant terms. This was seen across various k values (i.e. the number of terms extracted) in both English and Russian, with precision diminishing as more terms were considered. SketchEngine exhibited similar characteristics to TBXTools. ChatGPT presented a distinct advantage in the extraction process due to its tailored approach to terminology extraction based on the input prompts. This resulted in higher precision, particularly evident when the terms were directly linked to the input context, reducing the inclusion of irrelevant terms significantly. The performance of ChatGPT was robust across different k values, suggesting its effectiveness in handling extensive datasets without a significant loss in term relevance or accuracy. An error analysis focusing on the top 50 terms from each tool further underscored the differences in output quality. ChatGPT's outputs were particularly notable for their relevance and contextual accuracy, reflecting the model's strength in generating pertinent content based on nuanced understanding of the domain-specific texts.

Overall, the study suggests that while traditional tools like TBXTools and Sketch Engine can efficiently capture a wide range of terms, they require improvements in precision to reduce noise. While capable of extracting numerous candidate terms, the lists produced are simply too long in many contexts and include too many words that are not real terms. This may hinder the process of terminology creation. ChatGPT, on the other hand, demonstrates a strong capacity for generating both high-quality terminology lists and accurate definitions, making it a compelling choice for stakeholders in the fashion industry looking for reliable and context-aware terminology extraction solutions. However, the lists of candidates produced by ChatGPT are significantly shorter which means they may miss important terms. This analysis highlights the importance of selecting the appropriate tool based on the specific needs and contexts of terminology extraction tasks, especially in specialised fields such as fashion where accuracy and relevance are crucial.

References

Afzaal, M., Naqvi, S. B. & Qiang, G. (2023), Language, Corpora, and Technology in Applied Linguistics, Frontiers Media SA

Chodkiewicz, C., Bourigault, D. and Humbley, J. (2002), ‘Making a workable glossary out of a specialised corpus’, *Lexis in Contrast: Corpus-based Approaches* 7, 249

Corpas Pastor, G. & Seghiri, M. (2007), ‘El concepto de representatividad en la lingüística del corpus: aproximaciones teóricas y metodológicas. Technical document.’, Málaga: University of Málaga. pp. 373–382

Giguere, J. and Iankovskaia, A. (2023) Leveraging Large Language Models to Extract Terminology. In *Proceedings of the First Workshop on Natural Language Processing tools and resources for translation and interpreting applications.* 54 – 57

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. (2014), ‘The sketch engine: ten years on’, *Lexicography* 1(1), 7–36.

Levenshtein, V.I. (1966) Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10, 707-710.

Lew, R. (2023) ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications* **10**, https://doi.org/10.1057/s41599-023-02119-6

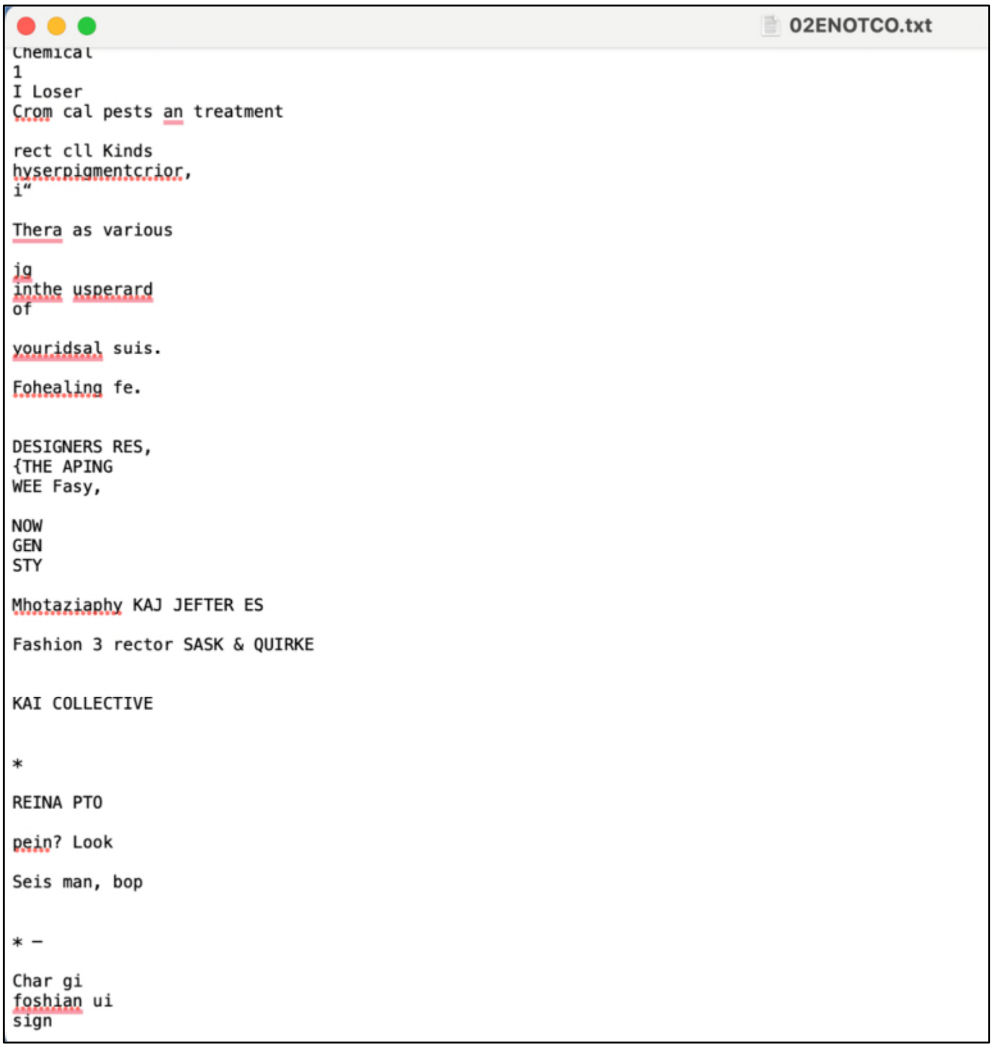
Massion, F., 2024. Terminology in the Age of AI: The Transformation of Terminology Theory and Practice. *Journal of Translation Studies* 4, 67–94. https://doi.org/10.3726/JTS012024.04

Muegge, U. (2023) *Terminology Extraction for Translation and Interpretation Made Easy: How to use ChatGPT and other low-cost, web-based programs to create terminology extraction lists and glossaries quickly and easily*. ISBN 9798989304301

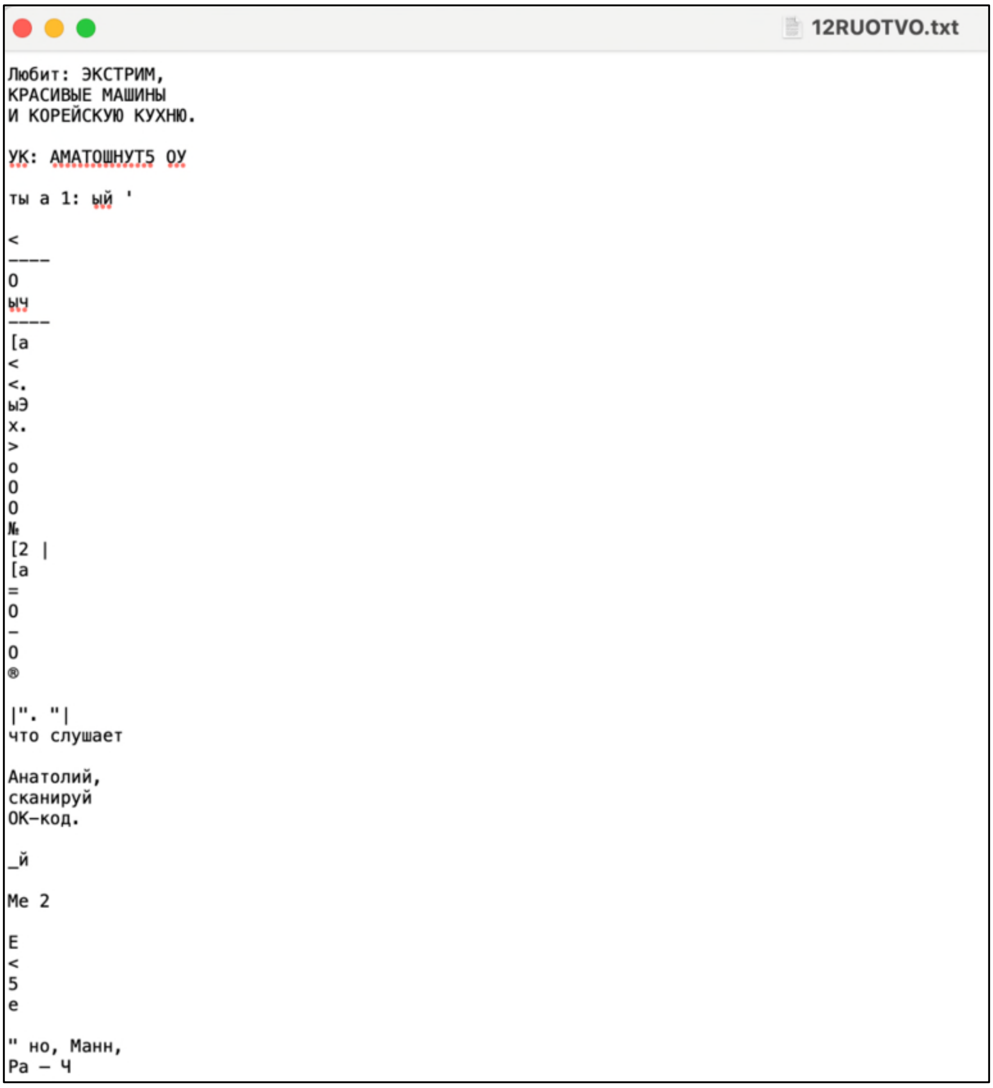
Oliver, A. and Vàzquez, M. (2015), Tbxtools: a free, fast and flexible tool for automatic terminology extraction, in ‘*Proceedings of the international conference Recent Advances in Natural Language Processing*’, pp. 473–479

Appendix

Examples of noisy output in English and Russian produced when the PDF files were converted to TXT format.



```
Chemical
1
I Loser
Crom cal pests an treatment

rect cll Kinds
hyserpigmentcrior,
i“

Thera as various

jg
inthe usperard
of

youridsal suis.

Fohealing fe.


DESIGNERS RES,
{THE APING
WEE Fasy,

NOW
GEN
STY

Mhotaziaphy KAJ JEFTER ES

Fashion 3 rector SASK & QUIRKE


KAI COLLECTIVE


*

REINA PTO

pein? Look

Seis man, bop


* —

Char gi
foshian ui
sign
```



```
Любит: ЭКСТРИМ,
КРАСИВЫЕ МАШИНЫ
И КОРЕЙСКУЮ КУХНЮ.

УК: АМАТОШНУТ5 ОУ

ты а 1: ый '

<
----
0
ыч
----
[a
<
<.
ыЭ
х.
>
о
0
0
№
[2 |
[a
=
0
-
0
®

|". "|
что слушает

Анатолий,
сканируй
ОК-код.

_й

Ме 2

E
<
5
e

" но, Манн,
Pa - Ч
```

Enhanced multilingual speech-based post-editing based on user feedback with COPECO-SPEECH

Jeevanthi U. Liyana Pathirana

Faculty of Translation and Interpreting,
University of Geneva

Jeevanthi.Liyana@etu.unige.ch

Perrine Schumacher

F.R.S.-FNRS Postdoctoral Researcher,

University of Liège

p.schumacher@uliege.be

Pierrette Bouillon

Faculty of Translation and Interpreting,
University of Geneva

Pierrette.Bouillon@unige.ch

Jonathan Mutal

Faculty of Translation and
Interpreting, University of Geneva

jonathan.mutal@unige.ch

Abstract

We present our experience in a work-in-progress: a free-to-use speech-enabled web-based translation/post-editing workbench. We integrated spoken post-editing commands for English and French, the result of continuous feedback received through multiple user tests which will also be continued in the future. In our current version of this platform, we have two speech recognition engines that can be used: Google Web Speech and Azure speech recognition services. Speech and Post-Editing commands for English and French are integrated. Currently we design user tests on typing and spoken post-editing modes through controlled experiments. Quantitative analysis will be performed on the translated texts, speech post-editing commands used, time taken to post-edit and number of keystrokes used when necessary. The design aims to perform qualitative analysis based on questionnaires provided with a Likert Scale approach and gather user opinions as well on how to improve the workbench behavior for better usability. We aim to use our findings to improve COPECO-SPEECH for different languages and text types, add more speech commands, and let interested users try our platform for speech-based translation and post-editing.

1 Introduction

Translators can benefit from Automatic Speech Recognition (ASR) by reducing their typing time and effort, and thus increasing their productivity. Previous studies have shown how ASR, when combined with machine translation (MT), can improve translators' workflows and ergonomics, as well as translation quality (Ciobanu, 2014, 2016). Some commercial CAT tools have integrated ASR systems to offer speech-based translation and post-editing: for example, memoQ (memoQ, 2023) with Apple speech, MateCat with dictation (MateCat, 2023a, 2023b) and Dragon Naturally Speaking (Nuance Communications, 2023). However, these solutions are either proprietary or require licenses and specific software or hardware configurations. Previous research (Mesa-Lao, 2014) explored the use of ASR to refine MT outputs. Our recent study showed that translators in international organizations are open to using speech for post-editing (Liyanapathirana et al., 2019), and a further study demonstrated improved translation quality when combining ASR with MT systems (Liyanapathirana and Bouillon, 2021). We also found that various ASR tools used alongside typing can boost productivity in professional settings

(Liyanapathirana and Bouillon, 2022), though these are based on proprietary technology, which limits their accessibility.

Multimodal interfaces have been studied for their ability to improve translation processes, incorporating ASR as one of the components. Studies have assessed multimodal interfaces with voice and touch features (Teixeira et al., 2019; Zapata, 2014), web-based platforms where ASR usage was not fully investigated (MateCat, 2023a), and the multi-modal post-editing (MMPE) CAT tool, which integrated ASR commands, text input, and touch reordering to post-edit but required unique hardware setups for testing (Herbig et al., 2020a). In summary, prior studies on speech-based translation and post-editing involved methods that needed special equipment, separate applications, or commercial software such as Dragon (Nuance Communications, 2023) and Trados Studio (Trados, 2023). These constraints inspired/prompted the development of a free, web-based translation workbench to facilitate speech-based post-editing and translation, which is being incrementally improved and will be accessible to all users upon account creation.

2 COPECO-SPEECH Design

We chose COPECO (Mutal et al., 2020), a tailor-made, open-source PE platform as our foundation platform. COPECO is a collaborative effort between Geneva University and Liège University, aiming to collect student post-edits alongside teacher corrections, create an open-source post-editing corpus, and establish a systematic approach for annotating translation errors. Learner corpora are valuable resources for educators, helping them analyze student errors and refine course content. Additionally, COPECO provides an online platform for annotating and reviewing student post-editing tasks, facilitating a hands-on learning experience.

When designing COPECO-SPEECH, COPECO was integrated with ASR engines via an application programming interface (API), expanding the approach suggested in (memoQ, 2023). This initiative builds on prior research into the potential of speech recognition tools for post-editing and translation, as well as an evaluation of how commercial speech recognizers function in post-editing tasks. While such tools can be effective, they are often costly or lack customization options to meet specific user needs.

COPECO-SPEECH addresses these limitations, offering an accessible and tailored alternative. It allows us to implement and test different ASR engines if necessary, in the future. Current ASR solutions support speech recognition but lack specific commands for post-editing. We thus took the initiative to develop/create a set of custom speech commands for post-editing tasks in each target language.

Research comparing spoken post-editing, spoken translation, and traditional typing translation further supports the benefits of integrating speech-based methods into translation workflows. For example, previous research found that both translation dictation (spoken translation) and post-editing were generally faster than traditional typing, though each modality presented unique challenges (Morita et al., 2016). Post-editing often involved scattered typing behavior, while dictation resulted in more coherent text production, highlighting the efficiency trade-offs of each method. Similarly, research on the COPECO platform demonstrated that using speech recognition tools for post-editing could reduce the physical effort and time associated with typing, thereby enhancing productivity (Liyanapathirana et al., 2023a, 2023b)). However, these studies also emphasized the importance of addressing challenges such as ensuring accuracy in word order and managing repeated phrases. These findings demonstrate the potential of COPECO-SPEECH to provide a flexible solution that emphasizes the strengths of spoken and typed modalities, offering a tailored approach to meet diverse user needs while optimizing efficiency and output quality.

The use of spoken post-editing (PE) commands in COPECO-SPEECH will be analyzed to see if it brings benefits compared to traditional typing or generating full translations from scratch using Automatic Speech Recognition (ASR) systems (e.g. Matecat). In spoken post-editing, spoken commands are designed to streamline specific post-editing tasks, such as correcting isolated errors or making small adjustments to phrasing, rather than re-translating entire segments. This approach can potentially save time by reducing the need for typing or repeating large amounts of text, particularly in cases where only minor edits are required. However, challenges such as managing word order or selecting specific repeated phrases within a segment are acknowledged, as these tasks may require greater precision and could be less intuitive with voice commands alone.

COPECO-SPEECH thus also allows for a combined modality approach, where users can seamlessly switch between spoken commands and traditional typing as needed. This flexibility ensures that users can choose the most efficient method for each task. For example, while spoken commands might be ideal for quick fixes or navigating through a segment, typing may be preferred for resolving complex structural issues or nuanced edits. The tool is designed to integrate both modalities, enabling users to activate or switch between them fluidly based on the nature of the task, thereby optimizing both speed and output. This hybrid approach aims to enhance user productivity and adaptability while catering to different preferences and workflows.

COPECO-SPEECH allows translation trainers to assign text-based tasks to students/users, with machine translation suggestions. The students can choose to translate from scratch or post-edit: using typing, speech, or both.

The microphone setup can be configured to be "ON" for all segments within the task or activated on demand. The system is designed to support both verbal commands and traditional typing, allowing users to choose a mixed modality approach. This flexibility ensures that users can make use of strengths of each input method to optimize their workflow and improve efficiency. Once the task is complete, students can return it to trainers for correction using predefined or customized annotation schemes. The platform also displays the corpus, including translations, corrections, speech commands, error annotations, and reference translations (if any). Additionally, it generates an open-source post-editing corpus from the post-edits by students and trainers. This data can be shared anonymously.

As a novel feature, an additional page has been included in COPECO-SPEECH, which allows translation trainers to view all speech commands used by a student. Currently, this feature is used for testing purposes enabling us to check and refine the functionality of speech commands. However, in real/practical use, this feature can serve as a valuable resource within COPECO-SPEECH, providing trainers with greater insight into student behavior during speech-based translation tasks.

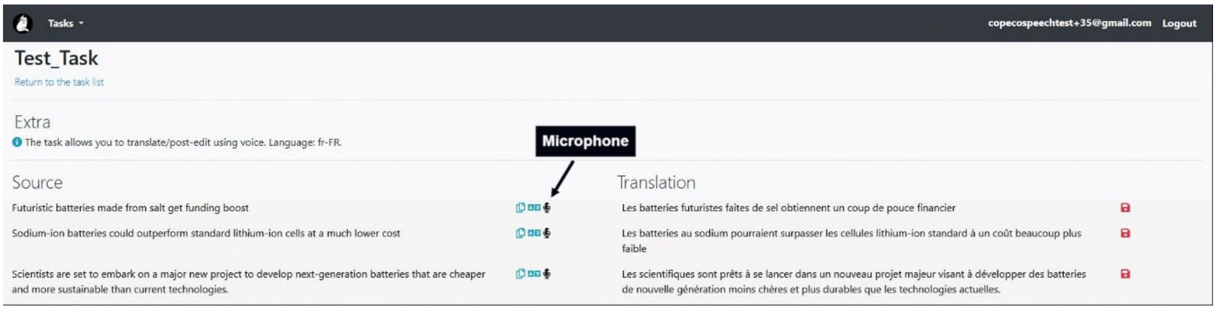
COPECO-SPEECH will enable translators to work using speech technology for translation and post-editing, while offering translation trainers a better understanding of the common challenges students encounter when using speech (commands). The platform can also offer a useful aid for analysing speech-based translation and post-editing patterns, collecting valuable data on these behaviours.

The COPECO-SPEECH tool is primarily designed for educational purposes, building on the foundation established by the COPECO project. It introduces a new area of exploration, focusing on how spoken post-editing can support translators and their perception of its advantages. The main goal of COPECO-SPEECH aligns with COPECO's emphasis on translation education, targeting both teachers and students. It seeks to train students in post-editing using multimodal approaches, equipping them with skills for adapting to emerging technologies in professional settings.

Although COPECO-SPEECH is primarily designed for use in education, it has potential applications in professional contexts. However, these would require additional features, such as enhanced security and integration with other tools. Ultimately, COPECO-SPEECH serves as a pedagogical platform for investigating how speech-based methods can support post-editing and translation workflows.

3 Current Status and Future Work

COPECO-SPEECH is now functional as a speech-enabled, web-based translation/post-editing workbench and is undergoing continuous improvement (Figure 1). It is also in the process of being deployed as a university service, allowing any user to freely access and explore its features.



Test_Task

Return to the task list

Extra

The task allows you to translate/post-edit using voice. Language: fr-FR.

Source	Translation
Futuristic batteries made from salt get funding boost	Les batteries futuristes faites de sel obtiennent un coup de pouce financier
Sodium-ion batteries could outperform standard lithium-ion cells at a much lower cost	Les batteries au sodium pourraient surpasser les cellules lithium-ion standard à un coût beaucoup plus faible
Scientists are set to embark on a major new project to develop next-generation batteries that are cheaper and more sustainable than current technologies.	Les scientifiques sont prêts à se lancer dans un nouveau projet majeur visant à développer des batteries de nouvelle génération moins chères et plus durables que les technologies actuelles.

Figure 1. A task in COPECO-Speech workbench, with a microphone that can be used for spoken translation/post-editing.

In the current version of the platform, we can use either Google's Web Speech API or Azure's speech recognition service as the ASR engine, with integrated speech and post-Editing commands for English and French. These commands are the result of multiple user tests involving both translation students and professionals. These tests provided insights at various levels, including browser, language, usability and hardware levels, which contributed to refining the workbench's functionality.

For example, some post-editing commands developed for English include: "clear segment“," "move cursor to x position", "select word/phrase", selecting and replacing words/phrases, "delete word/phrase", insert punctuation marks, "previous segment" and "next segment".

During these user tests, we also identified additional user-friendly commands, such as "undo" which provides greater flexibility for the user.

These findings were further expanded when developing commands for the French language. Feedback from the professional translator was incorporated to improve the design and functionality of our platform for French. Improvements were made to HTML rendering and speech recognition accuracy via additional steps accomplished by integrating regular expressions, homophone detection, browser rendering improvements and fuzzy matching via programming approaches. Some French post-editing commands developed are able to select a text/phrase, delete the selected text/phrase, undo the previous change, replace one phrase with another and insert punctuation marks.

Command in French	**Description**
selectionner/z *A B C*	Selects/Highlights phrase A B C if it exists in the text
supprimer/z ceci	Deletes any highlighted text
annuler/z choix	Undoes highlighting in the text
annuler	Undoes the previous change
remplacer *A B C* par *D E F*	Replaces A B C phrase (if it exists) by D E F
effacer *A B C*	Delete A B C phrase if it exists
effacer segment	Delete the segment
point d'interrogation	Inserts a question mark

Table 6: A selected set of French post-editing commands

3.1 Quantitative and Qualitative analysis/evaluation of COPECO-SPEECH modalities

We are currently designing an experiment where selected users (professional translators) test COPECO-SPEECH using both typing and spoken post-editing modes.

Following the approach outlined in Herbig et al., 2020c, we perform a structured/controlled test for each modality. Instead of running an MT system, we manually introduce errors into the reference set to ensure that each segment contains only one or two major errors. Participants can correct a set of sentences using a pre-chosen modality. Within each modality, we try to capture slightly different edit cases, such as deleting single words or a group of words, removing duplicates, replacing one word or group of words with another. For initial tests, participants will be provided with the exact correction to apply for each segment, as well as the modality to use. This approach standardizes editing behaviors across participants, enabling comparable subjective ratings, feedback, and time measurements. To mitigate ordering effects, participants complete the tasks in counter-balanced order, with the modalities in random order. For each task, the post-edited text, number of keystrokes, duration, and speech commands used will be recorded for quantitative analysis.

Upon completing the experiment, participants will rate each modality using 7-point Likert scales [5] ranging from “strongly disagree” to “strongly agree”, assessing whether the interaction “is a good match for its intended purpose”, “is easy to perform”, and “is a good alternative to the current mouse and keyboard approach”. In addition, we will also collect user satisfaction ratings, preferences, and suggestions through questionnaires and interviews.

We will use our findings to improve COPECO-SPEECH for different languages and text types, add more speech commands, and let interested users try our platform for speech-based translation and post-editing. The findings from this experiment will guide further improvements to COPECO-SPEECH, including the addition of speech commands, support for different languages and text types, and expanded access for users interested in experimenting with speech-based translation and post-editing.

References

Ciobanu, Dragoș. 2014. Of Dragons and Speech Recognition Wizards and Apprentices. In Tradumatica 2014, pages 524-538.

Ciobanu, Dragoș. 2016. Automatic Speech Recognition in the Professional Translation Process. In Translation Spaces. A Multidisciplinary, Multimedia, and Multilingual Journal of Translation, volume 5, pages 124-144.

Nuance Communications. 2023. Dragon Speech Recognition Solutions. https://www.nuance.com/dragon.html.

Herbig, Nico, Santanu Pal, et al. 2020. Multi-modal Approaches for Post-editing Machine Translation. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, page 231. ACM.

Herbig, Nico, Santanu Pal, Tim Düwel, Raksha Shenoy, Antonio Krüger, and Josef van Genabith. 2020. Improving the Multi-modal Post-editing (MMPE) CAT Environment Based on Professional Translators' Feedback. In Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation, pages 93-108.

Herbig, Nico, et al. 2020. MMPE: A Multi-modal Interface for Post-editing Machine Translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Liyanapathirana, Jeevanthi, Pierrette Bouillon, and Bartolomé Mesa-Lao. 2019. Surveying the Potential of Using Speech Technologies for Post-editing Purposes in the Context of International Organizations: What Do Professional Translators Think?. In Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks, pages 149-158.

Liyanapathirana, Jeevanthi, and Pierrette Bouillon. 2021. Integrating Post-editing with Dragon Speech Recognizer: A Use Case in an International Organization. In Translating and the Computer 43, pages 55-67.

Liyanapathirana, Jeevanthi, and Pierrette Bouillon. 2022. Exploring Different Speech Recognizers for Post-editing Translation Outputs: A Pilot Study in an International Organization. In Translating and the Computer.

Liyanapathirana, Jeevanthi, Pierrette Bouillon, and Jonathan David Mutal. 2023. "A Pedagogical Platform for Spoken Post-Editing (PE): The Integration of Speech Input into COPECO." In *Proceedings of the International Conference on Human-Informed Translation and Interpreting Technology (HiT-IT 2023)*, pages 195–202.

memoQ. 2023. memoQ Homepage. https://www.memoq.com/.

Liyanapathirana, Jeevanthi, Pierrette Bouillon, and Jonathan Mutal. 2023. Towards a Free, Web-Based Workbench for Speech-Enabled Translation and Post-Editing: Speech Integrated COPECO. In Translating and the Computer 45.

Mesa-Lao, Bartolomé. 2014. Speech-enabled Computer-aided Translation: A Satisfaction Survey with Post-editor Trainees. In Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation, pages 99-103.

Mutal, Jonathan D., Pierrette Bouillon, Pascale Schumacher, and Johanna Gerlach. 2020. COPECO: a Collaborative Post-Editing Corpus in Pedagogical Context. In North American Component of the International Association for Machine Translation. 1st Workshop on Post-Editing in Modern-Day Translation.

MateCat. 2023. MateCat Homepage. https://site.matecat.com/.

MateCat. 2023. MateCat Guide. https://guides.matecat.com/translate-1.

Morita, Tadashi, Noriko Kawai, and Koichiro Ishikawa. 2016. "ENJA15: Comparative Analysis of Post-editing, Translation Dictation, and Manual Typing in English-to-Japanese Translation." In Proceedings of the 22nd Annual Meeting of the Association for Natural Language Processing, pages E7-3.

Teixeira, Carlos S., Joss Moorkens, Daniel Turner, Joris Vreeke, and Andy Way. 2019. Creating a Multimodal Translation Tool and Testing Machine Translation Integration Using Touch and Voice. In Informatics, volume 6, page 13. MDPI.

Trados. 2023. Trados Studio Homepage. https://www.trados.com/products/trados-studio/.

Zapata, Julian. 2014. Exploring Multimodality for Translator-Computer Interaction. In Proceedings of the 16th International Conference on Multimodal Interaction, pages 339-343.

Evaluating 24 language models for a Speech-to-Text tool at the European Parliament

Francesco Fernicola

European Parliament

francesco.fernicola@europarl.europa.eu

Abstract

The European Parliament's translation service (DG TRAD) is working on a speech-to-text and machine translation (MT) tool that can automatically transcribe and translate parliamentary multilingual debates in real time, covering the 24 official languages of the EU. The purpose of this development is to increase accessibility for the deaf and hard-of-hearing people, who currently have no access to the plenary debates. We analysed the automatic speech recognition (ASR) and the machine translation (MT) components separately. While latency and word error rate (WER) scores were analysed for ASR, the MT evaluation delved deeper into the human annotations, using the Multidimensional Quality Metrics (MQM) framework combined with automatic metrics (COMET). We benchmarked the quality of the tool's internal MT engine against the EU's eTranslation tool. For significance testing, we employed Cohen's Kappa and Matthews Correlation Coefficient (MCC) for interannotator agreement, and performed correlation analysis using Spearman's correlation coefficient to uncover patterns between MQM and COMET scores. Using the use-case of the real-time speech-to-text and MT tool developed by the European Parliament, this study underscores the challenges in multilingual translation quality assessment. It emphasises the need for consistent annotation and robust evaluation frameworks, putting human evaluation at the centre of the process.

1 Introduction

The European Parliament's translation service (DG TRAD) is working on a speech-to-text and machine translation (MT) tool that can automatically transcribe and translate parliamentary multilingual debates in real time, covering the 24 official languages of the EU. The purpose of this development is to increase accessibility for the deaf and hard-of-hearing people, who currently have no access to the plenary debates.

For the scope of the project, we distinguish between the evaluation of the Speech Models (ASR), which is performed using only automatic metrics based on human golden standards, and the evaluation of the Translation Models (MT) implementing a mix of human annotation and automatic metrics. For the ASR Models we employ the standard metric Word Error Rate (WER). It is important to remember that WER only shows improvements based on a word-by-word, non-case sensitive comparisons, i.e. matching a human-made transcription with the automatically generated transcription. It does not take into account other types of improvements that only a human evaluator can assess, like correct punctuation, capitalisation of letters and better sentence segmentation (Woodard & Nelson, 1982).

Hence, for the scope of this paper, we only perform a deeper analysis of the evaluation of the MT component.

2 Methodology

2.1 Evaluation Metrics

All human evaluations were carried out by trained Intercultural and Language Professionals (ILPs), following the Multimodal Quality Metrics framework developed by Lommel et al. (2014). A total of two evaluators performed the annotation for each language independently. The error labels of the MQM framework that the annotators worked with were the following:

1. **Terminology**: errors arising when a term does not conform to normative domain or organizational terminology standards or when a term in the target text is not the correct, normative equivalent of the corresponding term in the source text. In this category we also included named entities, acronyms, numbers, dates, etc. that are incorrect.

2. **Accuracy**: errors occurring when the target text does not accurately correspond to the propositional content of the source text, introduced by distortion, omission, or additions to the message.

3. **Punctuation**: includes punctuation errors (punctuation substitution, addition or deletion).

4. **Grammar**: errors that occur when a text string (sentence, phrase, other) in the translation violates the grammatical rules of the target language.

5. **Register**: errors that occur when a text uses a level of formality higher or lower than required by the specifications or by common language conventions.

6. **Other**: errors that do not fall into the previous categories

7. **Unintelligible**: refers to segments that contain so many errors that it is not possible to list all of them. It should be used when there are more than five major errors in a segment.

For each category, a severity level was assigned to each marked error: neutral, minor or major. The 0-25 scale was based on Chapter 3 of Freitag et al. (2023). Table 1 shows a representation of the weights that were used for the errors in our analysis. As a measure of interpretability, we note that higher score equals worse quality, i.e. a score of 1 would indicate that for that language - on average - it is possible to find one minor error per segment.

Severity	Category	Weight
Major	Non-translation	25
	all others	5
Minor	Fluency/Punctuation	0.1
	all others	1
Neutral	All	0

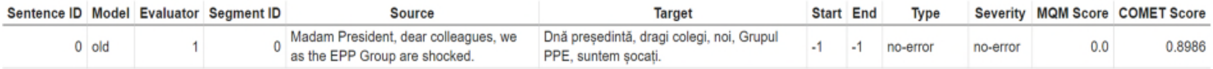
Table 1. Google's MQM error weighting

In addition, we compute the quality estimation metric COMET to offer a comparison between human evaluation and state-of-the-art automatic metrics in the domain of parliamentary debates. We use the default model wmt22-comet-da, which is accessible in version 2.0.2 of the Unbabel/COMET framework. At the time of the development of the corpus, this was the most accessible version of the model, although now a more recent version has been published in the form of the COMETKIWI-22 model, which ranked first in the WMT tasks at both sentence-level and word-level prediction tasks and the fine-grained error span detection task, while reaching word-, span- and sentence-level granularity (Rei et al., 2022; Guerreiro et al., 2023).

2.2 Corpus

The annotated corpus consists of a collection of speeches delivered by Members of the European Parliament (MEPs) in their original language, accompanied by their respective machine translations into a target language. We collected five speeches per language, always bidirectional starting from English (EN<>XX), with sentence level segmentation and alignment, resulting in ten minutes of manually evaluated speech translation data per language.

Figure 1 shows an excerpt from the corpus, with data collected following the 2024 Quality Estimation Shared Task corpus structure (Blain et al., 2023). The additional column labelled “Model” was only used to distinguish between internal model versions.



Sentence ID	Model	Evaluator	Segment ID	Source	Target	Start	End	Type	Severity	MQM Score	COMET Score
0	old	1	0	Madam President, dear colleagues, we as the EPP Group are shocked.	Dnă președintă, dragi colegi, noi, Grupul PPE, suntem șocați.	-1	-1	no-error	no-error	0.0	0.8986

FIGURE 6. CORPUS ENTRY EXAMPLE

2.4 Statistical analysis

We computed the Inter-Annotator Agreement (IAA) score using MCC (Matthews correlation coefficient). The formula for MCC is the following (Matthews, 1975):

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

The scores were calculated by treating one of the two annotators as the gold annotation to obtain True Positives (TP), False Positives (FP), True Negatives (TN), False Negatives (FN). In cases where this could result in an undefined MCC, the score was assigned as either 0 or 1 based on given values. We weigh by the number of annotated errors to counteract the effect of undefined values in each category.

We computed the MCC three times during our analysis. Firstly, we compared error types solely at the segment level. Subsequently, we focused on error spans within segments, focusing on overlaps of errors within these segments. Lastly, we also explored this at the token level, emphasizing the detection of token-level error overlaps within the segments. Table 2 shows the different levels of annotation.

	Segment-wise MCC	**Token-wise MCC**	**Span-wise MCC**
Evaluator 1	Madam President , dear colleagues	Madam President , dear colleagues	Madam President , dear colleagues

Evaluator 2	Madam President , dear colleagues	Madam President , dear colleagues	Madam President , dear colleagues

Table 2. Example Annotation Comparison for a Speech Segment

Then, we employed Spearman (**ρ**) correlation coefficient to assess the degree and direction of the linear relationship between the human MQM annotation scores with respect to the COMET metric.

3 Results

All error scores presented here were weighted by their severity according to the MQM framework as described in Chapter 1 and normalized by the number of annotated segments for each language pair. The results are provided in Table 3 with all scores reported as MQM points per segment. We report that the **accuracy** category has the strongest influence on the final error score, averaging a total of 1.2, with all language pairs above 0.4, except for PT (0.23). Unintelligible errors follow with an average of 0.59, even though they are not present in 7 of the 22 languages analysed. Due to the relatively high weight of unintelligible errors and due to their high frequency, especially in RO, their impact on the total score is still rather strong. Terminology errors account for 0.52, with values again fluctuating between 0.05 (SL) and 2.01 (EL). Together with **accuracy**, **unintelligible**, and **terminology**, the **grammar** category completes the list of categories strongly affecting the overall MQM score. With an average of 0.27, all languages are within the standard deviation (0.27) except for MT (1.2). The **other** category contributes an average of 0.09 with scores consistently below 0.18 except for MT (0.64). Both **punctuation** and **register** only seem to have a minor influence on the total score due to the relatively low frequency of those categories and the decreased weight of punctuation errors (see Table 1).

However, the distribution of error categories did not only differ among language pairs. In many cases, high discrepancies between the two annotators within a language pair could be observed. This can be clearly seen in the MCC scores for agreement between the annotators. For 10 of the languages, the agreement is negligible (i.e. lower than 0.2), 6 languages had a weak positive relationship between annotators (i.e. lower than 0.3), and 6 languages showed a moderate positive relationship (i.e. lower than 0.4) between evaluators. In languages with exceptionally low agreement between annotators, it was often the case that certain error categories were almost exclusively used by one of the annotators.

The correlation between COMET score and MQM shows correlation strength ranges between negligible and moderate. Since higher COMET scores and higher quality correlate with lower MQM error scores (i.e. lower amounts of errors), Spearman's Correlation Coefficient is negative throughout all languages. Previous research has found similar correlations for EN-RO (Rios Gaona et al. 2023).

At the single evaluator level, we report discrepancies of at least 0.1 between annotators in 8 of the 22 languages analysed, the largest discrepancies being present in RO, SK, and EL. On the other side of the spectrum, there were also 10 language pairs with very similar correlations between MQM points per segment and COMET and differences below 0.5.

Interestingly, these (dis-)agreements do not always correspond with the previously mentioned inter-annotator agreement. While LV has a relatively high IAA (0.36), the degree of correlation between COMET and MQM points per segment differs between evaluators. This could be due to the fact that the IAA calculation does not take into account different weights for different error severities. Spearman's correlation coefficient on the other hand cannot incorporate error types because COMET does not provide them.

4 Conclusion

While both systems exhibit comparable translation quality, discernible advantages emerge in certain contexts. Notably, eTranslation demonstrates a marginally superior performance in High-Resource Language settings, while RWS exhibits slightly greater efficacy in Low-Resource Language scenarios. The fine-grained evaluation also revealed similar tendencies in the error annotation for both systems, with accuracy and unintelligible being the most commonly annotated error types. It should be noted however that in the case of RWS, the statistical difference is caused by the accuracy errors, whereas the remaining categories show very similar distributions. This finding underscores the importance of tailoring system selection to the specific linguistic landscape of the application domain.

A crucial aspect affecting the reliability of the evaluation pertains to the Inter-Annotator Agreement (IAA). While FR has a relatively low IAA (0.18), the degree of correlation between COMET and MQM points per segment does not differ a lot between evaluators (0.03). This could be because the IAA calculation does not take into account different weights for different error severities. Spearman's correlation coefficient, on the other hand, cannot incorporate error types because COMET does not provide them. Consequences of this are especially evident in the negligible correlation found for both EL and LV, both suspiciously very high quality languages, close if not surpassing DE, although they belong to the low-resource category. Such disparities highlight the need to ensure consistent annotations and robust evaluation frameworks, by furthering research in standardised evaluation procedures to ensure consistency and reliability across systems. Furthermore, although exceptionally useful to have a different perspective on the quality of MT, state-of-the-art automatic metrics such as COMET should still be handled with the utmost care, by offering a comparison with human annotations in a human-in-the-loop system, keeping in mind the potential pitfalls that come with out-of-domain data and low-resource languages.

Acknowledgements

We wish to thank all the DGTRAD ILPs who helped us annotate the corpus in 24 languages, having the patience to learn an error annotation framework tailored to MT, constantly giving us feedback and overall contributing to a project which is essential for the European Union.

We also thank the team at the Commission's DGT who helped us set up the eTranslation API and the various colleagues who helped this project to succeed.

Finally, we wish to thank the wonderful trainees who helped to process, collect and analyse the data over the course of the last year: Daniel Exler, Artur Bondarenco, Alexia Mathiopoulou and Nils Bouckaert.

References

Blain, F., Zerva, C., Rei, R., Guerreiro, N. M., Kanojia, D., de Souza, J. G., ... & Martins, A. F. (2023). Findings of the WMT 2023 shared task on quality estimation. In Proceedings of the Eighth Conference on Machine Translation (pp. 629-653).

Freitag, M., Mathur, N., Lo, C.-k., Avramidis, E., Rei, R., Thompson, B., . . . Foster, G. (2023). Results of WMT23 Metrics Shared Task: Metrics Might Be Guilty but References Are Not Innocent. Proceedings of the Eighth Conference on Machine Translation (pp. 578-628). Singapore: Association for Computational Linguistics.

Guerreiro, N. M., Rei, R., van Stigt, D., Coheur, L., Colombo, P., & Martins, A. F. (2023). xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection. https://arxiv.org/abs/2310.10482v1.

Lommel, A. R., Burchardt, A., & Uszkoreit, H. (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. Tradumàtica: Tecnologies de la Traducció, pp. 455-462.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure, 405(2), 442-451.

Rei, R., Treviso, M., Guerreiro, N. M., Zerva, C., Farinha, A. C., Maroti, C., Martins, A. F. (2022). CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. Proceedings of the Seventh Conference on Machine Translation (WMT) (pp. 634-645). Abu Dhabi: Association for Computational Linguistics.

Rios Gaona, M. A., Chereji, R.-M., Secara, A., & Ciobanu, D. (2023). Quality Analysis of Multilingual Neural Machine Translation Systems and Reference Test Translations for the English-Romanian language pair in the Medical Domain. Proceedings of the 24th Annual Conference of the European Association for Machine Translation (pp. 355-364). Tampere: European Association for Machine Translation.

Woodard, J. P., & Nelson, J. T. (1982). An information theoretic measure of speech recognition performance. In Workshop on standardisation for speech I/O technology, Naval Air Development Center, Warminster, PA.

Language	Accuracy	Grammar	Other	Punctuation	Register	Terminology	Unintelligible	Avg.
BG	0.54	0.52	0.02	0.08	0.00	0.53	1.82	3.52
CS	1.06	0.45	0.11	0.04	0.00	0.69	0.00	2.35
DA	1.19	0.22	0.01	0.03	0.04	1.21	0.56	3.26
DE	0.49	0.20	0.08	0.00	0.00	0.08	0.00	0.84
EL	1.35	0.07	0.17	0.11	0.01	2.07	0.00	3.79
ES	1.45	0.60	0.13	0.13	0.07	0.63	0.00	3.01
ET	1.56	0.18	0.04	0.01	0.01	0.33	0.06	2.19
FI	0.42	0.17	0.03	0.00	0.01	0.34	1.19	2.16
FR	1.96	0.06	0.01	0.07	0.00	0.27	0.08	2.44
HR	0.75	0.11	0.15	0.00	0.00	0.09	0.16	1.25
HU	5.59	0.44	0.03	0.02	0.10	0.69	0.00	6.87
IT	0.94	0.15	0.04	0.00	0.03	0.18	1.90	3.23
LT	0.56	0.15	0.16	0.00	0.00	0.33	1.27	2.47
LV	1.85	0.31	0.01	0.01	0.04	0.29	1.62	4.13
MT	1.25	1.20	0.64	0.03	0.00	1.15	0.58	4.86
NL	1.46	0.35	0.07	0.00	0.10	1.05	0.00	3.03
PL	0.70	0.09	0.09	0.00	0.01	0.09	0.13	1.11
PT	0.23	0.04	0.03	0.00	0.00	0.05	0.00	0.36
RO	1.14	0.45	0.12	0.02	0.00	0.22	2.77	4.72
SK	0.67	0.01	0.00	0.00	0.00	0.30	0.40	1.37
SL	0.71	0.03	0.01	0.00	0.00	0.04	0.32	1.11
SV	0.53	0.13	0.02	0.00	0.04	0.80	0.16	1.68
Avg.	1.20	0.27	0.09	0.03	0.02	0.52	0.59	2.72

Appendix

Table 3. Average MQM Points per Segment by Language, and Error Type (RWS)

Language	Accuracy	Grammar	Other	Punctuation	Register	Terminology	Unintelligible	Avg.
BG	0.97	0.32	0.06	0.01	0.02	0.68	0.25	2.35
CS	0.68	0.40	0.05	0.03	0.00	0.51	0.40	2.09
DA	0.65	0.27	0.18	0.00	0.00	1.82	0.00	2.94
DE	0.20	0.13	0.51	0.00	0.00	0.09	0.00	0.95
EL	0.15	0.00	0.00	0.00	0.00	0.04	0.01	0.21
ES	1.01	0.05	0.29	0.00	0.00	0.40	0.19	1.96
ET	0.95	0.07	0.05	0.00	0.00	0.03	0.00	1.12
FI	0.83	0.25	0.09	0.03	0.03	0.30	3.07	4.62
FR	0.49	0.26	0.19	0.07	0.09	0.50	0.57	2.20
GA	0.82	0.60	0.01	0.00	0.02	0.37	0.59	2.43
HR	0.43	0.05	0.31	0.03	0.00	0.30	2.16	3.30
HU	1.56	0.96	0.91	0.00	0.05	0.23	3.12	6.86
IT	1.31	0.13	0.03	0.00	0.00	0.14	0.07	1.71
LT	0.91	0.12	0.01	0.03	0.00	0.02	1.01	2.13
LV	0.40	0.18	0.11	0.00	0.04	0.33	0.01	1.10
MT	0.39	0.28	0.01	0.00	0.00	0.02	0.34	1.05
NL	0.50	0.04	0.00	0.00	0.00	0.96	0.07	1.59
PL	1.22	0.15	0.07	0.00	0.04	0.27	0.43	2.19
PT	0.39	0.09	0.00	0.00	0.00	0.05	0.00	0.55
RO	0.72	0.08	0.03	0.00	0.00	0.01	0.00	0.86
SK	1.67	0.13	0.33	0.01	0.00	0.43	0.34	2.93
SL	1.51	0.53	0.31	0.10	0.08	1.79	0.91	5.25
SV	1.58	0.17	0.00	0.00	0.00	0.30	0.15	2.22
Avg.	0.84	0.23	0.15	0.01	0.01	0.42	0.59	2.29

Table 4. Average MQM Points per Segment by Language, and Error Type (eTranslation)

Integrating automatic speech recognition into remote healthcare interpreting: A pilot study of its impact on interpreting quality

Shiyi Tan

University of Surrey, UK

s.tan@surrey.ac.uk

Constantin Orăsan

University of Surrey, UK

c.orasan@surrey.ac.uk

Sabine Braun

University of Surrey, UK

s.braun@surrey.ac.uk

Abstract

This paper reports on the results from a pilot study investigating the impact of automatic speech recognition (ASR) technology on interpreting quality in remote healthcare interpreting settings. Employing a within-subjects experiment design with four randomised conditions, this study utilises scripted medical consultations to simulate dialogue interpreting tasks. It involves four trainee interpreters with a language combination of Chinese and English. It also gathers participants' experience and perceptions of ASR support through cued retrospective reports and semi-structured interviews. Preliminary data suggest that the availability of ASR, specifically the access to full ASR transcripts and to ChatGPT-generated summaries based on ASR, effectively improved interpreting quality. Varying types of ASR output had different impacts on the distribution of interpreting error types. Participants reported similar interactive experiences with the technology, expressing their preference for full ASR transcripts. This pilot study shows encouraging results of applying ASR to dialogue-based healthcare interpreting and offers insights into the optimal ways to present ASR output to enhance interpreter experience and performance. However, it should be emphasised that the main purpose of this study was to validate the methodology and that further research with a larger sample size is necessary to confirm these findings.

1 Introduction

Since the introduction of simultaneous interpreting (SI) through electro-acoustic sound transmission systems, technological advances have continuously shaped the world of interpreting. They have given rise to new forms of interpreting, including technology-mediated interpreting or distance interpreting, technology-supported interpreting or computer-assisted interpreting (CAI), and even technology-generated interpreting or machine interpreting (Braun, 2019).

Currently, one of the most promising technologies used to support interpreting workflows is automatic speech recognition (ASR) which converts human speech signals into a sequence of words using computer programmes (Jurafsky and Martin, 2008). In the context of interpreting, ASR was initially utilised to automate the process of querying glossaries and retrieving information in CAI tools (Fantinuoli, 2017). Driven by classic machine learning technologies including Hidden Markov Models and more recently by deep learning technologies, ASR now shows increasingly robust performance and can directly support the interpreting process by providing real-time transcripts of source speeches. Researchers were thus motivated to explore the practical application of ASR in interpreting. Improved accuracy for the rendition of “problem triggers” such as numbers, specialised terms and proper names in SI was achieved with ASR output (e.g., Desmet *et al.*, 2018; Defrancq and Fantinuoli, 2021). This initial focus on ASR’s role in addressing specific stumbling blocks resulted in little attention being paid to its impact on overall interpreting quality. Few studies put the spotlight on ASR in consecutive interpreting (CI) as interpreters often can rely on notetaking as a memory aid and face less time pressure in CI compared to SI.

ASR output generally is neither entirely error-free, nor fully synchronised with the acoustic signal, possibly causing inaccuracies, delays and distractions for interpreters. Consequently, the investigation of how ASR can impact interpreting quality as a whole, is equally important for its wider adoption and development. The integration of ASR into CI-based public service interpreting also deserves attention, particularly in the contexts of healthcare and legal interpreting where accuracy can be a matter of “life and death”. These types of interpreting tasks often feature obscure terminology and frequent use of numbers, units and dates which require correct rendition.

As an extra source of information, ASR output also competes for interpreters’ cognitive resources with other processing tasks during interpreting such as comprehension and production (Seeber, 2011). The number of studies that have explored this problem is limited (e.g., Yuan and Wang, 2024; Li and Chimel, 2024), largely due to the complexity and challenges involved in examining cognitive performance.

The current study, as a pilot study for a larger research project, explores the integration of ASR into dialogue-based healthcare interpreting in an attempt to understand whether ASR helps or hinders interpreters. The rest of the paper is structured as follows. We begin by revisiting previous studies on the intersection of ASR and interpreting, followed by a description of the methodology in Section 3. Section 4 presents the results of the pilot study and Section 5 is a discussion of preliminary findings. The paper concludes with a summary of insights and limitations as well as an outline for future work.

2 Literature review

This section reviews previous explorations into ASR in interpreting from two aspects: the assessment of ASR systems’ performance in interpreting and its impact on interpreting quality. Following a summary of research gaps, three research questions are proposed.

1 Assessment of ASR systems’ performance in interpreting

Certain criteria need to be met by ASR systems to be applied in interpreting. An ASR system should be speaker-independent, able to manage continuous speech, support large-vocabulary recognition and provide the option to add specialised terms for improved recognition (Fantinuoli, 2017). A low word error rate (WER), a metric measuring transcription errors, and a low real-time factor (RTF), a metric assessing transcription speed are also expected in ASR systems (Fantinuoli, 2017).

Using three English texts containing 119 terms and 11 numerals, Fantinuoli (2017) tested that Dragon Naturally Speaking, an ASR engine integrated into the CAI tool InterpretBank 4, reached an accuracy of approximately 95% for term transcription after importing a list of English specialised terms from a bilingual glossary and 100% for numeral transcription. Student interpreters can maintain accuracy and fluency in SI with a 3-second latency in an automatic suggestion system for numbers (Fantinuoli and Montecchio, 2022). Using the Google Cloud Speech-to-Text API for ASR, InterpretBank demonstrated low latency and high precision (96%) in number transcription (Defrancq and Fantinuoli, 2021). In Fritella's study (2022), the latency of SmarTerp, an ASR-integrated CAI tool, was 2 seconds in transcribing name entities, specialistic terms and numbers.

In relation to the format of the transcribed text, research by Defrancq and Fantinuoli (2021) noted that the running transcript was a distraction for some students and preferences were divided regarding what aspects of figures to be displayed, such as only numbers or both numbers and units, and how they should be displayed on the screen.

2 Impact of ASR on interpreting quality

Currently, most research has investigated the impact of ASR on the rendition of numbers and specialised terms. As a result of displaying the numbers on slides, the accuracy rate of number interpreting rose from 56.5% to 86.5% (Desmet *et al.*, 2018). Defrancq and Fantinuoli (2021) found that the interpreting accuracy rates of nearly all number types were enhanced when ASR was available, a finding echoed by Pisani and Fantinuoli (2021), who reported a significant decline in the error rate of number renditions. A difference between the two studies lies in the way the transcribed numbers were presented. Numbers were embedded and highlighted in the entire transcript in the former study, while in the latter, numbers were shown in isolation. With Zoom live captioning, the error rates in interpreting interest periods containing numbers and proper names saw a 30% reduction (Yuan and Wang, 2023).

To date, only a few studies have examined the effectiveness of ASR in relation to overall interpreting quality with various quality assessment frameworks being adopted. Cheung and Li (2022) found that the presence of captions in a video enhanced accuracy but reduced fluency among student interpreters, based on two scoring sheets for each measure. A significant improvement in overall interpreting performance with live captions was also observed among trainees, using quality assessment criteria from the researchers' institution (Yuan and Wang, 2024). In an experiment with professional interpreters, Rodríguez González *et al.* (2023) reported a notable decline in the total number of interpreting errors with ASR support, although style-related errors increased, as assessed through the NTR model (Romero-Fresco and Pöchhacker, 2017). However, all these studies pertained to simultaneous interpreting.

In relation to consecutive interpreting, Chen and Kruger (2022) introduced a computer-assisted consecutive interpreting (CACI) mode that integrates ASR technology with machine translation (MT). Different from studies that employed ASR as a supplementary tool during interpreting, this study required interpreters to listen to the source speech and respeak it into iFLYTEK, an ASR system generating textual output, which was subsequently translated by an MT system. The interpreters then produced a target speech by consulting both the ASR-generated text and the MT output. In CACI, overall Chinese-to-English interpreting quality was enhanced, and fluency was improved in both directions. With a similar research design, Wang and Wang (2019) found that the accuracy of CI was enhanced with ASR-supported MT reference being provided, but no clear conclusion was reached regarding fluency. However, it is important to be aware that in these studies, the differences observed resulted from the combined effects of ASR and MT, making it impossible to draw conclusions about ASR alone.

Research gaps in ASR-supported interpreting

Several research gaps have emerged from the reviewed studies. First, as most previous studies focused solely on using ASR to support the interpreting of “problem triggers”, such as numbers and terms, the impact of ASR on overall interpreting quality remains underexplored.

Second, most ASR systems used are off-the-shelf software, leaving little leeway to adjust transcription accuracy or customise output format in experiments. Generally, these ASR systems can be classified into four types (Table 1).

Third, diversity was observed in the presentation of ASR-generated text, ranging from only numbers (e.g., Fantinuoli and Montecchio, 2022; Desmet *et al.*, 2018) to entire transcripts with or without numbers being highlighted (e.g., Defrancq and Fantinuoli, 2021; Rodríguez González *et al.*, 2023; Saeed *et al.*, 2023), from chunked segments (Cheung and Li, 2022) to scrolling captions (Yuan and Wang, 2023). Some studies divided the interface into distinct sections to display different types of transcribed text, such as numbers with units of measurement, proper names and specialised terms (Fantinuoli *et al.*, 2022), named entities, terms and numbers (Frittella, 2022), and terminology and numerals (Fantinuoli, 2017). With these variations, no agreement was reached on the optimal way of presenting ASR output, and no study tested varying types of ASR output.

Fourth, most of the reviewed research engaged student interpreters, with only a few studies involving professional interpreters (e.g., Frittella, 2022; Rodríguez González *et al.*, 2023; Li and Chmiel, 2024). Although trainee interpreters are more accessible than experienced interpreters for experimental and pedagogical purposes, the significance of involving professional interpreters is crucial, especially concerning the application of ASR in authentic tasks.

Types of ASR systems	Specific tools and Key studies
Simulated ASR systems	Slides (Desmet *et al.*, 2018) Video (Fantinuoli and Montecchio, 2022)
CAI tools with ASR features	InterpretBank (Fantinuoli, 2017; Defrancq and Fantinuoli, 2021; Pisani and Fantinuoli, 2021) SmarTerp (Frittella, 2022) KUDO Interpreter Assist (Fantinuoli *et al.*, 2022)
Stand-alone ASR engines	iFLYTEK (Chen and Kruger, 2022) Dragon Anywhere (Wang and Wang, 2019)
Platforms with captioning features	Zoom captioning (Yuan and Wang, 2023) YouTube subtitles (Li and Chmiel, 2024)

Table 7. ASR systems used in previous studies

Last, current studies also vary in two research design-related factors that presumably impact results: whether participants received ASR training prior to the experiments and which quality assessment framework was implemented.

Given these gaps, we proposed three research questions for the full research project:

1) Is there a significant difference in overall interpreting quality between interpreting with varying types of ASR support and without ASR support?

2) Does the interpreting quality vary across different types of ASR output?

3) How do interpreters interact with different types of ASR output?

In a mixed-methods approach, we conducted experiments complemented by post-experiment retrospective reports and semi-structured interviews in a pilot study, to tentatively explore answers to these questions.

3 Methodology

This section describes the participant information, interpreting materials, experiment design and procedure as well as data analysis methods.

Participants

In the pilot study, four trainee interpreters (all females, mean age = 27.5, range = 24-31, *SD* = 3.51) were recruited within the guidelines of the ethics committee. They were recent graduates from a one-year master's programme in interpreting at a university in England, where they all had completed four compulsory modules on CI and SI. Three of them held a bachelor's degree in English or Translation. All spoke Chinese as their mother tongue and English as their second language. Their average IELTS score was 7.0 (range = 6.5-7.5, *SD* = 0.41). Their prior use of ASR software was limited to classroom demonstrations.

Materials

The interpreting materials used in this study were four scripts adapted from authentic medical consultations provided by a private hospital in London. The scripts are four consecutive consultations between a nephrologist and a patient with renal disease. The difficulty of the four scripts (Table 2) was controlled to ensure comparability based on word count, duration, speed and the Flesch reading ease index, which measures how difficult a text is to understand. A score between 60 and 70 indicates that the text is written at a standard level of readability and can be easily understood by individuals aged 13 to 15. The scripts were recorded into videos by an English native speaker portraying the doctor and a Chinese native speaker acting as the patient.

The Microsoft Azure Speech Service API (Microsoft) was called to generate bilingual transcripts of the consultations, chosen for its relatively high accuracy and low latency. No domain customisation was applied. The word error rates of the four scripts ranged from approximately 15% to 20%.

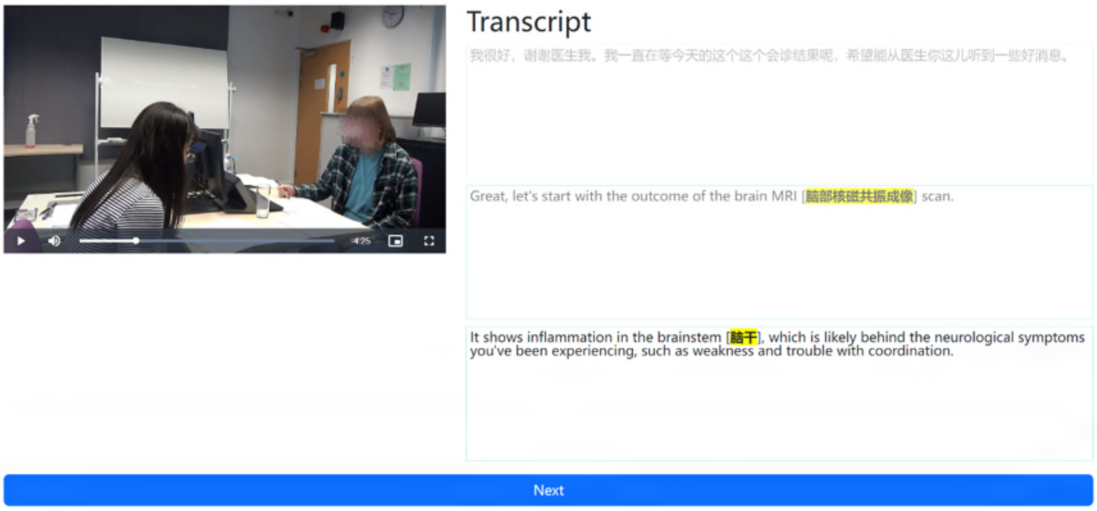
	Word count (words)	**Duration (minutes)**	**Speed (wpm)**	**Flesch reading ease index**	**Word Error Rate (English utterances)**
Script 1	756	6'07''	124	70.2	19.80%
Script 2	772	5'28''	141	66.1	14.62%
Script 3	742	5'33''	134	62.0	17.04%

Script 4	779	5’55’’	132	61.3	19.68%

Table 2. Difficulty control and word error rate of the scripts

Apparatus

To provide different types of ASR output, this study opted not to use CAI tools with ASR functions or platforms with captioning features. An interface (Figure 1) was designed by the research team for interpreters to carry out video remote interpreting tasks under various conditions. In the top left, a video player is displayed, while the right side features an ASR section with three text boxes. The current utterance appears in the bottom text box and gradually moves up as the next utterance is transcribed. The transcript related to an utterance was programmed to automatically appear immediately upon completion of the utterance. At that point the video was automatically paused to enable the interpreter to interpret. A blue “Next” button at the bottom allows the interpreter to listen to the next utterance either by pressing the spacebar on the keyboard or by clicking the mouse.



Transcript

我很好，谢谢医生我。我一直在等今天的这个这个会诊结果呢，希望能从医生你这儿听到一些好消息。

Great, let's start with the outcome of the brain MRI [脑部核磁共振成像] scan.

It shows inflammation in the brainstem [脑干], which is likely behind the neurological symptoms you've been experiencing, such as weakness and trouble with coordination.

Figure 1. Interface for remote video interpreting with ASR

Experiment design

To test how various types of ASR output affect interpreting quality, a baseline condition without ASR support was first devised, followed by three conditions with different types of ASR support. The conditions are as follows (Figure 2):

Condition 1: interpreting without ASR support

Condition 2: interpreting with partial ASR support (including the transcription of specialised terms and numbers and their translations)

Condition 3: interpreting with full ASR support (including the transcription of entire dialogue with the translations of numbers and specialised terms)

Condition 4: interpreting with ASR-fed ChatGPT summary (including a bullet-point summary with the translations of numbers and specialised terms). In this condition we provided ChatGPT with the following prompt to generate summaries:

There is the output of an ASR system which transcribed a doctor-patient conversation.

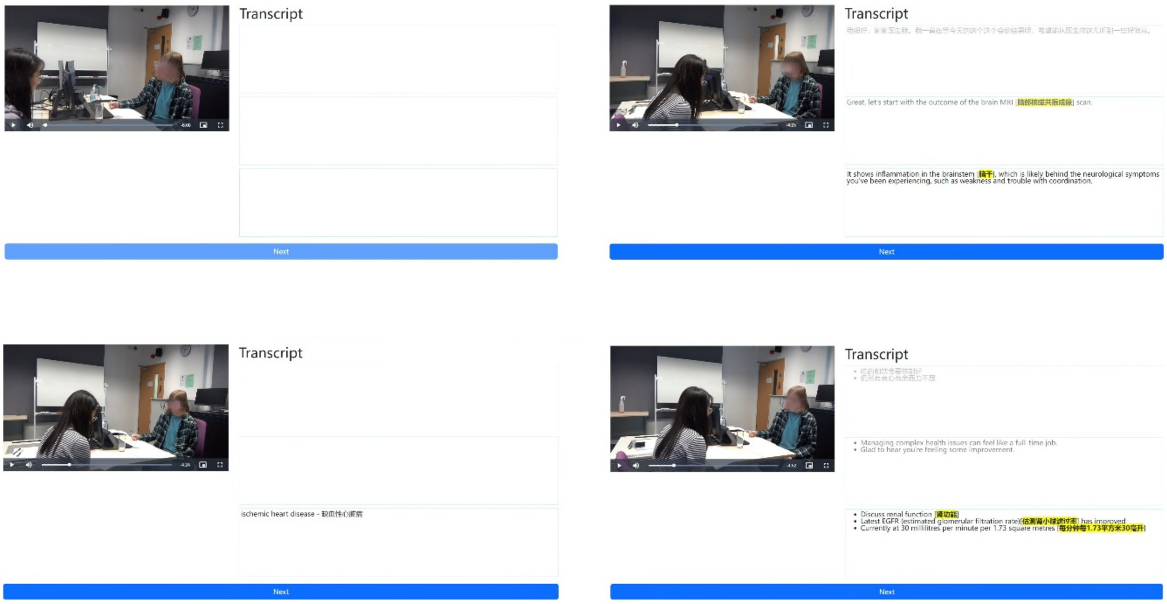
The output is used by an interpreter to support their interpretation.

Shorten the output to about half length making sure that the important information is kept.

Make sure you keep the important information.

This short version will be shown to the interpreter to help them interpret the conversation.

Present the output using bullet points.



我很好，谢谢医生我。我一直在等今天的这个这个会诊结果呢，希望能从医生你这儿听到一些好消息。

Great, let's start with the outcome of the brain MRI [脑部核磁共振成像] scan.

It shows inflammation in the brainstem [脑干], which is likely behind the neurological symptoms you've been experiencing, such as weakness and trouble with coordination.

ischemic heart disease – 缺血性心脏病

- 吃药和饮食要控制好
- 仍然有点心有余而力不足

- Managing complex health issues can feel like a full-time job.
- Glad to hear you're feeling some improvement.

- Discuss renal function [肾功能]
- Latest EGFR (estimated glomerular filtration rate)[估测肾小球滤过率] has improved
- Currently at 30 millilitres per minute per 1.73 square metres [每分钟每1.73平方米30毫升]

Figure 2. Interface without ASR support (top left), interface with partial ASR support (bottom left), interface with full ASR support (top right) and interface with ASR-fed ChatGPT summary (bottom right)

	Task 1	Task 2	Task 3	Task 4
Participant 1	Condition 1	Condition 2	Condition 3	Condition 4
Participant 2	Condition 2	Condition 3	Condition 4	Condition 1
Participant 3	Condition 3	Condition 4	Condition 1	Condition 2
Participant 4	Condition 4	Condition 1	Condition 2	Condition 3

Table 3. Randomised experiment design

Table 3 shows a randomised Latin square experiment design to ensure that each interpreter experiences every condition equally.

Procedure

This pilot study was conducted from 5th to 9th August 2024. Before the experiment, all participants completed a pre-experiment questionnaire concerning demographic information, English proficiency, interpreting experience, prior experience with using ASR for interpreting tasks and vision conditions via Qualtrics.

On the experiment day, participants came to our interpreting lab. They were first briefed on the theme of the consultations and the names of the interlocutors. To closely simulate authentic medical interpreting, they were allotted only 15 minutes to prepare. Before each task, they received a 9-point eye calibration to ensure accurate tracking of eye movements using EyeLink 1000 Plus eye tracker (SR Research). The eye tracker was employed to measure interpreters' eye movement behaviours indicative of cognitive effort across different conditions, another key focus of our study. They filled out the NASA Task Load Index (Hart and Staveland, 1988) immediately after each task, to self-assess the workload they perceived using the NASA TLX iOS app. The eye-tracking data and self-assessment results will be analysed and reported in a future study. Each interpreting task lasted around 15 minutes on average. Their interpreting output was audio-recorded and transcribed verbatim for data analysis.

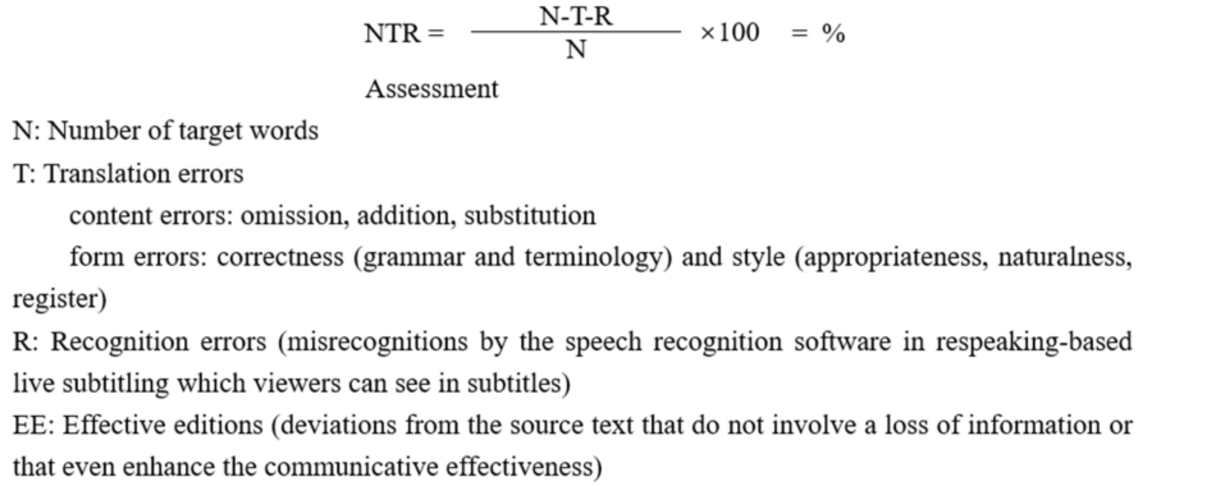
A retrospective session started after the completion of all tasks and a short break. Participants verbally reported their interactions with ASR in the recently performed interpreting tasks with some cues provided in a paper. The duration of this session varied between 7 and 19 minutes. In addition, participants partook in a semi-structured interview to share their overall attitudes towards using ASR in healthcare interpreting and offer suggestions for improvement. This part lasted between 19 to 49 minutes. The last two sessions were both audio-recorded and transcribed verbatim.

Data analysis

The interpreting quality was analysed using an adapted version of the NTR model (Romero-Fresco and Pöchhacker, 2017), an error-based framework originally developed to evaluate accuracy in interlingual subtitling. It has recently been adapted for assessing interpreting quality (e.g., Korybski *et al.*, 2022; Rodríguez González *et al.*, 2023) for its identification of translation errors.

The original NTR model consists of a formula and an overall assessment (Figure 3). In interpreter-mediated conversations, interlocutors usually only hear the interpreter's output. Recognition errors, therefore, did not apply to the interpreting workflow in this study and were not considered when using this formula. The formula calculates the accuracy rate, while the overall assessment comprises the accuracy rate, comments on issues not covered by the formula, such as effective editions, the speed, delay and overall flow of the interpreting output, and a final conclusion (Romero-Fresco and Pöchhacker, 2017, p.159). Ultimately, it is the overall assessment that indicates the quality.

The NTR model adopts a three-level grading system to classify errors by severity: “minor errors”, “major errors” and “critical errors”, deducting 0.25, 0.5 and 1 points respectively. As a meaning-focused model, it evaluates errors based on the “idea unit”, defined by Chafe (1985) as a “unit of intonational and semantic closure”, which typically encompasses a verb phrase along with a noun, prepositional or adverbial phrase. Minor errors cause largely insignificant deviations, major errors often result in isolated information loss, and critical errors produce an utterance with an entirely new meaning (Romero-Fresco and Pöchhacker, 2017, p.152).



$$\text{NTR} = \frac{\text{N-T-R}}{\text{N}} \times 100 = \%$$

Assessment

N: Number of target words

T: Translation errors

content errors: omission, addition, substitution

form errors: correctness (grammar and terminology) and style (appropriateness, naturalness, register)

R: Recognition errors (misrecognitions by the speech recognition software in respeaking-based live subtitling which viewers can see in subtitles)

EE: Effective editions (deviations from the source text that do not involve a loss of information or that even enhance the communicative effectiveness)

Figure 3. The NTR model (Romero-Fresco and Pöchhacker, 2017)

To conduct the assessment, all 16 interpreting outputs were transcribed verbatim into text, segmented into idea units and manually aligned with the source material in the NTR sheets. Although all participants performed bidirectional interpreting tasks, our current analysis addressed only the quality of English-to-Chinese interpreting. This focus was driven by the greater complexity of the doctor’s utterances, which often included technical terms and numbers, coupled with the listening challenge of interpreting from a second language. To ensure evaluation consistency and reduce rating subjectivity, each output was analysed by two evaluators who received training on NTR evaluation before carrying out the task. When discrepancies arose, the evaluators engaged in discussions to reach an agreement.

Participants’ retrospective reports were analysed to answer the third research question. Specifically, participants’ reflections on the specific types of information they sought from ASR support, the way they used ASR and their preferences for the presentation of ASR output were examined.

4 Results

Table 4 presents the results of the interpreting quality assessment. Compared with the baseline condition (no ASR support), the average scores of interpreting quality under the three ASR-supported conditions all increased, by 0.52, 2.2 and 2.12 points respectively. Among the three conditions, interpreting with full ASR support yielded the highest mean score, while interpreting with partial ASR support had the lowest mean. The difference in interpreting quality between using full ASR transcripts and ChatGPT summaries was very small.

The breakdown of each participant's scores revealed that for three participants, their interpreting quality improved with ASR regardless of the types of ASR output and scored highest with full ASR support. Only Participant 4 had the lowest score with partial ASR support. The interpreting quality of Participant 1 and Participant 2 improved steadily as the amount of source text provided increased from partial ASR support to ASR-fed ChatGPT summaries, and finally to full ASR transcripts. In contrast, Participant 3 had her best performance with the ChatGPT summary.

Despite the limited sample size, an attempt was made to address the first two research questions by running inferential statistical tests using IBM SPSS Statistics (Version 26). Shapiro-Wilk test (Elliott and Woodward, 2007, p.25) confirmed that all data under each condition conformed to a normal distribution. One-way repeated measures ANOVA tests (Elliott and Woodward, 2007, p.175) were administered to all conditions. Mauchly's test of sphericity (Keppel and Wickens, 2004, p.376) was not violated (p = .419 >.05). A significant difference was found in interpreting quality across the four conditions, F (3, 9) = 48.271, p = .000<.01, partial η^2 = .942. Therefore, post-hoc pairwise comparisons using Bonferroni correction (α = 0.05/6 = .0083) (Elliott and Woodward, 2007, p.9) were conducted to find which pairs were significantly different (Table 5).

	C1 (no ASR)	**C2 (partial ASR)**	**C3 (full ASR)**	**C4 (ASR-fed ChatGPT summary)**
P1	96.49	96.88	98.98	98.78
P2	96.63	97.15	98.44	98.29
P3	95.83	97.11	98.52	98.76
P4	97.45	97.34	99.27	99.03
Mean	96.60	97.12	98.80	98.72
Range	95.83-97.45	96.88-97.34	98.44-99.27	98.29-99.03
Standard Deviation	.665	.189	.392	.309

Table 4. Descriptive statistics for interpreting quality per participant under each condition

Pairwise comparison (by condition)	**Mean difference**	**Standard error**	**Sig.**
1 vs 2	-.520	.287	.168
1 vs 3	-2.203	.227	.002*
1 vs 4	-2.115	.315	.007*
2 vs 3	-1.682	.197	.003*
2 vs 4	-1.595	.161	.002*
3 vs 4	.087	.111	.487

Note: * for p<.05

Table 5. Results of post-hoc comparisons (repeated measures ANOVA)

To answer the first research question, the results revealed that the interpreting quality in the full ASR transcript condition (*M* = 98.80, *SD* = .392) and the ASR-fed ChatGPT summary condition (*M* = 98.72, *SD* = .309) was significantly higher than that in the condition without ASR (*M* = 96.60, *SD* = .665). However, no significant difference was observed between the partial ASR condition (*M* = 97.12, *SD* = .189) and the no ASR condition.

To answer the second research question, compared to the condition with partial ASR support, the interpreting quality was significantly higher with full ASR transcript and ASR-fed ChatGPT summary. There was no significant difference in the interpreting quality between the full ASR condition and the ChatGPT summary condition.

It should be noted that the power analysis showed a low statistical power of .141 (Elliott and Woodward, 2007, p.8), suggesting a limited capacity to detect true effects within the current sample. The inferential results may not be reliable due to insufficient power and therefore, should be interpreted with caution. However, a large effect size (*f* = .728) (Kelley and Preacher, 2012, p.147) was yielded by the sensitivity analysis using G*Power, indicating the substantial differences in interpreting quality across the conditions and the practical significance of the findings despite the low power.

As the NTR model allows us to access specific error types, the distribution of error types across various conditions (Figure 4) and per participant (Figure 5) was also examined.

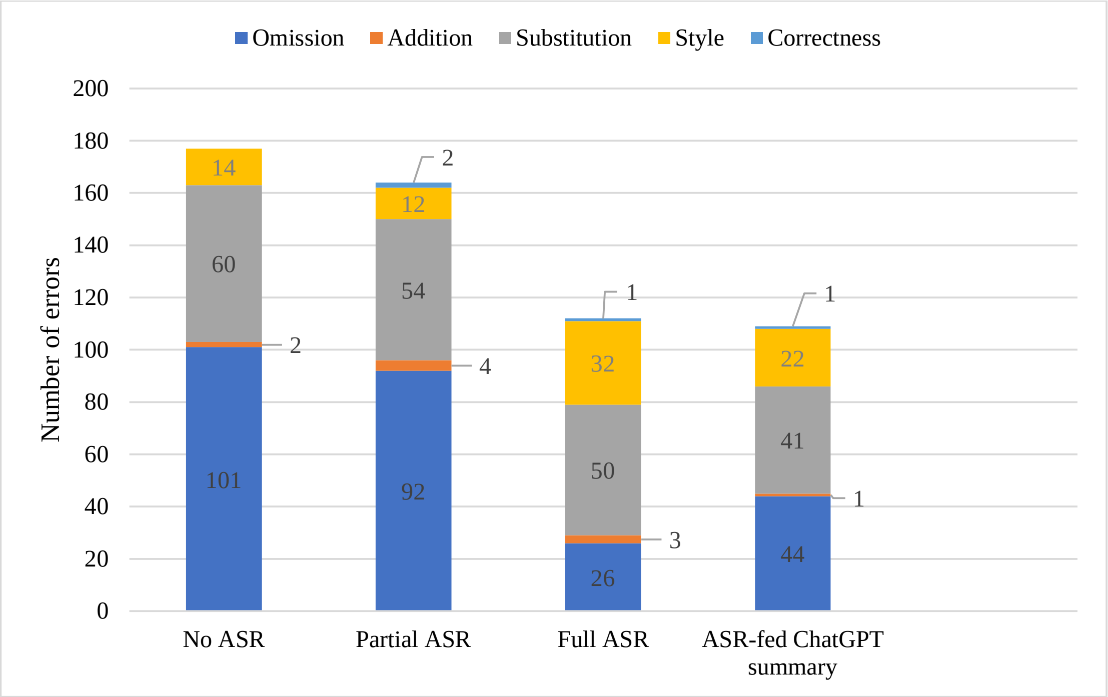


Figure 4. Error type distribution across conditions

Compared to the condition without ASR, the full ASR condition achieved the biggest reduction in omission errors, decreasing by 74.26%, followed by a 56.44% drop in the ChatGPT summary condition. However, the full ASR condition showed the highest increase in style errors, with the ChatGPT summary condition close behind. The ChatGPT summary condition witnessed the largest reduction in substitution errors, by 31.67%. The error type distribution between the no ASR condition and the partial ASR condition showed only modest differences.

The error distribution per participant under each condition is visualised in Figure 5. It shows that all participants made the fewest omission errors when assisted by full ASR transcripts, followed by ChatGPT summaries. Participant 2 was a major contributor to style errors observed in the full ASR condition (16 of 32 errors) and the ChatGPT summary condition (9 of 22 errors). This highlighted the need for a deeper examination of the errors made by individuals.

The third research question pertained to participants' interaction with ASR technology. Three participants shared that they relied on ASR support for medical terms, medicine names, dosages and units. One participant noted that she used ASR when she had difficulties understanding the interlocutor. When asked whether they used ASR support consistently throughout the task or only as needed, three participants believed that they primarily counted on their own listening skills and comprehension abilities, only resorting to the transcript when they struggled to understand the original utterance. One mentioned that after finishing interpreting an utterance, she occasionally reviewed the transcript to verify the accuracy of her delivery. Conversely, one admitted constantly using the provided transcript during interpreting and also expressed that the condition with only terms and numbers was distracting. However, preliminary analysis of interpreting errors (see Appendix A) and eye tracking data suggests that these participants frequently referred to ASR output during interpreting, evidenced by the reproduction of ASR errors in their interpreting output and patterns observed in eye fixation positions.

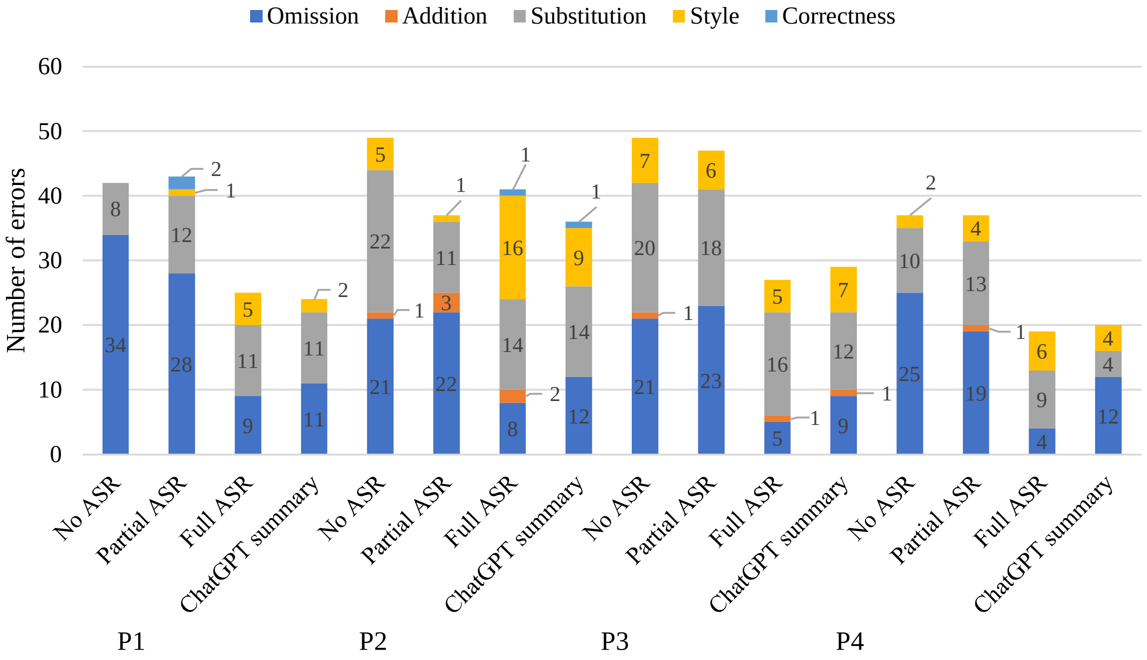


Figure 5. Error type distribution per participant

All participants expressed a preference for full ASR support if given the option. One explained that the information not provided by the partial ASR and ASR-fed ChatGPT summary conditions could be exactly what an interpreter might miss. Another participant mentioned that when full ASR output was available, she could allocate less effort to listening and just turn the task from interpreting into sight translation. Also, one would prefer to have the entire ASR transcript if the domain was foreign to her. However, she cautioned that in cases of familiar topics, the availability of a full ASR transcript could potentially interfere with interpreting by making interpreters overly dependent on it.

5 Discussion

The descriptive statistics indicated an increase in overall interpreting quality with varying types of ASR support. The integration of ASR into dialogue-based remote healthcare interpreting may have a positive impact on interpreting quality. Through inferential statistical analysis, we found that when the full ASR transcript or ASR-fed ChatGPT summary was provided, interpreting quality was enhanced significantly. This initial result was consistent with previous findings on the impact of captions on SI quality among student interpreters (e.g., Cheung and Li, 2022; Yuan and Wang, 2024) and professional interpreters (e.g., Rodríguez González *et al.*, 2023; Li and Chmiel, 2024). However, there was no significant difference in interpreting quality between the condition without ASR support and the partial ASR condition (numbers and terms). This contradicted previous research that revealed significant improvements in the accuracy of number renditions with numbers being provided (e.g., Desmet *et al.*, 2018; Defrancq and Fantinuoli, 2021; Pisani and Fantinuoli, 2021; Yuan and Wang, 2023). One explanation for this discrepancy is that previous studies primarily assessed the quality of numbers or terms only. Although our finding was embedded in the context of dialogue interpreting, it echoed Fritella's (2022) view that an ASR-integrated CAI tool may not necessarily facilitate number renditions in SI unless critical variables in the test speech were considered, for instance, the complexity of the speech, and a holistic assessment approach rather than concentration on the isolated numeral.

When examining the quality difference across varying ASR output conditions, the descriptive results showed that both the full ASR support and the ASR-fed ChatGPT summary were associated with a substantial improvement in interpreting quality, with only minimal differences detected between them. The provision of transcripts with only terms and numbers led to the smallest increase in mean interpreting quality scores. Inferential statistics suggested that interpreting quality was notably lower under the partial ASR condition compared to other ASR conditions, while no statistical difference in interpreting quality was found between the full ASR and the ChatGPT summary conditions. This initial finding implied that both full ASR transcripts and ChatGPT summaries were effective in improving interpreting quality. However, it did not suggest that partial ASR support should be ruled out when offering ASR solutions to interpreters. This result may be attributed to participants' limited experience in integrating ASR into healthcare interpreting. Professional interpreters may benefit from this type of ASR output as it may serve as a solution for "problem triggers" and reduce superfluous distraction. Given the lack of previous studies comparing the impact of varying types of ASR output, this finding provided a starting point for exploring the optimal ASR presentation that benefits interpreters most.

When taking a closer look at the error type distribution, it was found that the availability of full ASR transcripts helped all participants deliver more complete renditions by effectively addressing omission errors. Although nearly half of the stylistic errors in both the full ASR and the ChatGPT summary conditions were attributable to one participant, the high frequency of style-related errors in both conditions may suggest an association between these ASR outputs and less satisfactory stylistic appropriateness in interpreting. This issue was mainly manifested as disfluency phenomena, including filled fillers and silent pauses in participants' delivery which was consistent with Rodríguez González *et al.*'s (2023) findings in ASR-supported remote SI.

Regarding the interaction between interpreters and the technology, participants reported how they applied ASR output to the tasks and their preferences for ASR output presentation. They reported selectively using ASR support, mainly for specialised terms and comprehension issues. Three participants believed that they relied more on themselves than ASR. All participants preferred having access to a full ASR transcript, mentioning benefits such as improved completeness of information delivery, reduction of listening and analysis demands, and assistance with unfamiliar topics. However, as these results were based on self-reports from a sample of four, they should be interpreted with prudence.

6 Conclusions and future plans

In this pilot study, we explored the impact of ASR on remote dialogue interpreting in healthcare settings. The preliminary findings suggested that the availability of full ASR transcripts or ASR-fed ChatGPT summaries improved interpreting quality. However, access to transcripts of numbers and terms did not contribute to better interpreting quality. Participants' self-reported interactions with ASR were generally consistent, including the selective use of ASR output and a preference for full ASR transcripts.

These findings should be treated with caution, as they were exploratory and based on a very limited sample size. A number of limitations should also be acknowledged. First, the findings only reflected trainee interpreters' performance and experience with ASR support. They cannot be generalised to professional interpreters who will be the focus in our main study. Second, to ensure accurate eye tracking, note-taking was prohibited during the experiments. This may have had an impact on the interpreting quality, especially in the condition where no ASR was provided. To minimise the impact of confounding variables, common turn-taking issues, such as overlapping speech and interruptions in interpreter-mediated conversations were avoided in the simulated consultations. Moreover, the video remote interpreting interface with ASR support designed by our research team may be less familiar to the participants than those common commercial platforms with captioning features. These factors posed a risk of reducing the study's ecological validity. Third, as this study only evaluated English-to-Chinese interpreting output, these findings may not fully represent the quality of the entire bidirectional dialogue interpreting. Finally, given the observed limitations in applying the NTR model to assess healthcare interpreting quality, modifications may be necessary to adapt the model more effectively to this interpreting scenario.

To conclude, this pilot study successfully validated the methodology. In the next phase of our study, we will analyse the eye-tracking data to investigate how participants allocated their cognitive effort when different types of ASR output were provided. A detailed comparison between ASR transcription errors and interpreting errors will be performed to explore how they may have used ASR support. We will also refine our research design and conduct the main study with a larger sample of interpreters who are experienced in healthcare interpreting.

References

Braun, Sabine. 2019. Technology and interpreting. In O'Hagan Minako (Ed.). *The Routledge handbook of translation and technology*. Routledge, pages 271-288.

Chen, Sijia, 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: 399-420.

Cheung, Andrew. K.F., and Tianyun Li. 2022. Machine aided interpreting: An experiment of automatic speech recognition in simultaneous interpreting. *Translation Quarterly*, 104: 1-20.

Defrancq, Bart, and Claudio Fantinuoli. 2021. Automatic speech recognition in the booth: Assessment of system performance, interpreters' performances and interactions in the context of numbers. *Target*, 33: 73-102.

Desmet, Bart, Mieke Vandierendonck, and Bart Defrancq. 2018. Simultaneous interpretation of numbers and the impact of technological support. In Claudio Fantinuoli (Ed.). *Interpreting and technology*. Language Science Press, pages 13-27.

Elliott, Alan C., and Wayne A. Woodward. 2007. *Statistical analysis quick reference guidebook: With SPSS examples*. Sage Publications.

Fantinuoli, Claudio, and Maddalena Montecchio. 2022. Defining maximum acceptable latency of AI-enhanced CAI tools. *arXiv preprint arXiv*:2201.02792.

Fantinuoli, Claudio, Giulia Marchesini, David Landan, and Lukas Horak. 2022. Kudo interpreter assist: Automated real-time support for remote interpretation. *arXiv preprint arXiv*:2201.01800.

Fantinuoli, Claudio. 2017. Speech recognition in the interpreter workstation. In *Proceedings of the Translating and the Computer 39*, pages 25-34.

Frittella, Francesca Maria. 2022. CAI tool-supported SI of numbers: A theoretical and methodological contribution. *International Journal of Interpreter Education,* 14: 32-56.

González, Eloy Rodríguez, Muhammad Ahmed Saeed, Tomasz Korybski, Elena Davitti, and Sabine Braun. 2023. Assessing the impact of automatic speech recognition on remote simultaneous interpreting performance using the NTR Model. In *Proceedings of the International Workshop on Interpreting Technologies - SAY IT AGAIN 2023*, pages 1-8.

Hart, Sandra G., and Lowell. E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati (Eds.). *Human mental workload*. North-Holland, pages 139-183.

Jurafsky, Daniel, and James H. Martin. 2009. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall, 2nd edition.

Kelley, Ken, and Kristopher J. Preacher. 2012. On effect size. *Psychological methods,* 17: 137-152.

Li, Tianyun, and Agnieszka Chmiel. 2024. Automatic subtitles increase accuracy and decrease cognitive load in simultaneous interpreting. *Interpreting,* 26: 253-281.

Romero-Fresco, Pablo, and Franz Pöchhacker. 2017. Quality assessment in interlingual live subtitling: The NTR Model. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, 16: 149-167.

Seeber, Kilian G. 2011. Cognitive load in simultaneous interpreting: Existing theories—new models. *Interpreting*, 13: 176-204.

Wang, Xinyu, and Caiwen Wang. 2019. Can computer-assisted interpreting tools assist interpreting? *Transletters. International Journal of Translation and Interpreting*, 3: 109-139.

Wickens, Thomas D., and Geoffrey Keppel. 2004. *Design and analysis: A researcher's handbook*. Upper Saddle River, NJ: Pearson Prentice-Hall.

Yuan, Lu, 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.

Yuan, Lu, and Binhua Wang. 2024. Eye-tracking the processing of visual input of live transcripts in remote simultaneous interpreting: Preliminary findings. *FORUM*, 22: 118-144.

Appendix A: Examples of Interpreting Errors

Condition 1: No ASR support

Original script	ASR output	Interpreting output	Back translation
Excellent, that's important for managing stomach acidity effectively.	N/A	很好，这非常重要。	Great, that's very important.
Error type(s): omission error **Analysis: The interpreter left out the reference to stomach acidity management.**			

Condition 2: Partial ASR support

Original script	ASR output	Interpreting output	Back translation
And your serum creatinine has decreased to 1.6. Your urea levels are also better at 50 down from 70 last check.	serum creatinine - 血清肌酐 urea levels - 尿素水平 1.6 - 1.6 50 - 50 70 - 70	还有这个血清肌酐以及呃这个尿素水平也是…是 1.6, 50 和 70 这个数据。	And your serum creatinine and um urea levels are also…are 1.6, 50 and 70.
Error type(s): omission error **Analysis: The interpreter omitted phrases indicating changes and instead only stated the numbers.**			

Condition 3: Full ASR support

Original script	ASR output	Interpreting output	Back translation
To reduce the inflammation, you will start with a high dose of corticosteroids, specifically Prednisone at 60 mg daily.	To reduce the inflammation, you will start with a high dose of corticosteroids [皮质类固醇], specifically Prednisone [泼尼松] at 60 minutes daily.	要想减少炎症，首先呃你要服用高剂量的皮类…皮质类固醇，尤呃尤其是泼尼松，大概每天 60 分钟。	To reduce the inflammation, first um you will start with a high dose of cor…corticosteroids, esp um especially Prednisone, for about 60 minutes per day.
Error type(s): substitution error **Analysis: The interpreter followed ASR's transcription error.**			
I can see that you don't have oedema, your chest is clear, and your abdomen is soft and not tender.	I can see that you don't have. Your chest is clear and your abdomen is soft and not tender [一碰就痛].	呃我…我看到了你没有过敏原。你的…呃胸部是…胸腔是很干净的，你的腹部很柔软，并没有一碰就痛。	Um I…I can see you don't have allergies. Your…um your breast is…chest is clear, and your abdomen is soft and not tender.
Error type(s): substitution error **Analysis: ASR omitted the term "oedema," which the interpreter then incorrectly substituted with "allergies."**			
From the test results, it looks like your diabetes has not been well controlled, which could be contributing to your symptoms.	From the test results, it looks like your diabetes has not been well controlled, which could be contributed to your symptoms.	呃根据检查结果，呃看起来你的糖尿病呃已经得到了很好的控制，呃这些可以导致你现在的这个状况…症状的。	Um from the test results, um it looks like your diabetes has been well controlled, um which could be contributing to your current situation... symptoms.

Error type(s): substitution and style errors **Analysis: The ASR transcription was correct, but the interpreter reversed the meaning and frequently used filled fillers like "um."**			
You also have a congestive heart failure and ischemic heart disease, which are affecting your circulation.	You also have a congested portfolio and ischemic heart disease [缺血性心脏病], which are affecting your circulation.	呃同时你有去做一些检查以及心脏的缺血…有缺血性心脏病，这些也有可能会影响到你的循环。	Um at the same time you did some checks and the ischemia in the heart … there is ischemic heart disease, which might also affect your circulation.
Error type(s): addition and omission errors **Analysis: The interpreter added, "Um at the same time you did some checks." ASR mis-transcribed "a congestive failure" as "a congestive portfolio," but the interpreter did not follow this error and instead omitted it.**			

Condition 4: ASR-fed ChatGPT summary

Original script	ASR output	Interpreting output	Back translation
To reduce the inflammation, you will start with a high dose of corticosteroids, specifically Prednisone at 60 mg daily.	• To reduce inflammation, start with a high dose of corticosteroids [皮质类固醇], Prednisone [泼尼松] at 60 mg daily [每天 60 毫克].	我们呃现在是需要呃解决你，减少你这个炎症，我们是需要用到类固醇，呃还有泼尼松，每天需要有 60 毫克。	We um currently need to um solve you, reduce your inflammation. We will use steroids, um also Prednisone at 60 mg daily.
Error type(s): omission error **Reason: The ChatGPT summary corrected the ASR transcription error by changing "60 minutes" to "60 mg," which helped avoid any misunderstanding for the interpreter. Here is only a minor omission of the phrase "a high dose of."**			
Yes, your test results indicate you have anaemia, diabetic nephropathy, hypertension, a history of myocardial infarction and stroke.	• Test results indicate: Anaemia [贫血] Diabetic nephropathy [糖尿病肾病] Hypertension [高血压] History of myocardial infarction [心肌梗死] Stroke	嗯，[哎]…是的，你的检查结果显示你有贫血，还有这个糖尿病，你高血压、心肌梗死、中风。	Um, [sigh]…yes, your test results indicate you have anaemia, also diabetes, you hypertension, myocardial infarction and stroke.
Error type(s): substitution and style errors **Reason: The interpreter incorrectly substituted "diabetic nephropathy" with "diabetes." The style error was noted due to the interpreter's use of fillers like "um" and even a sigh, along with the listing of conditions by repeating the ChatGPT summary verbatim without any transitional words, as in "you hypertension … ."**			

Innovation Test Lab – an Innovation Hub at the European Parliament's Directorate-General for Translation

Francesco Rossi

European Parliament

francesco.rossi@europarl.europa.eu

Abstract

In today's rapidly evolving digital landscape, advancements in translation technologies such as neural machine translation (NMT) and artificial intelligence (AI) are transforming the way organisations deliver language services. To harness these advancements effectively, the European Parliament's Directorate-General for Translation (DG TRAD) has established the Innovation Test Lab. The Innovation Test Lab is a virtual platform that provides a dedicated environment where emerging technologies can be rigorously tested and evaluated before integrating them into the organisation's operations. The goal is to drive efficiency, optimise workflows, and enhance service quality by systematically incorporating cutting-edge tools and processes. This paper will explore how the Innovation Test Lab operates, from the initial submission of candidate technologies to their assessment, testing, and final integration. It will highlight the participatory approach that ensures both top-down strategic direction and bottom-up contributions from users. Furthermore, the paper will examine the expected outcomes and long-term impact of this platform on DG TRAD's mission to maintain a leading position in translation services.

1 Introduction

The Innovation Test Lab is a highly sophisticated virtual environment that mirrors DG TRAD's existing IT infrastructure. This replication allows the testing of technologies in conditions that are identical to the real operational environment. One of the Lab's greatest strengths is its ability to offer accurate assessments of a new tool's performance once it is integrated into the complex ecosystem of DG TRAD, ensuring that any potential issues are identified and resolved early in the testing phase.

The Lab serves for the testing of several types of technologies. It functions as a comprehensive testing ground for a wide range of solutions from software tools aimed at enhancing translation quality to innovative platforms for project management or AI-driven translation processes. It also facilitates collaboration across DG TRAD, engaging different units in the exploration and validation of new tools.

Practical Implementations of the Test Lab

As said above, the Innovation Test Lab is a virtual environment that mirrors DG TRAD's existing IT infrastructure. The Test Lab is provided solely in the form of a virtual machine and therefore functions as a "virtual copy" of a DG TRAD computer, replicating exactly the conditions in which a user is accustomed to work and its technical specifications.

The use of a virtual environment has at least three evident advantages:

- a certain degree of flexibility in terms of delivery time, which is a particularly relevant feature when Lab requests come with tight time constraints;
- the possibility to create multiple instances of the Test Lab, enabling DG TRAD to provide several tens of machines in a timely manner and for as long as they are required;
- the easy creation and termination of each Test Lab, due to the Lab's virtual nature.

In general, DG TRAD's policy is to adopt a very rigorous approach: each creation of a machine (or a set of) must be focused on the specific purpose of testing. Once the test is completed and the results are documented, the machines are terminated and no data stored on them is kept.

Moreover, the establishment of the Test Lab marks a substantial advancement over prior methods, as no dedicated environment for testing innovative solutions previously existed in this form.

2 The Process: the Innovation Test Lab Workflow

Candidate technologies submitted to the Innovation Test Lab need to go through the workflow detailed below. DG TRAD can run tests on any kind of software in the Test Lab, the sole requirement being that they are in line with the business needs and respect the security regulations.

Technology Submission

The process begins with the submission of candidate technologies, which can occur through two distinct channels:

- Top-Down submissions: key business actors, such as senior management, the Business Analysis Cell (BA Cell), or the Applications & IT Systems Development (DAS) Unit may identify a specific technology that aligns with DG TRAD's strategic goals. These stakeholders bring the technology to the Strategy and Innovation Unit (SIU) and the Information Technology & IT Support (ITS) Unit for further evaluation;
- Bottom-Up submissions: DG TRAD fosters an inclusive approach by allowing all users to propose technologies. These submissions often arise from user participation in working groups, innovation platforms, or even informal channels.

Initial Assessment

Once a technology is submitted, the SIU performs an initial evaluation to assess its relevance and potential value to DG TRAD. At this stage, input from other actors, such as the DAS unit, ITS unit, or Euramis PreTranslation (PreTRAD)[1] Unit, may be requested to provide a more in-depth analysis of the technology's implications. If the technology requires a financial

[1] Euramis PreTranslation Unit: Unit in the Directorate for Technology of DG TRAD, whose main activities are: automatic document pre-treatment, maintenance of the Euramis interinstitutional databases for translation re-use, verification of the technical conformity of documents, as well as a number of supporting tasks such as helpdesk for Language Units (LUs) and external translation providers, contribution to the IT developments and to various training activities.

investment, the SIU, in collaboration with the requester and the DAS unit, assesses the budgetary requirements.

Depending on the complexity of the technology and the potential impact it may have on DG TRAD's operations, a dedicated forum may be convened to further discuss and evaluate the technology's viability. This collaborative approach ensures that all key stakeholders have an opportunity to provide input on the assessment.

Technical Workflow

After the initial assessment, a technical workflow is established to guide the testing phase. The SIU, in cooperation with the ITS unit and the requester, examines several key elements, including:

- the nature of the software (purpose, benefits, risks, other configurations and settings);
- the required number of machines, environments and users access;
- security, installation, integration and interoperability.

The ITS unit plays a crucial role in addressing any security concerns and ensuring that the technology can be safely integrated into DG TRAD's infrastructure. The SIU finalises the technical setup, preparing the Test Lab for the testing phase.

Test Execution

Once the technical environment is configured, the actual testing of the candidate technology begins. The requester takes the lead in conducting the tests, gathering data on the tool's performance, usability, and alignment with DG TRAD's needs. Depending on the kind of technology examined, the Innovation Test Lab evaluates various aspects, taking into consideration different and variable criteria. The results are documented and analysed[2], with input from other relevant units when necessary.

This phase is crucial as it provides insights into the feasibility, desirability, and added value of the technology. The results are reviewed by the SIU, which coordinates with other stakeholders to evaluate the tool's overall effectiveness.

Formal Recommendation and Integration

Once a technology passes the testing phase and receives positive evaluations, a formal recommendation is submitted to DG TRAD's IT governance structures. The governance bodies review the proposal, taking into account both the technical assessments and strategic objectives of DG TRAD. Upon approval, the ITS unit proceeds with the Software Selection Procedure, formally integrating the technology into the organisation's operations.

[2] results of the evaluations are for internal use only

3 A Systematic Approach to Innovation

The structured process of testing technologies in the Innovation Test Lab mirrors DG TRAD’s broader strategy for fostering innovation. Each new technology undergoes a thorough evaluation across multiple dimensions: technical feasibility, user experience, cost-benefit analysis, and long-term value. This approach ensures that DG TRAD remains agile and responsive to technological advancements while minimising risks associated with the adoption of new tools.

4 Embedding Innovation into DG TRAD’s Culture

One of the key goals of the Innovation Test Lab is to embed a culture of innovation within DG TRAD. By providing a platform for both top-down and bottom-up submissions, the Lab encourages all users to engage with new technologies and contributes to the organisation’s digital transformation. This participatory method empowers users to take ownership of innovation, driving creativity and fostering an openness to change.

The Lab not only facilitates the testing of technologies but also serves as a hub for ongoing collaboration and knowledge-sharing. Users, coordinators, and working groups are actively involved in the testing and evaluation processes, ensuring that a wide range of perspectives are considered when making decisions about technology adoption.

5 Expected Outcomes and Long-term Impact

The Innovation Test Lab is expected to deliver several key outcomes that will benefit DG TRAD in the long run. These include:

- **Efficiency gains**: by introducing technologies that streamline workflows and automate time-consuming processes;
- **Service improvements**: the Lab enables the identification and adoption of tools that enhance translation accuracy and quality, leading to better service delivery for DG TRAD’s clients;
- **Cost savings**: through rigorous testing and evaluation, the Lab ensures that only the most cost-effective and beneficial technologies are integrated. Moreover, by evaluating and testing a candidate technology in the Test Lab, we can enhance predictability and mitigate the risk of potential issues arising.

In addition to these tangible benefits, the Innovation Test Lab also helps DG TRAD cultivate an innovative mindset throughout the organisation. By embedding a systematic process for exploring, testing, and adopting new technologies, the Lab fosters a proactive approach to digital transformation and ensures that DG TRAD remains on the cutting edge of translation services.

6 Success Stories and Use Cases

The Test Lab was used to support the Proofs of Value (POVs) carried out within the Single Digital Workflow Tool project. This project focused on finding tools that could combine the different systems used by DG TRAD into one single solution to manage its complex workflow. The project manager set up five POVs and the Test Lab was configured to meet the needs of these tests. Various user profiles were created: test user profiles were assigned to specific testers, and a few administrator accounts were set up to oversee the process. The project ran smoothly, completing its evaluations quickly and moving on to the next stages.

In the near future, one of the possible uses for the Test Lab will be to test a new podcast script creator, which was provided earlier in 2024 by the Directorate-General for Innovation and Technological Support (DG ITEC) to DG TRAD. This tool is designed to help DG TRAD create Clear Language content, particularly for podcasts. DG TRAD produces podcasts in all 24 official languages of the European Union, as well as in Ukrainian, and the demand for tools that can support this work is steadily increasing.

In the medium-long term, the Test Lab could be used to evaluate the performance, accuracy and quality of AI-generated translations. This initiative is still at in the early stages and is expected to be followed-up in the coming months.

7 Conclusion

The Innovation Test Lab is a cornerstone of DG TRAD’s efforts to remain a leader in the translation industry. Its structured and participatory approach to testing and adopting new technologies ensures that DG TRAD can harness the latest innovations in neural machine translation, artificial intelligence, and other language technologies. By engaging users at all levels of the organisation and providing a realistic testing environment, the Lab not only drives technological advancement but also fosters an inclusive innovation culture that will support DG TRAD’s digital transformation for years to come.

References

Ammann, Paul, Offutt Jeff. (2008). Introduction to Software Testing.
Chauhan, Naresh. (2010). Software Testing: Principles and Practices.
Crispin, Lisa, Gregory Janet. (2009). Agile Testing: A Practical Guide for Testers and Agile Teams.
Dustin, Elfriede, Rashka Jeff, Paul John (1999). Automated Software Testing: Introduction, Management and Performance.
Humble, Jez, Farley, David (2010). Continuous Delivery: Reliable Software Releases through Build, Test, and Deployment Automation.
Kaner, Cem, Bach James, Pettichord Bret (2001). Lessons Learned in Software Testing.
Kaner, Cem, Falk Jack, Nguyen Hung Quoc. (1999). Testing Computer Software.
Lewis, William E. (2004). Software Testing and Continuous Quality Improvement.
Mathur, Aditya. P. (2008). Foundations of Software Testing. Continuous Delivery: Reliable Software Releases through Build, Test, and Deployment Automation.

Standards and Guidelines

IEEE Std 1008-1987: IEEE Standard for Software Unit Testing

ISO/IEC/IEEE 29119 - Software Testing Standards

ISTQB Advanced Level Syllabi

Machine translation (MT) quality evaluation: What do MT experts have to learn from translators and translation scholars?

Ting Liu 劉婷

National Research Council of Canada

School of Translation and Interpretation, University of Ottawa

tliu109@uottawa.ca

Rebecca Knowles

National Research Council of Canada

rebecca.knowles@nrc-cnrc.gc.ca

Chi-kiu Lo 羅致翹

National Research Council of Canada

chikiu.lo@nrc-cnrc.gc.ca

Elizabeth Marshman

School of Translation and Interpretation, University of Ottawa/OLST

elizabeth.marshman@uottawa.ca

Abstract

Evaluating the quality of machine translation (MT) output is critical as MT continues to gain traction in the translation industry. Unfortunately, MT evaluation is challenging due to the limitations of automatic metrics and the complexity of human assessments. In this position paper, we argue that Translation Studies (TS) scholars and language professionals have essential contributions to make towards the design and implementation of valid and useful MT human evaluation protocols. Drawing on the literature and our own experience, we firstly argue that a functionalist perspective on translation quality is the most appropriate option for evaluating MT. Secondly, we assert that providing a broad understanding of context will help to clarify the expectations of the translation and facilitate (and stabilize) human evaluation. Finally, unless the context entails other needs, we encourage the choice of professional translators or subject matter experts over bilinguals and crowd workers as evaluators, given the greater sensitivity they have been shown to display in assessment tasks. Given these recommendations, we argue that increased exchange and collaboration between the TS and MT scholarly communities (and the language industry) are necessary to ensure effective MT human evaluation in the future.

1 Introduction

Quality assessment/evaluation plays a pivotal role in both professional translation and machine translation (MT) contexts. In a professional context, quality evaluation ensures that translation products meet standards, informs employment decisions, and serves as a framework for criticism and discussion (Colina, 2015). It also assesses translator competence, facilitates self-monitoring and peer feedback, and contributes to ongoing professional development. In the context of MT, quality evaluation is crucial as it helps assess the accuracy and reliability of machine-generated translations and guides improvements in MT systems (Koehn, 2020). It is also essential to determine when and how to implement systems and whether human intervention is needed. Just as with human translation, these evaluations ensure MT output meets functional and contextual needs, informing decisions on its practical use. However, it seems we have entered a phase of exaggerated excitement, with some researchers claiming that MT has achieved human parity (Hassan et al., 2018). This claim quickly became the object of criticism, in large part due to challenges of defining translation quality and to methodological limitations that may bias results (e.g., Läubli et al., 2018; Toral et al., 2018; Krüger, 2022).

In this position paper, we explore how Translation Studies (TS) scholars and language professionals can contribute to improving MT evaluation protocols. Specifically, we seek to address challenges in human evaluation design for MT researchers and developers. Our central research question is: What do MT researchers have to learn from translators and translation scholars? To answer this, we propose a contextualized, functionalist perspective on translation quality as the most appropriate foundation for MT evaluation (inspired by Nord 1997/2014). We argue that providing evaluators with adequate contextual and co-textual information, alongside a functionalist definition of quality, may help clarify translation expectations and lead to more consistent evaluations.

MT serves various user groups, from professional translators engaging in post-editing tasks to other users who employ MT for day-to-day activities. Each of these groups has different expectations from MT output. Professional translators tend to have higher standards and expectations for MT quality, particularly when the translation serves functions like cultural preservation or information dissemination (Bowker, 2021). In contrast, non-professional users, or those operating in time- and cost-sensitive contexts (such as business environments), may be more inclined to accept translations that are less than perfect (Bowker, 2019). Given their demonstrated sensitivity in assessment tasks (Toral et al., 2018; Freitag et al., 2021), we recommend professional translators (rather than bilinguals or crowd workers) as evaluators, unless the context entails other needs. For example, in certain domains, the expertise of subject-matter experts may be essential to evaluate the acceptability of translations for a given purpose or audience. In some cases, combining expert evaluations with end-user feedback may provide valuable insights, revealing potential gaps between professional standards and user satisfaction. While this approach may increase evaluation costs and complexity, potential gains in evaluation reliability and precision may justify the investment, particularly when high-quality output is required or when claims of human parity are made.

Overall, this paper presents a philosophical framework with proposed practical applications; however, further research is needed for a full-scale practical implementation to ensure a pragmatic, user-centred approach to MT evaluation.

2 Background and context

The use of computers for translation dates to the early days of computing, with MT emerging as an early application of computer science to language (Nirenburg et al., 2003). Neural machine translation (NMT), now a leading technology in translation, outperforms earlier approaches through deep learning architectures that enhance both accuracy and fluency. Furthermore, the architectural underpinnings of NMT have also contributed to the development of large language models (LLMs), which are now widely used for various natural language tasks. As NMT and LLMs achieve impressive translation quality, rigorous evaluation methods are essential to accurately assess the reliability and precision of their output, especially in professional contexts where high standards are required.

2.1 MT quality evaluation

In research, evaluation campaigns like those at the Conference on Machine Translation (WMT) provide a platform for comparing MT systems through shared tasks (Kocmi et al., 2023). These campaigns are key to advancing the field by highlighting the strengths and weaknesses of various models, guiding further research and development (Koehn, 2020). On the industry side, language service providers (LSPs) are increasingly incorporating NMT into their workflows to enhance productivity, maintain competitive pricing, and manage large-scale projects (Lommel et al., 2014; Bowker, 2019). However, the challenge remains in ensuring that these productivity gains do not compromise translation quality.

2.2 MT quality evaluation methods

MT quality evaluation methods, including both automatic metrics and human assessments, are employed to monitor and maintain the desired level of translation quality (Koehn, 2020). An automatic metric is a quantitative method of evaluating machine translation without human intervention. This is commonly done using reference-based metrics, which evaluate translation quality of the MT output by comparing it to one or more human-generated reference translations. To address the lack of high-quality human reference translations, recent reference-free metrics estimate MT output quality directly against the source text. Earlier metrics, such as BLEU (Papineni et al., 2002), are mostly based on exact word matching, while some newer metrics, such as COMET (Rei et al., 2020), are machine learning-based and finetune massively multilingual language models with human evaluation data (see below) to predict translation quality scores. Every year, the WMT Metrics shared task (Freitag et al., 2023) assesses how well metrics correlate with human evaluators and explores the challenges of automatic MT evaluation metrics.

Human evaluation is considered the gold standard in MT quality evaluation (Kocmi et al., 2023; Freitag et al., 2023; Knowles and Lo, 2024). In Liu et al. (2024), we cluster human evaluation methodologies into three general categories: manual scoring, semi-automatic, and task-based approaches. Manual scoring involves an evaluator directly assigning a score or ranking to one or more systems using numerical scales. For example, evaluators may assign scores indicating the degree to which the translated text retains the original meaning (adequacy) and its grammatical correctness and natural form in the target language (fluency) (Koehn, 2020). Direct Assessment (DA) is commonly used for scoring MT quality, especially in WMT tasks since 2016 (Kocmi et al., 2024). Metrics like DA+SQM, which combine DA with Scalar Quality Metrics, aim to enhance annotator consistency (Kocmi et al., 2023). Semi-automatic approaches use human annotations, with automated systems generating MT scores from them. The Error Span Annotation (ESA) approach (Kocmi et al., 2024) blends the continuous rating approach of DA with the error severity marking of MQM, positioning it between manual and semi-automatic evaluation methods. Task-based evaluations require annotators to utilize MT output to complete specific tasks, such as template filling, question answering, or semantic parsing. These evaluations assess the utility of the translation by asking whether the MT output is sufficient for completing the task, rather than directly judging its quality. Scores are used to rank the effectiveness of different MT systems, providing valuable insight into which systems or system versions are the most promising for further development or deployment. In research, these rankings inform decisions about improving specific models, while in industry, they guide the selection of systems that can best meet real-world needs, such as productivity gains, cost reduction, or handling specialized content.

2.3 Challenges for automatic metrics

MT evaluation is far from a simple task (Koehn, 2020). Earlier automatic metrics, such as BLEU, while widely used due to their efficiency and cost, are increasingly recognized as limited in scope and often poorly correlated with real-world perceptions of MT's usefulness, as they tend to emphasize surface-level similarities rather than capturing deeper semantic accuracy or fluency (Freitag et al., 2021). Although newly proposed machine learning-based metrics correlate better with human evaluators in judging the quality of MT output, the performance of these metrics in low-resource languages and/or low-quality MT systems is usually not as good (Lo et al., 2023a) because they require a large volume of training data to build the underlying language representation. Recent automatic metrics also struggle to provide an intuitive interpretation to show how their scores correspond to different levels of translation quality (Lo et al., 2023b).

2.4 Challenges for human evaluation

While human-driven MT evaluation takes various forms, it faces the challenge of subjectivity. Inter-annotator variation arises from differences in how annotators interpret meaning preservation and linguistic quality, resulting in inconsistencies in scoring. This lack of standardization in evaluation criteria exacerbates the issue, as annotators often rely on personal judgment (Koehn, 2020). For example, adequacy-alone evaluations (such as those used in Hassan et al., 2018) are prone to subjectivity because annotators may apply different criteria for what constitutes an “adequate” translation, influenced by factors like their linguistic background, personal biases, or the specific context of the task. Error classification methods (e.g., MQM), while more systematic and detailed, still rely heavily on the judgment and experience of annotators when categorizing translation errors. In both cases, the variability in human judgment often leads to poor inter-annotator agreement (Specia et al., 2021; Al Sharou and Specia, 2022), undermining the reliability and comparability of the evaluation outcomes. These approaches face the common challenge of balancing subjective human interpretation with the need for consistent and objective evaluation standards.

Some other key challenges of human evaluation include their time-consuming and labour-intensive nature and the resulting substantial financial and resource costs, especially when expertise is needed to apply complex and rigorous error typology frameworks like MQM, which demands experts trained in its protocol and error classification (Gladkoff et al., 2022). Task-based evaluations are even more complex and costly to design and implement compared to manual scoring methods.

3 Insights from translators and translation scholars

Although it is impossible to comprehensively address all the issues in MT evaluation, we can draw upon insights from translation scholars and language professionals to make targeted improvements in specific areas. Building on both the literature and our own experience, we explore potential ways to address some of the challenges.

3.1 Proposal 1: Adopt a functionalist perspective on translation quality

Human MT evaluation is first challenged by the need to select an appropriate definition of quality. As Krüger (2022) highlighted, the definition of translation quality used in the context of MT assessment may bias evaluations towards either human or machine translation. A review of the TS literature reveals diverse perspectives on the concept of quality (Liu et al., 2024). Among these perspectives, in this paper, we argue that a functionalist view of translation quality offers valuable insights for MT research on quality evaluation.

Functionalism in TS, especially in practical contexts, emerged as a response to the limitations of equivalence-based approaches, emphasizing the purpose of translation (Nord, 2014). Reiss (1971) introduced the concept of a functional category in translation criticism, linking text types to translation methods. Reiss and Vermeer (cited in Nord 2014, 29) further developed this idea with Skopostheorie, which prioritized the intended purpose (Skopos) of the translation over strict adherence to the source text. Skopos refers to the overarching goal of the translation, which must be fulfilled through intentional actions by the translator. Vermeer (1989) explained that all actions, including translation, are goal-oriented and involve a choice between different methods based on the intended result. Here, the translation must serve its purpose within the context in which it is used, meaning the translator should ensure that the text functions appropriately for its target audience, i.e., is “fit-for-purpose” (Bowker, 2019).

In an ideal scenario, the translation brief from the client outlines the Skopos and directs the translation process. The brief should include details regarding the intended functions of the text, the target audience, the anticipated time and place of reception, the medium through which the text will be delivered, and the purpose behind its creation or reception (Nord, 1997/2014). The brief helps the translator make informed decisions about how to approach the text, guiding choices around tone, style, and content adaptation. It also clarifies the desired outcome of the translation, ensuring that it serves its intended function effectively, whether that be to inform, persuade, entertain, or fulfill another purpose in the target context.

However, in MT, the concept of agency and intention becomes less clear. In specialized MT or production settings, the client assumes the role of the actor, directing intentional actions to align the MT output with the desired objectives. In shared tasks like those in WMT, the task description functions as a form of (often less detailed) brief, serving a similar role in outlining the expected outcomes for the MT system.

From this functionalist perspective, prioritizing purpose over strict linguistic equivalence, evaluation of translations should also be guided by a description of the task’s objectives. Quality should be assessed by how effectively the translation fulfills its intended function in the target context. This does not exclude the possibility that strict linguistic equivalence may be called for by the specific purpose of a translation, but acknowledges that it is not always the optimal strategy.

However, the concepts of purpose and function in translation remain underexplored in much of the existing MT literature. This gap has been identified as a critical area for growth, particularly in human-centered MT evaluation (Liebling et al., 2022). To address this gap, incorporating functionalist principles into MT evaluation, especially to define clear, purpose-driven goals for systems, offers a valuable approach. This involves not only focusing on the linguistic adequacy of translation quality but also aligning with reasonable expectations and identifying the intended users.

MT human evaluation often focuses on adequacy and fluency[1] (Koehn 2020). Adequacy is understood as the degree to which the translated text conveys the original meaning. However, from a functionalist perspective, adequacy could consider how well the translation fulfills the purpose outlined in the translation brief. This broader view links adequacy to the translation's Skopos, expanding beyond strict meaning transfer. To adopt this functionalist approach, we suggest evaluating adequacy through components tied to the translation's specific function. For example, in legal contexts, adequacy must not only ensure accurate meaning transfer but also uphold legal obligations and enforceability.

By broadening the definition of adequacy to include functional considerations, we aim to provide a more consistent framework for annotators, which would help mitigate the subjectivity that arises from varying interpretations of meaning preservation. Instead of focusing solely on meaning transfer, annotators would assess whether the translation meets the purpose of the text in its specific domain (e.g., legal, medical, creative). This fit-for-purpose approach ensures that MT evaluation reflects the practical needs of different types of texts, leading to more consistent and purpose-driven assessments.

3.2 Proposal 2: Pay attention to a broad understanding of context

Assessing function requires careful consideration of the context in which the translation will be used. Here, we will first establish a broad understanding of context, and then discuss contextual analysis, with a focus on situational context. Finally, we will explore the value of using an evaluation brief (Liu et al., 2024) with a functional-componential approach (Colina, 2008) to analyse contextual factors.

In TS, context plays a crucial role in ensuring that meaning is accurately conveyed. However, defining context in translation practice has been a longstanding challenge. As translation tasks became more complex—particularly in commercial and government sectors—it became clear that a broader, more nuanced understanding of context was necessary to guide translators (Melby and Foster, 2010). From the functionalist perspective, this context is closely tied to the translation brief. Melby and Foster (2010) split context into five essential components: co-text, chron-text, rel-text, bi-text, and non-text. These factors help translators understand the source text and produce an appropriate target text.

[1] For a discussion of fluency, see Section 3.2.

Castilho and Knowles (2024) used this framework to emphasize the growing need for MT researchers to expand their understanding of context. Typically, when MT researchers discuss adding context, they are referring to within-document context, i.e., shifting from sentence-level translation to paragraph- or document-level. However, context could also include other related documents, real-world knowledge, and audience expectations. These factors help to improve consistency and reduce ambiguities. They suggested that context continue to take centre stage in addressing key issues in MT.

However, there are alternative ways of conceptualizing context. House (2006) approaches context from broader philosophical and sociolinguistic perspectives, exploring how texts are situated within their cultural, temporal, and social environments. She highlights the dynamic and shifting nature of context as texts move between environments. House (2023) further identifies seven standards of textuality that define what makes a text coherent and meaningful. These standards are: cohesion (lexicogrammatical relations that link parts of a text), coherence (conceptual consistency that ties content together), intentionality (the text producer's purpose), acceptability (the reader's ability to accept the text as coherent based on their socio-cultural background), informativity (the newness of the information presented), situationality (the text's relevance to its socio-temporal context), and intertextuality (the relationship between a text and other related texts).

While Melby and Foster’s (2010) framework focuses on the external resources translators need to access during the translation process—such as co-text, rel-text, and non-text—it does not fully address how translations are received and interpreted by their audiences. This is where House’s (2023) seven standards of textuality add depth. By incorporating intentionality and acceptability, House ensures that translations resonate with the target audience’s cultural and linguistic expectations. Additionally, situationality and intertextuality are crucial for producing translations that are coherent and meaningful within their broader cultural context. In this way, House’s seven components broaden our understanding of context by connecting it directly to how users will perceive and engage with the translation.

For MT researchers, House’s view of context is essential because it emphasizes the need to consider how translations are received and understood, beyond just linguistic accuracy. Intentionality and situationality challenge MT systems, as they lack access to the text producer’s purpose or the socio-temporal context. While human translators can interpret intent and adapt to cultural nuances, MT systems rely on algorithms and predefined data. However, intentionality in MT can be applied through the user’s request—such as specific instructions on formality or tone. Additionally, MT systems can be set up to operate within user-defined constraints, allowing organizations or end-users to specify the translation’s style and function, further aligning the system’s output with its intended purpose.

Given the complexity of context, it is essential to focus on effective methods for analysing it. To properly assess NMT output, we need to break context down into manageable components to consider how various contextual elements impact the translation's effectiveness. A key approach for this is situational context analysis, which divides context into three interrelated components: field, tenor, and mode (House 2015, 127). Drawing on register analysis (Halliday and Hasan, 1989), these components connect texts to their situational contexts. Field refers to the subject matter or events discussed (what is happening), tenor examines the relationships between participants (author, reader, translator), and mode considers the medium of communication (written, spoken, digital). Together, these components ensure that the translation captures not only the meaning itself but also the social and communicative functions of the text in its context.

Applying this analysis approach to evaluating aspects like adequacy and fluency could lead to enhancements. First, adequacy can be viewed through two distinct lenses: *semantic adequacy*, which assesses meaning accuracy and domain-specific knowledge (field), and *formal adequacy*, which examines how well the translation aligns with the social roles and relationships (tenor) and communication medium (mode). By separating these dimensions, we ensure that the evaluation of adequacy does not conflate meaning accuracy with social and formal appropriateness, which require distinct considerations.

Second, fluency assessments can be refined to account for the situational context analysis. Just as adequacy is divided into distinct dimensions, fluency can be evaluated through three lenses: *grammatical fluency* ensures that the translation follows the structural norms of the target language (*field*); *stylistic fluency* examines the appropriateness of tone and formality, ensuring that the translation reflects the relationships between participants (e.g., formal vs. informal communication), aligning with the social roles and relationships (*tenor*); *idiomatic fluency* evaluates whether expressions and phrases sound natural and appropriate for the specific medium of communication, such as written or spoken discourse (*mode*). For example, fluency would consider whether the translation reflects the correct formality level in a business email versus a casual text message, or whether the rhetorical style of a political speech is preserved in the translation.

Additionally, *fitness for purpose* can serve as an overarching principle that encompasses both adequacy and fluency, ensuring that the translation meets the specific goals and expectations of the task at hand. This principle ensures that, beyond being accurate and fluent, the translation effectively fulfils its intended communicative function, whether in legal, medical, or creative contexts. By applying this principle within a functional framework, we ensure that translations are evaluated in line with their practical, real-world objectives.

To operationalize these ideas, Liu et al. (2024) suggests the use of an evaluation brief, analogous to and aligned with a translation brief and designed to guide the evaluation task. Implementing a functional-componential approach (Colina, 2008), the evaluation brief should provide annotators with essential background information, including details such as the source and target languages, the intended audience, the purpose of the text, and relevant style guidelines. In addition to these, we propose that the evaluation brief should also include situational context—field, tenor, and mode—to ensure annotators fully grasp the broader communicative setting in which the translation operates. These situational elements provide a structure to account for how social roles, subject matter, and communication channels influence the text's meaning and purpose. Based on this contextual foundation, Colina’s functional-componential approach can be applied to break down the evaluation into manageable components, ensuring that each component is evaluated in relation to the translation’s specific function, leading to more targeted and reliable assessments. This includes focusing on both the overall function of the text, as defined by its purpose and audience, and the specific contextual elements (field, tenor, mode) that contribute to fulfil that function.

Let's take an example of evaluating a translation of a legal contract. The first step is to create an evaluation brief, which provides annotators with critical contextual information such as the source and target languages (e.g., English to French in a Canadian context), the purpose of the text (e.g., to create a legally binding agreement between two business entities), the intended audience (e.g., legal professionals), and style guidelines (e.g., the formal tone typical of legal contracts, which demands precision, neutrality, and clarity). Additionally, the brief should outline the situational context, including field (legal terminology and concepts), tenor (the formal relationships between parties—such as business partners or client and service provider), and mode (the structured format of a legal document). In the next step, a functional-componential approach can be used to assess whether the translation fulfils its function (e.g., legal enforceability) and whether all components align with the text's purpose.

3.3 Proposal 3: Consider the expertise of annotators to align human evaluation with users' perspectives

In MT, annotators vary widely in terms of expertise, including crowd workers, bilingual speakers, and professional translators. Perhaps due in part to this range of participants, annotation is a highly variable process. Low agreement among annotators—especially on critical errors—has been a persistent issue (Specia et al., 2021). This lack of agreement can stem from a variety of factors, including task complexity and a lack of clear understanding of what constitutes a critical error in the context of the translation's purpose. Without a shared understanding of the functional and contextual requirements of the translation, annotators may weigh errors differently, leading to inconsistent assessments (Al Sharou & Specia, 2021). The MT community has recognized the need to recruit evaluators with the skills and knowledge to improve the consistency and reliability of annotation.

The first two proposals above highlight the importance of adopting a functionalist-contextual approach to translation quality. Together, they reinforce the idea that translation is not an isolated activity, but one deeply embedded in its functional and contextual environment. In this proposal, we argue that professional translators are well-suited for the annotation role, particularly in contexts where a deeper understanding of function and context is necessary for accurate assessment. While MT serves a broad range of users, including subject matter experts and non-language professionals who interact with MT output in various ways, professional translators are uniquely equipped to evaluate whether the translation aligns with the original text's purpose and functions appropriately within its intended socio-cultural and communicative context.

The European Master's in Translation (EMT) competence framework (2022) lists a comprehensive set of competences (e.g., language and culture, technology and personal and interpersonal) to prepare translators for professional workplaces. For example, EMT graduates are trained to recognize and navigate language variants—whether social, geographical, or historical—and to apply the appropriate language conventions in their translations. They are also trained to identify cultural elements, values, and allusions in both written and spoken texts, ensuring that translations are not only accurate but also culturally appropriate. In this sense, professional translators are expected to be equipped to navigate the evaluation brief, as they are trained to interpret critical elements such as the intended audience, purpose, and style guidelines, along with situational factors like field, tenor, and mode. Their ability to synthesize this information allows them to approach evaluation with a clear understanding of the translation's communicative goals and the broader context in which it functions. Professional translators are also adept at managing risk, ensuring that critical errors are avoided or minimized (Krüger, 2022). Although, like other human annotators, professional translators may vary somewhat in their perceptions of errors, their preparation and ability to assess errors' severity allows them to provide more reliable and consistent evaluations, contributing to higher inter-annotator agreement (Toral et al., 2018).

Professional translators possess the heightened linguistic sensitivity needed to identify subtle errors or inconsistencies that others may overlook. Their expertise allows them to recognize issues that may seem minor but could have long-term impacts. This reinforces the importance of having professional translators, particularly in contexts where translation serves to preserve cultural identity, linguistic integrity, or political equity—areas that MT alone cannot adequately address.

However, in cases where highly specialized domain knowledge is required—such as in legal or medical translations—bilingual subject matter experts (SMEs), such as jurilinguists, may be brought in to ensure accurate translation of domain-specific terminology and knowledge. In the case of evaluating the translation of a legal contract, as discussed in Section 3.2, employing SMEs is essential due to the complexity of legal systems and language. An annotator with legal expertise is crucial to ensure the translation reflects the conceptual, cultural, and linguistic complexities in legal language accurately, as failing to do so could lead to serious consequences.

To conclude, we emphasize the necessity for MT researchers to select annotators whose expertise is aligned with the functional and contextual needs of target users. Given the challenges in MT evaluation, such as discrepancies in annotator judgments and the variability in error severity weighting, professional expertise offers a promising approach to help address these issues. Professional translators—or in specific cases, bilingual SMEs—bring valuable insights to the evaluation process.

4 Concluding remarks

In concluding this paper, we stress the importance of integrating both functional and contextual understanding in MT evaluation practices. By aligning evaluation briefs with the communicative purpose of translations and the specific needs of their target users, incorporating essential context, and selecting annotators with the right expertise, MT evaluation can evolve to better reflect the complex nature of translation. This user-centric, context-driven and functionalist approach allows for more precise evaluations, which may contribute to the refinement of MT systems to meet the demands of real-world applications.

Given these recommendations, we argue that increased exchange and collaboration between the TS and MT scholarly communities, the language industry, and other relevant disciplines—

such as the Social Sciences—is essential for improving MT evaluation practices. Involving experts with diverse perspectives, including those familiar with questionnaire design, can help ensure that evaluation frameworks are grounded not only in theoretical insights and practical expertise but also in robust methodologies for data collection and user feedback. As Kay (1980) observed, translation is a deeply human task, and effective MT requires the integration of both human expertise and machine capabilities. This aligns with our call for purpose- and expertise-driven, user-centric evaluation processes that ensure MT output serves real-world needs effectively.

References

Al Sharou, Khetam, 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*, 171–180. Ghent, Belgium: European Association for Machine Translation.

Bowker, Lynne. 2019. Fit-for-purpose Translation. In *The Routledge Handbook of Translation and Technology*, edited by Minako O'Hagan, 453–469. Routledge.

Bowker, Lynne. 2021. Can Machine Translation Meet the Needs of Official Language Minority Communities in Canada? A Recipient Evaluation. *Linguistica Antverpiensia, New Series – Themes in Translation Studies* 8. https://doi.org/10.52034/lanstts.v8i.248.

Castilho, Sheila, and Rebecca Knowles. 2024. A Survey of Context in Neural Machine Translation and Its Evaluation. *Natural Language Processing*, 1–31.

Colina, Sonia. 2008. Translation Quality Evaluation: Empirical Evidence for a Functionalist Approach. *The Translator* 14 (1): 97–134.

Colina, Sonia. 2015. *Fundamentals of Translation*. Cambridge University Press.

European Master's in Translation Group. 2022. *European Master's in Translation Competence Framework 2022*. European Master's in Translation Group. https://commission.europa.eu/system/files/2022-11/emt_competence_fwk_2022_en.pdf [last accessed 30 September 2024]

Freitag, Markus, George Foster, David Grangier, Vikas Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics* 9: 1460–1474.

Freitag, Markus, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster.2023. Results of WMT23 Metrics Shared Task: Metrics Might Be Guilty but References Are Not Innocent. In *Proceedings of the Eighth Conference on Machine Translation*, 578–628. Association for Computational Linguistics.

Gladkoff, Serge, Irina Sorokina, Lifeng Han, and Alexandra Alekseeva. 2022. Measuring Uncertainty in Translation Quality Evaluation (TQE). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1454–1461, Marseille, France. European Language Resources Association.

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, 33–41. Sofia, Bulgaria: Association for Computational Linguistics.

Halliday, M. A. K., and Ruqaiya Hasan. 1989. *Language, Context, and Text: Aspects of Language in a Social Semiotic Perspective*. 2nd ed. Oxford University Press.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv preprint*arXiv:1803.05567. https://arxiv.org/abs/1803.05567.

House, Juliane. 2006. Text and Context in Translation. *Journal of Pragmatics* 38 (3): 338–358. https://doi.org/10.1016/j.pragma.2005.06.021.

House, Juliane. 2015. *Translation Quality Assessment: Past and Present*. Routledge.

House, Juliane. 2023. Looking at Translation from Different Perspectives. In *Translation: The Basics*, 2nd ed., 23–35. Routledge. https://doi.org/10.4324/9781003355823.

Kay, Martin. 1980. The Proper Place of Men and Machines in Language Translation. *Research Report CSL-80-11*, Xerox Palo Alto Research Center. Reprinted in *Machine Translation* 12 (1997): 3–23. https://doi.org/10.1023/A:1007911416676.

Knowles, Rebecca, and Chi-kiu Lo. Calibration and Context in Human Evaluation of Machine Translation. *Natural Language Processing*, 2024, 1–25. https://doi.org/10.1017/nlp.2024.5.

Kocmi, Tom, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova, and Jun Suzuki. 2023. Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet. In *Proceedings of the Eighth Conference on Machine Translation*, 1–42. Association for Computational Linguistics.

Kocmi, Tom, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation. *arXiv preprint* arXiv:2406.11580. https://doi.org/10.48550/arXiv.2406.11580.

Koehn, Philipp. 2020. Evaluation. In *Neural Machine Translation*, edited by Philipp Koehn, 41–64. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108608480.005.

Krüger, Ralph. 2022. Some Translation Studies Informed Suggestions for Further Balancing Methodologies for Machine Translation Quality Evaluation. *Translation Spaces*, March. https://doi-org.proxy.bib.uottawa.ca/10.1075/ts.21026.kru.

Liebling, Daniel J., Kathryn Heller, Samantha Robertson, and Wesley Hanwen Deng. 2022. Opportunities for Human-Centered Evaluation of Machine Translation Systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, edited by Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruiz, 229–240. Seattle, United States: Association for Computational Linguistics.

Läubli, Samuel, Rico Sennrich, and Martin Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-Level Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4791–4796. Brussels, Belgium: Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1512.

Liu, Ting, Chi-kiu Lo, Elizabeth Marshman, and Rebecca Knowles. 2024. Evaluation Briefs: Drawing on Translation Studies for Human Evaluation of MT. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, 190–208. Chicago, USA: Association for Machine Translation in the Americas.

Lommel, Arle, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica* 1 (12): 455–463.

Lo, Chi-kiu, Samuel Larkin, and Rebecca Knowles. 2023a. Metric Score Landscape Challenge (MSLC23): Understanding Metrics' Performance on a Wider Landscape of Translation Quality. In *Proceedings of the Eighth Conference on Machine Translation*, 776–799. Singapore: Association for Computational Linguistics.

Lo, Chi-kiu, Rebecca Knowles, and Cyril Goutte. 2023b. Beyond Correlation: Making Sense of the Score Differences of New MT Evaluation Metrics. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, 186–199. Macau SAR, China: Asia-Pacific Association for Machine Translation.

Melby, Alan K., and Christopher Foster. 2010. Context in Translation: Definition, Access and Teamwork. *The International Journal for Translation & Interpreting Research* 2.

Nirenburg, Sergei, Harold L. Somers, and Yorick A. Wilks. 2003. *Translation*. In *Readings in Machine Translation*. MIT Press.

Nord, Christiane. 1997/2014. *Translating As a Purposeful Activity: Functionalist Approaches Explained*. New York: Routledge. http://ebookcentral.proquest.com/lib/ottawa/detail.action?docID=1666958.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania: Association for Computational Linguistics.

Rei, Ricardo, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*

(EMNLP), edited by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, 2685–2702. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213.

Reiss, Katharina. 1971. *Möglichkeiten und Grenzen der Übersetzungskritik: Kategorien und Kriterien für eine sachgerechte Beurteilung von Übersetzungen*. 1st ed. München: M. Hueber.

Reiss, Katharina. 1983. Adequacy and Equivalence in Translation. *The Bible Translator (Technical Papers)* 3: 301– 308.

Specia, Lucia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 Shared Task on Quality Estimation. In *Proceedings of the Sixth Conference on Machine Translation*, 684–725. Online: Association for Computational Linguistics.

Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, 113–123. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6312.

Vermeer, Hans J. 1989. Skopos and Commission in Translational Action. In *Readings in Translation Theory*, edited by Andrew Chesterman, 173–187.

The Evolution of Smart Translation: CAT + NMT + GenAI + MTPE + HT ≼ Translation?

Marija Omazić, Blaženka Šoštarić, Romana Čačija

Faculty of Humanities and Social Sciences, University of Osijek, Croatia

momazic@ffos.hr
bsostaric@ffos.hr
rcacija@ffos.hr

Abstract

Recent advancements in artificial intelligence (AI) have opened up numerous possibilities for incorporating it into the translation process and enhancing translation quality. This paper examines the effectiveness and implications of integrating AI into translation workflows, focusing on the innovative Adaptive Generative Translation (AGT) technology by memoQ, which leverages Microsoft's Azure OpenAI service. The study compares AGT's performance with traditional neural machine translation (NMT) systems—specifically ModernMT, Google Translate, and Microsoft Translator— used within the memoQ translation management system. Key research questions address whether AGT substantially improves translation quality metrics, compares with Microsoft Translator as its baseline, and how effectively it utilizes existing translation resources. The methodology involves analysing translations of a 13,457-word administrative text from English to Croatian using automatic and qualitative assessment methods, including metrics like BLEU, ChrF, and COMET. The study aims to determine AGT's impact on terminological accuracy, consistency, tone, and style, especially in morphologically rich languages like Croatian. Results will provide insights into the potential of AI-enhanced translation tools to surpass traditional NMT systems and bring them closer to human translation, offering a forward-looking perspective on the evolution of smart translation technologies.

1 Introduction

In the latest European Language Industry Survey (ELIS, 2023), artificial intelligence (AI) and machine translation (MT) were found to “dwarf” all other industry trends across all operational segments, eliciting both positive and negative sentiments and forecasting that by 2025, MT will completely dominate the translation industry, becoming something of a *new lingua franca*. This may mainly be true of adaptive neural machine translation (NMT), such as ModernMT and RWS Language Weaver, which surpass the quality of traditional NMT providers (such as Google Translate and Microsoft Translator) due to their adaptive nature but still hold a relatively small share of the MT market.

The recent advancements in large language models (LLMs), deep learning, and generative artificial intelligence (GenAI) have undoubtedly opened up numerous possibilities for incorporating those advanced solutions into different stages of the translation process, affecting translator roles and workflows. AI can assist in translation quality assurance (TQA), help with terminology work, and support automatic AI-assisted pre and post-editing and error correction with justifications. For instance, AI-driven TQA tools can identify and correct errors that human translators might overlook, like Unbabel's Quality Intelligence API[1], ensuring higher quality translations.

Furthermore, AI can automatically perform machine translation quality estimation (MTQE) to help users choose the best provider for their project. For example, MTQE is available in the Phrase CAT tool, automatically selecting and supplying the best-rated MT. MemoQ's AIQE (Artificial Intelligence-based Quality Estimate) offers the same functionality, which runs a pre-translation with multiple engines and lets AIQE pick the best translation segment by segment. Trados also offers Language Weaver MTQE, an AI model designed to automatically assess the quality of Language Weaver output based on real-world post-editing scenarios. UnbabelMT[2] lets you test several MT providers simultaneously and provides comparative results to allow you to choose the best provider for your project.

AI, like ChatGPT[3] and TowerLMM, can also provide zero-shot AI translation, which has yet to match the quality of adaptive NMT. However, if coupled with RAG (retrieval-augmented generation), AI translation has been shown to come close to, or surpass, adaptive neural machine translation (NMT) quality for specific language pairs and domains[4]. This approach allows for translations without the need for prior exposure to the particular language pair, making it highly versatile and adaptable.

The most advanced developments, however, can enhance NMT output with the capabilities of LLMs and generative AI solutions. The integrated NMT, LLM, and GenAI technologies can potentially address some challenges NMT or zero-shot AI systems face, such as maintaining terminological accuracy, consistency, style, and fluency in morphologically rich languages. Such hybrid solutions are, for example, Unbabel's multilingual TowerLLM (Alves et al. 2024), which captures the power of Generative AI and optimizes it for translation in particular, or memoQ's new adaptive generative translation (AGT) technology. The rise of such advanced solutions will impact translation workflows and the role and training of translators.

2 Rationale

In this paper, we will critically examine this new stage in the evolution of "smart translation," focusing on the integration of advanced translation technologies—specifically, computer-assisted translation (CAT) tools and their resources (such as translation memories, termbases, and aligned corpora)—with neural machine translation (NMT) and Generative AI. This convergence within a unified translation management system (TMS) operated by human translators represents a significant shift in translation workflows. The objective was to create a

[1] It can be tested at https://qi.unbabel.com/, last accessed on 1 October 2024

[2] https://mtdemo.unbabel.com/, last accessed on 1 October 2024

[3] For zero-shot ChatGPT translation compared to NMT translation by Google Translate, Microsoft Translator, Phrase Translate and ModernMT, see Omazić and Šoštarić (2023). In that study, ChatGPT was shown to significantly underperform compared to all other providers in the en-hr language pair.

[4] The Tower LMM performance results for different languages and domains are presented here: https://tinyurl.com/mpph6u5p, last accessed on 1 October 2024

highly automated yet human-enhanced translation process that blends the speed and computational efficiency of AI with the nuanced expertise of human translators.

By seamlessly integrating these technologies, the potential for synergy is enormous, with the promise of faster, more accurate translations that benefit from AI precision and human quality control. This paper will assess the output of the memoQ AGT integrated hybrid system, comparing it with leading neural machine translation (NMT) solutions and traditional human translation to gauge its current performance. Through this comparative analysis, we aim to uncover the strengths and limitations of the system, providing insights for further refinement and optimization.

This investigation will not only highlight current capabilities but also identify key areas for future development, ultimately paving the way for more sophisticated, efficient, and reliable "smart" translation systems that better serve the needs of global communication in the digital age. Furthermore, final thoughts on the impact on translator training will be provided.

3 Background: memoQ’s Adaptive Generative Translation

A notable development in this area was the introduction of the patent-pending generative AI-based translation automation solution called Adaptive Generative Translation (AGT) technology by memoQ in November 2023, with the release of memoQ 10.4. AGT performs domain-adapted machine translation, combining the in-context learning ability of large language models (LLMs) with the text retrieval functionality of a translation management system (TMS). It builds on memoQ’s existing technologies, TM+, TB, and LiveDocs, which allow users to add their content to memoQ TMS resources in large amounts and a wide variety of formats and make them ready for instant reuse. Using the generative power of an LLM, memoQ AGT creates translations tailored to the customer’s existing language resources and user-provided data. It achieves this without retraining or fine-tuning the LLM itself. As a result, both the language data and the control over the translation process remain protected. It combines MT and AI capabilities, leveraging Microsoft Translator as the baseline machine translation and Microsoft's Azure OpenAI service, a limited-access enterprise service, offering the same generative language models as OpenAI and providing the security capabilities of Microsoft Azure. Azure OpenAI service does interact with any service operated by OpenAI, such as ChatGPT or OpenAI API, meaning that data is not shared with other customers, OpenAI, Microsoft, or other third-party products or services, and it is not used to train or improve large language models[5].

AGT is designed to automatically adapt to specific translation domains without adjusting or training new models. AGT automatically generates a prompt sent to the LLM segment by segment, requesting it to take into account the translator's existing resources, such as locally stored translation memories, the translator's termbases, as well as aligned and non-aligned supporting documents that can be created as a helpful reference resource in memoQ LiveDocs. It should be noted that MemoQ AGT is expected to perform more effectively with ample resources and larger translation volumes. It has been designed primarily for enterprises with

[5] AGT Terms of Service are available at https://tinyurl.com/47kywaww, accessed on 1 October 2024

ample linguistic resources and LSPs who work with projects with substantial background resources (TMs, terminology, or parallel texts).

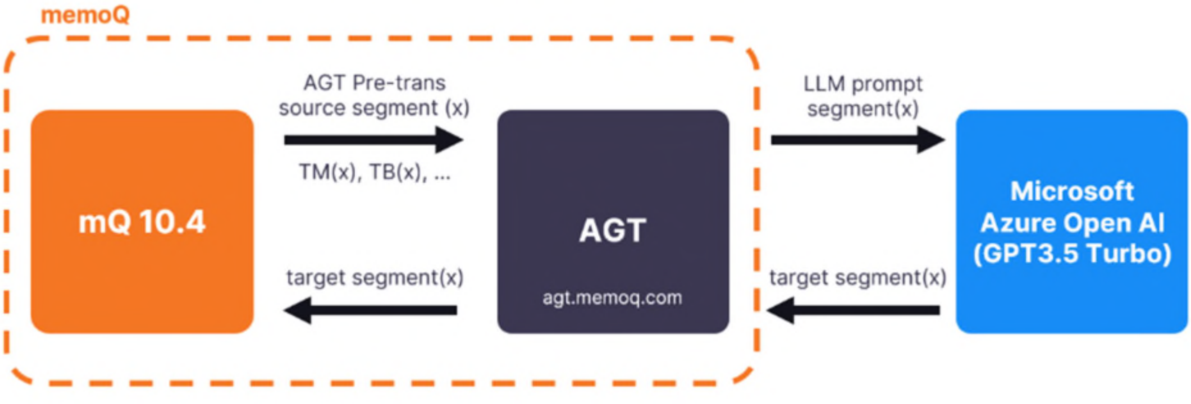


Figure 1. AGT integration in memoQ 10.4[6]

Initially, memoQ AGT was only made available to a select group of registered users invited to test drive it. The waitlist to test it is still open, and access to new users is gradually being granted[7], as AGT is still in the introductory and experimental stages. The service is free during the testing phase, and after you use the assigned quota, you should start paying for the service. MemoQ AGT is integrated and can only be used in the memoQ TMS as a machine translation plugin with a generated API key. It works with translation resources within memoQ.

The unique selling proposition of memoQ AGT is that it excels at using low fuzzy matches at rates as low as 50 or 60%, which are rarely considered by human translators, and adapts them to the source text. It guarantees higher quality translations compared to relying solely on machine translation or translation memories. It arguably maintains consistency with the terminology and language of previous translations, reducing the need for extensive editing. This should lead to faster, more accurate translations while minimizing time and resource requirements.

There are currently no research papers published on the performance of AGT that we could review, the only research conducted so far is Aladrović's MA thesis (2024), prepared within the scope of this project and under our supervision.

4 Research questions

This paper aims to explore the claims stated above and assess the effectiveness, advantages, and potential implications of using AGT in the translation workflow in memoQ by comparing it with several selected traditional neural machine translation (NMT) outputs and human

[6] source: https://docs.memoq.com/helpcenter/Guides/AGT%20tool/01%20Introduction.htm?Highlight=agt

[7] memoQ AGT currently supports the following languages: Albanian, Arabic, Bosnian, Bulgarian, Chinese, Croatian, Czech, Danish, Dutch, English, Estionian, Finnish, French, German, Greek, Hebrew, Hungarian, Indonesian, Italian, Japanese, Korean, Latvian, Lithuanian, Macedonian, Norwegian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Turkish, Ukranian, Vietnamese.

translation using the same memoQ translation management system environment and the same resources. The study seeks to answer the following research questions:

1) Does memoQ Adaptive Generative Translation (AGT) significantly improve translation quality metrics over other traditional NMT systems and adaptive NMT systems?

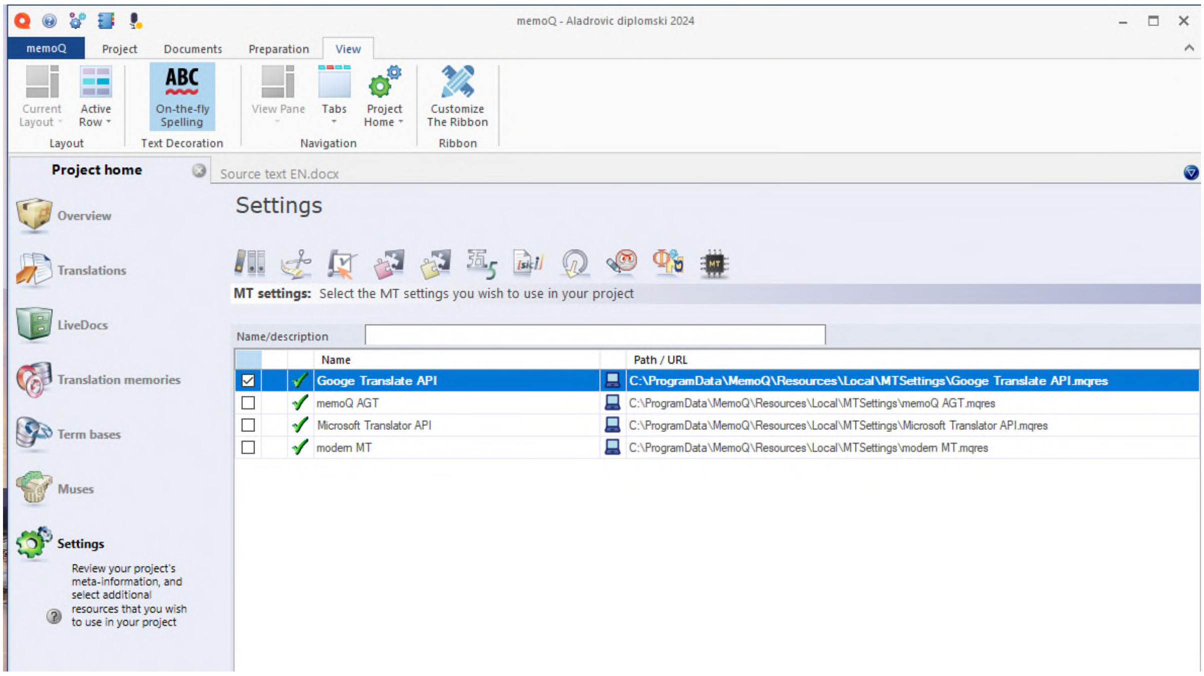
2) How does the AI-enhanced output of AGT differ from the baseline output of Microsoft Translator, and what are the implications of these differences for the quality of translation in specialised domains?

3) How effectively does memoQ Adaptive Generative Translation (AGT) utilise existing translation memories and reference documents compared to standard NMT outputs when working within the memoQ translation management system?

4) How does Adaptive Generative Translation (AGT) compare to standard neural machine translation (NMT) systems and human translation in terms of terminological accuracy, consistency, tone, and style when translating administrative texts from English to Croatian?

5 Methodology

To answer those research questions, the translations by three selected NMT providers who provide en-hr translation, namely ModernMT, Google Translate, and Microsoft Translator, as well as AGT translation, were analysed against each other, as well as against human translation, using automatic machine translation quality assessment methods and comparative qualitative analysis performed by five evaluators with extensive translation experience. All NMT systems were integrated into memoQ via plugins and API keys.



Settings

MT settings: Select the MT settings you wish to use in your project

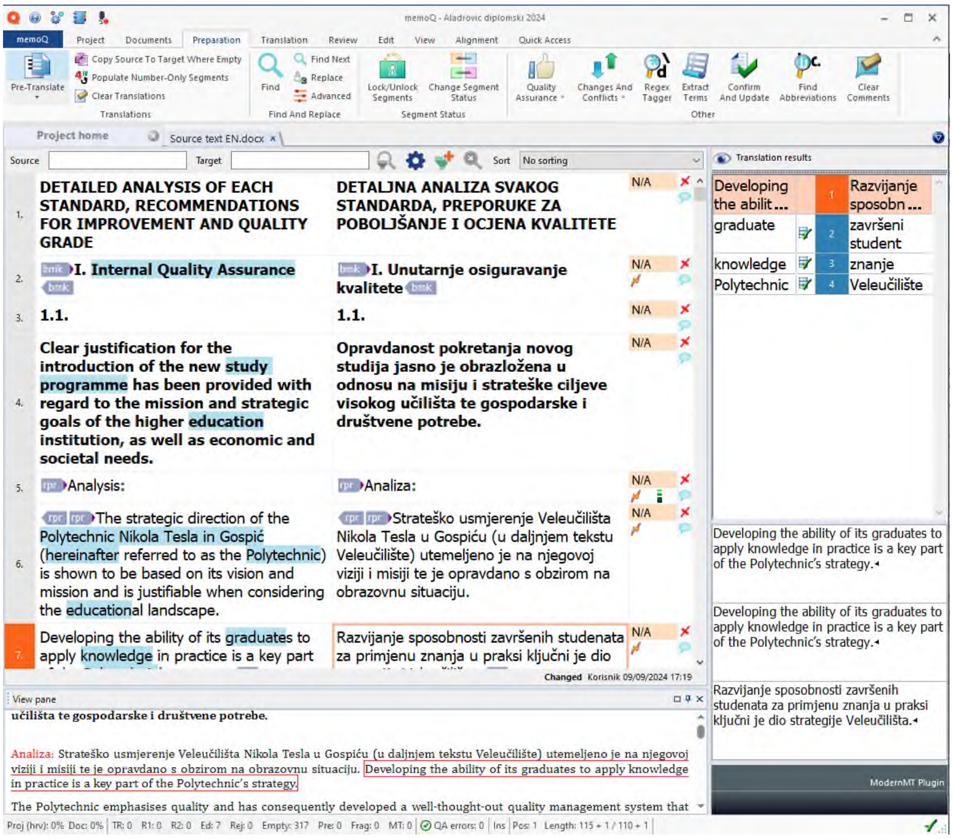
Name/description

		Name		Path / URL
☑	✓	Googe Translate API		C:\ProgramData\MemoQ\Resources\Local\MTSettings\Googe Translate API.mqres
□	✓	memoQ AGT		C:\ProgramData\MemoQ\Resources\Local\MTSettings\memoQ AGT.mqres
□	✓	Microsoft Translator API		C:\ProgramData\MemoQ\Resources\Local\MTSettings\Microsoft Translator API.mqres
□	✓	modern MT		C:\ProgramData\MemoQ\Resources\Local\MTSettings\modern MT.mqres

Figure 2. MT plugins and APIs used to pre-translate the source text in memoQ

The source text was an original 13,457-word administrative report on quality assurance in higher education in Croatia written in the English language. The report was drafted by panel members who are non-native speakers of English, which may have affected the quality of the source text and, subsequently, the quality of machine-translated outputs. A separate project was set up in memoQ for this research. The project included a sizeable en-hr translation memory, containing 37855 translation units provided by the client who exported it from the CAT tool of their choice (Trados), plus a secondary local translation memory in memoQ created by the freelance translator containing 9615 entries from previous similar translation projects for the same client. It also included a clean and well-maintained en-hr term base containing 1115 verified terms on quality assurance in higher education and a LiveDocs corpus with five reference files provided by the client for the current project. The reference files included the site-visit protocol, study programme proposals in English and Croatian, the decision to appoint the Expert Panel and the report form in both English and Croatian, containing the main headings and subheadings as they should appear in the translation.

The source text was then pre-translated in memoQ using four NMT providers: Google Translate, Microsoft Azure AI Translator plugin, modernMT, and memoQ AGT. The pre-translation was done on the entire document at once.



	Source	Target	
1.	DETAILED ANALYSIS OF EACH STANDARD, RECOMMENDATIONS FOR IMPROVEMENT AND QUALITY GRADE	DETALJNA ANALIZA SVAKOG STANDARDA, PREPORUKE ZA POBOLJŠANJE I OCJENA KVALITETE	N/A
2.	bmk I. Internal Quality Assurance bmk	bmk I. Unutarnje osiguravanje kvalitete bmk	N/A
3.	1.1.	1.1.	N/A
4.	Clear justification for the introduction of the new study programme has been provided with regard to the mission and strategic goals of the higher education institution, as well as economic and societal needs.	Opravdanost pokretanja novog studija jasno je obrazložena u odnosu na misiju i strateške ciljeve visokog učilišta te gospodarske i društvene potrebe.	N/A
5.	rpr Analysis:	rpr Analiza:	N/A
6.	rpr rpr The strategic direction of the Polytechnic Nikola Tesla in Gospić (hereinafter referred to as the Polytechnic) is shown to be based on its vision and mission and is justifiable when considering the educational landscape.	rpr rpr Strateško usmjerenje Veleučilišta Nikola Tesla u Gospiću (u daljnjem tekstu Veleučilište) utemeljeno je na njegovoj viziji i misiji te je opravdano s obzirom na obrazovnu situaciju.	N/A
7.	Developing the ability of its graduates to apply knowledge in practice is a key part	Razvijanje sposobnosti završenih studenata za primjenu znanja u praksi ključni je dio	N/A

Changed Korisnik 09/09/2024 17:19

Developing the abilit...	1	Razvijanje sposobn...
graduate	2	završeni student
knowledge	3	znanje
Polytechnic	4	Veleučilište

Developing the ability of its graduates to apply knowledge in practice is a key part of the Polytechnic's strategy.

Developing the ability of its graduates to apply knowledge in practice is a key part of the Polytechnic's strategy.

Razvijanje sposobnosti završenih studenata za primjenu znanja u praksi ključni je dio strategije Veleučilišta.

učilišta te gospodarske i društvene potrebe.

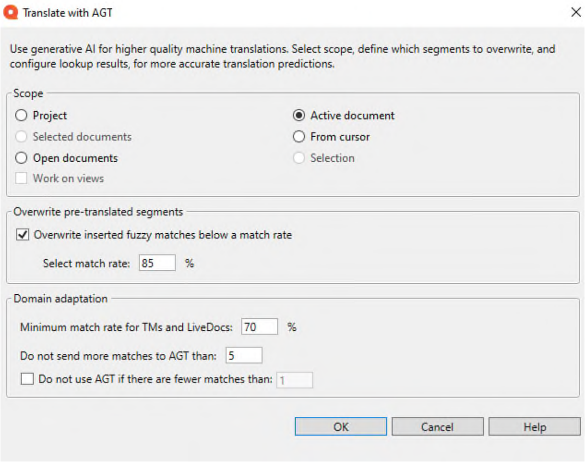
Analiza: Strateško usmjerenje Veleučilišta Nikola Tesla u Gospiću (u daljnjem tekstu Veleučilište) utemeljeno je na njegovoj viziji i misiji te je opravdano s obzirom na obrazovnu situaciju. Developing the ability of its graduates to apply knowledge in practice is a key part of the Polytechnic's strategy.

The Polytechnic emphasises quality and has consequently developed a well-thought-out quality management system that

Proj (hrv): 0% Doc: 0% | TR: 0 R1: 0 R2: 0 Ed: 7 Rej: 0 Empty: 317 Pre: 0 Frag: 0 MT: 0 | QA errors: 0 | Ins | Pos: 1 Length: 115 + 1 / 110 + 1

Figure 3. Pre-translation result for ModernMT

AGT was set up to overwrite inserted fuzzy matches below the 85% match rate, and for the domain adaptation, the minimum match rate for TMs and LiveDocs was set at 70%, and not more than five matches were sent to AGT.



Use generative AI for higher quality machine translations. Select scope, define which segments to overwrite, and configure lookup results, for more accurate translation predictions.

Scope

○ Project ● Active document
○ Selected documents ○ From cursor
○ Open documents ○ Selection
□ Work on views

Overwrite pre-translated segments

☑ Overwrite inserted fuzzy matches below a match rate

Select match rate: 85 %

Domain adaptation

Minimum match rate for TMs and LiveDocs: 70 %

Do not send more matches to AGT than: 5

□ Do not use AGT if there are fewer matches than: 1

Figure 4. Pre-translation setup for memoQ AGT

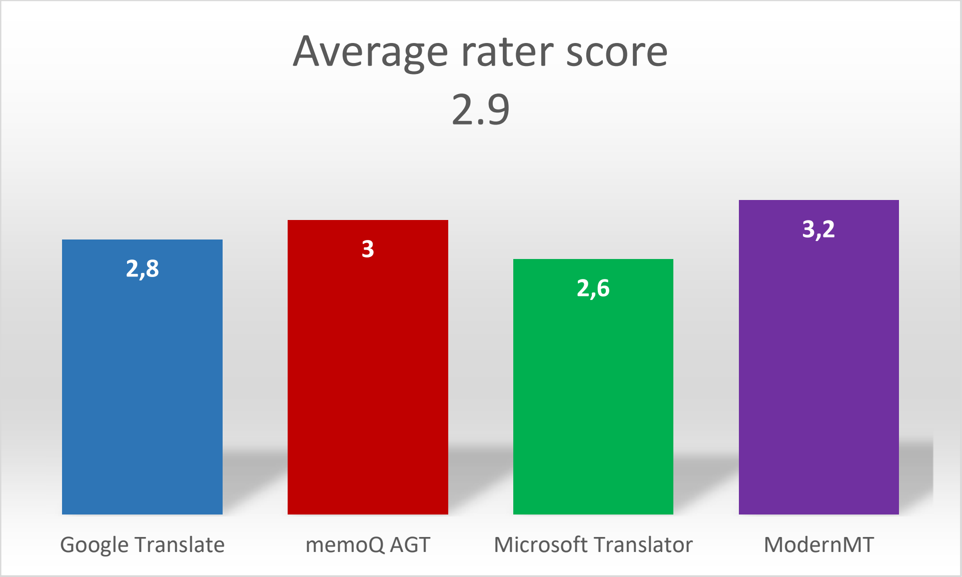
The automatic evaluation of the four versions was conducted using the MATEO evaluation platform[8], using the BLEU, ChrF, and COMET metrics (Vanroy, Tezcan and Macken, 2023). As the source text was large, we divided it into five smaller random samples that did not include the pre-defined sections from the report form, took the matching translations from all providers, and prepared the documents for evaluation so there were no empty lines. They had to be reformatted to the UTF8 format. The source file, human translation as the reference file, and four translations obtained from the four NMT providers were uploaded to the MATEO platform.

The qualitative analysis of selected examples also assessed AGT's claims of surpassing neural machine translation regarding terminological accuracy, consistency, tone, and style by leveraging existing translation memories and reference documents while considering broader contextual factors. For the qualitative analysis, we created worksheets containing 100 sentences from each translation. The sentences were collected randomly from report sections which were not part of the standard report form, making sure that there was no selection bias. We asked the evaluators to grade them on a scale from 1 to 4, with 4 being the highest. The raters were also asked to flag the main issues they spotted and make any observations they thought were relevant regarding the use of terminology, gender or syntax, style, fluency, technical matters, formatting, etc. They were instructed to pay special attention to the comparison between Microsoft Translator and AGT output, as it was expected that AGT would enhance Microsoft Translator's baseline through AI-driven refinements. This comparison allowed us to assess the real impact of AI enhancements on MT translation quality in this specialised domain. By comparing the Microsoft Translator’s baseline with the AGT-refined output, we can see AI's changes and improvements.

6 Findings

The average MT quality scores by human evaluators are shown in Figure 5.

[8] https://mateo.ivdnt.org/Evaluate [last accessed on 30 August 2024]



Instant domain adaptation placed AGT in a favourable position compared to Google Translate and Microsoft Translator; however, notable limitations remain in terms of fluency, stylistic coherence, and contextual understanding, which was reflected in the issues flagged by raters in their comments. With AGT, it is still impossible to tell which of the resources was prioritised, as in some instances it was evident that the termbase was used as a primary resource, whereas in others, a term in the termbase was ignored and other resources were given priority, resulting in a less acceptable translation. This is still a marked improvement over engines like Google Translate and Microsoft Translator that do not reference TB at all, as observed in this analysis. Overall, ModernMT was rated as the best by all evaluators, with AGT following closely. While Google Translate and Microsoft Translator have demonstrated their utility, their performance was insufficient to be favourably compared to AGT or ModernMT.

The qualitative analyses also involved the comparison of different solutions and flagging potential issues. Several examples are provided in the tables below.

ST[9]	The external stakeholders met by the Panel provided further evidence of the necessity for a programme of this nature in the region.	TB: external stakeholders – vanjski dionici; the Panel – Stručno povjerenstvo	Rater grade (1-4)	Rater comments
AGT	Vanjski dionici s kojima se Stručno povjerenstvo susrelo pružili su dodatne dokaze o potrebi za ovom vrstom studijskog programa u regiji.		4	word order, TB
MdMT	Vanjski dionici s kojima se Stručno povjerenstvo susrelo pružili su dodatne dokaze o potrebi za ovom vrstom studijskog programa u regiji.		4	word order, TB
MT	Vanjski dionici s kojima se sastao panel pružili su dodatne dokaze o potrebi za takvim programom u regiji.		2	

[9] Source text

GT[10]	Vanjski dionici s kojima se susreo Odbor pružili su dodatne dokaze o potrebi za programom ove prirode u regiji.		2	
TT	Vanjski dionici s kojima se Stručno povjerenstvo susrelo pružili su dodatne dokaze o potrebi za ovom vrstom studijskog programa u regiji.			

Table 1. Comparative tables showing ST, TT and different MT outputs

ST	Learning outcomes are aligned with the level of CroQF.	TB: learning outcomes – ishodi učenja CroQF – HKO	Rater grade	Rater comments
AGT	Ishodi učenja su usklađeni s razinom CroQF-a.		2	TB? word order?
MdMT	Ishodi učenja usklađeni su s razinom CroQF-a.		3	
MT	Ishodi učenja usklađeni su s razinom HKO-a.		4	
GT	Ishodi učenja usklađeni su s razinom CroQF-a.		3	
TT	Ishodi učenja usklađeni su s razinom HKO-a.			

Table 2. Comparative tables showing ST, TT and different MT outputs

The raters also flagged the frequent issues with incorrect tag placement by all NMT providers. They praised the AGT's use of terms from the term base but noted that although it outperformed other providers, it was still inconsistent.

Considering all the examples, we can conclude that AGT still has room for improvement before it can be regarded as the best. ModernMT has an advantage in producing a more fluent and meaningful translation when the target language is Croatian. AGT, on the other hand, struggles with that even though it provides better solutions for specific terms when compared to GT and MT. However, although rare, there are certain cases where GT or MT provided the preferred output. Nevertheless, they struggle the most when it comes to translating terms that require contextual awareness and have difficulty adapting the translation to sound fluent and natural in the target language.

The raters agreed that AGT shows considerable potential, and it is plausible to hypothesize that it could yield better results in different language pairs. AGT occasionally struggled to provide accurate translations, unlike ModernMT, which excelled by leveraging a broader contextual understanding of the entire document rather than isolated sentences. Based on the rater evaluation, Adaptive Generative Translation (AGT) demonstrates clear improvements over standard NMT engines under the same conditions and with the same resources. However, it still falls short of the overall quality achieved by adaptive NMT systems such as ModernMT.

The results of the automatic MT quality assessment of all four translations are presented in Figure 5., showing the BLEU, ChrF, and COMET scores.

[10] Google Translate

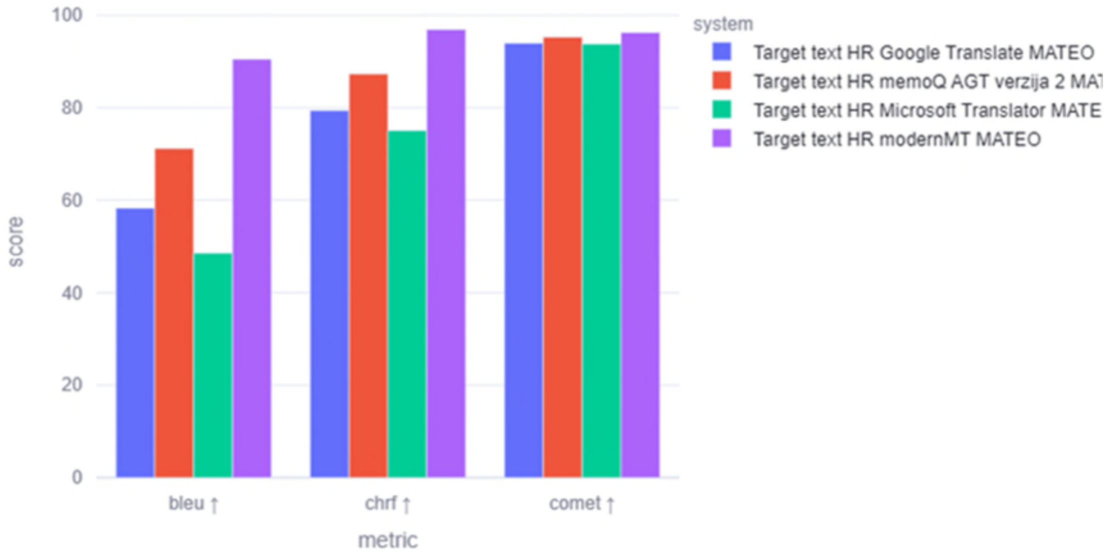


Figure 5. Automatic MT quality assessment

A higher metric score correlates with better translation quality. However, we can see that some variation exists between specific metric outputs. This is because BLEU and ChrF measure string overlap between the MT outputs and reference translations, whereas COMET assesses translation quality more broadly. COMET employs machine learning for evaluation, moving beyond superficial text comparisons typically used by traditional metrics. By accounting for fluency, meaning preservation, and adequacy, COMET proves most effective in translation cases requiring a more nuanced understanding of quality. Conversely, BLEU measures the correspondence of phrases between a translation and its reference by focusing on word matches, making it particularly suited for evaluating translations where word order and exact phrase matching are crucial. ChrF evaluates machine translation by calculating the similarity between MT output and reference translation using character n-grams rather than word n-grams. Comparative analysis of the scores reveals that the most significant variation in individual scores occurs within the BLEU metric, while ChrF generally shows higher scores with less deviation. COMET, in contrast, achieves the highest scores with minimal variance across specific translation outputs.

These metric results show that AGT represents an advancement compared to standard NMT engines like Google Translate or Microsoft Translator. AGT demonstrates performance improvements over Microsoft Translator across all metrics, as AGT uses MT as its baseline but introduces significant enhancements. However, ModernMT consistently outperforms all the other systems tested, achieving the highest scores across all metrics. In the English-Croatian language pair, AGT does not yet seem to match the results of adaptive NMT (ModernMT), but it shows clear improvement over standard NMT systems like Google Translate and Microsoft Translator. This may be attributed to the limited resources available during testing. The results could potentially differ with an even larger translation memory (TM) or different AGT settings.

The comparison between human raters and automatic metrics highlights some interesting patterns. Human evaluators rated ModernMT the best overall, with AGT closely following, outperforming Google Translate and Microsoft Translator. However, AGT exhibited fluency and contextual understanding issues, though it showed improvement by referencing a termbase (TB), something Google and Microsoft lacked. The raters flagged errors such as incorrect tag placement and inconsistencies in AGT's use of TB terms. In contrast, automatic metrics—BLEU, ChrF, and COMET—painted a slightly different picture. COMET, which uses machine learning for a broader evaluation of translation quality, provided the highest scores with minimal variance, indicating superior fluency and meaning preservation in ModernMT's translations. BLEU and ChrF focused more on phrase and character overlaps, with BLEU showing the most variation. AGT performed better than Google and Microsoft across all metrics but still lagged behind ModernMT, reinforcing human raters' assessments that AGT, while an improvement, has room for growth before matching the adaptive capabilities of ModernMT.

The study has certain limitations; firstly, the size of resources used. AGT is mainly designed to cater for institutions, enterprises, and LSPs with substantial and well-maintained background resources. We did include a large TM and a well-maintained TB, but we assume it may perform better with even more input it can adapt and use for translation generation. AGT performance in this study was tested using only an administrative text in the en-hr language pair, so all findings and conclusions refer to that particular case only.

7 Conclusion

This paper explored the performance and potential implications of using AI in the translation workflow, focusing on comparing memoQ AGT with selected traditional neural machine translation (NMT) outputs within the memoQ translation management system. Based on the research conducted and considering the limitations of this research in terms of size, resources used, and language combination, the following answers to the research question can be proposed.

The qualitative and quantitative quality assessment aimed to establish how Adaptive Generative Translation (AGT) compares to standard neural machine translation (NMT) systems and human translation regarding terminological accuracy, consistency, tone, and style when translating administrative texts from English to Croatian. The first question dealt with whether AGT provides significant improvements over traditional NMT systems regarding improvements in translation quality metrics. Our human and automatic assessment results indicate that AGT in our research setting shows some, but not necessarily significant, improvements over traditional NMT systems such as Google Translate and Microsoft Translator. However, it still does not match the quality of adaptive NMT provided by ModernMT or human translation for the en-hr language pair.

Furthermore, both qualitative and quantitative analyses have shown that the AI-enhanced output of AGT differs from the baseline output of Microsoft Translator and is assessed more favourably in both human and automatic evaluations, which clearly illustrates that AGT offers a certain added value in comparison to its baseline. It clearly excels over standard NMT in utilizing the terminology stored in the term base it can access. However, it has not been consistently accurate in drawing terms from the termbase in all cases.

The quality improvements noted in the output of AGT and ModernMT, which utilize more of the translator’s existing resources, have clear implications for the quality of translation in specialised domains. Their translation is assessed as less generic, more aligned with the client’s requirements, style, tone, and voice, and requires less post-editing effort, which may shorten turnaround times. Even though the improvements in AGT output are evident compared to standard NMT, and Microsoft Translator in particular, it is not easy to discern which resources are prioritized in the process: termbases, translation memories, or reference documents, as the outcomes are sometimes erratic, and termbase solutions are sometimes overridden by solutions from reference documents, which are not always accurate. AGT still does not allow setting priority ranking for resources, which would possibly improve its performance.

Overall, AGT's performance in our research setting can be rated as one step above traditional NMT like Google Translate or Microsoft Translator, and one step below adaptive NMT like ModernMT. All systems are shown to still be inferior to human output and require human oversight. This has implications for translator training, which needs to evolve to keep pace with the evolution of the translation process towards AI-enhanced, smart translation, and move more in the direction of tool selection, quality assessment, and post-editing.

Keeping in mind that AGT was tested with somewhat limited resources in the early stages of the tool’s development in the academic and freelance settings, it would be interesting to test it with much bigger resource sets kept by large international organizations, such as EU institutions and UN organizations, where it could draw on their large data sets as resources and potentially perform even better as a result.

Future research could also focus on different types of discourse and language pairs to better understand AGT’s performance, cost-effectiveness, and efficiency in various settings and conditions.

The results of this research also have implications for including AI in translator training, namely the need to carefully balance academic integrity and rigour with language service demands and market readiness. The focus should be on adaptability and metacognitive skills for the evolving role of translators in hybrid, automated workflows. AI should be included in translator training carefully as it is essential, acknowledging the risks of over-reliance on these tools versus their productivity benefits. Thoughtful integration of AI can boost students’ digital literacy, vigilance, and resilience, preparing them for the complex challenges in the competitive translation industry.

Finally, the integration of Adaptive Generative Translation (AGT) into translation workflows has significant implications for translator agency, competence, and required skill sets. While AGT and similar AI-driven tools can enhance productivity and accuracy, they also necessitate a shift in the translator's role, demanding new competencies and a revised understanding of agency within the translation process. Translator agency traditionally refers to the translator's autonomy and decision-making power in rendering a source text into a target language. The introduction of AI tools like AGT can be perceived as potentially undermining this agency by automating certain aspects of the translation process. However, a more nuanced perspective recognizes that AGT can actually redefine and enhance translator agency, but more research is needed to test the direction in which it is developing. Rather than being replaced by AI, translators can leverage AGT to streamline repetitive tasks, freeing up time and cognitive resources to focus on the more creative and complex aspects of translation, where human expertise remains irreplaceable. This shift allows translators to exert their agency in new ways, such as customising AGT settings to align with specific project requirements and client preferences, curating and managing linguistic resources (translation memories, termbases, reference documents) to ensure the accuracy and consistency of AI-generated output, critically evaluating and refining AI-generated translations, ensuring fidelity to the source text and adherence to quality standards. In terms of translator competence, the integration of AGT requires a broadening of the traditional skill set. While linguistic proficiency and subject matter expertise remain essential, translators must also develop new competencies to effectively collaborate with AI, such as advanced translation literacy, quality asssessment, post-editing proficiency and data management. The integration of AGT does not diminish the role of the translator but rather reshapes it, requiring a shift in focus and the acquisition of new skills.

Looking ahead, several potential technological advancements could enhance AI-assisted translation systems like memoQ AGT, particularly for applications in specialised domains. One limitation of AGT is its occasional struggle with fluency and understanding broader context compared to adaptive NMT systems. Addressing this issue would be crucial for wider adoption. Possible advancements include developing more sophisticated language models with a deeper understanding of linguistic nuances, syntax, and semantics. Current AI translation models primarily focus on sentence-level analysis. Developing algorithms that can analyse and incorporate context from the entire document could improve the accuracy and coherence of translations, especially in handling complex technical concepts and ensuring consistent terminology usage. The noted inconsistency in how AGT prioritises resources, sometimes overriding TB entries with potentially less accurate translations from reference documents could be resolved by improved resource management, which would enhance user confidence and trust in the system, such as user-defined resource ranking, i.e. allowing users to specify priority levels for different resources (TMs, TBs, reference documents), which would give translators greater control over how AGT leverages these resources, ensuring alignment with specific project requirements and client preferences. Furthermore, a more user-friendly, intuitive interface with robust customisation features could make AGT more appealing to a wider range of users. Also, providing translators with more fine-grained control over the style and tone of AI-generated output would allow for better adaptation to different target audiences and client requirements. Integrating real-time feedback mechanisms and collaborative editing tools would facilitate interaction between translators and AI, enabling a more seamless and efficient translation workflow.

By addressing these areas and continually incorporating advancements in artificial intelligence and natural language processing, developers can create AI-assisted translation systems that are more accurate, reliable, and user-friendly. Such improvements would likely generate greater interest and accelerate the adoption of these technologies in the translation industry, particularly within specialised domains that demand high levels of precision and consistency.

References

Aladrović, Josip. 2024. Comparative Analysis of MemoQ Adaptive Generative Translation (AGT) with Conventional Machine Translation (MT) Engines. MA thesis. Osijek. University of Osijek.

Alves, M. Duarte, João, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza and André F. T. Martins. 2024. Tower: An Open Multilingual Large Language Model for Translation-Related Tasks. ArXiv, abs/2402.17733.

Cieselski, Jourik. 2024. Neural Machine Translation Versus Large Language Models. Which technology will drive the future of automated translation? *MultiLingual Magazine*. https://multilingual.com/magazine/june-2024/neural-machine-translation-versus-large-language-models/ [last accessed on 30 August 2024].

The European Language Industry Survey 2023. https://elis-survey.org/wp-content/uploads/2023/03/ELIS-2023-report.pdf [last accessed on 30 August 2024].

Kornacki, Michał, & Pietrzak, Paulina. 2024. Hybrid Workflows in Translation: Integrating GenAI into Translator Training (1st ed.). Routledge. https://doi.org/10.4324/9781003521822

Omazić, Marija and Blaženka Šoštarić. 2023. New Resources and Methods in Translating Legal Texts: Machine Translation and Post-Editing of Machine-Translated Legal Texts. In Kordić, Lj. (ed.) Language(s) and Law. Osijek: Pravni fakultet. available at https://www.pravos.unios.hr/wp-content/uploads/2023/09/Publication-Languages-and-Law.pdf

Vanroy, Bram, Arda Tezcan, and Lieve Macken 2023. "MATEO: MAchine Translation Evaluation Online." In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Trarindu Ranasighe, Eva Vanmassenhove, Sergi Vidal Alvarez, Nora Aranberri, Mara Nunziatini, Carla Parra Escartin, Mikel Forcada, Popović Maja, Carolina Scarton and Helena Moniz (Eds.) *Proceedings of the 24th Annual Conference of the European Association for Machine Translation* (pp. 499–500). Tampere, Finland: European Association for Machine Translation (EAMT).

Using Large Language Models to build an efficient Adaptive Machine Translation System

Andrzej Zydroń
XTM International
azydron@xtm.cloud

Rafał Jaworski
XTM International
rjaworski@xtm.cloud

Szymon Kaczmarek
XTM International
skaczmarek@xtm.cloud

Abstract

The last two decades have seen significant advances in machine translation (MT). Driven by access to very large volumes of bilingual data and access to ever more powerful computer resources computer scientists have gradually improved the quality and fluency of MT from the initial experiment with Statistical MT through Neural MT and finally Large Language Model (LLM) based MT. The goal of MT has always been to achieve human level quality in terms of accuracy, voice and fluency. Nevertheless, all MT is ultimately restricted by the material available for training the engine/model. If no material is available a priori then by inference no training is possible. Training is also very expensive and time-consuming and the MT engine/LLM model is out of date as soon as training starts as it will not include any subsequent data. This problem is further amplified when dealing with poorly resourced languages.

We would like to propose an alternative approach when using Large Language Models for MT, one which allows the system to 'learn on the fly'. Building on prior research, this work proposes an enhanced method of using LLM models to create a complete Adaptive MT system using advanced LLM-based Retrieval Augmented Generation combined with dynamic Translation Memory (TM).

1 Introduction

Adaptive MT is a form of MT that uses continuous feedback during the course of a translation to learn from the translation to date and to translate subsequent text segments using the previously translated segments as examples (Moslem et al., 2023). Extensive research and experiments conducted by the ADAPT Centre Department of Computing have shown that such an approach results in significantly higher translation quality over both zero shot and trained Neural MT and LLM-based generated translation. Adaptive Machine Translation using fuzzy matches, (Moslem et al., 2023) significantly improves translation quality and the purpose of this paper is to propose an integrated production workflow architecture that provides an effective and efficient environment for implementing such a system.

2 Large Language Models

The introduction of Large Language Models (LLMs) has significantly changed the field of artificial intelligence, especially in how computers understand and generate human language (Zydroń et al. 2023). At the heart of large language models (LLMs) lies the transformer architecture, a revolutionary framework introduced by Vaswani et al. (2017), which has become foundational for most leading language processing systems. The transformer's distinctive capability to process data sequences in parallel through self-attention mechanisms has enabled remarkable advancements in the generation, comprehension, and interpretation of human language at scale (Zydroń et al. 2023).

3 Vectorization and Embeddings

Vectorization is a fundamental process in the operation of large language models (LLMs), wherein textual data is transformed into numerical representations—specifically, vectors—allowing machine-learning algorithms to process and understand natural language efficiently. This process is critical for enabling LLMs to grasp linguistic patterns, semantic relationships, and contextual nuances. Below is a detailed description of vectorization in the context of LLMs, supported by academic references.

3.1 Text Representation

- Word and Contextual Embeddings: Traditional methods of vectorization involve generating word embeddings, where each word is represented as a dense vector in a high-dimensional space. This technique, exemplified by Word2Vec and GloVe, captures semantic relationships by positioning similar words closer together in the vector space (Mikolov et al., 2013; Pennington et al., 2014). More advanced models, such as BERT and GPT, utilize contextual embeddings that adjust the representation of a word based on its surrounding context, thereby addressing polysemy and providing richer semantic understanding (Devlin et al., 2018).

3.2 Tokenization

- Process of Tokenization: Before vectorization, input text undergoes tokenization, breaking it down into smaller units, such as words or subwords. This is crucial for managing diverse vocabularies efficiently. Byte Pair Encoding (BPE) is a common approach that segments words into subwords, allowing models to better handle rare or compound words (Sennrich et al., 2016). The tokenization process enables LLMs to represent various linguistic constructs in a manageable way.

3.3 Position Encoding

- Incorporating Sequential Information: In transformer architectures, positional encodings are integrated with token embeddings to retain information about the order of tokens in a sequence. Since transformers lack a built-in mechanism for processing sequential data, these positional encodings are essential for capturing the relationships between tokens (Vaswani et al., 2017). This aspect is vital for understanding context and maintaining coherence in language generation tasks.

3.4 Dimensionality and Computational Efficiency

- High-Dimensional Vectors: The vectors produced through vectorization typically have high dimensionality, often consisting of hundreds or thousands of dimensions. This high-dimensional representation facilitates the capture of complex linguistic features but also poses challenges regarding computational resources (BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018). The use of efficient matrix operations in LLMs allows for rapid processing of these vectors, enabling scalability and real-time language applications.

3.5 Impact on Model Performance

- Quality of Representations: The efficacy of vectorization has a direct impact on the overall performance of LLMs. High-quality vector representations enhance the model's ability to comprehend context, semantics, and intricate relationships within the language, leading to improved accuracy in various natural language processing tasks (Brown et al., 2020).

3.6 Applications and Use Cases

- Diverse Applications: Vectorization underpins numerous applications of LLMs, including sentiment analysis, chatbots, machine translation, and more. By converting text into numerical vectors, LLMs can effectively analyze, generate, and respond to human language in a coherent and contextually relevant manner (Radford et al., 2019).

In summary, vectorization is an essential process for transforming text into numerical formats that LLMs can understand, allowing them to perform a wide range of natural language processing tasks effectively.

4 Vector Stores

Vector stores, also known as vector databases or vector search engines, are specialized storage systems designed to manage and retrieve high-dimensional vector representations of data efficiently. These systems are increasingly important in machine learning and natural language processing applications, particularly in the context of large language models (LLMs), where embeddings play a crucial role. Below is a detailed description of vector stores, accompanied by relevant academic references.

4.1 Definition and Purpose

Vector stores are optimized for storing, indexing, and querying vectors, which are numerical representations of data points in a continuous vector space. They enable fast similarity search and retrieval operations based on distance metrics, such as Euclidean distance or cosine similarity. The primary purpose of vector stores is to facilitate efficient handling of high-dimensional data, especially in applications involving embeddings from machine learning models (Lin et al., 2021).

4.2 Architecture

- Data Structure: Vector stores typically use specialized data structures like KD-trees, ball trees, or approximate nearest neighbor (ANN) algorithms to organize and index vectors. These structures enable efficient querying, allowing systems to quickly locate vectors that are similar to a given input vector (Kleinberg, 2003).
- Scalability: Modern vector stores are designed to scale horizontally, accommodating large volumes of data and high-dimensional vectors. They can handle millions or even billions of vectors, making them suitable for large-scale applications (Zhang et al., 2020).

4.3 Indexing Techniques

- Approximate Nearest Neighbor Search: given the high dimensionality of vectors, exact nearest neighbor searches can be computationally expensive. Vector stores often implement ANN algorithms to provide fast, approximate results with a trade-off in accuracy. Techniques such as locality-sensitive hashing (LSH) and Product Quantization (PQ) are commonly used for this purpose (Indyk & Motwani, 1998; Jegou et al., 2011).
- Hierarchical Clustering: some vector stores utilize hierarchical clustering methods to group similar vectors, which helps improve search efficiency by reducing the number of comparisons needed during retrieval (Berkhin, 2006).

4.4 Use Cases

Vector stores are particularly useful in applications involving:

- Semantic Search: by storing embeddings from LLMs, vector stores enable semantic search capabilities, allowing users to retrieve documents or data points that are contextually relevant rather than merely keyword-based (Karpukhin et al., 2020).
- Recommendation Systems: vector stores can power recommendation systems by storing user and item embeddings, facilitating fast retrieval of similar items based on user preferences (Wang et al., 2018).
- Image and Video Retrieval: in computer vision, vector stores can manage feature embeddings from images or videos, enabling efficient content-based retrieval (Babenko et al., 2014).

4.5 Integration with Machine Learning Models

Vector stores are often integrated with machine learning pipelines, where embeddings generated by models (such as BERT or other LLMs) are stored for downstream tasks. This integration allows real-time querying and retrieval of relevant data based on user inputs or model outputs, facilitating interactive applications (Rusu et al., 2019).

4.6 Challenges and Future Directions

While vector stores provide significant advantages, they also face challenges such as:

- Dimensionality Curse: as the dimensionality of vectors increases, the performance of traditional distance metrics can degrade, leading to less effective retrieval (Bellman, 1961).
- Storage Efficiency: managing storage requirements for large volumes of high-dimensional vectors is crucial for maintaining performance and cost-effectiveness (Zhang et al., 2020).
- Integration with Other Data Types: future developments may focus on enhancing the capabilities of vector stores to integrate with structured data, enabling more comprehensive data management solutions.

Conclusion
In summary, vector stores are specialized systems designed for the efficient management and retrieval of high-dimensional vector data. Their integration with machine learning models and applications in semantic search, recommendation systems, and multimedia retrieval highlight their importance in modern data processing and analysis. As the field evolves, addressing challenges related to dimensionality and storage efficiency will be crucial for the continued advancement of vector databases.

5 LLMs as MT engines

The best LLM Models perform very well in translation tasks, equalling the quality of output from the best Neural MT engines (Inten.to 2024). LLM MT can also be significantly cheaper that Neural MT, although it can also be much slower, presently, at generating output. In addition, LLM models do not exhibit any tendency to hallucination when doing translation tasks.

Comparing the machine translation capabilities of Neural Machine Translation (NMT) systems with those of Large Language Models (LLMs) reveals both distinct advantages and disadvantages for each approach. Below, I provide an overview of both technologies, their comparative capabilities, and the pros and cons of each, supported by academic references.

5.1 Neural Machine Translation (NMT)

NMT refers to a class of machine translation systems that utilize deep learning techniques, specifically neural networks, to translate text from one language to another. The most notable architectures in NMT are encoder-decoder models with attention mechanisms.

Capabilities:

- Contextual Understanding: NMT systems, particularly those using attention mechanisms, excel at capturing contextual relationships between words in a sentence, allowing for more fluent and coherent translations (Bahdanau et al., 2015).

- End-to-End Training: NMT models are trained end-to-end on large parallel corpora, optimizing the translation process directly without the need for intermediate linguistic rules (Koehn & Knowles, 2017).

Pros of NMT:

- Fluency and Coherence: NMT systems tend to produce more fluent and natural-sounding translations compared to earlier statistical methods (Wu et al., 2016).
- Handling of Long-Range Dependencies: attention mechanisms allow NMT to manage long-range dependencies effectively, improving the quality of translations for complex sentences (Vaswani et al., 2017).
- Adaptability: NMT models can be fine-tuned on domain-specific corpora, which enhances their performance in specialized fields (Freitag & Al-Onaizan, 2016).

Cons of NMT:

- Data Dependency: NMT systems require large amounts of high-quality parallel data for effective training, which can be a limitation for low-resource languages (Koehn & Knowles, 2017).
- Inability to Generalize Beyond Training Data: NMT may struggle with phrases or contexts not seen during training, leading to less accurate translations for novel sentences, the so-called 'out of context word (OOCW) problem. This may lead to the system either continually repeating the OOCW or picking a word at random for the translation depending on the system settings (Bertoldi & Federico, 2012).

5.2 Large Language Models (LLMs)

LLMs are advanced models trained on vast corpora of text data to perform a variety of tasks, including translation. They use architectures like transformers, which enable them to learn contextual relationships across vast amounts of data.

Capabilities:

- Contextual and Cross-Domain Understanding: LLMs can understand and generate text across different domains and languages, leveraging their training on diverse datasets (Brown et al., 2020).
- Few-Shot and Zero-Shot Learning: LLMs can perform translation tasks with minimal or no task-specific training, adapting to new languages or styles based on prompt engineering (GPT-3) (Radford et al., 2019).

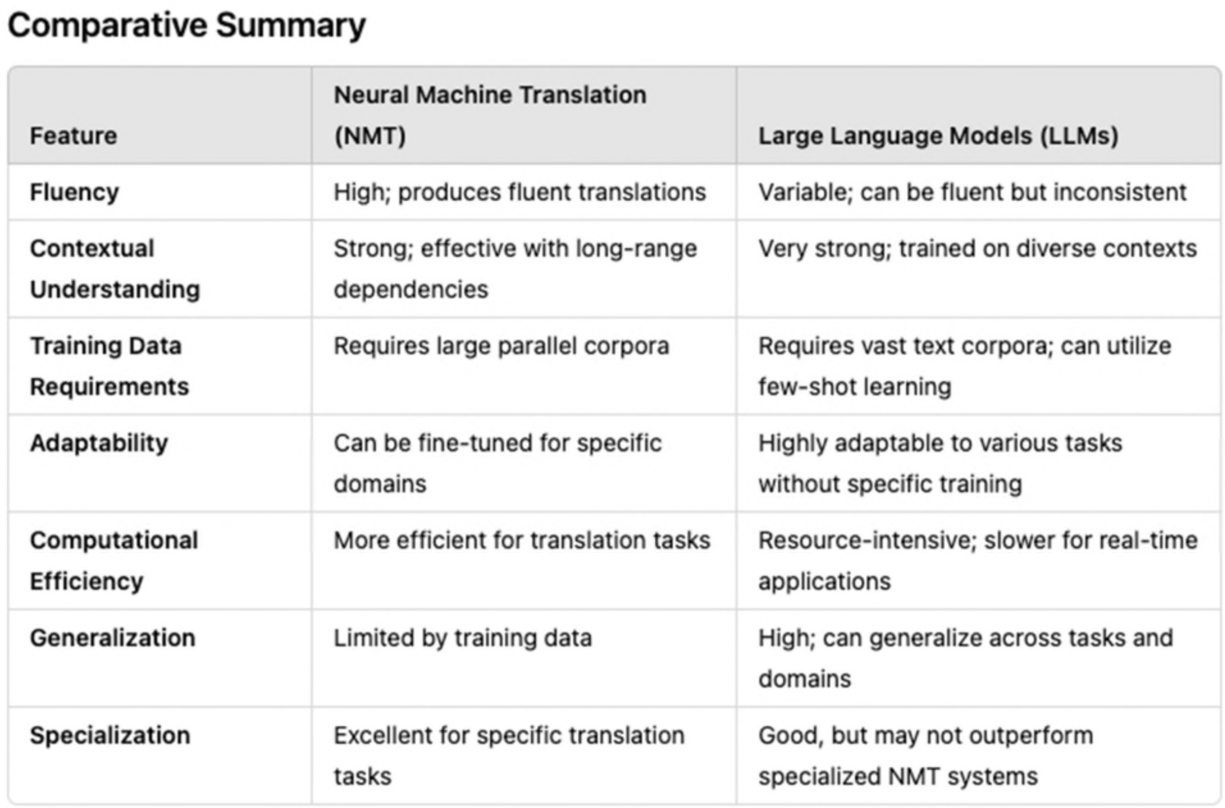
Pros of LLMs:

- Versatility: LLMs are not limited to translation; they can perform a wide range of language tasks, including summarization, question answering, and content generation (Brown et al., 2020).

- Large-Scale Knowledge: The extensive training data allows LLMs to incorporate world knowledge, cultural nuances, and idiomatic expressions into translations (Bhatia et al., 2021).
- Reduced Data Requirement: LLMs can leverage few-shot learning to perform translations even with limited examples, making them useful for low-resource languages (Brown et al., 2020).

Cons of LLMs:

- Inconsistent Quality: While LLMs can produce high-quality translations, they can also generate erratic or nonsensical outputs, especially for complex sentences or specific contexts (Webb et al., 2022).
- Resource Intensive: Training and running LLMs require significant computational resources, which can be a barrier for deployment in some settings (Kaplan et al., 2020).
- Lack of Fine-Tuning for Specific Tasks: While LLMs can generalize well, they may not perform as effectively as specialized NMT systems when fine-tuned for specific translation tasks (Liu et al., 2021).



Comparative Summary

Feature	Neural Machine Translation (NMT)	Large Language Models (LLMs)
Fluency	High; produces fluent translations	Variable; can be fluent but inconsistent
Contextual Understanding	Strong; effective with long-range dependencies	Very strong; trained on diverse contexts
Training Data Requirements	Requires large parallel corpora	Requires vast text corpora; can utilize few-shot learning
Adaptability	Can be fine-tuned for specific domains	Highly adaptable to various tasks without specific training
Computational Efficiency	More efficient for translation tasks	Resource-intensive; slower for real-time applications
Generalization	Limited by training data	High; can generalize across tasks and domains
Specialization	Excellent for specific translation tasks	Good, but may not outperform specialized NMT systems

Figure 1, Comparative summary of Neural Machine Translation and Large Language Models tasked with translation

In summary, both Neural Machine Translation systems and Large Language Models have their strengths and weaknesses. NMT systems excel in producing fluent translations for specific domains when ample parallel data is available, while LLMs offer versatility and the ability to adapt to various tasks with fewer constraints on training data. The choice between the two approaches largely depends on the specific requirements of the translation task, the available resources, and the desired outcomes.

6 Improving LLM MT performance

Several different techniques can be used to improve the quality of LLM MT output:

- Custom models
- Few-shot learning
- Retrieval Augmented Generation (RAG)

6.1 Custom Models

Parallel bilingual corpora can be used to train a specific version of a general LLM model. This process can be very time consuming and expensive and the custom model can be expensive to deploy. The output quality does not appear to provide a significant improvement in the quality of the translation without further fine tuning. The fundamental problem with this approach is that the data is out of date the moment the training begins because it does not include any subsequent translation data.

6.2 Few-Shot Learning

Few-shot learning enhances the capabilities of large language models (LLMs) by providing them with multiple examples, which helps the models grasp the underlying patterns or rules of a task more effectively. This approach leverages the extensive pre-trained knowledge of LLMs, enabling them to perform specific tasks with minimal additional input, thereby improving deployment speed and efficiency.

The effectiveness of few-shot learning relies heavily on the selection and arrangement of examples, which need to be representative and informative (Zhao et al., 2021; Lu et al., 2022). However, the model’s accuracy can vary based on the prompt used, as it tends to favor recent tokens and may repeat answers from the end of the prompt. A recommended practice is to place content-free input last in the prompt to mitigate this bias.

Additionally, LLMs exhibit Majority Label Bias, where they favor frequent responses in the prompt due to an unbalanced training set. A study (Min et al., 2022) indicated that randomly assigned labels in training examples do not adversely affect performance in classification and multiple-choice tasks; instead, the distribution of the input text and the overall format of the sequence are more critical for effective outcomes.

6.3 Retrieval Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) is an innovative approach in the field of natural language processing that combines the strengths of retrieval-based methods with generative models. This technique is particularly relevant for large language models (LLMs), as it enhances their ability to generate accurate and contextually relevant responses by integrating external knowledge sources. Below is a detailed description of RAG in LLMs, supported by academic references.

6.3.1 Definition of RAG

Retrieval-Augmented Generation (RAG) refers to a hybrid framework that combines a retrieval component with a generative model. In this approach, relevant information is retrieved from a large external knowledge base or corpus before generating a response. This allows the model to produce more informed and contextually accurate outputs, particularly for tasks that require factual knowledge or specialized information (Lewis et al., 2020).

6.3.2 Architecture of RAG

RAG typically involves two main components:

- Retrieval Component: this component is responsible for searching and retrieving relevant documents or passages from a large corpus based on the input query. This process often employs techniques such as dense retrieval methods, which use embeddings to find semantically similar documents quickly.
- Generative Component: after retrieving relevant information, the generative model (often based on transformer architecture) processes this information along with the original query to generate a coherent and contextually appropriate response. The generative model is fine-tuned to incorporate the retrieved context effectively, allowing it to produce accurate and informative outputs (Karpukhin et al., 2020).

6.3.3 Advantages of RAG

- Enhanced Knowledge Utilization: by incorporating external knowledge, RAG allows LLMs to overcome limitations associated with their pre-trained knowledge, particularly when dealing with facts or details not present in the training data. This capability is especially beneficial for domains requiring up-to-date or specific information (Guu et al., 2020).
- Improved Accuracy and Contextuality: the integration of relevant documents enables the model to ground its responses in factual data, reducing the likelihood of hallucinations—where models generate inaccurate or misleading information (Garncarek et al., 2022).
- Efficiency: RAG can improve the efficiency of LLMs by allowing them to focus on generating responses based on relevant information rather than relying solely on their internal knowledge, which may be outdated or incomplete.

6.3.4 Applications of RAG

RAG models have been applied to various tasks, including:

- Question Answering: RAG significantly enhances question-answering systems by retrieving relevant passages that inform the generated answers, leading to more accurate and context-aware responses (Guu et al., 2020).
- Dialogue Systems: in conversational AI, RAG can provide more informative and relevant responses by retrieving contextually appropriate data, improving the overall user experience (Karpukhin et al., 2020).
- Content Generation: RAG can be utilized for generating articles or reports by retrieving relevant data and synthesizing it into coherent text, thereby enhancing the richness and accuracy of the generated content.

6.3.5 Challenges and Future Directions

While RAG offers significant advantages, it also faces challenges:

- Retrieval Quality: the performance of RAG models heavily depends on the quality and relevance of the retrieved documents. Poor retrieval can lead to inaccurate or irrelevant responses (Karpukhin et al., 2020).
- Scalability: as the size of the knowledge base grows, efficient retrieval methods become increasingly important to maintain responsiveness in real-time applications.
- Integration Complexity: balancing the retrieval and generation components effectively requires careful tuning and architecture design to ensure that the model can integrate external information seamlessly (Lewis et al., 2020).

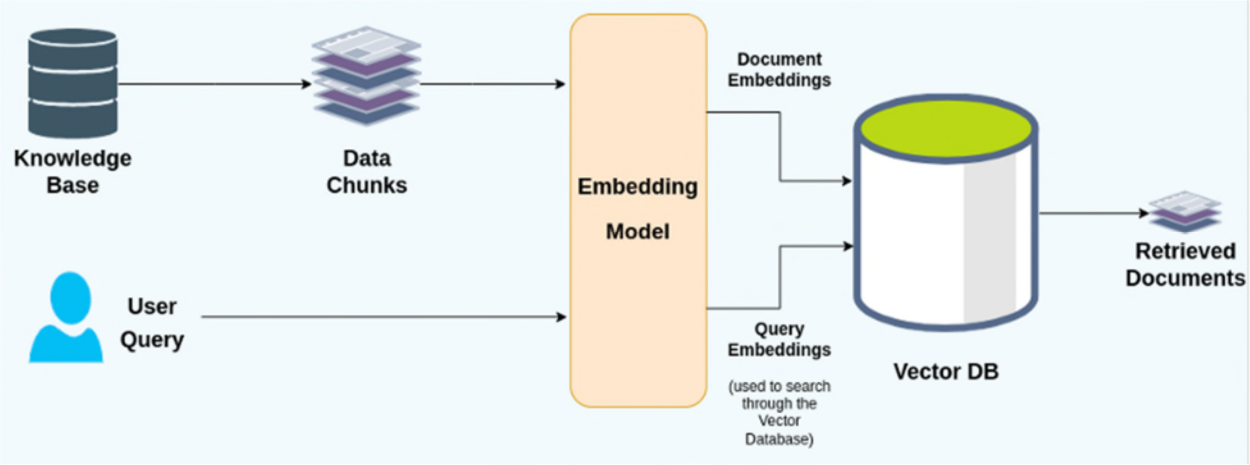


Figure 2, RAG using Vector store.

Conclusion

Retrieval-Augmented Generation (RAG) represents a powerful approach that combines the strengths of retrieval and generative models, enhancing the capabilities of large language models in producing accurate and contextually relevant outputs. By effectively integrating external knowledge, RAG models improve performance across various NLP tasks, paving the way for more informed and responsive applications in real-world scenarios.

7 Translation Memory

Translation Memory (TM) is a critical tool in the field of translation and localization, designed to improve efficiency, consistency, and accuracy in the translation process. It is essentially a database that stores previously translated segments (e.g., sentences or phrases) alongside their source text counterparts. The primary goal of a translation memory is to aid translators by suggesting translations for new content that matches or closely resembles existing entries in the TM.

Key Features of Translation Memory

1. Segmentation: TM systems typically break down source text into smaller segments, such as sentences or clauses, which can be easily matched against existing translations (Saldanha & O'Brien, 2013). This segmentation is crucial for effective retrieval of translations.

2. Fuzzy Matching: TM tools employ fuzzy matching algorithms to find partial matches between the source text and previously translated segments. This allows for suggestions even when the new text is not identical to any existing entry, enhancing the translator's productivity (Hansen, 2010).

3. Consistency and Quality: by reusing previously validated translations, TMs promote consistency across projects, particularly important in technical fields where terminology and phrasing must remain uniform (Doherty & Moorkens, 2017).

4. Cost-Effectiveness: utilizing a TM can significantly reduce translation costs and time, as translators can focus on new or modified segments rather than re-translating familiar content. This efficiency is particularly beneficial in large-scale localization projects (Bowker & Pym, 2006).

Applications of Translation Memory

Translation memory systems are widely used in various domains, including legal, technical, and medical translations, where accuracy and consistency are paramount. They are also essential in localization efforts for software and websites, helping to maintain coherent language use across different platforms and languages (Fowler, 2017).

8 Bilingual Terminology Databases

Bilingual terminology databases are specialized resources that store terms and their equivalents in two languages, facilitating the translation process and promoting consistency in terminology usage across various fields. These databases serve as essential tools for translators, terminologists, and localization professionals, particularly in technical, legal, and scientific domains.

Key Features of Bilingual Terminology Databases

1. Structured Organization: bilingual terminology databases typically organize entries in a structured format, allowing for easy retrieval of terms along with their definitions, context, and usage examples. This structure aids users in understanding the nuances of each term (Garrido et al., 2013).

2. Standardization: these databases promote standardization of terms across languages, ensuring that specific terminology is used consistently. This is particularly crucial in fields where precise language is essential, such as medicine, law, and engineering (Drouin, 2003).

3. Contextual Information: many bilingual terminology databases provide contextual information, including grammatical details, synonyms, and usage notes. This additional context helps translators choose the most appropriate term based on the specific context in which it is used (Pérez & Ceballos, 2020).

4. Integration with Translation Tools: bilingual terminology databases can be integrated with translation memory systems and computer-assisted translation (CAT) tools, enhancing translators' workflows by providing instant access to relevant terminology during the translation process (Rogers, 2019).

Applications of Bilingual Terminology Databases

Bilingual terminology databases are widely used in various sectors, including translation and localization services, academic research, and multilingual communication. They play a crucial role in ensuring accuracy and consistency in translations, especially for specialized content where terminology may vary significantly between languages (Baker, 1992).

Conclusion

In summary, bilingual terminology databases are vital resources that enhance the quality and efficiency of translation processes. By providing standardized, context-rich terminology, these databases support translators in producing accurate and consistent translations across languages.

9 Putting it all together: LLM based Adaptive MT

Moslem et al., 2023, describes how using an LLM model and few-shot examples with an appropriately formulated prompt can produce MT output that surpasses that of the very best Neural MT engines. In this paper we build on this work to propose a fully automated and integrated Adaptive MT system that can be used in a production environment.

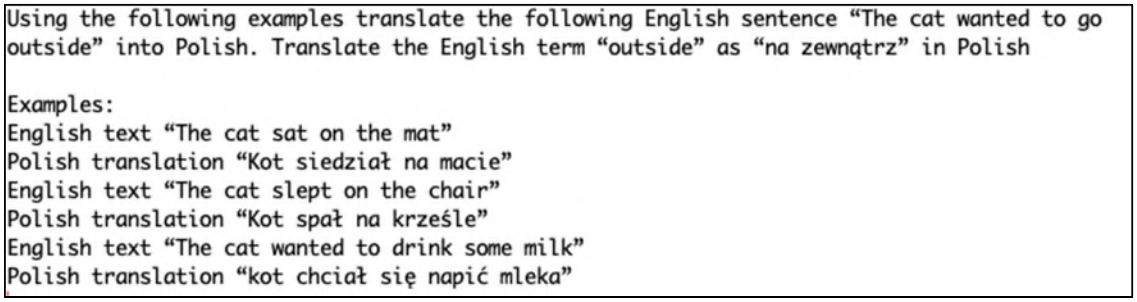
We have described the key components that can be used to create an Adaptive MT system:

- Large Language Model (LLM)
- Vectorization and Embeddings
- Vector Stores
- Translation Memory (TM)
- Bilingual Terminology

These components can be integrated into a complete system.

The proposed Adaptive MT system functions as follows:

1. All segments in the Translation memory are vectorized and the embedding values are stored in a Vector Store and linked to the Translation Memory record.
2. All new translations are immediately stored in both the Translation Memory and the Vector Store.
3. When a new source segment is presented to the system check to see:
    a. if a direct match already exists in the Translation Memory. If it does then use the Translation Memory match: STOP
    b. If a similar match, using Levenshtein distance on the source segments, is found then create an appropriate prompt using the ‘fuzzy’ source and target segments and the new segment and submit to the LLM for translation: STOP
    c. If a close semantic match is found in the Vector Store, then retrieve all of the close matches and any terminology found in the Terminology database is noted, create an appropriate prompt that includes the translation examples and the required term translation and submit to the LLM for translation (effective RAG): STOP
    d. If no close semantic matches are found then any distant or, failing that, random examples are retrieved and any terminology found in the Terminology database is noted, create an appropriate prompt that includes the translation examples and the required term translation and submit to the LLM for translation instructing the model to mimic the style and voice of the few-shot examples (effective RAG): STOP



Using the following examples translate the following English sentence “The cat wanted to go outside” into Polish. Translate the English term “outside” as “na zewnątrz” in Polish

Examples:
English text “The cat sat on the mat”
Polish translation “Kot siedział na macie”
English text “The cat slept on the chair”
Polish translation “Kot spał na krześle”
English text “The cat wanted to drink some milk”
Polish translation “kot chciał się napić mleka”

Figure 3, few-shot RAG example prompt for translation

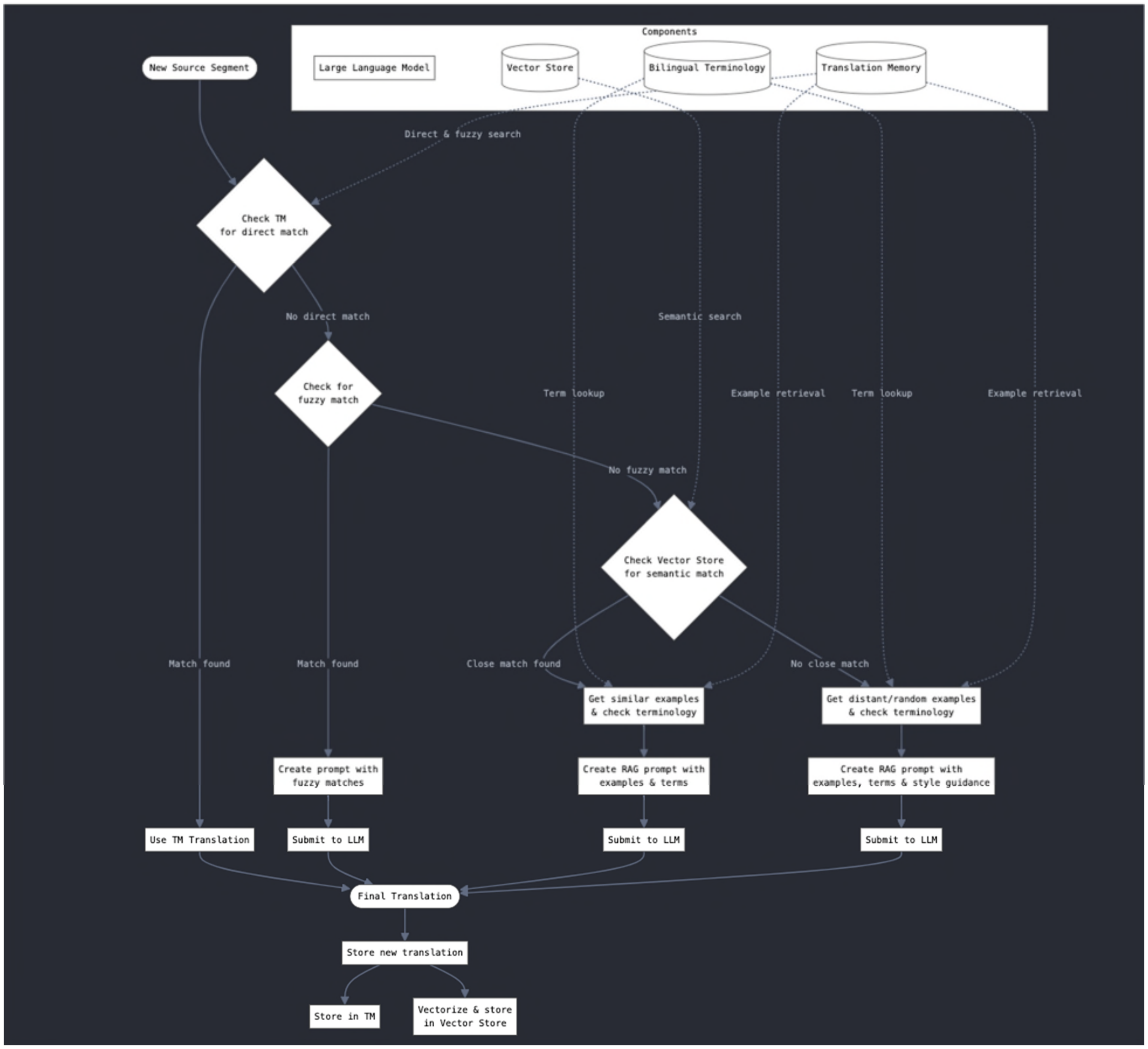


Figure 4, Adaptive MT flow diagram

References

Babenko, A., Slesarev, A., Chigorin, A., & Yan, T. (2014). Neural Codes for Image Retrieval. *arXiv preprint arXiv:1412.2306*.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*.

Baker, M. (1992). *In Other Words: A Coursebook on Translation*. Routledge.

Bellman, R. (1961). Adaptive Control Processes: A Guided Tour. *Princeton University Press*.

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. *In Grouping Multidimensional Data*, 25-71.

Bertoldi, N., & Federico, M. (2012). Domain Adaptation for Statistical Machine Translation. *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, 163-172.

Bhatia, A., Gupta, N., & Kaur, S. (2021). A Survey on the Impact of Transformer Models on Machine Translation. *Journal of Ambient Intelligence and Humanized Computing*, 12(3), 2801-2814.

Bowker, L., & Pym, A. (2006). *Data-Driven Translation: A Translation Memory Approach*. In *Translation Technology and its Role in Language Services*. *Translators' Journal*, 151(1), 127-141.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Doherty, S., & Moorkens, J. (2017). The Impact of Translation Memory on Translation Quality. *Translation Studies*, 10(2), 189-203.

Drouin, P. (2003). Terminology and Translation: A Multidisciplinary Approach. *Translation Studies*, 5(1), 57-73.

Fowler, C. (2017). *Translation Memory and Its Role in Localization: An Overview*. *The Journal of Internationalization and Localization*, 4(2), 176-192.

Freitag, M., & Al-Onaizan, M. (2016). Fast, Accurate Domain Adaptation for Neural Machine Translation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 10-15.

Garncarek, P., Stokes, J., & Stojanovic, J. (2022). Analyzing Hallucinations in Generative Models: A New Approach to Language Model Evaluation. *arXiv preprint arXiv:2212.00551*.

Garrido, M. J., et al. (2013). Creating a Bilingual Terminology Database for Specialized Translation. *Terminology*, 19(1), 49-65.

Guu, K., Khatri, C., Pasupat, P., & Tung, Z. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Proceedings of the 37th International Conference on Machine Learning*, 119, 3921-3930.

Hansen, I. (2010). *Translation Memory: The State of the Art*. In *Translation and Interpreting Studies*, 5(2), 253-272.

Indyk, P., & Motwani, R. (1998). Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. *Proceedings of the 30th Annual ACM Symposium on Theory of Computing*, 604-613.

Inten.to: The State of Machine Translation 2024, https://inten.to/machine-translation-report-2024

Jegou, H., Delaunay, J., & Grangier, D. (2011). Product Quantization for Nearest Neighbor Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11), 2172-2182.

Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*.

Karpukhin, V., Oguz, B., Yin, W., & Yih, W. T. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *arXiv preprint arXiv:2004.04906*.

Karpukhin, V., Oguz, B., Yin, W., & Yih, W. T. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *arXiv preprint arXiv:2004.04906*.

Kleinberg, J. (2003). An Impossibility Theorem for Clustering. *In Advances in Neural Information Processing Systems*, 15.

Koehn, P., & Knowles, R. (2017). Six Challenges for Neural Machine Translation. *Proceedings of the First Workshop on Neural Machine Translation*, 28-39.

Lewis, P., Liu, T., Goyal, N., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive Tasks. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 955-964.

Lin, J., Yang, K., Zhao, Y., & Wang, Y. (2021). A Survey of Vector Database: A New Paradigm of Big Data Management. *IEEE Transactions on Big Data*, 8(2), 364-377.

Liu, C., Hu, Z., & Gao, X. (2021). Fine-tuning Pre-trained Language Models: Weight Initializations, Data Orders, and Early Stopping. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1855-1865.

Lu et al., (2022). Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity, arXiv:2104.08786v2 [cs.CL]

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, 26.

Min et al., (2022). Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?, arXiv:2202.12837v2 [cs.CL]

Moslem et al., (2023). Adaptive Machine Translation with Large Language Models, arXiv:2301.13294v3 [cs.CL]

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543.

Pérez, E., & Ceballos, M. (2020). The Role of Bilingual Terminology Databases in Enhancing Translation Quality. *Journal of Language and Translation Studies*, 7(2), 93-107.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI GPT-2 Report*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI GPT-2 Report*.

Rogers, M. (2019). The Integration of Terminology Management and Translation Memory Systems: Current Practices and Future Trends. *International Journal of Translation Studies*, 31(3), 213-230.

Rusu, A. A., et al. (2019). Incremental Learning with Support Vector Machines. *In Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2736-2742.

Saldanha, G., & O'Brien, S. (2013). *Research Methodologies in Translation Studies*. *Routledge*.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1715-1725.

Vaswani, A., Shard, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Kaiser, Ł. (2017). Attention is All You Need. *Advances in Neural Information Processing Systems*, 30.

Vaswani, A., Shard, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Kaiser, Ł. (2017). Attention is All You Need. *Advances in Neural Information Processing Systems*, 30.

Wang, H., et al. (2018). The Role of Neural Networks in Recommendation Systems. *In Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 3-8.

Webb, S., et al. (2022). An Empirical Study of the Performance of Large Language Models in Translation. *In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 101-111.

Wu, Y., et al. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*.

Zhang, H., et al. (2020). A Survey of Vector Databases: A Novel Paradigm for Similarity Search. *IEEE Transactions on Knowledge and Data Engineering*, 33(5), 2336-2354.

Zhao et al., (2021). Calibrate Before Use: Improving Few-Shot Performance of Language Models, arXiv:2102.09690v2 [cs.CL]

Zydroń, A., Jaworski, R., Kaczmarek, S. Large Language Models – origin, growth and capabilities, ASLING TC45 2023

Bios
Keynote speakers, Authors and Presenters

Anwar Alfetlawi is a PhD student at the University of Ottawa, School of Translation and Interpretation. He is also an experienced freelance translator and an ESL instructor. His main research interests include simultaneous interpretation, translation technology, and the integration of educational technology in ESL classes.

Maria Aretoulaki has been working in AI and Machine Learning for the past 30 years: NLP, NLU, Artificial Neural Networks, Speech Recognition, Machine Translation, Text Summarisation, Voice and Chat Conversational Experience Design. She has focused mainly on natural language conversational human-machine interfaces, mainly for Contact Centre applications across all the main verticals for organisations worldwide. She has designed, built and optimised conversational Speech IVRs, Voicebots, Chatbots, Voice and Digital Assistants in several languages for large Enterprises worldwide and Government. In 2018, Maria coined the term “Explainable Conversational Experience Design”, which later morphed to “Explainable Conversational AI” and more recently – with the explosion of LLMs and the ChatGPT hype – to “Explainable Generative AI” to advocate for transparent, responsible, design-led AI bot development that keeps the human in the loop and in control.
Having run her own company for 14 years, Maria joined GLOBALLOGIC / HITACHI in 2022, where she has helped shape and leads the company group’s Explainable Conversational AI and Responsible Generative AI Design strategy. She focuses on Language and Prompt Engineering, Prompt and Chatbot Design, Ontology Design, Knowledge and Content Engineering, as well as AI Ethics and AI Law. She has been working on Hitachi’s GenAI Platform and the Hitachi AI Policy Committee and is also a Hitachi Design Community Ambassador.
Maria holds a Post-Doc in Conversational AI, a PhD in Neural Networks for Text Summarisation, an MSc in Machine Translation and a B.A. Hons in Linguistics & English.

Ahmed Elhuseiny Bedeir is a PhD candidate at the University of Ottawa, School of Translation and Interpretation, and an English-Arabic translator certified by the American Translators Association and the Association of Translators and Interpreters of Ontario. His research interests include translation technology teaching and the use of translation in foreign language classes.

Haifa Ben Naji is a PhD candidate at the University of Ottawa, School of Translation and Interpretation. Her research interests include terminology management in commercial environments, localization, and translation technology teaching.

Anastasiia Bezobrazova has recently graduated from the prestigious European Masters in Technologies for Translation and Interpreting (EM TTI) Programme at University of Malaga and New Bulgarian University with a dissertation about the use of large language models (LLMs) for terminology extraction. She has a bachelor degree in linguistics from Moscow Pedagogical State University where her dissertation focused on the difficulties of translating English terms in the field of design and fashion. Anastasiia’s current research interests are related to the use of LLMs for making texts more accessible for particular categories of users.

Bhavani Bhaskar has worked as a Japanese-English translator for five years and holds a Master’s degree in Computational Linguistics from the University of Stuttgart, where she specialized in machine learning and language processing technologies.
For the past six years, she has been a researcher and developer with the European Commission, focusing on cutting-edge advancements in NLP and AI. Her work has been instrumental in machine translation and generative AI projects, particularly within the high-performance computing (HPC) systems of EuroHPC. She has played a key role in the development and training of AI models, contributing to the future of multilingual communication and AI-driven solutions.

Romane Bodart is a teaching and research assistant in English-to-French translation at UCLouvain (Belgium), where she teaches legal translation at master’s level. Her PhD thesis focuses on post-editing (PE), with special emphasis on PE quality, and PE training in translator education. She contributed to the POST EDIT ME! pedagogical project and the development of MTPEAS (Machine Translation Post-Editing Annotation System).

Pierrette Bouillon has been Professor at the Faculty of Translation and Interpreting(FTI), University of Geneva since 2007. She is currently Director of the Department of Translation Technology (referred to by its French acronym TIM) and Dean of the FTI. She has numerous publications in computational linguistics and natural language processing, particularly within speech-to-speech machine translation, accessibility and pre-editing/post-editing.

Cristian Brașoveanu is a team leader for AI-related projects at the European Commission's Directorate-General for Translation (DGT). DGT is the provider of the Commission's flagship eTranslation neural machine translation service, and of new services such as eBriefing, eSummary, and eReply.
He has coordinated a number of projects, including the development of new services based on large language models for generative AI and the new supercomputing projects aimed at leveraging DGT's high quality multilingual data and the EuroHPC network of European supercomputers to contribute to better multilingual AI for Europe.
He also coordinates the AI@EC Network, the European Commission's internal corporate AI community. It is a community of both business users and technical experts, focusing on identifying AI needs, facilitating AI explorations and projects, and sharing and building up AI knowledge across the Commission.
Before, he worked as policy and programme officer at DG CONNECT (Communications Networks, Content and Technology), including in the area of AI and robotics. He was involved in work relating to initiatives such as the first Commission Communication on Artificial Intelligence for Europe, the High Level Expert Group on Artificial Intelligence, and the European AI Alliance. In the past, he also worked as a translator at DGT, in the English Language Department. That work involved translating into English a broad range of types of document covering all areas of EU policy. He also promoted and contributed to the development of new tools to facilitate the translation workflow in the Department.

Sabine Braun is Professor of Translation Studies and Director of the Centre for Translation Studies at the University of Surrey, UK, a Research-England funded Centre of Excellence. She also serves as a Co-Director of Surrey's Institute for People-Centred AI. She specialises in human-machine integration in translation and interpreting to improve access to information, digital content and public services. For over a decade, she spearheaded a European-funded research programme investigating video-mediated interpreting in legal proceedings to improve language access in the justice sector (AVIDICUS 1-3; 2008-16), while contributing her expertise in video interpreting to other justice sector projects (e.g. QUALITAS, 2012-14; Understanding Justice, 2013-16; VEJ Evaluation, 2018-20). Subsequently she advised justice sector institutions on the use and risks of video-mediated interpreting, delivered training, developed European guidelines, and co-authored a DIN standard. She has also explored the use of video and virtual reality platforms for training interpreters and users of interpreting services (IVY, 2011-3; EVIVA, 2014-15; SHIFT, 2015-18; EU-WEBPSI, 2021-24) and is currently involved in projects investigating the application of communication technologies and AI-enabled language technologies to multilingual health communication (MHealth4All, 2021-24; Interpret-X, 2021-24). Furthermore, she conducts research on audio description and other translation modalities related to accessible communication. In the Horizon 2020 project MeMAD (2018-21), she explored the potential for (semi-)automating audio description to enhance digital media inclusion. In 2024, she launched a Leverhulme Trust-funded Doctoral Training Network on AI-Enabled Digital Accessibility (ADA). Her overarching interest centres on fairness, transparency, and quality in the use of technology in language mediation.

Romana Cacija Since 2000, I have taught at the Faculty of Humanities and Social Sciences in Osijek (Croatia), where I work to advance students' understanding of English and translation studies. I instruct undergraduate courses in Contemporary English Language and a graduate course in Literary Translation. In my Contemporary English Language classes, I emphasize advanced linguistic structures, practical language application, and cultural context. At the graduate level, I developed the Literary Translation course, guiding students through the nuances of translating literary texts. In addition to key linguistic challenges, this course equips students with the skills to navigate cultural nuances, figurative language, and stylistic variation, preparing them for the complexities of professional literary translation.
My research focuses on figurative language and cultural specificity in translation, particularly within young adult fiction. As a doctoral researcher in Linguistics, I study the challenges of translating figurative language, including slang, colloquial and other contemporary expressions, from English to Croatian. The research explores how translators adapt figurative language to engage Croatian youth, balancing accessibility with the text's style and tone. Additionally, it investigates the role of anglicisms and other loanwords, analyzing when and why they are retained or adapted. My

research assesses the impact of these language choices on the authenticity and readability of the translated text, contributing to a broader understanding of translation as a bridge between linguistic and cultural contexts.
To stay at the forefront of translation studies, I attend a number of conferences and workshops, both locally and internationally. The Translating Europe Forum in Brussels and the PETRA-E Network (Plateforme Européenne pour la Traduction Littéraire) have been instrumental in shaping my approach to translation in an increasingly globalized world. These events, along with other conferences on linguistics and literary translation, provide valuable insights into emerging trends and best practices. I integrate these insights into my teaching and research, ensuring that students benefit from the latest developments in translation studies.
Beyond academia, I am an active literary translator, collaborating with Croatian publishers on various literary genres, with a recent focus on young adult fiction. Over the years, I have translated 15 fiction and non-fiction titles, continually refining my translation skills. This professional work allows me to apply academic principles in real-world contexts, balancing fidelity to the source text with a nuanced approach to style and cultural resonance. Additionally, I work with a number of agencies and institutions on diverse translation projects, further broadening my expertise beyond the literary field.
With advanced proficiency in English and a specialized expertise in translation, research, and education, I bring a unique blend of linguistic insight and practical experience to every project. My extensive background in both linguistics and literary translation enables me to provide nuanced and culturally sensitive translation services across diverse contexts. Looking ahead, I am passionately committed to advancing the field of literary translation through both my research and teaching. By making impactful literature accessible to Croatian audiences and inspiring the next generation of translators, I aim to contribute meaningfully to the evolving landscape of translation studies, strengthening cultural connections across languages and bridging societies through the power of translated words.

Joke Daems is assistant professor human-computer interaction in empirical translation & interpreting studies at Ghent University. They are a member of the EQTIS research team (Empirical and Quantitative Translation and Interpreting Studies) and the LT3 Language and Translation Technology team. Their research focuses on the impact of translation technology (such as machine translation) on translation (process and product), translators (attitudes), and society (e.g., gender bias). They obtained a PhD in Translation Studies in 2016, based on a comparative study of manual translation and the post-editing of machine translations by students and professional translators. In 2017, the thesis was awarded the CIUTI PhD Award. Later work focused on interactive, adaptive MT and the potential of MT for literary translation. Although mainly involved in research, they have taught numerous classes on translation technology and machine translation post-editing on a master's and postgraduate level.

Paula Diez-Ibarbia graduated in English Studies and holds a MA in Language Analysis and Processing. Currently, she is a researcher at the Ontology Engineering Group. Her present research is focused on the representation and conversion of terminological resources into formats compliant with Semantic Web standards, with the objective of enhancing data interoperability.

Aletta G. Dorst is an Associate Professor in Translation Studies and English Linguistics at Leiden University. Her research focuses on metaphor variation, metaphor translation, style in translation, literary machine translation, and machine translation literacy. She recently led an NRO Comenius Senior Fellow project on "The value of machine translation in the multilingual academic community" and was the lead researcher for the work package on metaphor identification and translation on the ZonMW Memorabel project "Dementia in metaphors". At Leiden University Centre for Linguistics she teaches a range of courses in the Minor Translation and the MA Translation, including courses on Translation Studies, Translation Technology, Multimodal Translation and Subtitling.

Christos Ellinides is currently the Director-General for the Directorate-General for Translation (DGT) at the European Commission. He is also the Chairman of the Management Board of the Translation Centre for the Bodies of the European Union (CdT).
He joined the Commission in 2006 as Director for corporate digital solutions and services in the Directorate-General for Informatics and was appointed later to the post of Deputy Director-General in DGT. Prior to this, Mr Ellinides has been responsible for the introduction and effective use of information and communication technologies in various organisations operating mainly in Europe, for more than 2 decades.
His career has evolved through posts and assignments in varying industries, and he has acquired broad and extensive experiences mostly in a management capacity including the functions of both a CEO and a CIO. His profile is marked

by advanced organizational, coordination, and managerial skills, coupled with profound expertise in IT system platforms, collaborative tools, and telecommunications. He has been instrumental in spearheading innovative corporate infrastructure solutions for the Commission's information systems and language technologies, leveraging the latest technological advancements, including artificial intelligence.

He has delivered speeches, lectures, and presentations on business, IT and linguistic matters on an international and pan-European level. He is an active member in several professional bodies. He has been a director and board member in a number of organisations with substantial international exposure.

Christos Ellinides is a Chartered Engineer (CEng), Chartered IT Professional (CITP) and a Fellow of the British Computer Society (FBCS). He holds a M.Sc. in Business Systems Analysis and Design from the City University of London (UK) and a B.Sc. in Business and Computing from the Nova University of Miami (USA).

Michael Farrell is an untenured lecturer in post-editing, machine translation, and computer tools for translators at the International University of Languages and Media (IULM), Milan, Italy, the developer of the terminology search tool IntelliWebSearch, a qualified member of the Italian Association of Translators and Interpreters (AITI), and a Council member of Mediterranean Editors and Translators. Besides this, he is also a freelance translator and transcreator. Over the years, he has acquired experience in the cultural tourism field and in transcreating advertising copy and press releases, chiefly for the promotion of technology products. Being a keen amateur cook, he also translates texts on Italian cuisine. He spoke at TC36 on solving terminology problems with IntelliWebSearch, at TC39 on building a custom machine translation engine as part of a postgraduate university course, at TC40 on Raw Output Evaluator, a freeware tool for manually assessing raw outputs from different machine translation engines and at TC44 on how translators incorporate machine translation into their workflow.

Francesco Fernicola began working in the Speech-to-Text Unit of the European Parliament's translation service (DG TRAD) in September 2023. He started as a Computational Linguist and is now serving as a Project Manager and Terminology Coordinator. Holding a Master's degree in Specialized Translation from the University of Bologna, with a focus on Machine Translation Evaluation (both automatic and manual), he has participated in various projects within the fields of Corpus Linguistics, Computational Linguistics and Natural Language Processing. His work has focused on Quality Estimation techniques for Machine Translation, Misogyny Identification, as well as Sentiment and Emotion Analysis.

Amal Haddad Haddad is a lecturer at the Department of Translation and Interpreting of the University of Granada (Spain). She is a member in LexiCon Research Group. She studied English and Spanish Language and Literature at the University of Jordan; and Translation and Interpreting at the University of Granada. She holds an MA in Translation and Interpreting and a PhD in Translation and Terminology from the University of Granada. Her research interests lie in the areas of the Translation, Terminology, Corpus Linguistics, as well Translation Technologies. She has different publications in national and international journals and publishers.

Ildikó Horváth has been the director of the Translation Centre for the Bodies of the European Union since 1 February 2022. Ms Horváth has a versatile background, she was Vice Dean for International Affairs as well as Head of Department and Associate Professor in the Faculty of Humanities of ELTE University, Budapest, Hungary. She was also the Director of its Institute of Language Mediation and the President of the European Masters in Conference Interpreting (EMCI) Consortium.

With wide-ranging experience as a translator and conference interpreter as well as a PhD in Translation Studies and Applied Linguistics, Ms Horváth has a solid understanding of the Translation Centre's core business. In her different roles, she has developed sound management and networking skills, and has overseen the organisation of translation and interpreting studies at both MA and PhD level. She is also the author of numerous articles and books on translation and interpreting, with a recent focus on artificial intelligence and the digital transformation of the language services market.

Rafał Jaworski, PhD, works as a Linguistic AI Expert at XTM International. He is an academic lecturer and scientist specialising in natural language processing techniques. He develops robust AI algorithms for the needs of computer assisted translation. These include, among others, automatic lookup of linguistic resources and automatic post editing. At XTM International he leads a team of young and talented AI specialists who put his visions and ideas into practice.

Szymon Kaczmarek, MSc, holds the position of Junior Linguistic AI Specialist at XTM International. As a computer science and linguistics enthusiast, Szymon merges both domains to create pioneering natural language processing solutions. With a keen interest and expertise in large language models and deep learning, he strives to enhance and optimise computer-assisted translation tools. Collaborating with the accomplished AI team at XTM International, Szymon actively engages in pushing the boundaries of linguistic AI technologies by implementing state-of-the-art ideas into reality.

Rebecca Knowles is an Associate Research Officer in the Multilingual Text Processing team at the Digital Technologies Research Centre (DT), National Research Council of Canada (NRC-CNRC). She specializes in machine translation and computer-aided translation, with a recent focus on methodologies for human evaluation of machine translation.

Maarit Koponen is Professor of Translation Studies at the University of Eastern Finland. Her research focuses on theoretical and practical aspects of translation technology, particularly the use of machine translation both in professional translation workflows and in non-professional settings. She has published various articles on the use of machine translation and post-editing in translation and subtitling workflows, quality of machine translation and the impact of machine translation use on copyright and authorship. She has also previously worked as a professional translator.

Foteini Kotsi is a multiskilled linguist. She obtained her bachelor's degree in Foreign Languages, Translation, and Interpreting from the Ionian University. She has been trained in translation within the European Commission. Foteini has obtained hands-on experience in the localisation industry having worked as both a linguist and a localisation project manager while specialising in CAT tools and translation technology.
In 2020, she was awarded a scholarship to the European Master's for Technology in Translation and Interpreting. Her hands-on experience in the language service industry fuelled her passion for research, searching for new ways to help professionals be more productive and optimise common localisation workflows.

Todor Lazarov holds a PhD degree in Computational linguistics and has a background in Linguistics. He has also specialized Artificial Intelligence in the University of Amsterdam. He has a diverse experience with CAT tools and has also established successful collaboration with different commercial MT providers. He is providing mentorship and education to freelance translators about current trends and translation technologies.

Marie-Aude Lefer is Associate Professor of Translation Studies and English-French translation at UCLouvain, Belgium, where she acts as Head of the Louvain School of Translation and Interpreting. Her current research interests include technology in translator education, machine translation post-editing training, post-editing and translation quality assessment, translation error annotation, corpus approaches to student translation and post-editing, post-editing pricing methods, and fair pay. She has co-edited nine volumes and special issues, such as Empirical Translation Studies: New methodological and theoretical traditions(De Gruyter, 2017), Extending the Scope of Corpus-based Translation Studies (Bloomsbury, 2022) and Learner Translation Corpus Research (Benjamins, 2023). Her most recent journal publications include The Machine Translation Post-Editing Annotation System (MTPEAS): A standardized and user-friendly taxonomy for student post-editing quality assessment (Translation Spaces) and Introducing MTPE Pricing in Translator Training: A Concrete Proposal for MT Instructors (The Interpreter and Translator Trainer).

Ting Liu is a third-year PhD candidate in the School of Translation and Interpretation of University of Ottawa. Her research interest includes human evaluation of machine translation and translation pedagogy in the digital era.

Jeevanthi Liyanapathirana is a PhD student at the Faculty of Translation and Interpreting, University of Geneva, where her research question lies on incorporating speech technologies for translation and post editing purposes. She has been a fellow in translation technology as well as a translation technologist in the World Intellectual Property Organization, Geneva and is currently working as a Document and Translation Technologies Specialist at World Trade Organization, Geneva, Switzerland. She holds a Masters of Philosophy in Computational Linguistics from the University of Cambridge, UK (MPhil in Computer Speech, Text and Internet Technology) and a Bachelor of Science (Computer Science Special Degree) from the University of Colombo, Sri Lanka. She has participated in multiple EU projects, Swiss National Science Foundation projects as well as South Asian Localization projects involving machine translation, speech recognition and Computational Linguistics in general. She has worked as a research intern in machine translation at Idiap Research Institute, Switzerland as well as at Language Technology Research Laboratory at University of Colombo where she worked as research assistant in Computational Linguistics. Currently, she is also a member of the Bibliomics and Text Mining Group at the University of Applied Sciences, Geneva.

Chi-kiu Lo 羅致翹 is a Senior Research Officer in the Multilingual Text Processing team at the Digital Technologies Research Centre (DT), National Research Council of Canada (NRC-CNRC). She specializes in machine translation quality evaluation and estimation based on structural and lexical semantics, with a recent focus on methodologies for human evaluation of machine translation.

Elizabeth Marshman is an Associate Professor at the University of Ottawa School of Translation and Interpretation, and a member of the Observatoire de linguistique Sens-Texte. Her research interests include user perspectives on translation technologies, technology teaching, and computer-assisted terminology.

Patricia Martín-Chozas works as a postdoctoral researcher in Artificial Intelligence at the Ontology Engineering Group and as Assistant Professor at Universidad Politécnica de Madrid. Her research has been oriented to the generation and representation of terminological resources by means of Semantic Web technologies. Her next research steps are focused on the exploitation of terminological resources published as Linked Data to improve the performance of Large Language Models.

Elena Montiel-Ponsoda is an Associate Professor of Applied Linguistics at Universidad Politécnica de Madrid (UPM), Spain, and a member of the Ontology Engineering Group at the same University. Her main research interests are in the common ground between Terminology and Ontology Engineering. Her research has focused on the development of models to enrich ontologies with multilingual information and to expose terminologies and other language resources as linked data. She is currently involved in several national research projects (INESData, TeresIA) that explore the sharing of language resources as linked data in the so-called "data spaces".

Jonathan David Mutal is a Research and Teaching Assistant at the Department of Translation Technology (referred to by its French acronym TIM). His research interests concentrate on neural machine translation, machine learning, natural language processing and evaluation. He is a strong advocate of producing research to bridge the gap between academia and business. Jonathan holds a BSc (5 years degree) in Computer Science and his master thesis consisted of an ongoing academia-industry collaboration that aims to integrate MT into the workflow of a big language service provider. The thesis describe the evaluations carried out to select an MT tool (commercial or open-source) and assess the suitability of machine translatio

Hana Nessakh is a PhD student at the University of Ottawa, School of Translation and Interpretation. Her research interests encompass translation technologies, including Artificial Intelligence and Neural Machine Translation, the long-term sustainability of the translation industry, and the ethical considerations of utilizing AI and NMT in this field.

Marija Omazić is a Full Professor at the Faculty of Humanities and Social Sciences, University of Osijek, Croatia. Dr. Omazić holds a PhD in Linguistics from the Faculty of Humanities and Social Sciences, University of Zagreb. Her academic journey was further enriched by a Fulbright scholarship at Northern Arizona University, where she specialized in corpus linguistics and phraseology.

Dr. Omazić was the founder and has been the Director of the MA Program in Translation and Interpreting Studies at her home institution since 2009, where she teaches courses in Simultaneous and Consecutive Interpreting, Terminology, Translation Technology, and Translation Practicum. Her research interests include translation and interpreting, phraseology, and cognitive linguistics.

In her professional service, Dr. Omazić has played a significant role in European research and academic quality assurance. She has been a reviewer and panel member for Horizon 2020 and COST projects and an evaluator for the European Master's in Translation (EMT) network. She also led WP7 Dissemination, Training, Awareness and Exchange on the FP7 project Mobility and Inclusion in Multilingual Europe. She was involved in the EU COST Action CA19102 LITHME as a Management Committee member and Jean Monnet Module LEULEX Languages and EU

Law Excellence project as a collaborator, as well as in several nationally funded research projects. She has participated in several academic quality assurance procedures across Europe.
Dr. Omazić's professional memberships include the European Society for Translation Studies (EST), the European Society of Phraseology (EUROPHRAS), and the Croatian Applied Linguistics Society (CALS). She is a member of the editorial boards of several scholarly journals, including Hieronymus, Jezikoslovlje and Strani jezici, and the advisory boards of journals ExELL and Latvijas intereses Eiropas Savienībā.
Her contributions to the academic community extend to her active engagement as a conference interpreter and translator.

Constantin Orasan is a Professor of Language and Translation Technologies at the Centre of Translation Studies, University of Surrey, UK and a Fellow of the Surrey Institute for People-Centred Artificial Intelligence. Before starting this role, he was a Reader in Computational Linguistics at the University of Wolverhampton, UK, and the deputy head of the Research Group in Computational Linguistics at the same university. He has over 25 years of experience in the fields of Natural Language Processing (NLP), Translation Technologies, Artificial Intelligence and Machine Learning for language processing. His recent research focuses on the use of Generative AI as a support tool for translators. His research is well known in these fields as a result of over 130 peer-reviewed articles in journals, books and international conferences.

Lucía Palacios-Palacios, graduated in Spanish Philology, obtained a MA in Language Analysis and Processing and is a first-year PhD student at the Ontology Engineering Group. Her research focuses on developing Word Sense Disambiguation (WSD) and Entity Linking/Matching techniques to facilitate the transformation of domain-specific terminologies into Linked Data formats.

Christine Pasquier is a Russian-French translation lecturer at UCLouvain, Belgium, respectively at the Faculty of Translation and Interpreting "Marie Haps" Saint-Louis – Brussels (Bachelor's degree) and at the Louvain School of Translation and Interpreting (Master's degree), specialized in the fields of scientific and technical translation, translation with regard to international public law, international affairs, geopolitics and geostrategic matters. She is also in charge of the revision and post-editin

Alicia Picazo-Izquierdo is a PhD candidate in the field of machine translation and specialized languages since 2021 at the University of Alicante. She holds a Degree in Translation and Interpreting (University of Alicante, 2020) and a Master's Degree in Teaching (University Francisco de Vitoria, Madrid, 2021). Her research interests are based on the fields of machine translation, computational linguistics, corpus studies, specialized translation, and translation quality. Her current research lines include corpus analysis of specialized languages, translation quality error typologies, and corpus annotation. She has participated in a CIUTI-funded research stay at the Institute of Translation and Interpreting of the Zurich University of Applied Sciences (Switzerland), where she cooperated with the project "Machine translation for crisis communication".
Her main publications are focused on neural machine translation error detection and language learning from a translational approach. Her professional career is mainly based in specialized translation, post-editing, and linguistic quality assurance. She has been working as an in-house translator and proofreader since 2021 at Traduloc, a Spanish language service provider company. She is a member of AELINCO, the Spanish Association of Corpus Linguistics."

Ilja Rausch is an AI engineer with a PhD in AI and Swarm Robotics. He is one of the principal investigators of projects focused on training Large Language Model (LLMs) on European high-performance computing (EuroHPC) infrastructure using large data volumes. He also contributes to the Commission-internal generative AI-based prototypes and the AI@EC network activities.
Prior to his current role, Ilja worked as a Senior Data Scientist at a Fortune 50 company, where he led a team of data scientists in a project applying generative AI in a business context. As a lead developer, he applied statistical modeling, machine learning, and AI to drive business value. He created models based on data analysis, statistics, data mining, and knowledge graphs.
Ilja worked with cloud-based platforms (Azure, Databricks) and on-premises infrastructure. He advised stakeholders and executives on the risks and benefits of innovative AI technologies and their potential to solve business problems. Ilja analyzed big data with PySpark and assisted in MLOps, helping stakeholders design solutions and unlock data-driven insights.

In addition to his technical expertise, Ilja has a strong background in knowledge transfer and mentoring. He has lectured on Graph Data Science, sharing his knowledge with students of computer science and engineering. Ilja has also assisted in tutorials on C++ and mathematics and supervised a dozen students and interns in both academic and business settings.

Francesco Rossi works at the European Parliament, Directorate General for Translation, for the Strategy and Innovation unit. He has considerable experience in IT, project management, digital accessibility and communication and is responsible and involved in various innovation projects, spanning from the fields of translation, AI and communication. Francesco's career began in journalism, working as a reporter, then it took a turn towards languages when he first joined the European Parliament as a trainee in 2013. He holds a Ph.D. in Information Technology, Communication, and Linguistics from the Università degli Studi di Salerno. Since October 2015, Francesco has been teaching as a lecturer at the University of Luxembourg. Francesco Rossi continues to be a driving force in the intersection of IT and linguistics, leveraging his extensive background to foster innovation within the European Parliament.

Francesco Saina is a multifaceted Italian linguist working as a translator and interpreter with English, French, and Spanish. He is also a university lecturer in translation, interpreting, and language technology, and collaborates on academic and industrial research projects on translation and interpreting technology and natural language processing. His works on the applications of digital technology to the language professions have been published in academic journals and presented at international conferences. His research activity focuses on computer-assisted translation and interpreting, applied linguistics, innovation, and training — at the intersection of theoretical investigation, professional practice, and instructional implementations.
With the Sapienza NLP research group, he developed DiBiMT, the first entirely hand-curated benchmark for the analysis and evaluation of semantic disambiguation biases in machine translation systems, which was awarded as the Best Resource Paper at the 2022 conference of the prestigious Association for Computational Linguistics (ACL)."

Leena Salmi is a University Lecturer in French and Translation Studies at the University of Turku. Leena has been involved in translator education for over 20 years and has 3 years' industry experience as translator and technical writer. Her PhD thesis (2004) dealt with the usability of computer user documentation and her current research interests relate to translator training, translation technology and translation quality assessment. Her teaching focuses on practical translation courses (French-Finnish), translation technology, and translation company simulation, as well as supervision of MA and PhD thesis. Leena is currently a member of the Board of the EMT network (since 2021) and the Chair of the Authorized Translators' Examination Board (certification of translators of official documents; since 2023). She has been involved in various research and professional activities in Finland such as the organization of the yearly KäTu Symposium on Translation and Interpreting Studies (since 2003) and different committees of the Finnish Association of Translators and Interpreters (2004-2009). She also has extensive contacts to translator associations and local translation companies.

Perrine Schumacher is currently working as a F.R.S.-FNRS postdoctoral researcher at the University of Liège (Belgium). She holds a PhD thesis in Languages, Letters and Translation Studies from the University of Liège and in Multilingual Information Processing from the University of Geneva. Her research interests focus on translation training and on the use of AI tools for translation purposes, particularly on machine translation post-editing.

Miriam Seghiri has BA in Translation and Interpreting (Spanish-English, French, Italian) and a PhD in Translation and Interpreting (with high honours) from University of Málaga (University's 2006 PhD Best Student Prize). She is currently Full Professor at the Department of Translation and Interpreting at the University of Málaga, Spain. Her research interests include specialised translation (scientific, technical and legal), corpus linguistics and ICTs. She has received the Translation Technologies Research Award (with Prof Gloria Corpas) in 2007 and the María Zambrano Award in 2013. At present, she is Pro Vice Chancellor for International Cooperation and Language Policy at the University of Malaga and Visiting Professor at Universidad Católica del Maule (Chile.) Her research has been presented in national and international academic publications.

Blaženka Šoštarić is a Senior Language Instructor at the Faculty of Humanities and Social Sciences, University of Osijek, Croatia, where she teaches Contemporary English Language and translation courses. Her teaching approach integrates practical translation and interpretation experiences, enabling students to gain real-world insights into language applications in both academic and professional contexts. Her earlier experience includes working as a language assistant for UN monitoring missions, such as UNCRO, UNPROFOR, and UNTAES, where her contributions facilitated essential communication in high-stakes, multilingual settings. Between 1998 and 2006, she worked as a Program and Language Assistant for the OSCE Mission to Croatia, performing translation and administrative tasks that supported international diplomatic initiatives. From 2005 to 2017, she served as an external translator for the Ministry of Foreign Affairs and European Integration in Zagreb, where she contributed to aligning Croatian legislation with EU standards.

Academically, Blaženka holds an MA degree in English and German Language and Literature and continues to deepen her expertise through various professional development programs, including training in translation technologies such as memoQ. Her professional interests include English for Academic Purposes (EAP), legal English, multilingualism, and translation studies, particularly in the context of translation technologies. She actively participated in the Jean Monnet Module LEULEX project on Language and EU Law Excellence, and was also involved in the EU COST Action CA19102 LITHME project, which examined the impact of technology on language in the human-machine era. Blaženka regularly participates in international conferences, including the Translating Europe Forum hosted by the European Commission, where she focuses on translation and the role of translators in maintaining quality and accuracy in the digital age, thereby remaining at the forefront of translation research and application.

Shiyi Tan is a second-year PhD student from the Centre for Translation Studies at University of Surrey. Her research interests include the cognitive process in interpreting, the use of technologies in interpreting and interpreter training.

Silvia Terribile holds a PhD in Translation and Intercultural Studies from the University of Manchester (UK), an MSc in Specialised Translation (Audiovisual) from University College London (UK), and a BA in Linguistic and Cultural Mediation from the University of Turin (Italy). Silvia's primary research interests are in the fields of translation technologies and localisation. Her PhD was a Collaborative Doctoral Award supervised by Prof Maeve Olohan, in partnership with the world-leading language service provider Toppan Digital Language (formerly TranslateMedia), and fully funded by the Arts and Humanities Research Council (AHRC) of UK Research and Innovation. Her doctoral project analysed real-world translation projects completed at Toppan Digital Language to investigate productivity in the post-editing of neural machine translation. Some of the main contributions of her research include: (1) the first large-scale investigation of translation and revision speed in human translation and post-editing, based on real-world data for 90 million words translated by 879 linguists across 11 language pairs, over 2.5 years; (2) the development of RECAP (Repetition, Error, Change, Action, Post-editing), a multi-layered typology to classify different types of edits to the machine translation output; (3) the application of RECAP to analyse edits in a small corpus of real-world English-to-Italian post-editing tasks that required different levels of post-editing effort; and (4) the development of AREA (Automating Repetitive Editing Actions), an algorithm that could automate up to 46% of repetitive edits in post-editing.

Silvia is currently co-organising the International Postgraduate Conference in Translation and Interpreting, which will take place at the University of Manchester in December 2024. She is also an active member of the Language in the Human-Machine Era EU COST Action, an international research network focusing on emerging language technologies. She has been teaching translation, focusing on translation technologies, for four years at the University of Sheffield, University of Manchester, and University of Roehampton. She has served as a Communications and Engagement Coordinator for the North West Consortium Doctoral Training Partnership of the AHRC, where she supported scholars in communicating the value of their research to academic and non-academic audiences in accessible formats. Silvia has previously managed the localisation of advertising campaigns for Nespresso at Hogarth Worldwide, as well as the localisation of websites, web games and mobile apps for Cartoon Network, Boomerang, Cartoonito, Boing and Toonix at Turner Broadcasting System (now part of Warner Bros. Discovery)."

Andrzej Zydroń is co-founder and CIO @ XTM International and technical architect of XTM Cloud; Andrzej Zydroń is one of the leading IT experts on Localization and related Open Standards and sits/has sat on the following Open Standard Technical Committees:

1. LISA OSCAR GMX
2. LISA OSCAR xml:tm
3. W3C ITS
4. OASIS XLIFF

5. OASIS Translation Web Services
6. OASIS DITA Translation
7. OASIS OAXAL
8. ETSI LIS

Zydroń has been responsible for the architecture of the word and character count GMX-V standard, as well as the revolutionary xml:tm. Zydroń is also head of the OASIS OAXAL technical committee.